Evaluating the information content of human microbiomes

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Benjamin Hillmann

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dan Knights

March, 2022

Acknowledgements

Thank you to everyone that enabled my development and research throughout my graduate experience. There are a lot of you, and unfortunately, I won't be able to mention everyone by name without writing another whole thesis.

To all of the members of the Knights Lab: you are all brilliant, driven, and you made working on research fun and exciting. Without all of your passion and insight, I would not have been successful. Thank you to Tonya Ward, Gabe Al-Ghalith, Robin Shields-Cutler, Abigail Johnson, Pajau Vangay, Jonathan Clayton, Jesse Kennedy, Shannon Zhou, Calvin Thoma, Suzie Hoops, Mo Houtti, and Jay (Ya-Fen) Wei.

To all of my personal, academic, and business mentors: thank you for believing in me and pushing my development and growth. Thank you to my Ph.D. committee Chad Myers, Rhan Blekhman, and Rui Kuang for generously giving their time. To my business mentors, Mark Albrecht and Paul Rice: your passion, enthusiasm, and belief in me prompted a visiting scholarship in industrial research. Without your encouragement and determination, I would not have had the life-changing experience of working with and around such driven and talented people.

To my Ph.D. advisor, Dan Knights: the amount of patience, enthusiasm, and advice you impart to everyone around you is unprecedented. Thank you for all of

your hard work and talent generously imparted and for elevating everything and everyone around you to greatness.

I feel incredibly privileged to revel in this accomplishment with my friends and family: my mother, Teri, for spending countless hours encouraging my academic growth and helping proofread all of my scholarly works from kindergarten until now; my father, Todd, for instilling in me a passion for computing at a young age; to my siblings, Toby, Nathaniel, and Katie, and to my friends, for always grounding me and reminding me to take time to enjoy my hobbies; and to my friend, Cassandra Roeder, for pushing the final steps towards completion.

To my life partner, Rae Hohle: I am eternally indebted to you for your partnership with me throughout this entire process. Thank you for pushing me to persevere at times and reminding me to take breaks at others.

Dedication

I dedicate this dissertation to all of the people working to advance science, especially those who are currently researching COVID-19.

Abstract

Microbes vastly outnumber all other organisms on earth and are integral to many aspects of the ecological fitness of the earth's soils, oceans, animals, and plants. Unfortunately, most of the microbes in these communities cannot be cultured, so to observe these communities' biological functions, we must study their DNA. After a researcher sequences a microbial community, they utilize informatics methods to correlate the taxonomic and functional profiles to their traits of interest. However, these methods assume that the underlying taxonomic and functional profiling are accurate. If procedures are developed to identify the profiles of a community more accurately, the increased precision will enable higher power testing of hypotheses and detection of these communities' causal roles. We propose novel, accurate, and data-efficient methods for taxonomic and functional profiles in shotgun metagenomic datasets.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Microbes vastly outnumber all other organisms on earth and drive many aspects of the ecological fitness of the earth's soils, oceans, animals, and plants [1]. Among many processes, microbial communities influence the rate of methane released from melting permafrost [2]. They are associated with numerous chronic human diseases [3]. They additionally regulate essential chemical balances in our soil, promoting the health of agriculture [4]. Unfortunately, most of these microbes cannot be cultured, so to observe the biological functions of these communities, we must study their DNA. The complete set of all of the DNA of a community of microbes sequenced together is the metagenome [5]. The metagenomic sequencing experiments to profile microbial communities generate extensive, complex, multi-dimensional data. As such, there is a surge in demand for scientists to precisely describe the microbes present within these communities to determine which microbes are beneficial and which are detrimental for different environments and applications.

We are still in the early stages of human microbiome research through studying metagenomes. With the advent of next-generation sequencing (NGS), we

have created methods to profile these microbial communities. There is room for improvements and discovery with the tools still evolving. In my short time researching this field, I have shown the human gastrointestinal microbiome's relation to gut dysbiosis in relationship to chemotherapy [6] and rheumatoid arthritis [7]. With the methodological improvements described herein, we are one step closer to unraveling the complex roles these microbes play in our everyday lives.

1.1 Background and Literature Review

1.1.1 The Complexity of the Human Microbiome

The number of human cells in the body are outnumbered at least ten times by resident microbes, with the majority of the microbes being *Bacteria* living in the gastrointestinal tract [8]. Given the number of cells and total genetic potential, the microbiome has many avenues to play an integral role in the health of humans. Furthermore, the makeup of the human microbiome is dynamic and can interact with its human host and change over time to fit its conditions [9]. However, the tendency of the microbiome to change over time varies from person to person [10]. Due to competitive ecological interactions, the microbiome has been shown to be quite robust against perturbations and the introduction of new organisms [11]. The first instance proving a microbe's engraftment long-term to a foreign microbiome required the strain to be cultured from a human host. Furthermore, the strain would likely only engraft to the new microbiome if a comparable organism does not already fulfill the microbe's ecological niche. This engraftment example proves both the microbiome's robustness to introductory microbes and shows the microbiome's ability to be altered if the desired community lacks an essential organism.

1.1.2 Microbiome as a Biomarker of Human Health

The human gastrointestinal microbiome has well-documented impacts on the health and wellbeing of its human host [12, 13]. Coupled with this relationship, the microbiome's composition is known to vary widely, and it is distinguishable across people in terms of age, geography, and sex [14]. Many microbiome researchers' objective is to precisely characterize the interaction between the microbiome and different diseases, hoping to help cure, diagnose, or treat those diseases. For example, researchers have shown the microbiome as a diagnostic tool for predicting a patient's risk of bacteremia infection before chemotherapy [15]. Researchers demonstrated the microbiome's response to diet by differentiating the microbiomes between patients that eat plant- or animal-based diets [16]. A direct, measurable way to influence the human microbiome, such as through probiotics or diet, would have profound impacts as a therapeutic role for human health.

1.1.3 Sequencing Techniques

Researchers have typically performed microbiome DNA analysis in one of two ways: using amplicon-based (16S) or whole-genome shotgun-based (WGS) sequencing methods [17]. Amplicon-based techniques typically amplify a highly variable region of the 16S ribosomal RNA gene for profiling *Bacteria*. Amplicon sequencing is affordable but often cannot distinguish between species due to sequence similarities and does not allow high-accuracy prediction of the functional repertoire. Because amplicon methods rely on DNA primers to amplify the region of interest, they are also subject to high bias levels. They can fail to capture organisms whose DNA sequence does not match the primers. As an alternative, WGS experiments sequence randomly selected fragments of all DNA present in a

microbiome. WGS directly measures the functional repertoire of the microbiome by capturing a snapshot of the total metagenomic content and allows up to strain-level characterization of microbiomes by mapping sequencing reads to the unique markers of strain reference genomes. WGS is typically ten times more expensive than amplicon sequencing due to its costly library preparation protocols and the additional costs of very deep sequencing. This forces researchers to choose between affordability with amplicon methods and accuracy with WGS. I focus the majority of the work within this thesis on improvements in cost efficiency and computational methods in WGS sequencing.

1.1.4 Sequencing Quality-Control Algorithms

Quality control of high-throughput, short-read sequencing data generated for microbiome datasets is essential for ensuring the proper analysis and profiling of the microbiome. The quality-control process is quite specialized depending on the sequencing methods utilized. Furthermore, the quality-control process is also quite labor-intensive due to the amount of unprocessed data created in a metagenomic sequencing experiment. Quality control of metagenomic sequencing is one of the first steps of analysis, and therefore all downstream rely on its properness. While the process is essential, it can be a barrier to proper metagenomic analysis due to the domain-specific knowledge required to complete it. In the case of WGS, each sequencing method can be error-prone where each of the nucleotide base pairs receives a quality score assigned to it by the sequencing machine. To alleviate the requirement for this intensive task within this thesis, I helped create and utilize the automated quality-control pipeline `shi7`, (pronounced "shiz -en") [18]. The computational pipeline `shi7`, specifically the *learning* mode, automatically

detects the pairing, barcodes, adaptors, base quality thresholds, and stitching parameters required for all datasets to ensure we use the highest-quality sequences for downstream profiling algorithms.

1.1.5 Microbiome Profiling Algorithms

When using metagenomics to study the human microbiome, researchers usually have two primary profiles of characterization: a profile of the microbes and their respective abundances and a profile of the potential functions and ecological roles the microbes impact. We organize the characterizations of the metagenome into tables. The tables have features for the rows, columns for the samples, and the entries are a quantification of relative abundance. There are two broad categories of tables: taxonomic and functional. Taxonomic tables quantify "who is there" in the microbiome and have features that can be taxonomic species, strains, or genomes. Functional tables quantify "what are they doing" in the microbiome and have features that can be the gene families or pathways.

Taxonomic Profiles

Several methods process quality-controlled sequences from WGS metagenomic sequencing experiments into a taxonomic profile. Three such approaches are alignment marker-gene, alignment full-genome, and k-mer full-genome. The alignment marker-gene approach uses a database of predefined marker genes and uses an alignment [19, 20] algorithm to map the query reads to the database [21]. The marker-gene approach simplifies some profiling processes since not as much memory and computation are required compared to using a full-genome database and alignment [22]. An alignment-free approach maps exact or partial matches of k-mers between the query sequences and full-length reference genomes [23, 24].

Some variation of the last common ancestor of all identified matched genomes from an individual query is reported as the taxonomy among all methods.

Functional Profiles

Functional profiles typically follow the same approach as creating taxonomic profiles, except the reference database is labeled with gene families, orthologies, or pathways instead of taxonomies [25]. This method of profiling matching query-reads to a gene database is known as direct observations or bag-of-genes approach. Another standard method, known as bag-of-genomes, can be taken to align to complete reference genomes instead. For every match to a genome, every gene present in the genome is summed and quantified as present [26].

1.1.6 Approaches for Analyzing a Profiled Microbiome

Researchers generally study a microbiome with a specific hypothesis they are interested in testing. For example, researchers characterized the gastrointestinal microbiome of patients to discover relationships between the metagenome and the risk of type-2 diabetes (T2D) [27]. In this study, after the researchers quantified the taxonomic profile, they were interested in the differences between the diversity and function of the microbiome for healthy subjects and those with T2D. Using alpha-diversity, with standard metrics being Shannon's [28], or Simpson's [29], we compare the within-sample diversity of the profiles on average between outcomes. Another interesting metric that research examines is the beta-diversity, or the between sample distance. One standard metric is to use the unweighted UniFrac [30] metric to compare distances between samples and to visualize the space using a Principal Coordinates Analysis (PCoA) [31]. Many other techniques exist to analyze the microbiome, including even machine learning applications to predict

outcomes [32]. A common problem is maintaining the entire collection of methods used in downstream analysis to maintain reproducibility and standardization across the field. Several computational pipelines that include common strategies in analysis and built-in tracking of all transformations done on the data have been implemented [33].

1.2 Thesis Statement, Specific Aims, and Approach

1.2.1 Thesis Statement

The overarching goal of this research is to establish practical, efficient, and accurate methods for taxonomic and functional profiling for studying the microbiome. We hypothesize that current processes are not optimized for metagenomic work: specifically, in the areas of depth of sequencing, computational efficiency, and genome presence detection. Overcoming the current shortcomings of amplicon and shotgun techniques presents significant challenges in metagenomics and prevents scientists from performing detailed, large-scale studies.

1.2.2 Specific Aims and Approach

Specific Aim 1

Optimize sequencing depth for shotgun human microbiome datasets in cost, data efficiency, and computational accuracy.

Hypothesis: Using sample multiplexing, lower-depth sequencing, and improved taxonomic profiling in "shallow" WGS metagenomic experiments will give

the cost-effectiveness of 16S sequencing with some of the benefits of WGS sequencing.

To test this hypothesis, we will:

- Bootstrap sample sequencing at various depths to evaluate the accuracy of taxonomic and functional profiles and establish a depth optimized for the accuracy and cost-effectiveness tradeoff.
- Evaluate, compare, and contrast the accuracy of shallow WGS, deep WGS, and 16S in terms of Bacterial taxa identified and abundance.

Specific Aim 2

Create an accurate and reproducible shotgun microbiome profiling pipeline.

Hypothesis: Best practices and incremental improvements in alignment algorithms and read disambiguation schemes packaged together into a single computational pipeline will promote more accurate and reproducible metagenomics experiments.

To test this hypothesis, we will:

- Research and develop a computational pipeline for taxonomic and functional profiling of WGS metagenomic datasets.
- Make the taxonomic profiler flexible to the computational capacity of the researcher.

Specific Aim 3

Evaluate methods for accurate genome presence-absence calling and addressing the long-tail of the taxonomic profiles by alleviating the false-positive biases of more extensive databases.

Hypothesis: By utilizing features of the alignment and the genomes within a database, a machine learning model can more accurately classify the presence-absence of genomes within a sample.

To test this hypothesis, we will:

- Curate a training and testing metagenomic WGS dataset with known genome presence-absence metadata.
- Engineer features and then train, optimize, and analyze a machine learning pipeline for presence-absence classification.
- Build and maintain a software pipeline so other researchers can utilize methodology.

1.3 Dissertation Layout

I have organized this dissertation into four manuscripts, each of which is either published or in the process of being published. It begins with a general introduction and background overview of the current practices and challenges in WGS metagenomic experiments. It follows with the research manuscripts described below. Finally, it concludes with a chapter discussing where this research sits in the field and future work for the area.

- Chapter 2: Evaluating the Information Content of Shallow-Shotgun Metagenomics is a research article outlining and evaluating the depth of WGS experiments and comparing them to 16S.
- Chapter 3: Modular Accurate and Scalable Framework for Microbiome

Quantification is a research article that outlines a novel taxonomic and functional profiling pipeline optimized for shallow- and deep-WGS experiments and compares them to other profilers in the field.

- Chapter 4: Classifying Reference Genome Presence" is a research article that outlines a machine learning pipeline that accurately identifies the presence-absence of genomes present in a profile after alignment.

- Chapter 5: Applications of Shallow-Shotgun Metagenomic Profiling Techniques is a technical report showing the earlier techniques applied to an end-to-end metagenomic experiment.

1.3.1 Conclusion

This thesis focuses on dramatic improvements in data efficiency profiling microbial communities using metagenomic sequencing data. After a researcher sequences a microbial community, they should utilize informatics methods to correlate the taxonomic and functional profiles to a trait of interest. However, these methods assume that the underlying taxonomic and functional profiling are accurate. If procedures are developed to identify the profiles of a community more accurately, the increased precision will cascade down every step along the metagenomics pipeline. The informatics methods will have more power to test hypotheses and detect these communities' causal roles with more detailed profiles. Accurate cost- and time-effective taxonomic quantification of environmental samples is essential. Given the weaknesses of the techniques mentioned earlier, we propose novel, highly data-efficient methods for taxonomic and functional profiles in shotgun metagenomic datasets.

Chapter 2

Evaluating the Information Content of Shallow-Shotgun Metagenomics[1]

2.1 Introduction

Although microbial communities are associated with many aspects of human, environmental, plant, and animal health, there exists no cost-effective method for precisely characterizing species and genes present in such communities. While deep whole-genome shotgun (WGS) sequencing provides the highest-level of taxonomic and functional resolution, it is often prohibitively expensive for large-scale studies. The prevailing alternative, high-throughput 16S rRNA gene amplicon sequencing (16S), often does not resolve taxonomy past the genus level and provides only moderately accurate predictions of the functional profile; thus, there is currently no widely accepted approach to affordable, high-resolution, taxonomic

[1] A version of this article has been published [34]

and functional microbiome analysis.

To address this technology gap, we evaluated the information content of shallow-shotgun sequencing with as low as 0.5 million sequences per sample as an alternative to 16S sequencing for large human microbiome studies. We describe a library preparation protocol enabling shallow-shotgun sequencing at approximately the same per-sample cost as 16S. We analyzed multiple real and simulated biological data sets, including two novel human stool samples with ultra-deep sequencing of 2.5 billion sequences per sample, and found that shallow-shotgun sequencing recovers accurate species-level taxonomic and functional profiles of the human microbiome. We recognize and discuss some of the inherent limitations of shallow-shotgun sequencing, and note that 16S sequencing remains a valuable and important method for taxonomic profiling of novel environments. Although deep WGS remains the gold standard for high-resolution microbiome analysis, we recommend that researchers consider shallow-shotgun sequencing as a useful alternative to 16S for large-scale human microbiome research studies.

2.1.1 Motivation for Shallow-Shotgun Sequencing

A common refrain in recent microbiome literature and scientific talks is that the field needs to move away from broad taxonomic surveys using 16S sequencing, and toward more powerful longitudinal studies using shotgun sequencing. However, performing deep-shotgun sequencing in large longitudinal studies remains prohibitively expensive for all but the most well-funded research labs and consortia, which leads many researchers to choose 16S sequencing for large studies, followed by deep-shotgun sequencing on a subset of targeted samples. Here we show that shallow or moderate-depth shotgun sequencing may be used by researchers to obtain species-level taxonomic and functional data at approximately the same cost

as amplicon sequencing. While shallow-shotgun sequencing is not intended to replace deep-shotgun sequencing for strain-level characterization, we recommend that microbiome scientists consider using shallow-shotgun sequencing instead of 16S sequencing for large-scale human microbiome studies.

2.1.2 Overview

Despite the close association of microbial communities with many aspects of human, environmental, plant, and animal health [8, 35, 36, 37] it is not currently possible to characterize precisely the species and genes present in a microbial community in a cost-effective manner. The microbial communities of human microbiomes are complex, multivariate and multidimensional, requiring large studies to power novel biomarker discovery and predictive modeling [8, 32]. Deep whole-genome shotgun metagenomics (WGS) provides highly resolved strain-level taxonomic and functional information, but is generally cost-prohibitive for large-scale studies. Many of the largest microbiome studies to date have been performed via 16S rRNA gene amplicon sequencing (16S), a cost-effective alternative, but 16S typically provides only genus-level taxonomic assignments [38] and rough estimates of the functional repertoire [39, 40], limiting the amount of information that can be learned from the data. The purpose of this paper is to evaluate shallow shotgun as a possible cost-effective alternative to 16S sequencing for large-scale biomarker discovery with improved taxonomic resolution and functional accuracy.

A major concern for the use of any microbiome assay is the ability to identify and quantify taxonomic and functional traits from within a complex community. Deep WGS has a number of advantages over 16S for microbiome profiling for these purposes in well-characterized environments; for example, deep WGS of mixed communities, such as the human gut microbiome, has been effective at

recovering strain-level polymorphisms and functional traits for abundant strains [41, 42, 43]. Both shallow- and deep-shotgun sequencing sequencing are also less subject to amplification bias than 16S because they do not rely on targeted primers to amplify a marker gene [44]. However, at the time of writing, WGS sequencing typically costs nearly an order of magnitude more per sample than 16S for library preparation and DNA sequencing.

Although extensive work has been done to characterize how many 16S reads are required for quantifying relevant biological signals, the same has not been done for taxonomic or functional profiling with shotgun metagenomics figure 2.1. The depth necessary for sequencing in a particular study depends on the purpose of the study; in many studies, the key goals are to understand which species and functions are present, and to identify biomarkers related to experimental groups or outcomes. To address the important question of how many reads are required to capture species-level taxonomic and functional assignments (e.g. KEGG Orthology groups [45]), we analyzed shotgun sequencing data at various depths. We found that shotgun sequencing can produce similar quality species and functional profiles to deep WGS using as few as 0.5 million sequences per sample, as demonstrated on deep whole-genome sequencing (WGS) samples from the Human Microbiome Project (HMP) dataset [8], the HMP mock community [44], a diabetes study [27], simulated human gut microbiomes, and two novel human stool samples on which we performed ultra-deep sequencing of 2.5 billion sequences per sample.

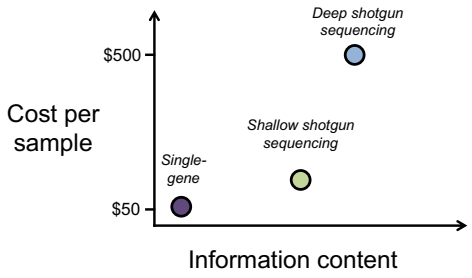

Figure 2.1: **The motivation for shallow-shotgun sequencing is that it provides more information than 16S amplicon sequencing, at only slightly higher cost.**

2.2 Methods

2.2.1 Alignment Algorithms Evaluated

Alignment was performed using several existing tools and algorithms, including Bowtie2 [20], Centrifuge [24], Kraken [23], an in-house k-mer based aligner for comparison with Kraken [46] and an accelerated adaptation of Needleman-Wunsch for exhaustive gapped semi-global alignment [47, 48].

2.2.2 Shotgun Species Profiling

After trimming sequences until quality score is above 20, and discarding trimmed sequences shorter than 80 bases or with average quality score of less than 30, query reads were mapped with several different alignment tools against representative and reference genomes from the RefSeq database version 82 [49] using a 95% identity threshold (also compared to 98% for precision and recall evaluation on

the simulated data). A read that mapped to a single reference genome is labeled with the NCBI taxonomic annotation. All reads that mapped to multiple reference genomes are labeled as the last common ancestor (LCA) of each label according to the NCBI taxonomy, and only species-level assignments are retained. We use a confidence-adjusted LCA that requires at least 80% of all tied best matches to agree for species annotation. All source code can be found at the GitHub repository: `https://github.com/knights-lab/SHOGUN` [50].

Additional analysis code used to generate figures and run tests for this manuscript can be found here: `https://github.com/knights-lab/analysis_SHOGUN`.

2.2.3 Shotgun Functional Profiling

Functional profiling was obtained using KEGG Orthology (KO) [45] annotations for RefSeq derived genes [49] from directly observed exhaustive gapped alignments in ultra-deep WGS sequencing. To improve the accuracy of the direct KO profiles for low-abundance genes, the KO profiles were separately predicted from reference genomes and the predicted profiles were used to augment the estimates of low-abundance KOs. Specifically, we identified those query reads with a 100% match to exactly one reference genome, and predicted the entire KO profile of that genome to be present in the sample, similar to a previously published approach [40]. This is similar to the PICRUSt algorithm for amplicon sequencing data [39], but without the intermediate steps of clustering short-read amplicons and identifying closely related reference genomes. The final KO profiles reported by `SHOGUN` are a weighted average of predicted and directly observed KO profiles. The predicted KO counts are weighted between 0.0 and 0.1 by a linear function of the coefficient of variation of the count for a given KO, estimated from the size of the binomial confidence interval for the observed count of a given KO, divided

by the count of that KO. The direct KO profiles receive the remainder of the weight, such that the direct KO profiles receive at least 90% of the weight for all genes, and the predicted KO profiles are trusted only for the lowest abundance genes where the expected variance in observed count is high. 16S sequences were aligned to Greengenes version 13.8 [51] at 98% identity with exhaustive gapped alignment [47, 48]. Where a query sequence aligned equally well to multiple reference sequences, the taxonomic assignment was made using the last common ancestor conserved across at least 80% of the set of references.

2.2.4 Human Microbiome Project data

We obtained deep WGS data from the Human Microbiome Project (HMP) [8] and sub-sampled the data to simulated shallow-shotgun sequencing depth. We annotated the deep WGS data using fully exhaustive gapped alignment for both taxonomy and functional profiles against all complete, representative bacterial genomes from the reference database RefSeq version number 82 [49]. We then rarefied these samples repeatedly to 1,000, 10,000, 100,000, and 1 million, and 10 million sequences per sample, and ran the `SHOGUN` pipeline to quantify species and gene profiles. The HMP mock community data are from runs SRR2726671 and SRR2726672 from NCBI accession SRX1342165 [44].

2.2.5 Simulated Human Metagenomes

The body sites analyzed from the HMP project were first grouped according to the broad stool, skin, and oral body sites. We calculated the average relative abundance of all samples within each group. The 100 most abundant species for each group were used for simulating communities. The reads were simulated

from a randomly selected strain belonging to each of those most abundant species according to the average proportion of that species in the respective body site group using the tool `dwgsim` [52]. The reads were simulated with default settings for Illumina single-end sequencing machines with a 5% mutation rate where 2% of mutations are indels and a maximum of ten ambiguous bases per query sequence.

2.2.6 Sequencing Library Preparation

Shotgun DNA sequencing was performed on the Illumina HiSeq platform. DNA was extracted using the Qiagen DNeasy PowerSoil kit, and was quantified using the Quant-iT PicoGreen dsDNA assay (Thermo Fisher). DNA sequencing libraries were prepared using one-quarter-scale NexteraXT reactions (Illumina). The resulting DNA libraries were denatured with NaOH, diluted to 8 pM in Illumina's HT1 buffer, spiked with 1% PhiX and a HiSeq 1x100 cycle v3 kit (Illumina) was used to sequence samples. Samples are barcoded and multiplexed on a HiSeq high-output run, with an expected output of at least 0.5 million total sequences per sample. For the ultra-deep-shotgun sequencing, 64 separate libraries were prepared as described above but using full Nextera reactions from a homogenized stool sample and were multiplexed on a HiSeq 3000 high-output run, using an entire run per sample.

2.2.7 Data Availability

The data for the ultra-deep WGS sequencing have been deposited in the European Nucleotide Archive with the accession code PRJEB24152.

2.3 Results

A comparison between deep- and shallow-shotgun sequencing in real and simulated biological data sets demonstrated that shallow shotgun provides nearly the same accuracy at the species and functional level as deep WGS sequencing for known species and genes in five key aspects of microbiome analysis: (1) beta diversity figure 2.2.B-C; (2) alpha diversity figure 2.2.D-E; (3) species composition figure 2.3.A,C; (4) functional composition figure 2.3.B,D; and (5) clinical biomarker discovery figure 2.4.

2.3.1 Alpha- and Beta-Diversity Profiling

We obtained deep WGS data from the Human Microbiome Project (HMP) [8] and sub-sampled the data to simulate shallow-shotgun sequencing depth across five body subsites representing skin, oral, and gut habitat microbiomes. Surprisingly little sequencing was needed to discover the same trends found in deep WGS data. To discover these trends, we annotated the deep WGS data using an accelerated version of fully exhaustive gapped Needleman-Wunsch alignment [47, 48] for both taxonomic and functional profiles, against all complete, representative bacterial genomes from the reference database RefSeq version number 82 [49]. Fully exhaustive alignment allowed us to identify any and all ties for best match for each input sequence according to sequence identity. Species relative abundance profiles were derived by tabulating the number of sequences with at least 80% of the best hits belonging to one species. This is similar to the direct mapping approach used commonly in k-mer-based approaches to shotgun metagenomics taxonomic profiling [23], but with higher sensitivity and recall due to the use of gapped sequence alignment (see simulated data analysis section 2.3.2). Using the fully exhaustive

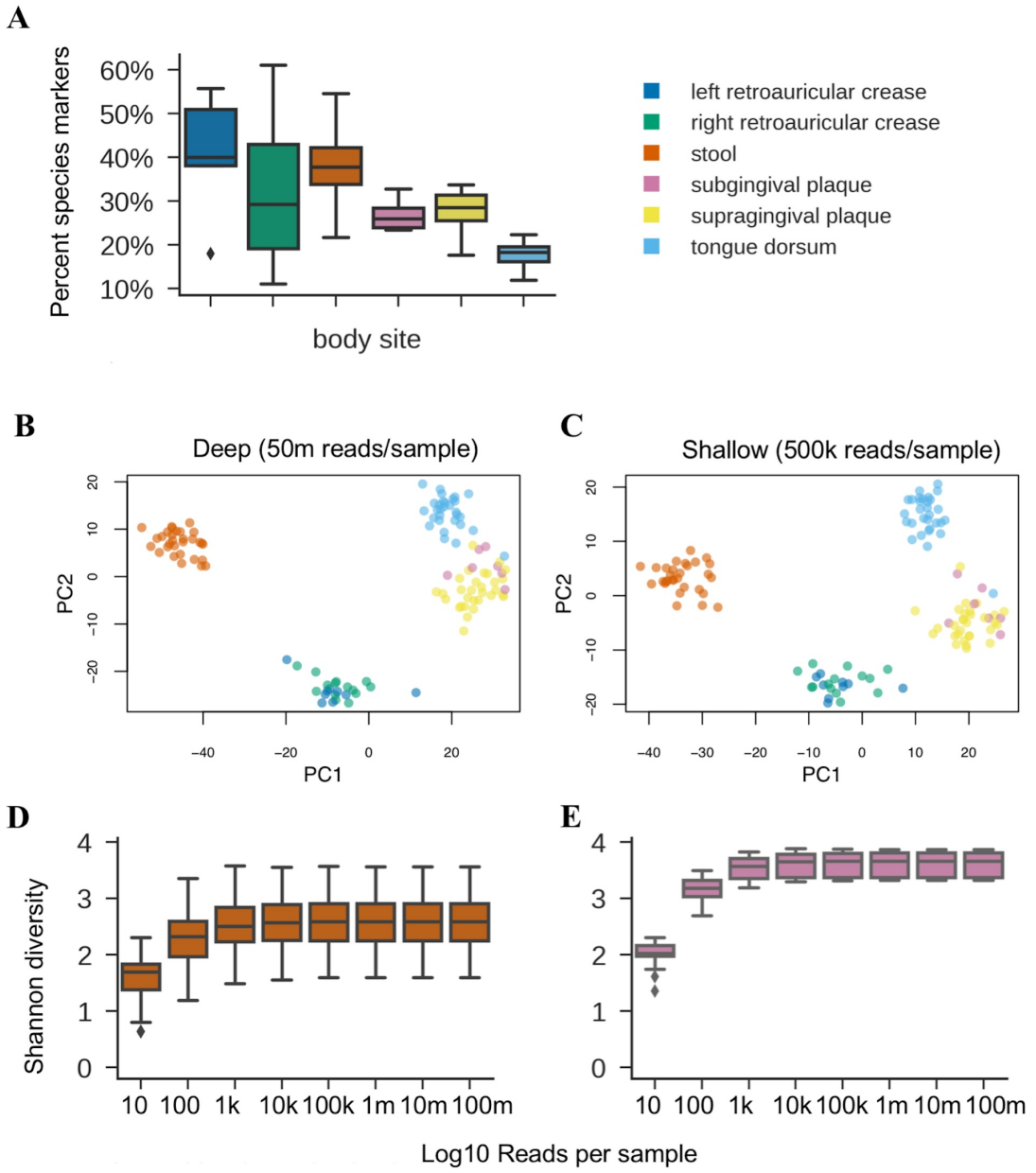

Figure 2.2: **Information content of deep and shallow-shotgun sequencing.** (**A**) Percent of raw shotgun DNA sequences that are unique to one bacterial species across different human body habitats ($n = 7$ distinct samples for plaque samples, $n = 30$ distinct samples for other body sites). (**B**, **C**) Principal coordinates analysis of Bray-Curtis beta diversity using deep (**B**) and shallow (**C**) sequencing (sample sizes as in [**A**] above). (**D**, **E**) Shannon diversity estimates at varied sequencing depths for human stool **D** and subgingival plaque microbiomes (**E**; sample sizes as in [**B**] above). Boxplots show minimum, first quartile, median, second quartile, and maximum, with outliers beyond 1.5 times the interquartile range plotted individually.

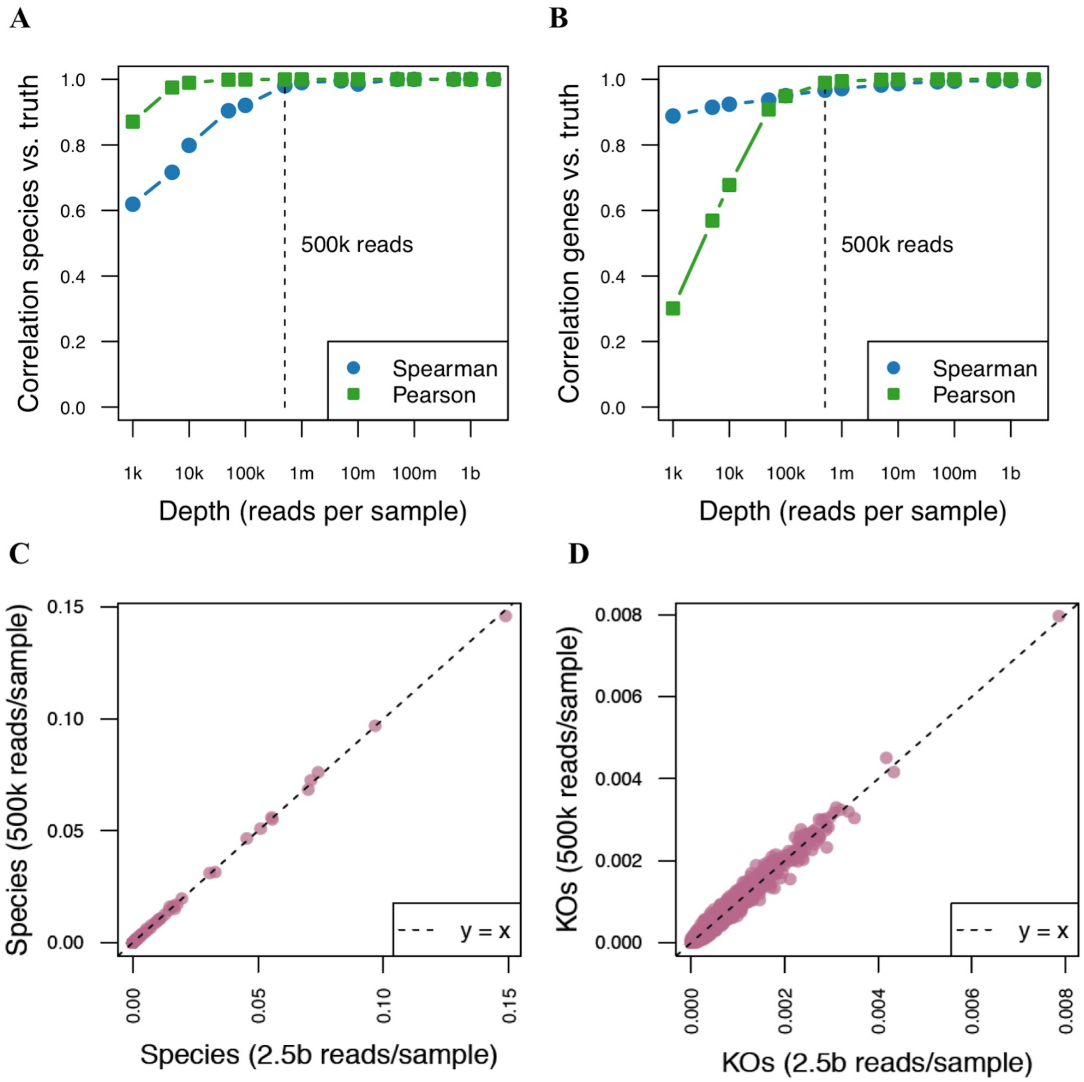

Figure 2.3: **Comparison of species and function profiles of shallow-shotgun with ultra-deep sequencing data.** (**A**, **B**) Correlation with ground-truth species (**A**) and KEGG Orthology group or KO (**B**) profile for known genes present in the reference database, at different sequencing depths, showing that as few as 0.5 million sequences recover nearly the full species and function profiles (ground truth based on 2.5 billion reads per sample; n = 4,394 genes and 694 species at each subsampling level from Subject 1 ultra-deep sequencing sample; comparable results from Subject 2 not shown). Gene and species profiles recovered from the ultra-deep data include only direct matches to genes and genomes present in the database; de novo assembly of novel genes and contigs from deep data is expected to yield additional uncharacterized gene content and is not possible with shallow-shotgun data. (**C**, **D**) Scatterplots of species (**C**) and KOs (**D**) at 0.5 million versus 2.5 billion reads per sample (same sample size as (**A**) and (**B**) above).

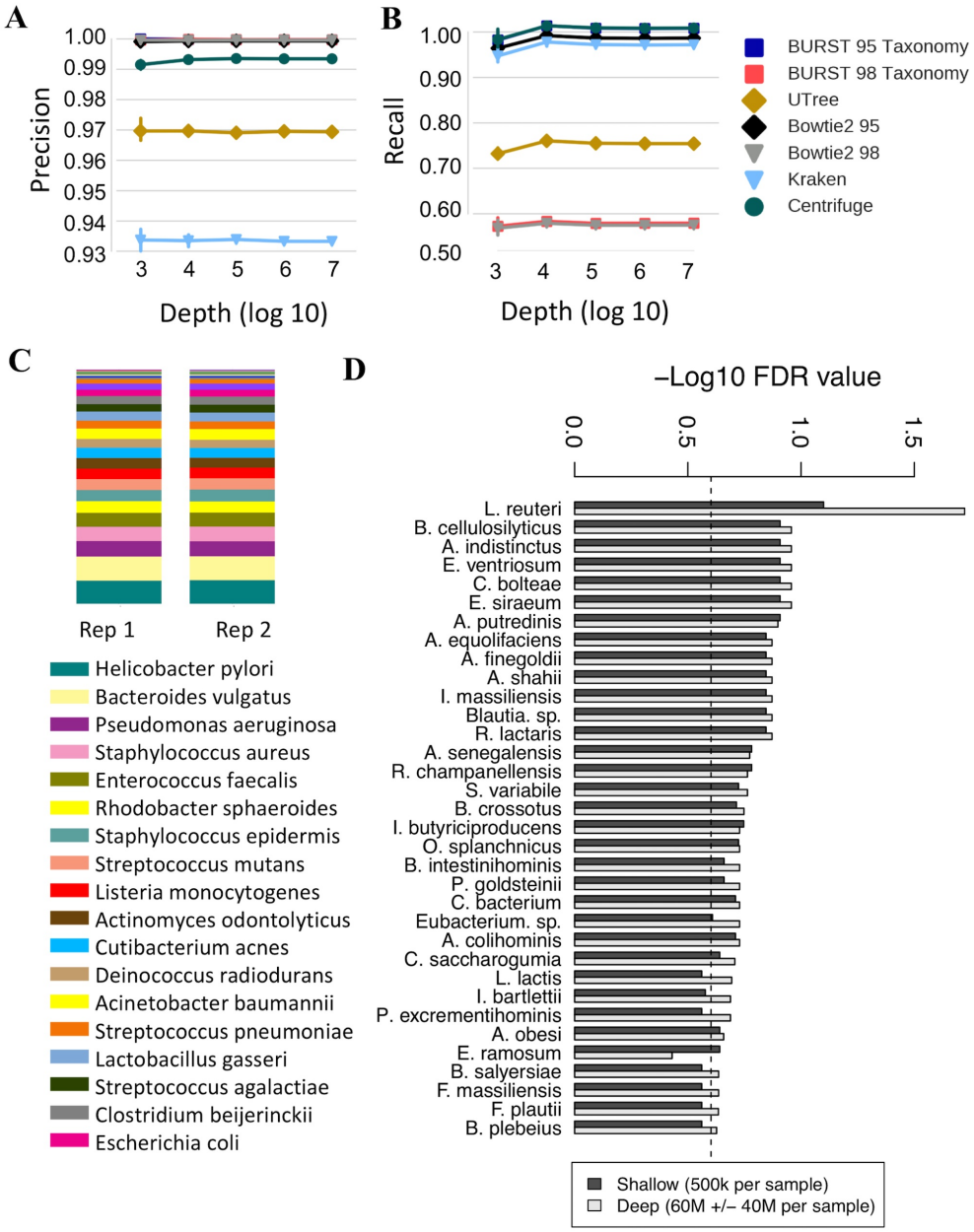

Figure 2.4: **Biomarker discovery using shallow-shotgun sequencing.** (**A**, **B**) Precision, recall for per-read species binning of different metagenomics analysis tools ("95" and "98" refer to the minimum alignment identity threshold used, n=5 distinct replicates per subsampling depth, error bars show standard deviation). (**C**) Stacked bar plot of species abundances recovered from HMP mock community shotgun sequencing data. (**D**) Negative log10 false-discovery-rate-corrected p-values using Mann-Whitney U tests for species associated with type 2 diabetes (17), compared between deep and shallow-shotgun sequencing ($n = 43$ healthy, $n = 53$ type 2 diabetes).

alignment approach, we found that at 20-40% of all sequences could be identified as species markers because they were uniquely present in only one species in the database figure 2.2.A. We then bootstrapped these samples repeatedly down to 10, 100, 1,000, 10,000, 100,000, and 1 million sequences per sample, and reran the analysis to quantify species-level alpha-diversity and beta-diversity profiles. In all cases, a depth of 0.5 million sequences was more than sufficient to recover the same alpha- and beta-diversity signals as with deep WGS figure 2.2.B-E.

2.3.2 Species and Functional Profiles

In order to compare the performance of shallow shotgun for species and functional profiling with ultra-deep sequencing, we obtained ultra-deep WGS sequencing of 2.5 billion sequences per sample on novel human stool samples from two individuals. At time of writing, to the best of our knowledge, this is the deepest sequencing performed on human stool microbiomes. Using these two novel samples, we measured the species profiles and functional profiles as described above using exhaustive gapped alignment against a full genome database for species and a database of genes annotated with KEGG Orthology groups (KOs) [45]. We performed this analysis at the full depth of 2.5 billion sequences, and then subsampled to lower depths. Species profiles at 0.5 million sequences per sample had an average correlation of 0.990 with ultra-deep WGS (Spearman correlation, $n = 112$, $p < 2.10 - 16$) across the two samples figure 2.3.A,C, and the average KO profile correlation was 0.971 (Spearman correlation, $n = 4,394$, $p < 2.10 - 16$) figure 2.3.B,D. For KO annotation we used direct gene observation for all but the lowest abundance genes, where we augmented the direct KO counts with counts of all KOs contained in observed reference strains in a similar manner to Piphillin [40]. We weighted the amount of augmentation according to the coefficient of

variation of a binomial distribution at the given observed proportion for a given gene with the augmented gene counts contributing at most 10% of the total counts for a given gene. In practice, this approach only affects the most rare genes with fewer than approximately 10 direct observations and offers a slight improvement in accuracy in Spearman correlation with virtually no change to Pearson correlation (see section 2.2). Our observed 97.1% Spearman correlation of the shallow and deep functional profiles is substantially higher than functional profiles predicted from 16S sequencing, which typically has 80-90% correlation with the directly observed functions [40, 39].

Follow the comparison of shallow-shotgun to ultra-deep sequencing of real biological samples, we also simulated deep WGS of complex metagenomes from a reference database to evaluate precision and recall of shotgun sequencing at different depths. Individual sequences were drawn at random from full reference genomes of selected species with a simulated 5% rate of sequencing error. Three different mixtures of species were selected from the database to match the average species-level composition of HMP samples from stool, oral, and skin body sites, respectively (see section 2.2)). Precision was defined as the fraction of simulated reads that were correctly assigned to their respective species divided by the total number of reads that mapped to the database. Recall was defined as the fraction of simulated reads that were correctly assigned to their respective species divided by the total number of simulated reads. We found similarly high precision rates of 0.985-0.995 when using exhaustive gapped alignment or Bowtie2 [20] at 95% or 98% alignment identity, or Centrifuge [24]; k-mer based methods including Kraken [23] and an in-house method for comparison [46] had lower precision figure 2.4.E. Recall was considerably higher when using 95% identity than 98% identity alignment with BURST or Bowtie2, likely due to the high error rate in the simulated

data figure 2.4.F. We also analyzed published shotgun data from the HMP mock community [44], recovering all expected species perfectly as the top 20 taxa, with the exception of *Bacillus cereus* which was recovered at the genus level due to highly overlapping species genomes in the genus *Bacillus* [53] figure 2.4.C.

2.3.3 Species-Level Biomarker Discovery

Finally, to assess the ability of shallow-shotgun sequencing to identify species-level biomarkers in a clinical study, we subsampled deep-shotgun sequencing data from a study of healthy individuals and individuals with type 2 diabetes (T2D) [54] to 0.5 million sequences per sample. We identified the species significantly associated with T2D in both the deep data and the shallow data using two-sided Mann-Whitney U tests, and found high concordance between the p-values for species down to 0.0005 relative abundance (average Spearman $\rho = 0.954$ across 10 subsampled replicates, $n = 94$, $p < 2.10{-}16$), indicating that 0.5 million sequences per sample enables discovery of species-level biomarkers with comparable power to deep-shotgun sequencing down to approximately 0.0005 relative abundance figure 2.4.D. Notably, this classification task contained a range of statistical signals ranging from very strong to marginally significant.

2.3.4 Comparison to 16S species profiles

As noted, 16S variable-region amplicon sequences often do not resolve taxa below the genus or family level, although some species can be identified [38]. To compare the overall concordance between 16S species profiles and shallow-shotgun sequenicng species profiles for pairs from the same sample, we calculated the Pearson's correlation R-squared value (coefficient of determination) of the 16S

and shallow-shotgun sequencing species profiles in each pair and found the average R-squared was 0.918. We then permuted the pairing of the 16S and shallow-shotgun sequencing profiles and repeated the average R-squared calculation to obtain a null distribution, showing that the R-squared between the true pairs of samples was better in all cases than the randomly assigned pairs (Monte Carlo permutation test $p < 0.001$) (figure 2.5.A). This demonstrated high overall concordance between 16S and shallow-shotgun sequencing species profiles relative to inter-subject differences. To compare the contributions to total relative abundance of observed species between 16S and shallow-shotgun profiles, we merged the species-level taxonomic profiles for paired 16S and shallow-shotgun analyses and measured the fraction of species attributed to 16S only, shallow shotgun only, or both. We found that there were many species only observed in the shallow-shotgun data, with some observed at high levels of abundance (figure 2.5.B,C), indicating that the 16S sequencing identified a subset of the dominant taxa at the species level.

2.4 Discussion

In this work we evaluated the information content of shallow-shotgun sequencing as a potential alternative to 16S in certain situations. We found that surprisingly few shotgun metagenomic sequences are needed to obtain reliable species and gene group profiles at approximately the same cost as 16S sequencing. We also compared shallow shotgun to deep shotgun on a number of biological data sets including samples from the HMP, a published deep-shotgun sequencing diabetes study, and simulated and mock communities, and found that we could recover similar trends in alpha and beta diversity, species profiles, and species biomarker

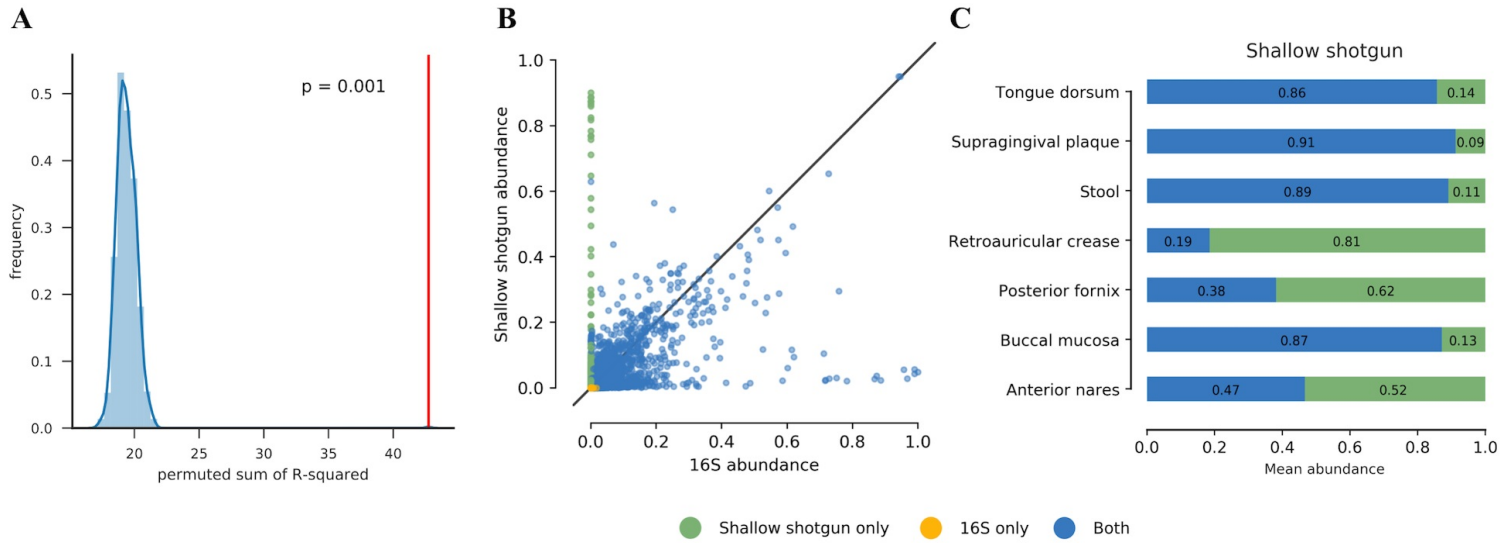

Figure 2.5: **Comparison of 16S and shallow-shotgun recovery of species-level taxa.** (**A**) Distribution of average Pearson correlation (R-squared) of species profiles between 16S and shallow-shotgun sequencing from the same HMP sample ($R^2 = 0.918$), compared to the permutation-based null distribution of R-squared values for random pairings ($p < 0.001$). (**B**) Scatterplot of relative abundance of species in shallow-shotgun sequencing versus 16S sequencing from the same HMP samples; species found only in one data type shown in different color. (**C**) Fraction of all observed species relative abundance accounted for by species found in 16S only, shallow shotgun only, or both.

discovery down to 0.05% relative abundance with as few as 0.5 million sequences per sample. We then analyzed two human stool samples with new ultra-deep-shotgun sequencing data at 2.5 billion reads per sample, the deepest sequencing coverage of any microbiome to our knowledge. We found that shallow sequencing recovers 97-99% correlated species and KEGG [45] Orthology group (KO) profiles when compared to the ultra-deep data.

We did not attempt to perform an exhaustive comparison of different sequence annotation tools as that was outside the scope of our investigation. Instead, we selected several tools representing different approaches to database search for comparison, including exhaustive semi-global gapped alignment [47, 48], heuristic gapped alignment using the Burrows-Wheeler transform [20], and k-mer-based search [23, 46]. Using simulated metagenomic data we found that tools using gapped alignment obtained higher precision and recall than tools using k-mer-based mapping. This result was expected as k-mer mapping requires exact matches of fixed-size k-mers whereas gapped alignment allows insertion of gaps at random to maximize overall sequence identity. In our simulated data, fully exhaustive end-to-end gapped alignment with a minimum threshold of 95% identity using an accelerated version of Needleman-Wunsch [47, 48] performed best in terms of recall. Several methods were approximately tied for highest precision. A potential advantage of gapped alignment over k-mer mapping is that current tools report the genomic coordinates of each match, allowing estimation of strain-level coverage, which may be useful for future work into novel algorithms that use strain-level coverage to further improve precision and recall for rare species.

We note a number of important limitations to shallow-shotgun sequencing (table 2.1). Shallow-shotgun sequencing may not be a viable replacement for 16S when characterizing blood or biopsy microbiomes, where there is likely to be more

host DNA contamination and relatively low bacterial biomass. Shallow shotgun also relies on whole-genome reference databases, and thus will require expansion of reference genomes to cover novel environments. When analyzing poorly characterized environments, researchers may consider combining 16S sequencing for identification of novel taxonomic groups with shallow shotgun for functional profiling. We have not attempted to compare deep or shallow-shotgun sequencing with 16S sequencing in environments with low representation of strains in the reference database, such as marine or soil samples. In these cases it is likely that shallow shotgun will still reveal useful functional profiles due to homology of some observed sequences to known genes and species, but we expect that 16S would provide superior profiling of novel taxa due to lack of available representative genomes covering endemic species, as has been observed for fresh water samples [55].

Advantages	**Disadvantages**
• Recovers similar alpha diversity, beta diversity, species biomarkers to deep shotgun • Similar cost to 16S • Sequencing libraries can be reused for deep shotgun	• Limited by database coverage • Cannot track strain variation/SNPs • Cannot assemble new genes/genomes • Limited recall and precision for species below 0.0005 abundance

Table 2.1: **Table that outlines the advantages and disadvantages of shallow shotgun sequencing.**

Shallow shotgun is not meant to be a replacement for deep WGS for strain-level resolution or tracking polymorphisms in strains, and cannot be used for novel

gene and genome assembly. For many of the metrics we examined, a depth of 0.5 million sequences per sample was sufficient, but deeper sequencing is warranted for detection of rare species below approximately 0.0005 relative abundance. For this reason, we recommend depths of one or two million per sample for increased sensitivity when possible. In addition, a general concern with any taxonomic annotation is that the boundaries of traditional species taxonomic labels do not necessarily reflect consistent entities at the genomic level when accounting for horizontal gene transfer and inaccurate annotations. These concerns can be alleviated to some extent using deep-shotgun sequencing and metagenomic assembly [42], co-abundance clustering [43], or proximity-based assembly [56], although de novo assembly of strains from complex microbiomes remains an active area of research.

We found that shallow sequencing of human stool microbiomes provides high-quality species and functional profiles of human microbiome samples, for little more than the cost of 16S amplicon sequencing when using a miniaturized library preparation protocol (see section 2.2). We have made available the gene and genome databases that we used together with a convenient Python-based wrapper script that allows users to compare several existing tools for performing both taxonomic and functional annotation (see section 2.2). Shallow shotgun has a number of important limitations and is not intended to replace deep whole-genome shotgun sequencing for strain-level analysis or novel gene and genome assembly. Nonetheless, shallow-shotgun analysis provides considerably more accurate functional profiles and more precise taxonomic resolution than 16S amplicon sequencing for human microbiome studies. Thus, shallow-shotgun sequencing is a viable alternative to 16S for researchers performing large-scale human microbiome studies where deep WGS may not be possible.

Chapter 3

SHOGUN: Modular, Accurate, and Scalable Framework for Microbiome Quantification[1]

3.1 Introduction

The rapidly decreasing cost of next-generation sequencing technology has led to a massive increase in the amount and rate of metagenomic data production, creating the potential to discover causal roles of microbes in many complex ecosystems. Currently microbiome DNA analysis is typically performed either using amplicon-based (16S) or whole-genome shotgun-based (WGS) sequencing methods. Techniques based on amplicon sequencing typically amplify a highly variable region of the 16S ribosomal RNA gene, although other genes may be targeted in particular cases. Amplicon sequencing is affordable but often cannot distinguish between species due to sequence similarities and does not allow high-accuracy prediction

[1] A version of this chapter has been published [50]

of the functional repertoire [57]. Amplicon methods rely on DNA primers to amplify the region of interest, and therefore are subject to high levels of bias and can fail to capture organisms whose DNA sequence does not match the primers. As an alternative, WGS sequencing randomly selects fragments of all DNA present in a metagenomic community. WGS sequencing directly measures the functional repertoire of the microbiome by capturing a snapshot of the total metagenomic content and allows for strain-level characterization of microbiomes by mapping reads to the unique markers of strain reference genomes. WGS is typically much more expensive than amplicon sequencing due to its use of costly library preparation protocols and the additional costs of deep sequencing.

3.1.1 Motivation

Overcoming the shortcomings of amplicon and shotgun techniques presents challenges in the field of metagenomics and prevents scientists from performing highly precise, large-scale studies. The critical barriers to quantifying microbiomes at the species level are affordability, accuracy, and reproducibility. After microbial communities have been sequenced, it is the objective of the researcher to utilize informatics methods to correlate the taxonomic abundance profiles of a sample to a trait of interest. However, these methods operate under the assumption that the underlying taxonomic profiles are accurate. If methods are developed to more accurately identify the profiles of a community, the increased precision will cascade down every step along the metagenomics pipeline. With more precise profiles, the informatics methods will have more power to test hypotheses and better ability detect the causal role these communities play. Accurate cost- and time-effective taxonomic quantification of environmental samples is essential. Often, the quantification and analysis of metagenomic data is carried out using custom, in-house

workflows, leading to redundant implementations of software and an inability to reproduce results across labs and studies [58]. To address these challenges, we propose the SHOGUN [59] pipeline, which assembles current best practices in the field into a single, easy to use, and flexible framework, to carry out taxonomic and gene abundance profiling of metagenomic WGS datasets.

3.2 Methods

A typical protocol for going from WGS sequences of a mixed metagenomic community to a taxonomic profile of abundances can vary widely depending on the computational taxonomic profile tool used. The focus of this research and the SHOGUN tool is to utilize the maximum amount of information given in WGS sequences to produce accurate taxonomic abundance profiles. We chose to evaluate tools against the taxonomic profiler SHOGUN [59] pipeline based on the following: ease of use and documentation, availability as an open-source tool, the ability to create a user defined database, the ability to summarize a taxonomic abundance profile at the species level, and the ability to scale to large datasets with multiple threads per process. As a result of these requirements, the tools Centrifuge [24], a longest common substring taxonomic profiler, and Kraken [23], an exact k-mer taxonomic profiler, were selected to evaluate alongside the SHOGUN pipeline.

3.2.1 SHOGUN Pipeline for Metagenomic Relative Abundance Estimation

SHOGUN is a command line interface (CLI) that can be installed with a single command, is open source and freely available, and is designed to be well documented

and easy to use. The SHOGUN pipeline for metagenomic taxonomic and gene abundance profiling is described in figure 3.1. The CLI was designed using a modular subcommand interface, so SHOGUN can be run in its entirety or each command can be run individually. All steps within the pipeline use the same database folder for consistency. The Python-language codebase is unit tested and every version of the pipeline receives a unique hash from GitHub that can be reconstituted for complete analysis reproducibility.

3.2.2 Taxonomic Abundance Profiling

The taxonomic profiling algorithm within SHOGUN is partitioned into three steps: sequence alignment, sequence taxonomic assignment, and rank-specific relative abundance estimation.

Sequence Alignment

There are three distinct alignment algorithms implemented for use within SHOGUN: Bowtie [20], a burrows-wheeler alignment algorithm, BURST [48], an optimal, exhaustive Needleman-Wunsch alignment algorithm, and UTree [46], a k-mer based alignment algorithm. Depending on the tool selected for sequence alignment, the downstream taxonomic assignment and rank-specific relative abundance estimation are automatically tuned so that the pipeline outputs comparable results for all of the algorithms. Each of the three different alignment modes were evaluated in terms of their accuracy and computational resources utilized.

Taxonomic Assignment

Taxonomic assignment is the process of assigning a taxonomy to a query sequence given the set of valid reference genome matches and match identities in the

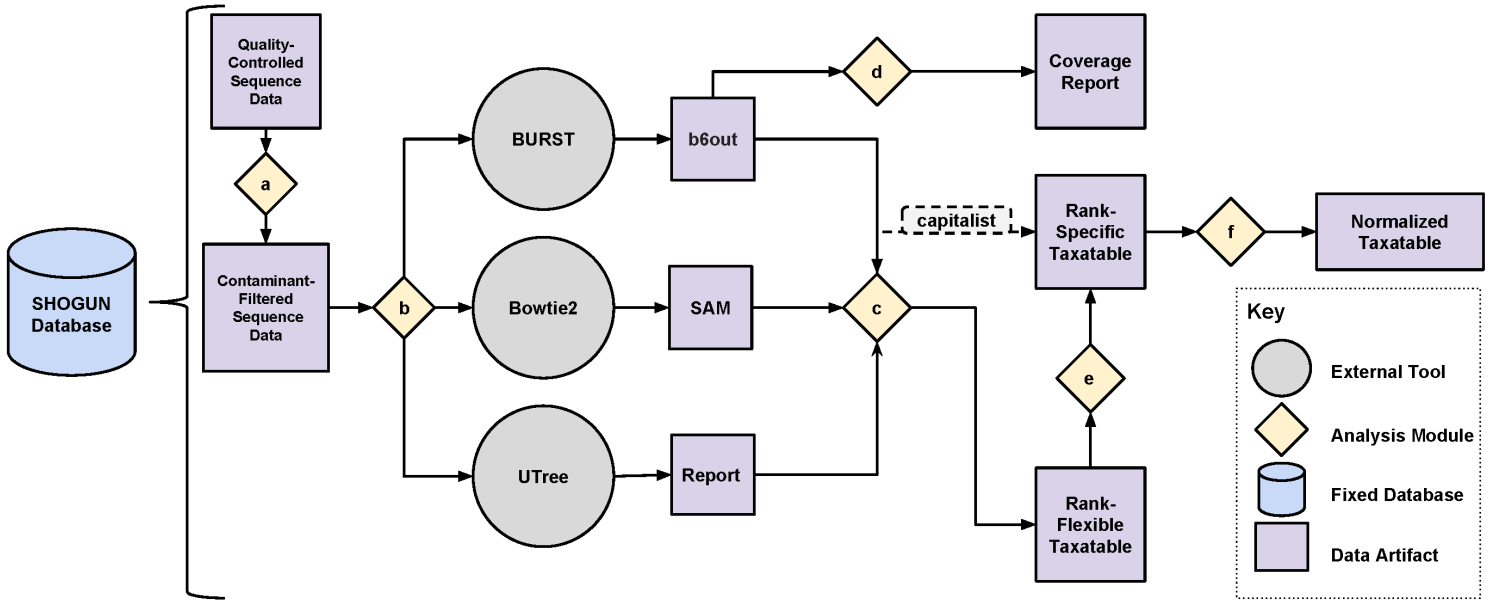

Figure 3.1: **Schematic overview of the computational pipeline** `SHOGUN`. For every step in the `SHOGUN` pipeline, the user must supply the pre-formatted `SHOGUN` database folder. To run every step shown here in a single command, the user can select the pipeline subcommand. Otherwise, the analysis modules can be run independently. (**A**) `filter` - The input quality-controlled reads are aligned against the contamination database using BURST to filter out all reads that hit human associated genome content. (**B**) `align` - The contamination-free reads are aligned against the reference genome database. The user has the option to select one or all of the three alignment tools BURST, Bowtie2, or UTree. (**C**) `assign-taxonomy` - Given the data artifacts from a `SHOGUN` alignment tool, output a Biological Observation Matrix (BIOM) format profile with the rows being rank-flexible taxonomies, the columns are samples, and the entries are counts for each given taxonomy per sample. The alignment tool BURST has two run modes, taxonomy and capitalist. If the capitalist mode is enabled, a rank-specific BIOM file is output instead. (**D**) `coverage` - The output from BURST can be utilized to analyze the genome coverage of each taxonomy across all samples in your alignment file. This can useful for reducing the number of false positive alignments by removing taxonomies below a minimum coverage score. (**E**) `redistribute` - The rank-flexible profile is summarized into a rank-specific profile. This summarizes both up and down the taxonomic tree. (**F**) `normalize` - Each sample in the profile is normalized to the median depth of all the samples for count based analysis tools that use BIOM tables.

database. Query sequences often match to multiple reference genomes equally well due to shared genomic regions across reference genomes in the database. When a query sequence matches multiple genomes, `SHOGUN` uses a confidence-weighted, last common ancestor (LCA) algorithm, referred to as *taxonomy* mode for the BURST, UTree and Bowtie2 aligners. The *taxonomy* mode read-disambiguation scheme results in a rank-flexible taxonomic profile, where the profile contains a mix of taxonomic levels; that is, some queries are assigned at the kingdom level, some at the phylum level, and so forth down to the highest resolution possible in the taxonomic tree. If all annotations are at the same level, preferably at a high taxonomic resolution such as the species level, the taxonomic profile is known as rank-specific. The BURST aligner can be run in the `SHOGUN` *taxonomy* mode, but it also comes packaged in `SHOGUN` with its own read-disambiguation scheme using a min-cut algorithm referred to as *capitalist* that returns rank-specific relative profiling.

BURST Optimal Sequence Alignment

The sequence alignment tool BURST is an exhaustive optimal aligner based on the Needleman-Wunsch algorithm [47]. By leveraging dynamic programming, BURST is able to rapidly identify and compute all tied best alignments above a specified percent identification cutoff. There are two primary taxonomic assignment modes utilized in `SHOGUN` from the BURST alignment tool, those being *capitalist* and *taxonomy* mode.

BURST's Rank-Flexible *taxonomy* Taxonomic Assignment

The *taxonomy* mode objective is to identify the most likely taxonomic annotation for a given query sequence. Taxonomy mode assigns the most likely taxonomic

annotation given a set of valid alignments for a query sequence using a confidence based last common ancestor algorithm. That is, for each query sequence, the set of taxonomic annotations from valid alignments above the user specified threshold is enumerated. The first voting round begins at the highest rank, typically at the kingdom level, where each of the alignments for a given query vote for their respective taxonomic rank. If a plurality of the unique taxonomic votes is determined to be above the user specified threshold (the default threshold being 70%), all children of that taxonomic rank are used in the next round of voting. Voting rounds continue until a taxonomic lineage leaf is reached or the plurality supporting vote does not exceed the threshold. If a round of voting does not exceed the specified support threshold, the interpolation stops and the last round's taxonomic lineage vote is returned for the query.

BURST's Rank-Specific *capitalist* Taxonomic Assignment

The *capitalist* algorithm enumerates all tied best hits above a user specified identity cutoff. The goal of *capitalist* algorithm is to return the minimal set of references that best explain all queries. The capitalist algorithm achieves this goal by translating the problem to finding the minimum set of query nodes in a bipartite graph such that each query sequence is connected to a single reference. This solution is very similar in nature to that of the minimum cut problem in prior computer science research [60]. In the bipartite graph, one partition of nodes is query sequences and the other partition of nodes is reference genomes. The edges in the graph are the valid best hits returned by the BURST alignment algorithm above the identify cutoff. The minimum set of references is found by greedily pruning edges until each query sequence retains only a single edge and therefore a single valid alignment. The order of the edges for pruning is set by a priority queue. The

edge priority in the queue is sorted in descending order of highest degree reference nodes. Every edge is placed into the priority queue and enumerated. For every edge in the priority queue, the edge is pruned if the degree of the connected query sequence is greater than one, translating to the query sequence and maintaining at least one valid alignment to a reference genome node. If an edge is pruned, the priorities of each other edge of the respective query sequence is updated in the priority queue with the new query degree. The pruning algorithm terminates once each query sequence node is degree one. In the final report, the edge and corresponding alignment reference node for each query sequence is reported.

Bowtie2 Burrows-Wheeler Alignment and Rank-Flexible Taxonomic Assignment

The alignment tool Bowtie2 [20] is a fast read alignment tool utilizing the Burrows-Wheeler transformation. The `bowtie2` command was modified from its default settings so that matches to ambiguous bases are penalized, sequential gaps are weighted differently than single gaps, and up to thirty-two valid alignments above the specified percent identification are reported. To save space, the header and unaligned sequences are set to be suppressed [61]. We do not recommend changing these settings to retain compatibility for downstream `SHOGUN` analysis. In order to assign a taxonomy to the query sequence with up to thirty-two alignments, the confidence-weighted, LCA of all the taxonomic lineages of each query sequences' set of alignments is reported in the same manner as BURST *taxonomy* mode.

UTree's *k*-mer Based Taxonomic Classification Scheme

The tool UTree maps sequencing reads to reference genomes using a k-mer indexing scheme similar to that of CLARK [62] and Kraken [23]. The purpose of

UTree is to be light on computational resources while still being able to search query sequences up to 16 megabase pairs in length against full sized reference genomes. It uses an efficient, unique, k-mer indexing scheme of reference taxonomic ranks through the use of a prefix-forest and a suffix binary tree. The database is built by stepping through each input reference sequence and determining whether each k letter window, known as a k-mer, is unique to that references taxonomic rank. When a k-mer is selected from the reference genome to be inserted into the database, it can be flagged as either unique or ambiguous to a specific taxonomic rank. In the case that the k-mer is unique to the lowest taxonomic rank, the k-mer is added to the suffix binary tree at the index retrieved from the prefix-forest with the associated taxonomic metadata. In the case where the k-mer is not unique, the existing k-mer's taxonomic rank within the database is demoted to the common ancestor of the current reference and the existing taxonomy. If the common ancestor is the root, it is excluded from the database when searching.

When searching a query sequence, all unambiguous k-mers in the query are searched for in the database. All k-mers that are found in the database are used to interpolate the taxonomic identity of the query in voting rounds in last-common ancestor algorithm similar to that of *taxonomy* but modified to suite k-mer taxonomic classification. The first voting round starts at the highest taxonomic rank; in this case it would be kingdom. A vote is placed for a taxonomic lineage for each uniquely identified k-mer from the query sequence. If a plurality of the unique k-mer votes is determined to be above the user specified threshold (the default threshold being 70%), all children of that taxonomic rank are used in the next round of voting. Voting rounds continue until a taxonomic lineage leaf is reached

or the plurality supporting vote does not exceed the threshold. If a round of voting does not exceed the specified support threshold, the interpolation stops and the last rounds taxonomic lineage vote is returned for the query.

3.2.3 Rank-Specific Relative Abundance Estimation

Converting higher level taxonomic annotations, such as the rank "family", to a lower desired rank annotation, for this example the rank "species", can produce erroneous results if one does not account for differences in genetic diversity across different levels of the taxonomic tree. The ideal case is that all the classifiers in `SHOGUN` report a rank-specific taxonomic profile at the species level so that the resulting profiles can be compared fairly. However, some of the alignment tools, such as Bowtie2 and BURST *taxonomy*, and UTree k-mer classification produce profiles that are rank-flexible and require summarizing to a specific level. In order to redistribute a rank-flexible profile to a rank-specific profile, `SHOGUN` implements the principle of empirical Bayesian redistribution of reads in a similar fashion to the Bracken [63] algorithm. The algorithm redistributes annotations at lower-resolution levels of the taxonomic tree to a specific higher-resolution according to each taxonomies uniqueness, number of assignments in the profile, and the median reference genome length.

The major difference between the original Bracken algorithm and the redistribution algorithm for `SHOGUN` is how the uniqueness probabilities U are calculated for each respective taxonomy i in the database. In Bracken, each genome in the database has a length in base-pairs L. Each genome is then sheared with a sliding window of size r and step size of $s = 1$ creating $(L_i + r + s)$ taxonomic annotations for each genome classified by Kraken [23]. In `SHOGUN`, we exchange the classification tool Kraken with the optimal alignment tool BURST. We set

BURST to recover all tied-best alignments above a 98% alignment identification, set the window size $r = 100$, and increased the step size $s = 50$ for the shearing of the genomes. When capitalist resolves its taxonomic assignments, we assigned the taxonomic annotation to be the last common ancestor of the genome that the sheared read was taken from and the genome that was classified. This holds the advantages that BURST can recover the locations of genome redundancy, obtain higher precision in classification of reads, and allow the relative abundance re-estimation to be provided through the tools used within `SHOGUN` to remain consistent.

3.2.4 Gene Abundance Profiling

Gene abundance profiles can be obtained two ways, either by prediction from the taxonomic profile or through direct metagenomic observation of annotated genes. Gene prediction profiles are obtained as designed by the PICRUSt [39] algorithm. For gene prediction, a mapping file is required that specifies the genes present and their respective copy-number for each genome in the database. The gene prediction algorithm also requires a normalized, rank-specific profile obtained through the `SHOGUN` taxonomic abundance pipeline described previously. The gene mapping is summarized to the same level as the taxonomic profile using the median number of each gene in the taxonomic clade. The genes mapping file is multiplied by the taxonomic table to create a gene abundance profile and normalized to sum to one as a relative abundance profile. The direct observation of genes pipeline requires a `SHOGUN` database of genes and a gene annotation mapping file. The `SHOGUN` taxonomic profiling pipeline is then used to map metagenomic reads to annotated genes in the database.

3.2.5 SHOGUN Database

To validate the performance of the pipeline, we selected representative genomes from bacteria, archaea, and viruses from the publicly available RefSeq nucleotide database version number 82 (Rep82) [49]. Explanations of the microbes present in this reference genome database are shown in table 3.1. We identified genes using UniProt [64] annotations obtained by running Prokka [65] on all the bacterial genomes and mapping them to Kyoto Encyclopedia of Genes and Genomes (KEGG) [45] annotations. The contaminate database was the human genome assembly obtained from the Genome Reference Consortium Human Build 38 (GRCh38) with no alternate scaffolds [66].

Kingdom	*Number of Genomes*	*Megabase pairs (Mbp)*
Archaea	238	627,101.26
Bacteria	4,884	19,308,087.26
Plasmid	614	198,476.94
Viroids	46	15.50
Viruses	7,194	253,668.36
Total	**12,976**	**20,387,349.32**

Table 3.1: **The number of strains and megabase pairs for each kingdom for all representative archaea, bacteria, and virus sequences in RefSeq version 82 (Rep82).** Each entry in the database is assigned a unique taxonomic string identifier at each level in the taxonomic tree down to strain. Each strain is a given a unique entry so that BURST capitalist properly disambiguates reads that hit multiple strains for taxonomic profiling and coverage analysis.

3.2.6 Simulated Human Microbiome Data

To test each internal alignment engine's accuracy of relative abundance estimation of metagenomic communities, we created a simulated community with known

species level taxonomy. The data were simulated according to abundances obtained from Human Microbiome Project using the top 100 most abundant species from each general body habitat according to the original study's results [67]. Reads were simulated from a strain of those species according to the average proportion of that taxonomy in their respective group using the tool dwgsim [52]. The reads were simulated with default settings for Illumina single-end sequencing machines with a 5% mutation rate where 2% of mutations were deletions and a maximum amount of ten ambiguous bases per query sequence. The two tools outside of the `SHOGUN` framework utilized were Kraken [23] and Centrifuge [24] and for a more complete benchmark of accuracy please compare the relative accuracies to the [68] evaluation. The alignment methods were evaluated using F1-scores, the average of precision and recall, of the known species assignment versus the identified species on a per query basis.

To validate the profilers on a community level, we rarefied each of the communities at various depths. This was done because the taxonomic profiles with a redistributed rank-specific profile can result in different profiles at different depths depending on the number of reads assigned to each taxonomy. To compare the profiled and simulated community, we used the Jaccard similarity and the Spearman correlation metrics. Each of the human oral, skin, and stool microbiomes were re-sampled with replacement ten times at each power of ten, starting at ten-thousand and ending at ten-million queries.

Simulated Timing Data

Each of the taxonomic assignment methods were also evaluated for speed and memory usage using the query sequences from the Kraken timing dataset and the

Rep82 reference database. The dataset contains a total of 10 million reads simulated for three different query lengths n to reflect current sequencing technology (n = 50, 100, 150).

Software Availability

The following benchmarked tools were installed from the Anaconda channel "knightslab" using versions `SHOGUN`=1.0.5, Utree=2.0rf, and BURST=0.99.7f. The rest of the tools were installed from the Anaconda channel "bioconda" with versions Bowtie2==2.3.4.1-0, Kraken=1.1-1, Centrifuge=1.0.3-2, and `dwgsim`=1.1.11-5.

3.3 Results

The purpose of the benchmark data was to measure the performance of the `SHOGUN` pipeline WGS data in comparison to other similar tools. The simulated timing data were run with each taxonomic profiler: `SHOGUN` BURST, `SHOGUN` Bowtie2, `SHOGUN` UTree, Centrifuge, and Kraken. Each classifier were evaluated for its precision and recall on a per read basis using the simulated human microbiomes, then averaged into its F1 scores as shown in figure 3.2. The alignment classifiers `SHOGUN` Bowtie2 and Burst were tuned to a percent identification threshold of $p = 0.95, 0.98$. The trade-off for lowering the percent identification threshold is increased alignment time and recall but decreased precision. However, in the simulated data, a percent id of $p = .95$ greatly increased each aligner's F1 score (`SHOGUN` BURST $p = 0.95, F1 = 0.969$, `SHOGUN` BURST $p = .98, F1 = 0.746$, `SHOGUN` Bowtie2 $p = 0.95, F1 = .938$, and `SHOGUN` Bowtie2 $p = 0.9, F1 = 0.739$).

The results of the community validation analysis in terms of Jaccard similarity and Spearman correlation are shown are shown in figure 3.3. Both Jaccard

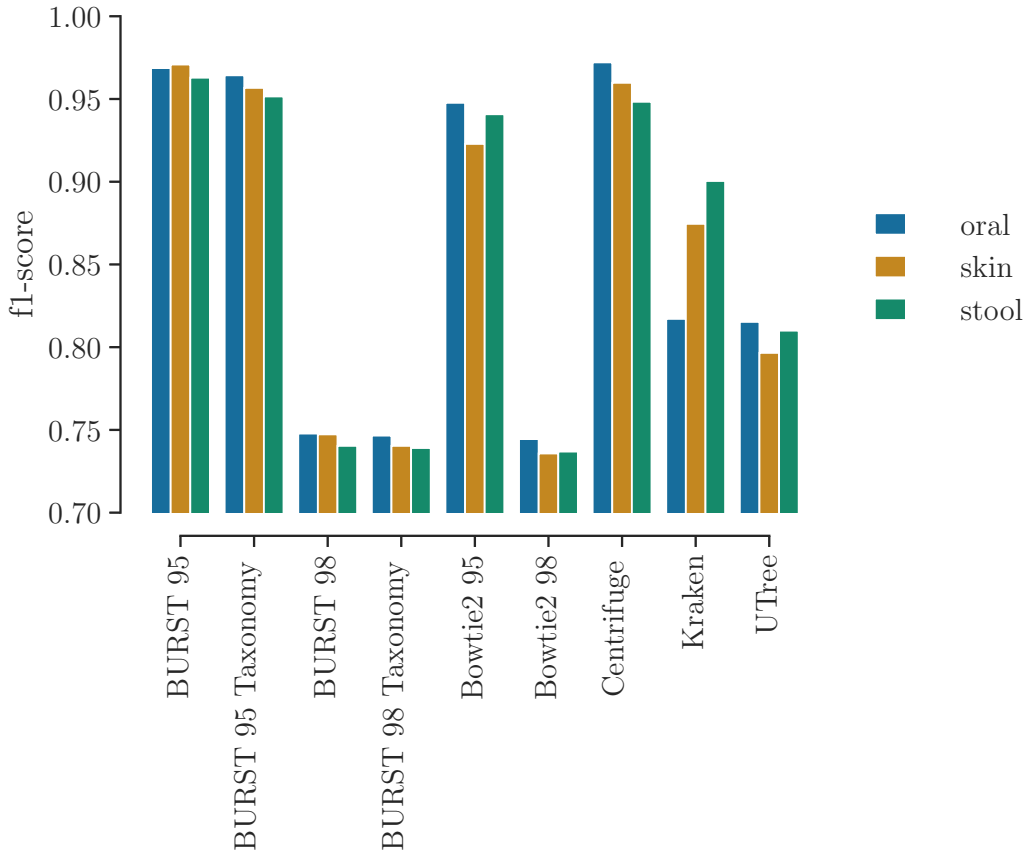

Figure 3.2: **The micro-averaged F1-score of each of the aligners on per read bases on the simulated stool, oral, and skin communities.** For the alignment methods, two different thresholds at 95% and 98% for alignment identification were set to account for recall bias in highly-divergent reads.

similarity and Spearman correlation were saturated between 10 thousand and 100 thousand taxonomy alignments per sample. Most of the profiling tools performed similar, however, the Kraken tool performed worse without the empirical redistribution of reads of the Bracken tool.

Each of the runtime and memory requirements for each profiler were evaluated. The memory usage of each profiler is shown in figure 3.4.A. The reads per minute with increasing number of threads per process is shown in figure 3.4.B.

3.4 Discussion

The results of the simulated communities showed that overall, with some fine tuning, each of the taxonomic profiling tools were able to properly recover the correct relative abundance communities. In terms of computational resources, the alignment free tool UTree was the fastest in terms of reads per minute and used significantly less memory than the other tools. In terms of accuracy, a properly tuned `SHOGUN` BURST pipeline had the highest overall accuracy. A summary of the findings for each of the tools is displayed in table 3.2.

3.4.1 Future work

Herein we proposed data-efficient methods for species-level resolution taxonomic profiles in shotgun metagenomic datasets. The objective of taxonomic profiling tools for analyzing shotgun data is to be as data efficient as possible of the query sequences. Below we outline some areas of exploration that have the potential to provide key discoveries in increasing the reproducibility and accuracy of WGS surveys with shotgun data.

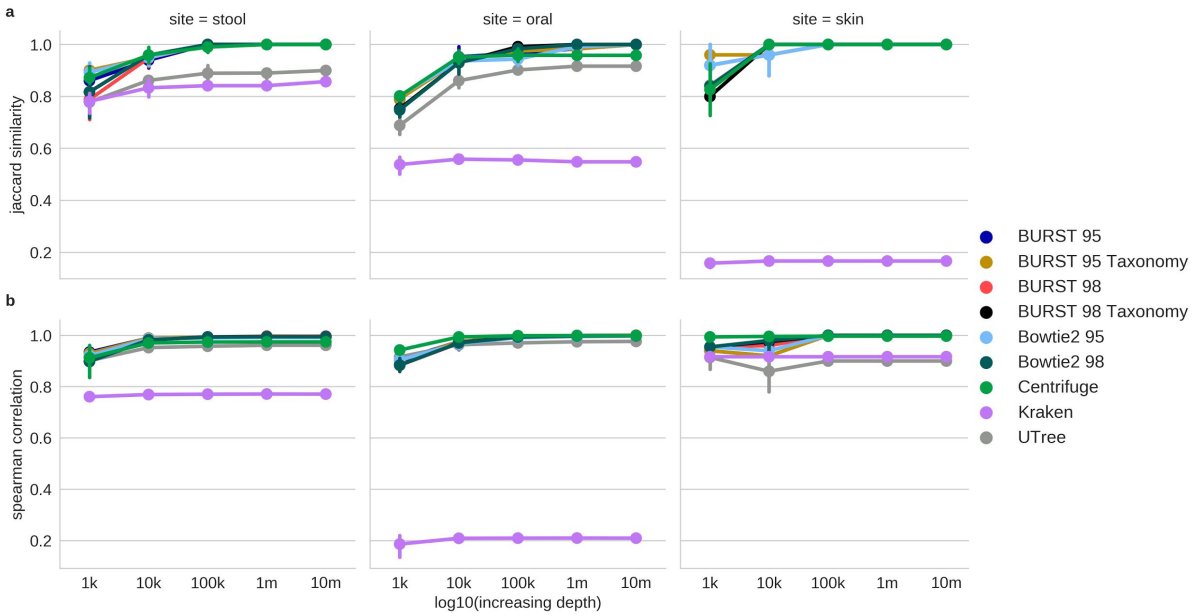

Figure 3.3: **Taxonomic profiling tools were validated on a community-level bases.** (**A**) Rarefaction curves of Jaccard similarity of known species and predicted species for each of the alignment tools on the simulated stool, oral, and skin communities. For most tools, all species are identified between ten-thousand and a hundred thousand reads. (**B**) The rarefaction curves of Spearman correlation of the tools predicted community with the known community for the simulated dataset. This shows that not only are the correct species identified, but they are also in the correct abundances.

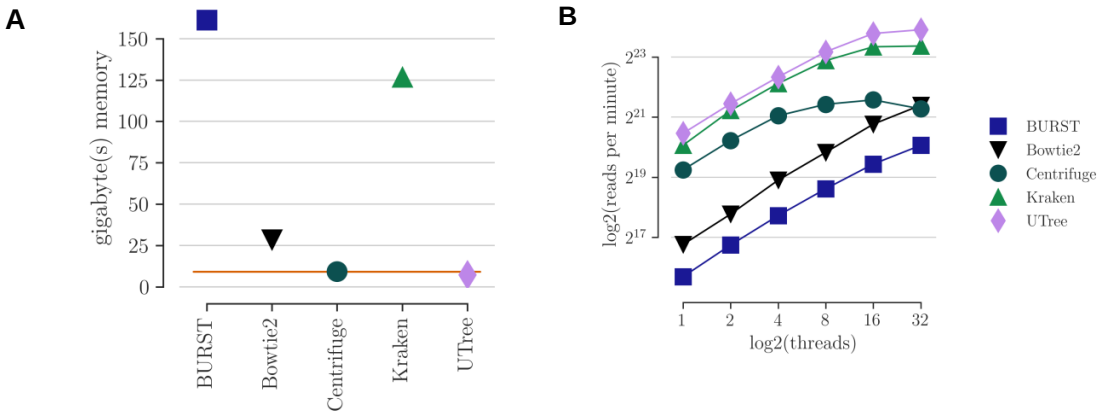

Figure 3.4: **The results of the speed and memory profiling showing that optimal alignment is within an order of magnitude of other sequence classifiers and scales efficiently across multiple cores.** (**A**) The maximum resident size set (RSS) in Gigabytes of each of the aligners with the Kraken timing dataset. The horizontal red line depicts the size of the original Rep82 database. (**B**) The scaling and efficiency in reads per minute of each of the aligners across many threads per process. Tools were selected based on an open-source codebase, ability to make a custom reference database, and an output file containing mappings per sequence. The two tools outside of the `SHOGUN` framework utilized were Kraken [23] and Centrifuge [24].

Approach	*Examples*	*Pros*	*Cons*
Last Common Ancestor Alignment	Bowtie2 *taxonomy* BURST *taxonomy*	Highest Precision	High RAM Slow No coverage analysis
Exhaustive Alignment	BURST Capitalist	Highest F1 Accuracy Highest Recall	High RAM Slow
Marker Gene[†]	Metaphlan	Low RAM Fast	Sparse Database
Exact *k*-mers		Fastest	No coverage analysis Sensitive to indels
	Kraken		High RAM Large Database
	Utree	Lowest RAM Smallest Database	
Longest Common Substring	Centrifuge	Fast Low RAM	No coverage analysis Sensitive to indels

[†] Marker gene taxonomic profilers were not evaluated in this study due to complexity of creating a database and low alignment rates.

Table 3.2: **Summarization of the tools tested for metagenomic profiling.** While this list is not comprehensive, we recommend using any tools in the approaches listed as legitimate options for analysis of shotgun data. From the tools surveyed, we found that BURST capitalist returns the most accurate results if properly tuned at the cost of high RAM usage and compute time. If computer resources is a concern, we recommend using either UTree on a computer with over 8GB RAM or Centrifuge on a computer with over 16GB RAM.

Percent identification tuning

The alignment algorithms only return valid hits above the percent identification specified by the user. Lowering the percent identification threshold will boost recall but result in higher running time and lower precision. Furthermore, the percent identification threshold is highly dependent on the reference database used and the metagenomic community sequenced. Exploring different percent identifications for different environments and databases is recommended for optimal taxonomic profiles in shotgun sequencing.

Reference database

The reference database that was used was a full genome reference database with manually annotated taxonomic distances. These genomes also do not currently contain the full set of annotated coding regions, rendering it complicated to predict the functional repertoire of genes. Furthermore, the taxonomic distance of genomes in the database is not evenly spaced, leading to large regions of taxonomic lineages that are interdependent. Exploration of different reference genome databases and their pros and cons for shallow-shotgun sequencing could lead to higher alignment rates and increased accuracy. Furthermore, annotating reference genomes in a database for their known functional genes present will allow `SHOGUN` to predict functional genes present in a metagenomic community.

Coverage per microbe analysis

Often times, false positive microbes are identified at very low abundance when using taxonomic profiling tools. These microbes are usually filtered out through prevalence or abundance thresholds. However, sometimes microbes may exist at

very low depths and their presence or absence is crucial for the microbial communities. Rather than an abundance or prevalence filter, rare microbes and spurious hits can be filtered out through coverage analysis of place of alignments on the genome. An evenly covered genome most likely means that the microbes does exist in the sample. However, a genome that only has a single gene completely covered is most likely not present but rather indicates a mutation in the gene of another microbe that is present. While this analysis is implemented in `SHOGUN`, coverage cutoffs were not properly identified due to lacking of labeled mock communities. Exploring coverage based microbes filtering could impact the precision of taxonomic profilers.

3.4.2 Conclusion

Taxonomic and gene abundance profiling is a standard step in quantifying the members of a microbial community. The pipeline `SHOGUN` performs these steps in an accurate, reproducible and efficient manner with metagenomic WGS data. The standardization that comes with `SHOGUN` will accelerate the scientific discovery extending our ability to understand the complexities of microbial communities and their influence across a broad range of ecosystems.

Chapter 4

Classifying Reference Genome Presence

4.1 Introduction

Metagenomic experiments characterize the microbial community of an environment using high-throughput DNA sequencing, allowing us to study the ecology of both culturable and unculturable biological communities. These microbial communities vary vastly in composition and include the microbes colonizing humans, animals, plants, and environments such as soils, oceans, and sediments [69]. High-resolution whole-genome shotgun sequencing (WGS) is one technique becoming increasingly popular to profile the communities of these various environments. Advancements in WGS sequencing technology for use in metagenomic experiments have many benefits, including decreased costs, increased scalability, and higher-resolution taxonomic profiles that can detect microbes from all domains of life [17]. However, with this new metagenomic-profiling technique comes many new quantitative challenges. Accurately identifying the presence or absence of reference

genomes present within metagenomic-profiled samples remains a key quantitative challenge.

4.1.1 Motivation

For many metagenomic experiments, one of the primary characterizations used in the analysis is a taxonomic profile of the taxonomies present and their counts or relative abundances in a sample. Taxonomic profiles are typically organized into large tables, where the columns represent taxonomic features, the rows represent samples, and the values represent the respective abundances, either as counts or relative abundances.

Many strategies exist for creating these taxonomic tables. A common strategy for quantifying a taxonomic table is to search the sample's DNA sequences against a database of taxonomically classified reference genomes [70, 71]. The taxonomic profiler then counts each query DNA sequence that matches uniquely to a reference genome. However, many of the query DNA sequences do not match uniquely to a single reference genome. This ambiguity is partly due to many shared genes and other regions of DNA between genomes in the reference databases and exacerbated when an observed species does not have an exact representation in the database. If a query DNA sequence matches multiple reference genomes, the profiler summarizes the taxonomic features at multiple ranks creating a rank-flexible taxonomic table. Several disambiguation schemes are known to create a rank-specific taxonomic table at the highest-resolution possible taxonomic rank, usually at the species- to strain-level for WGS datasets [63, 48, 72].

Once the profiler tallies each of the rank-specific taxonomic features, each sample's total counts can be normalized to account for the sampling process of sequencing, where many of the samples might have differing sequencing depths.

A conventional method is to rarefy each sample's sequencing reads to a standard depth and then normalize each sample's counts to sum to 1 as a relative abundance table [73].

Researchers utilize rank-specific taxonomic profiles to associate the features with an outcome variable of interest, such as Crohn's disease status in a human gut microbiome [74]. Researchers use a statistical differential-abundance test to detect microbes that are significantly associated with a disease state [75]. Differential-abundance tests are done for each taxonomic feature and are typically adjusted using a multiple-hypothesis correction for false-positive reductions. To help further filter the data and reduce false-positive associations, researchers often drop very rare taxonomies below a certain average relative abundance threshold [76, 77]. They will also drop taxonomic features with a low prevalence: taxonomic features that do not meet a minimum relative abundance in a certain percentage of their samples [78]. These taxonomic feature filtering steps effectively reduce some false-positive associations. However, they may fail to remove some spurious species and may also introduce false negatives by removing low-abundance species that are nonetheless genuinely present. For example, as shown in (figure 4.1), relative abundance can be misleading if there are two species from the same genus present in a sample, as there can be high relative abundance but only sporadic coverage of the genome due to shared DNA in the reference genomes.

To our knowledge, there exists no formal assessment of a standard minimum relative abundance threshold, so researchers choose the threshold at their discretion. An alternative approach to controlling false positive strain identification in metagenomic samples is to rely on a predetermined set of marker genes [21, 22]. This approach helps control false positives, but it only utilizes a small fraction of the input data because it ignores any DNA that does not match one of the marker

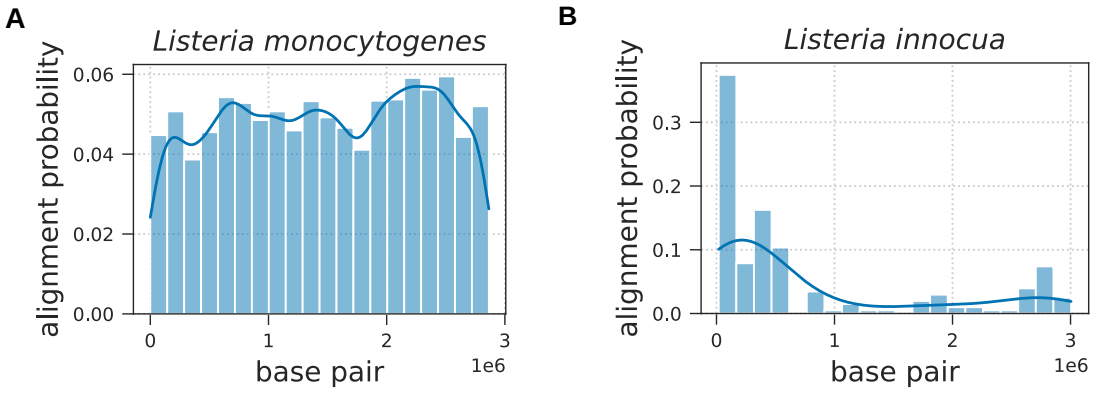

Figure 4.1: **Coverage distribution profile of two species within the genus *Listeria*.** These coverage profiles show the distribution of alignments (y-axis) along base-pairs of the genomes (x-axis). Both of these species of *Listeria* have a large number of alignments. However, the coverage distribution profiles show that only *Listeria monocytogenes* are likely to present in the sample. (**A**) A fairly uniform coverage of *Listeria monocytogenes* indicates that this species is most likely present. (**B**) Highly dispersed coverage distribution of *Listeria innocua* indicates that this species has shared genes with a species present, but this genome itself is not present.

genes.

Our goal in developing this algorithm was to enable improved false positive control in metagenomic sequencing while retaining the sensitivity of complete genome direct alignment methods. To address this challenge in taxonomic profiling, we propose a machine learning-based approach to classify the presence or absence of reference genomes in metagenomic samples. We hypothesized that a machine learning-based approach would accurately identify presence or absence in metagenomic sequencing taxonomic profiles. A machine learning approach would apply simplistic relative-abundance thresholds and learn their interactions with sample-specific features for each genome, such as smoothed genome coverage and database metadata features, such as the genome length and phylogenetic

distance, to neighbors of each reference. Using features more sophisticated than relative abundance should also have the benefit that increasing sequencing depth would increase information about the presence or absence of a given genome. If an algorithmic approach can confidently identify the presence or absence of reference genomes, the increase in accuracy will cascade down to the abundance aggregation, read disambiguation schemes, and lower false positives and negatives in differential abundance testing [79, 63].

4.2 Methods

We evaluated the primary outcome as creating a machine learning classifier to classify the presence or absence of a reference genome from an environmental WGS metagenomic dataset. In (figure 4.2), we outline the entirety of our approach, beginning from a set of WGS sequencing reads of an environmental sample and ending with a trained machine learning classifier.

4.2.1 Machine Learning Training and Testing Datasets

Mock Communities

For our purposes, mock communities are metagenomes that are a mixture of known microbial isolates or strains, which then have the DNA of the mixed community sequenced for benchmark purposes. Therefore, each microbe in the mock community is known and has a corresponding assembled reference genome. We selected several mock communities with publicly available reference genomes and raw isolate sequencing data, where the mock communities were sequenced on Illumina machines. These communities had varying isolate mixture distributions. We will refer to them as “even”, if they are equally mixed, and “staggered”, if they are

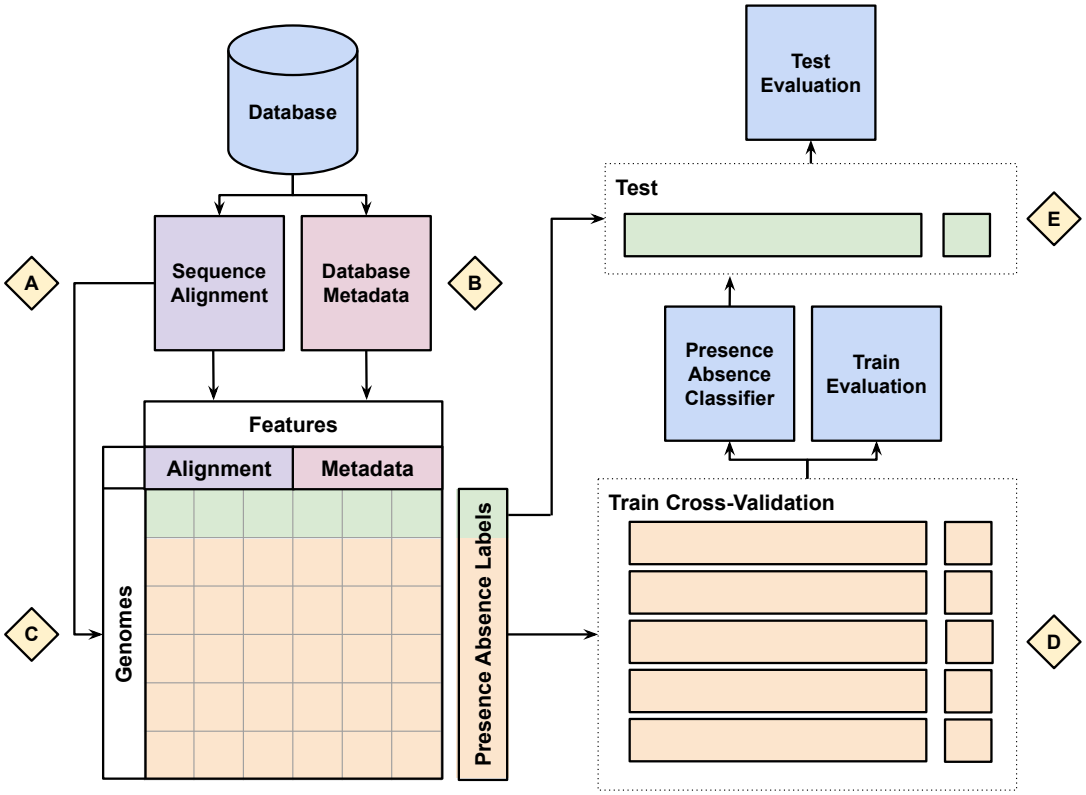

Figure 4.2: **Schematic for training the presence-absence machine-learning classifier.** (**A**) The pipeline calculates the features from the alignment of the query sequences to the database. (**B**) The pipeline calculates the database-specific metadata features, such as genome length, for each reference genome with an alignment. (**C**) The features, including the database, sequence alignment, and tree-derived features, are collated by the pipeline into a tidy machine learning dataset. This table has a row for each reference genome and sample combination, and a column for each derived feature. Every row is a genome that has a known presence-absence indicator. We then divide the dataset into a training and testing set. (**D**) We train a presence-absence machine-learning classifier in a cross-validation scheme on the training dataset. The pipeline utilizes the settings of the best performing machine learning classifier on the cross-validation to retrain a final classifier on the entire training dataset. (**E**) We use the best performing classifier evaluated on the training set and evaluate the performance of the pipeline on the held-out testing set.

unequally mixed. These mock communities for the training set are the dual-index community (Dual-Index) [80], the genome spike in the community (GIS20) [81], the human microbiome project communities (HMP) [67], and the Mock Bacteria ARchaea Community (MBARC26) [82]. The mock community for the testing set was the Zymo [83], chosen for the testing set because it contains both an even and staggered community. The complete list of mock communities, including the training and testing set split, as well as their respective data availability, are described in (table 4.1).

Simulated Communities

We simulated communities with similar diversity to human skin, oral, and gut microbiomes to enhance our training and test sets with the distribution of microbes as similar as possible to real-world microbiomes. We selected these simulated communities' DNA reads from Illumina shotgun assembly sequencing runs on isolate strains from the sequence read archive (SRA) included as a complete or reference genome in RefSeq release version 202 [49]. In total, we simulated 10 million DNA reads per microbiome, with differing numbers of reference genomes (20, 40, 100) and Shannon entropies (2.665, 2.067, 1.663) emulating saliva, skin, and stool microbiomes, respectively, as described in (table 4.1) [84]. We replicated each simulated community six times, divided into three for the training and three for the testing set. This approach allowed us to simulate mixed communities using the real-world sequencing profiles of the individual strains, thus incorporating all of the naturally occurring sequencing artifacts, such as biased sequencing coverage along each genome according to varying DNA guanine-cytosine (GC) content.

Real World Dataset

To demonstrate the utility of our machine learning pipeline, we downloaded and completed a full taxonomic profile analysis of a human gut metagenomic dataset (PRJEB40960) [85]. In our investigation, we tested for differentially abundant species in patients that developed neutropenic fever (NF). This dataset included 49 patients, of which 31 developed NF. To test for differential abundance, we used the analysis of the composition of microbes test (ANCOM) [75], correcting multiple hypotheses using a false discovery rate adjustment with a significance of 0.20 [86].

4.2.2 Machine Learning Pipeline

Features

In total, we evaluated 34 potential features for the machine learning pipeline. We categorized the features into three different subgroups depending on the origin of the calculation of the features. The first category, known as alignment-based features, is derived from aligning the query sequences to the reference database (GTDB version R95). These alignment-based features include number of alignments to the reference genome, expected coverage of the reference genome based on relative abundance and genome length, and Shannon entropy of the coverage distribution. The second category, known as database metadata features, is derived directly from the reference database independent of the query data. These database metadata features included the reference genome size, a genome completeness score, and the number of contigs of the genome. The final set of features, known as tree-based features, is derived from the nearest neighbor to a given reference strain on the taxonomic tree. These tree-based features included each of

the previous features for the nearest neighbor on the taxonomic tree. The motivation for these neighbor-based features was to provide information about how ambiguous the profile of matches was to a given reference strain. For example, if reference strain A has a decent alignment profile, but a closely related neighbor strain B has an even better alignment profile, that may decrease the likelihood that strain A is truly present.

Machine Learning Models

We performed predefined-split cross-validation, one for each of the datasets, on the training dataset to train the machine learning pipeline. We used an automated machine learning method, known as Tree-based Pipeline Optimization Tool (TPOT) [87], to narrow down our list of potential models, run feature selection, and optimize hyperparameters. We utilized most of the recommended configuration settings from TPOT, except that we modified each classifier to have their automated class imbalance handling setting, if possible. The best performing classifier from the cross-validation set, the extra-trees random forest classifier [88], was used to make predictions on the test set. For feature selection, we implemented a recursive-feature elimination schematic on the training dataset. Finally, to evaluate the current filtering practices and create a baseline model, a logistic regression model was trained, with the only feature being relative abundance.

Although our primary focus was acquiring accurate predictions from our machine learning models, we were also interested in understanding why our model made the predictions that it did. To evaluate the effects of features of the data on the classifier, we used the SHapley Additive exPlanations (SHAP) library [89, 90].

Evaluation Metrics

We were interested primarily in reducing the number of false-positive identified reference genomes in a taxonomic profile, with minimal impact on the false-negative rate. Because there are relatively few genomes present in a given sample when compared to a notable number of potential reference genomes present in the database, there was a large class imbalance in the ratio of positive (present, minority class) and negative (not present, majority class) cases per sample. To account for this class imbalance, we focused on precision and recall as our evaluation metrics, with our primary evaluation metric being mean average precision (AP) [91] on a precision-recall curve. The evaluation metric average precision is less sensitive to class imbalance than other evaluation metrics, such as the commonly used receiver-operator statistic area under the curve (ROC AUC). For example, let us suppose the minority class has few samples compared to the majority class, as in our case with the number of actual genomes. In that case, the ROC AUC tends to overestimate accuracy by overly weighting the true-positive and false-positive rates of the majority class. By contrast, the average precision metric will allow us to evaluate the classifier's performance focusing on the minority class.

4.2.3 Metagenomic Sequencing Processing

An Illumina machine sequenced every raw WGS metagenomic sample. The shi7 learning module ran quality control for each of the sequencing runs [18]. The BURST algorithm aligned the sequencing reads to the reference database [48]. We set a minimum alignment identity of 0.98 and multiple alignment disambiguation with the BURST *capitalist* mode. The *capitalist* mode reported all queries with a unique best alignment at 98% identities or above. If there were ties for the

best alignment, then the algorithm distributed these to the reference genomes to minimize the total number of reference genomes required to explain all alignments.

The database we used for all of the analyses was the Genome Taxonomy Database version R95 (GTDB) [92, 93]. This database has reference genomes taxonomically classified to the species level, and each species cluster must be at least 95% unique. Note that not all genomes in the mock and simulated metagenomes have an exact representative genome in the reference genome database. In order to get the actual presence-absence labels in these cases, we found the nearest species representative in the reference genomes using the GTDB-toolkit (GTDB-Tk) as outlined in figure 4.3 [94]. We required a minimum average nucleotide identity (ANI) of at least 95% for a genome to be considered a member of a given species.

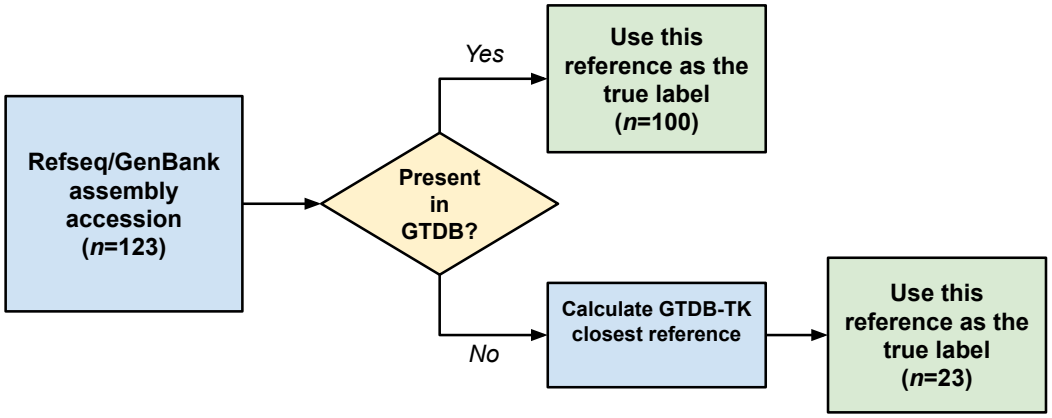

Figure 4.3: **Process for getting the nearest species representative for an assembled genome.** First, the assembly accession is searched for an exact match in the reference genome database. If no exact match is found, the nearest taxonomic cluster is assigned using the GTDB-Tk.

4.2.4 Software Availability

All code for training, testing, and applying the machine learning pipeline to new data, as well as any additional code required to reproduce the analyses in this manuscript, are available on GitHub at `https://github.com/knights-lab/type_1`.

4.3 Results

Our primary evaluation methodology was to determine the best classifier on the training set using cross-validation, train the classifier on the whole training set, and then evaluate the classifier according to AP on the testing set. Our secondary investigation was to determine the most predictive features for classifying presence or absence and investigating the interaction between coverage and number of alignments to the reference genome. We were also interested in the minimum coverage and number of alignments to a reference genome required for a true-positive association to be detected.

4.3.1 Extra Trees Classifier Outperforms Baseline Model

The automated machine learning evaluation showed us that the pipeline containing an extra trees classifier performed the best in terms of average precision ($AP = 0.97 \pm 0.03$) on the train set as shown in (figure 4.4) and (table 4.1). This model outperformed the baseline logistic regression model of ($AP = 0.80 \pm 0.13$). The pipeline's first step was a recursive feature selection schematic to eliminate correlated, low variance, and low predictive power features. The extra trees classifier utilizes the resulting filtered feature set. The extra trees pipeline also performed well on the test set ($AP = 0.94 \pm 0.04$), significantly outperforming the

baseline logistic regression model ($AP = 0.68 \pm 0.02$).

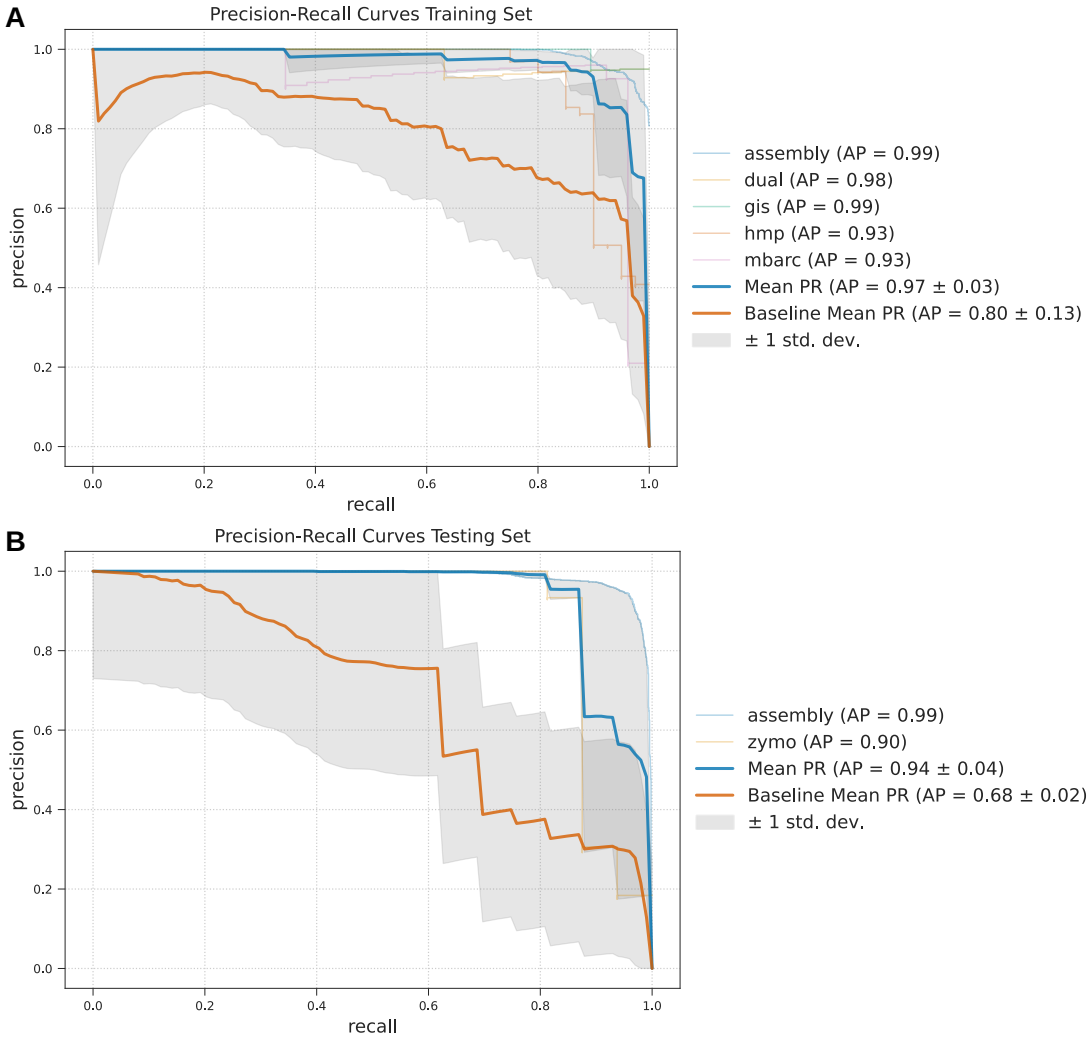

Figure 4.4: **Precision-recall curves for the the testing and training set.** (**A**) The highly-tuned classifier greatly outperformed the baseline model on all of the training datasets during cross-validation. (**B**) The precision-recall curves on the testing set show the highly-tuned classifier greatly outperforming the baseline model.

The recursive feature selection schema showed that using 11 out of the 34 candidate features had the best performance in terms of average precision, as

shown in (figure 4.5). The most noteworthy feature of SHAP predictive values was the largest gap with no alignments in the reference genome. An extensive coverage gap showed a positive correlation with the classifier classifying the genome as absent. The second most influential feature was the percent binned coverage of each genome. The feature aggregator calculated the percent binned coverage by dividing each genome into 10 thousand bins, and the number of bins with at least a single alignment is considered covered. The SHAP values indicated that if there was a high coverage of the reference genome, there was a significant positive impact in predicting the presence of a reference genome.

We were also interested in evaluating the machine learning model's performance on reference genomes with low coverage or a low number of query alignments. We conducted this analysis to understand the detection limit for rare species at a given depth of sequencing. The logistic regression model results showed that any reference genome with a relative abundance greater than 0.0006 would be considered present. The logistic regression ($AP = 0.031$) model accuracy is outperformed by the machine learning classifier ($AP = 0.865$) on reference genomes below ($n = 30,718$) the relative abundance threshold (relative abundance $<$ 0.0006) which the logistic regression model would always classify as absent. The accuracy of the logistic regression ($AP = 0.645$) was also outperformed by the machine learning classifier ($AP = 0.946$) on reference genomes above ($n = 2,389$) the relative abundance threshold. In order to find the thresholds for accurate classifications in terms of hits, relative abundance, and coverage of a reference genome, we investigated the average precision of reference genome samples above or below a specific feature threshold as shown in (figure 4.6). For example, in terms of accuracy above a feature threshold, references with at least 1,000 hits

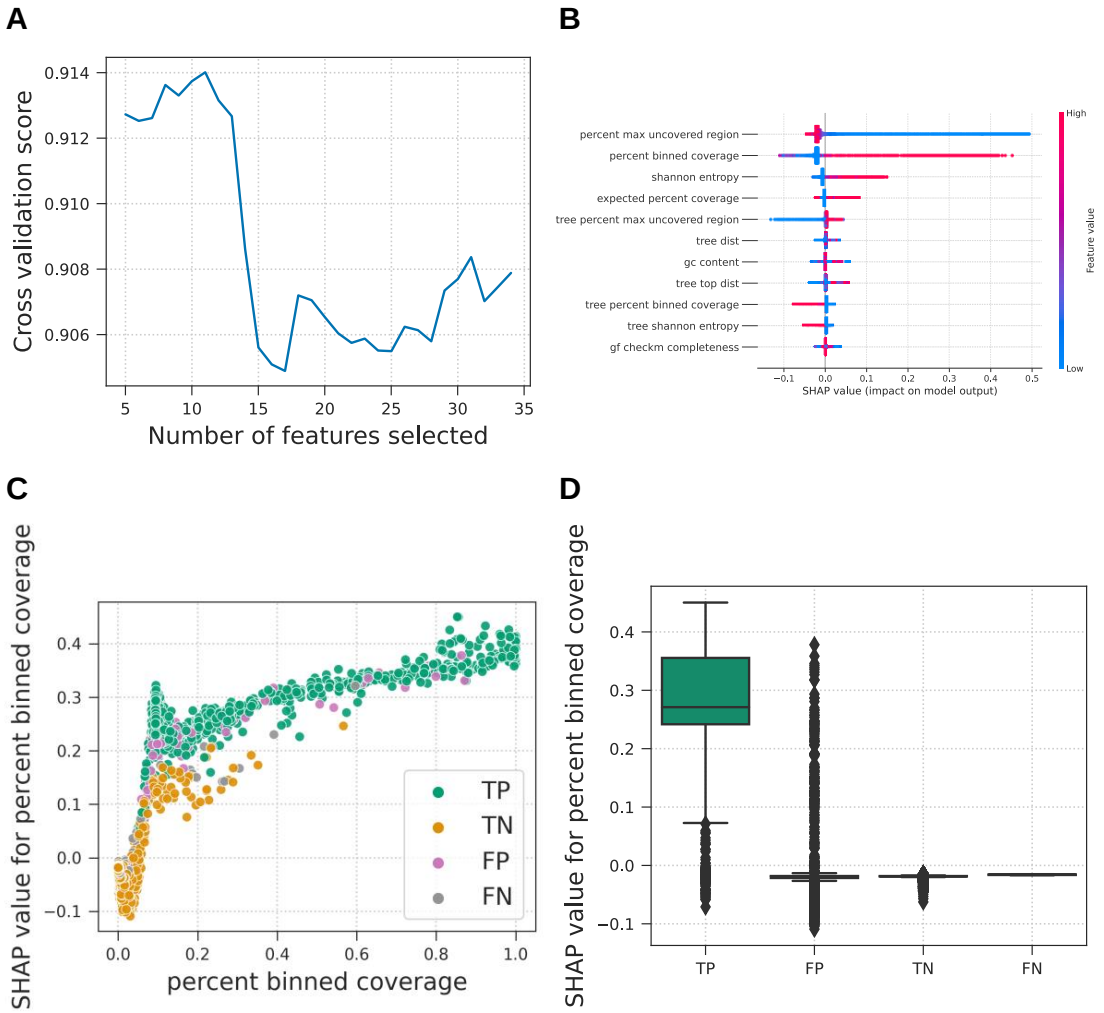

Figure 4.5: **Coverage features have the highest effect on classifier predictions.** (**A**) The line plot shows the performance of the model in terms of average precision over an increasing number of features. (**B**) A SHAP summary plot in order of features that have the most impact on the model's predictions. The percent uncovered region has the highest impact on the model. (**C**) Scatter plot showing the performance of the model across all values of the percent binned coverage feature. Higher values of coverage showed a high likelihood of a true-positive classification. (**D**) Box plot showing the distribution of performances according to SHAP impact on model classification.

to them ($n = 1,153$) were able to be accurately identified ($AP = 0.962$). Furthermore, the machine learning pipeline could accurately classify ($AP = 0.961$) presence and absence at a percent binned coverage of at least 0.11% ($n = 1,027$). The machine learning classifier achieved highly accurate classifications for values for a lower-bound threshold for hits, percent binned coverage, and relative abundance, indicating a significant amount of evidence obtained for classifying presence-absence for genomes with high values for these features. However, it required an upper-bound threshold of 1,204 alignments to a specific species reference genome to achieve at least 95% accuracy in terms of average precision of the entire cohort of reference genomes. This result shows that for a sequencing depth of 10 million reads, the machine learning classifier would accurately assess the presence of a specific microbial species if it has a relative abundance of at least 0.012%. In this example, to obtain targeted analysis for species below the 0.012% relative abundance threshold, it is advised to adjust sequencing for a higher depth.

In (figure 4.7), we show the results of the machine learning pipeline filtering the presence-absence of microbes on a metagenomic dataset. In each sample-genome combination in the entire dataset, there was a total of 131,133 presence absence-decisions to be evaluated. The machine learning pipeline classified 3,665 genomes as present with a presence classification rate of 2.7%. In the entire dataset, the machine learning pipeline identified a total of 695 unique genomes across all of the samples, 62 of which were not identified by the logistic regression model. The baseline logistic regression identified an additional 385 genomes as present that the machine learning model identified as absent, resulting in a total of 1,020 genomes present for the logistic regression model. Both of the filtering methods identified one species as being significantly different between the neutropenic fever outcomes. In addition, with the greater power awarded by multiple-hypothesis

with fewer hypothesis, the machine learning model was able to identify three more significantly different species between outcomes.

4.4 Discussion

In this research, we built a machine learning classifier as a solution for identifying the presence-absence of reference genomes in WGS metagenomic datasets. To our knowledge, this is the first machine learning classifier used as a solution to this problem. We found that our trained classifier could outperform a baseline classifier that emulated a common practice of creating a relative abundance filter. Our method showed that the machine learning classifier could take advantage of the complex relationship between number of alignments to a genome and statistics about the coverage distribution profile to classify presence-absence accurately.

Interestingly, despite prior studies often using relative abundance as the primary determinant of presence-absence, our pipeline did not select it as a feature in the final machine learning pipeline. We showed that our machine learning pipeline was able to achieve accurate predictions with as few as 800 aligned queries for a given genome, and with genomic coverage as low as 0.1%. These results show the performance of our classifier on metagenomic samples.

4.4.1 Future Work

There are some limitations to consider for the application of this study to real-life data sets. Importantly, we have not assessed the accuracy of this study at taxonomies below species level, namely to the strain- or sub-strain taxonomy classification. The GTDB imposes this limitation, as the GTDB clusters reference genomes as representative species at the 95% similarity level. Furthermore, there

is a lack of publicly available data from mock communities containing multiple closely related strains. The machine learning pipeline is currently limited to the GTDB database because several of the features rely on the metadata provided with that database. We utilized an optimal alignment method, so researchers using approximate or heuristic aligners may observe decreased performance. It would be interesting to vary the reference database and the method used for DNA alignment in future work. Finally, the inclusion of long-read sequencing technology such as PacBio or Nanopore shows promise in alleviating the false discovery alignment of genomes in metagenomic profiles. Combining long- and short-read sequencing runs in metagenomic datasets could be an alternative approach to solving the problems of false positive and false negative strain identification that we observe with only short-read sequencing.

4.4.2 Conclusion

In summary, we developed and created a novel machine learning pipeline for determining the true presence-absence of a reference genome in WGS metagenomic datasets. We rigorously trained and evaluated our pipeline on mock, simulated, and real-world WGS communities. We expect this model to reduce noise and false positives in future metagenomic studies by allowing researchers to determine with confidence which species are present in a given sample using a principled and data-driven approach.

Training

Name	True Number of Genomes	Number of Genomes with Alignments	Diversity	Archive
Dual-Index	20	771	Even	Link
GIS20	20	2,167	Staggered	PRJEB29139
MBARC	26	961	Staggered	PRJNA324704
HMP	44	3,223	Even & Staggered	SRR2822454
Saliva	508	12,984	2.665	
Skin	505	10,207	2.067	
Stool	501	7,865	1.663	
Total	1,624	38,178		

Testing

Name	True Number of Genomes	Number of Genomes with Alignments	Diversity	Archive
Zymo	20	1,644	Even & Staggered	PRJEB29504
Saliva	505	12,657	2.665	
Skin	504	9,716	2.067	
Stool	507	7,588	1.663	
Total	1,536	31,605		

Table 4.1: **Table outlining the samples in the training and testing datasets for the machine learning classifier.** There is a class imbalance for the True Number of Genomes in the communities compared to the Number of Genomes with at least a single alignment. Several of the genomes were unidentified with the nearest neighbor in the Genome Taxonomy Database (GTDB) due to them belonging to the Fungi domain.

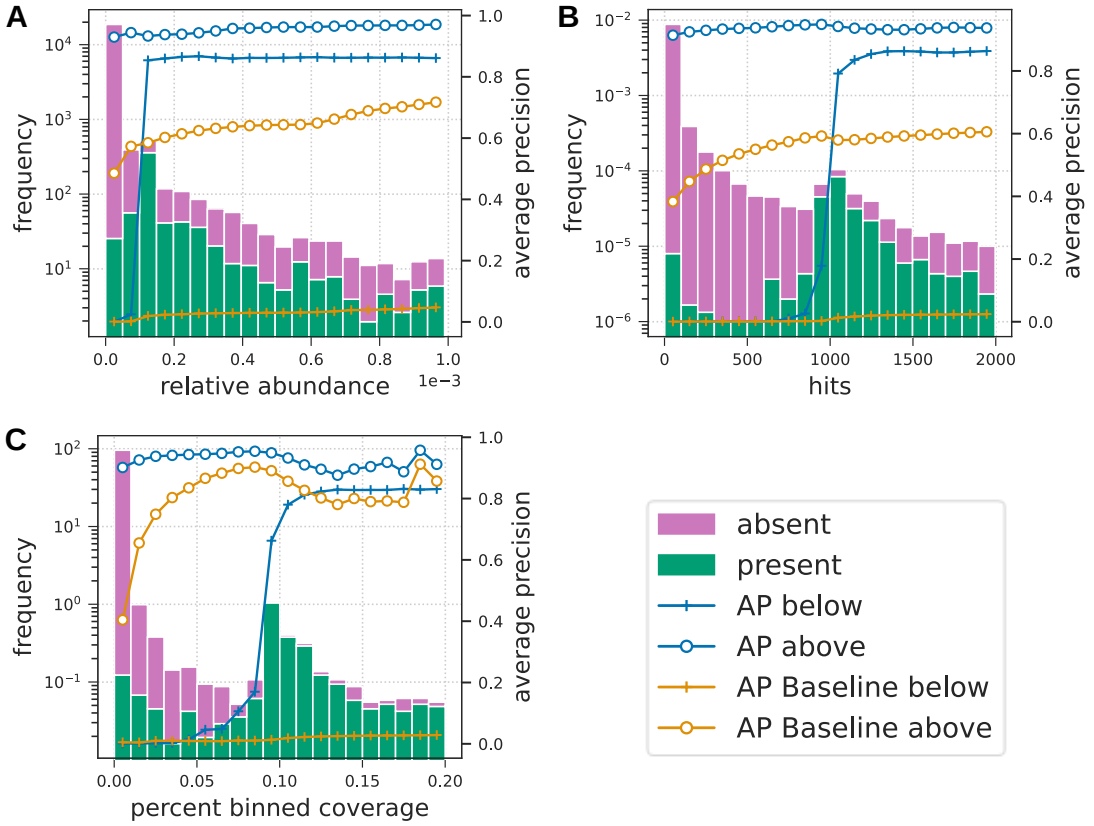

Figure 4.6: **Performance of the machine learning shows accurate classification across low number of alignments and coverage of genomes.** These plots highlight the ability of our machine learning pipeline to accurately classify low-abundant microbes, with increasing accuracy as the depth of sequence sampling increases. Stacked bars show the distribution of species presence-absence. Note that the distribution of the feature is log_{10} for clarity because, at low abundances and coverage, almost all reference genomes are not present depicted on the left y-axis. The x-axes are each a different feature in the dataset. There are four lines on each plot measured in terms of average precision on the right y-axis. The blue lines show the machine learning classifier, while the orange lines show the baseline classifier. Each of the lines reflects accuracy on samples either above or below the specified threshold. (**A**) The machine learning classifier accurately classifies low-relative abundance genomes. (**B**) This plot shows the accurate machine learning performance on a low number of query alignments hits to a reference genome. (**C**) This plot shows accurate machine learning classification accuracy at low percent binned coverage of the genomes.

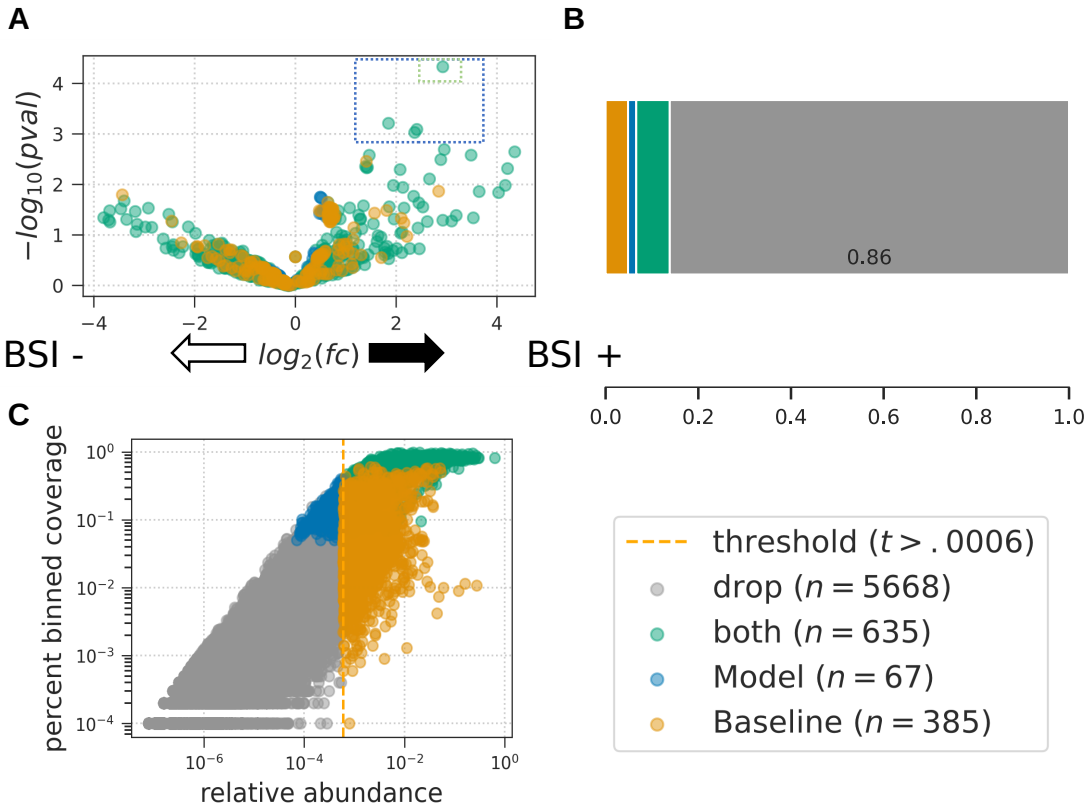

Figure 4.7: **The machine learning pipeline is able to determine significantly associated species not identified by the logistic regression model on the NF dataset.** (**A**) Scatter plot showing the differential abundance of reference genome species. The blue box depicts all reference species identified as significantly different at a false discovery rate of 0.2. (**B**) A stacked bar chart showing the proportion of reference species as present or absent by the baseline and machine learning model. (**C**) A scatter plot showing model and baseline classification of reference genomes stratified by percent binned coverage and relative abundance.

Chapter 5

Application of Shallow-Shotgun Metagenomic Profiling Techniques [1]

A healthy, adult human gastrointestinal tract is host to millions of microbes, and the collective taxonomy encompasses a diverse composition of archaea, bacteria, fungi, protozoa, and viruses [8]. The composition of the human gut microbiome is personalized, varying widely across individuals. A combination of environmental factors and host genetics drive the individuality of each human gut microbiome's composition and dynamics [9].

Many studies have shown that the human gut microbiota is a critical component in an individual's overall health [12, 96, 54]. Maintaining or restoring a healthy microbiome is essential for treating certain diseases. For example, restoring the human gut microbiome towards a healthy state has proven an effective

[1] A version of this article has been published [95]. BH aided in the software development, data analysis, writing, and editing of the original manuscript.

long-term treatment of *Clostridium difficile* colitis [97] and irritable bowel syndrome [98]. In other diseases, such as ulcerative colitis, microbiome therapy has proven to alleviate symptoms but not entirely prevent remission, suggesting that the microbiome is not the pathological source of the disease [99]. The conclusions of all the earlier studies call for more extensive cohorts to be warranted to solidify their results and suggest individualized therapy efficacy.

The microbes do not exist in isolation; instead, they form complex ecological interactions [100]. A common objective of metagenomic sequencing studies is the accurate prediction of these ecological microbe-microbe interactions from observational population abundance data [101]. The process of inferring a co-occurrence or correlation network from observational data is known as 'network inference.' Specifically, previous work shows that dynamic Bayesian networks (DBNs) can accurately infer microbe-microbe interactions and predict future changes to the microbiome community [102, 103, 104]. If researchers accurately infer an ecological network of underlying community ecosystems, it will provide a static snapshot of the community structure. Researchers can then use the ecological network to develop a dynamic microbial growth model, which would allow the inference of community stability, the effects of intervention and perturbation to the community, and population succession [105]. The main goal of these network methods is to produce accurate hypotheses of the overall health of a community, what the community structure looks like over time, and identify potential targets for intervention to establish the dynamics of the human gut microbiome.

5.0.1 Dietary Impact on the Human Gut Microbiome

Previous studies have shown that diet is a significant factor in explaining microbiome variation [76]. For example, researchers have shown dramatic shifts in the

microbiome composition after changing from an animal-based to a plant-based diet [16]. Several studies have demonstrated the effect of a strict dietary plan, summarized as Western, Mediterranean, or Gluten-free diets, and its dramatic shifts on microbiome composition [106]. Several targeted studies have also assessed the impact of the specific nutrients on microbial abundances [107]. These studies demonstrate the influence of diet on the microbiome, outlining a potential approach for microbiome mediation. However, the research on the effect of most foods and nutrients in shaping the microbiome is still ongoing.

The relationship between diet, the gut microbiome, and health suggests a method for modulating our health through an increased understanding of the dynamics of this relationship. To investigate the relationship between dietary intake and microbiome dynamics, we applied the methods described in this thesis to our previously published daily shallow-shotgun metagenomic dataset [95]. Each sample also contains a 24-hour dietary recall record, allowing us to explore the relationship of diet-microbiome dynamics through a DBN model systematically.

5.1 Methods

Our primary aim is to establish our presented tools' utility by building a more comprehensive reference database than at the time of original publication and applying our methods outlined in this thesis to generate highly accurate, taxonomic, and functional profiles. We outline all of our methods, and software, and data availability in table 5.1. For a full overview and in-depth details of the sample collection, please see the original paper as submitted by Johnson et al. [95]. To clearly delineate work that was done in the original study, we utilize the use of quotations. We outline all of the deviations from the original data analysis study

Reagent or Resource	Source	Identifier
Biological Samples		
Healthy, human microbiome samples	Johnson et al. [108]	ENA: PRJEB29065
Critical Commerical Assays		
PowerSoil DNA isolation kit	Mo Bio Laboratories, USA	Catalog No. 12888
Nextera XT DNA library preparation kit	Illumina, Inc	FC-131-1096
Deposited Data		
Genome Taxonomy Database	Parks et al. [92, 92]	Link version 95
Human Reference Assembly GrCH38	International Human Genome Sequencing Consortium [66]	ENA: PRJNA31257
Software and Algorithms		
BURST DNA aligner	Al-Ghalith and Knights	Link version 1.0 DB 15
`SHOGUN` metagenomic processing pipeline	Hillmann et al., chapter 3	Link version 1.0.8
The `shi7` quality control pipeline	Al-Ghalith et al. [18]	Link version 1.0.3
`bnlearn` bayesian network inference library	Scutari et al. [109, 110]	Link version 4.6.1
eggnog-mapper gene annotation pipeline	Cantalapiedra et al. [111]	Link version 2.1.6
FoodTree Methods	Vangay et al. [112]	Link
Diet analysis scripts	Johnson et al. [95]	Link
Custom analysis scripts	this paper	Link

Table 5.1: **Table outlining all data and software availability.**

here without quotations. Most of our added variations occur in the processing of taxonomic and functional profiling utilized.

Our secondary aim is to review diet's impact on microbiome variation and develop potential diet-microbe interactions. We are uniquely positioned with this dataset, as the cost-effectiveness of shallow-shotgun metagenomic sequencing allowed for dense longitudinal microbial sampling. Furthermore, the subjects' diets were unrestricted, with everyone being healthy, potentially allowing the discovery of any food-microbe interaction in a robust DBN approach in this healthy cohort. All work completed towards our second aim is novel.

5.1.1 Study Design

Our shallow-shotgun methods enabled the completion of a complete metagenomic survey. From Johnson et al. [95]:

> The study also included a 10-day, parallel, double-blind intervention trial to test the impact of medium chain triglycerides (MCTs) compared to long chain dietary triglycerides from extra virgin olive oil (EVOO) on microbiome composition. As there were only null findings from all tests for associations between MCT or EVOO supplementation and the microbiome, this manuscript focuses on the analysis of overall diet-microbiome covariation using all available samples.

Since the original findings were null, the original report focused on the analysis of overall diet-microbiome interaction, and that is the focus we continued on here. However, we do include the supplementation type within our network inference, as our updated database and methods may discover some novel interactions.

We were able to measure this relationship between habitual dietary intake and

daily microbiome variation using dense, longitudinal diet-microbiome data. From Johnson et al. [95]:

> To characterize the longitudinal relationship between diet and microbiome composition, we collected dietary intake data and fecal samples from 34 subjects for 17 consecutive days. Daily food records were collected using the automated self-administered 24-hour (ASA24) dietary assessment tool (2016, National Cancer Institute, Bethesda, MD, USA) [113].

We included all food profiles generated as described with the 3-day decaying weighted average in our analysis.

5.1.2 Microbiome Taxonomic Profiling

Johnson et al. [95] "submitted all human gut microbiome samples for processing at the University of Minnesota Genomics Center for DNA extraction, amplification, and sequencing." For our supplementary analysis, we learned the optimal quality-control parameters using the `shi7` [18] learning module and subsequently processed the single-end metagenomic shotgun reads for quality-control. We filtered all human reads using the `SHOGUN` filter commands by aligning to the GRCh38 human reference genome [66]. We then aligned the quality-controlled sequences to the Genome Taxonomy Database (GTDB) release version 95 [93]. We aligned all taxonomic profiles using the BURST [48] aligner at 97% identity in the *capitalist* mode within the `SHOGUN` pipeline (see chapter 3).

We preprocessed the taxonomy table from `SHOGUN` by dropping all samples with less than 23,500 alignments and dropped unlikely present species level identifiers with the `type_1` presence-absence classifier pipeline (see chapter 5). The

original study re-submitted some samples for sequencing, so we subsetted taxonomies to the intersection of taxonomies presented within both sequencing runs and preferred the re-ran sample as necessary. We further limited the number of species to those present within a subject for at least 25% of the study duration and found within 10% of the study subjects.

To correct for ploidy and genome length bias, we normalized by dividing the counts by the genome length and then multiplying by average genome length of all species in the profile [114]. To correct for the compositionality associated with microbiome sequencing data, we first imputed zeros and then applied the centered log-ratio transformation [115]. We assessed tree-based beta-diversity using unweighted UniFrac and the taxonomic tree provided by the GTDB. We used these normalized taxonomic profiles for all downstream processes.

5.1.3 Microbiome Functional Profiling

To find KEGG annotations for reference genomes in the GTDB 95, we used the `eggnog-wrapper` [111, 116]. We used the `SHOGUN` functional pipeline to obtain Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations for the taxonomic profile [45]. We corrected compositionality within the functional profiles with the imputation of zeroes and the centered log-ratio transformation.

5.1.4 Dynamic Bayesian Network Analysis

Bayesian Networks

Bayesian networks (BNs) are a concise representation of all the conditional dependencies between a set of random variables in the form of a directed acyclic graph

[110]. BNs are a class of probabilistic graphic models whose nodes represent random variables and edges occur between variables that are conditionally dependant on one another. If there is no edge between a pair of variables, it signifies that they are conditionally independent.

Dynamic Bayesian Networks

Dynamic Bayesian networks (DBNs) are a special case of BNs designed to model temporal data. In a temporal dataset, rather than having a separate BN for each point in time t, they model the transition from one point in time to the next. For example, suppose that we have a random variable T representing the relative abundance of a taxon. The relative abundance T was observed over multiple days $t = 1, 2, ..., n$. A DBN would model the change in relative abundance from one day to the next. More formally, the transition of $T_t \rightarrow T_{t+1}$ would be modeled using a stationary conditional distribution. Several studies have shown DBNs to be an effective model of the human gut microbiome [102, 104, 103].

Diet-Microbe Network Inference

We outline our approach to inferring a DBN for diet-microbe relationships in figure 5.1. We followed closely the methodology outlined by Lugo-Martinez et al. [103], adapted to our dataset with significant deviations listed here. Due to the time complexity of network structure learning, the taxonomic, functional, and food profiles were limited to the top fifty highest variant features of the taxonomic and functional profiles. For taxonomic profiles, we used species-level annotations for functional profiles, we used the KEGG-level annotations, and for diet profiles, we used the L3 level annotations from the food tree. Our clinical metadata included two discrete variables, Gender and Supplement, so the hybrid conditional

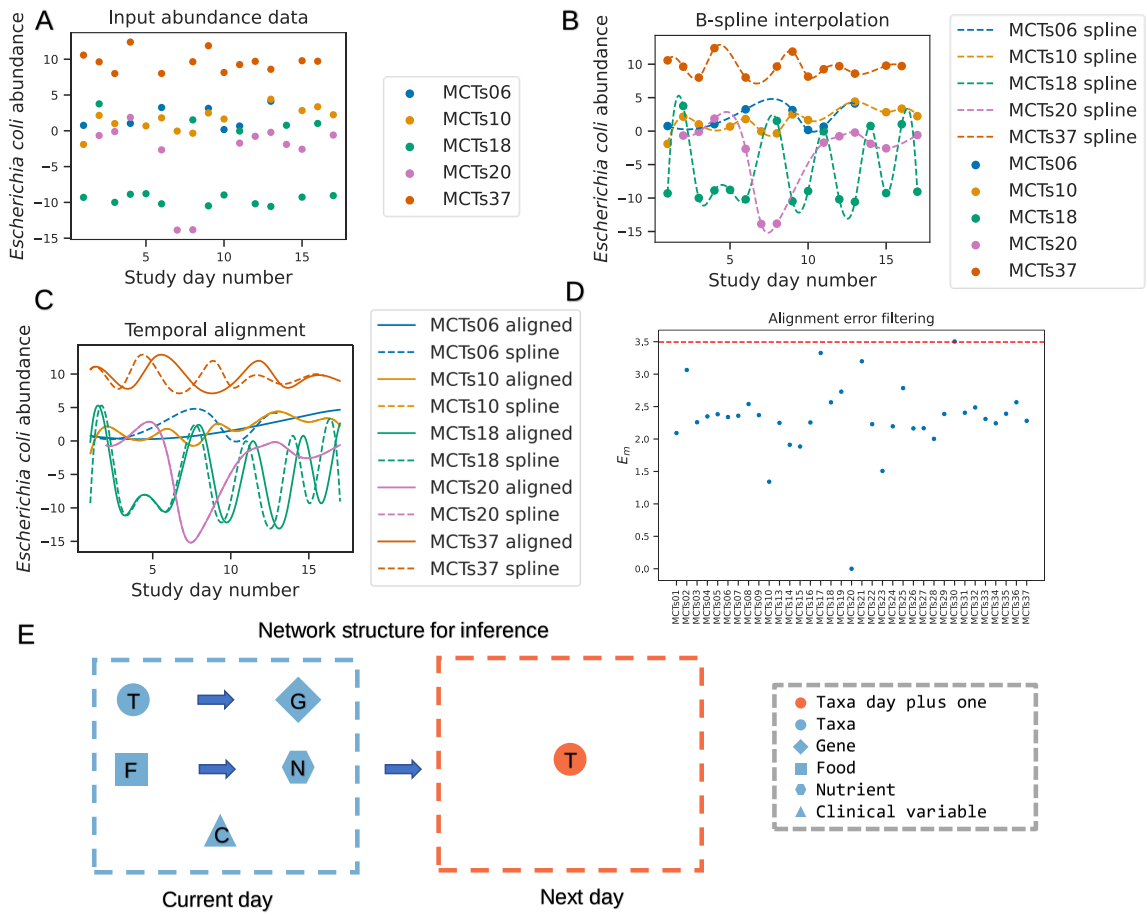

Figure 5.1: **Overview of the methodology for diet-microbe interaction network inference.** We only show the microbial taxon *Escherichia coli* at each step in the pipeline from a subset of samples. (**A**) Depiction of the centered-log ration with zero imputation timeseries. (**B**) A cubic B-spline curve is fit for each individual sample to standardize sampling rate and impute missing timepoints. (**C**) Temporal alignment of all taxa of each individual against the optimal reference subject (subject MCTs20 in magenta). (**D**) Postalignment filtering of sample MCTs30 with a higher alignment error than a predefined threshold (**E**) Constraints placed on learning of the network. All edges are allowed to taxa on taxa day plus one T_{t+1}.

independent mutual information was used for network structure inference [110]. We did not limit the number of incoming edges to the next day microbes T_{n+1}. The R library `bnlearn` [110, 109] was used for DBN structure learning using the semi-interleaved HITON-PC algorithm [117]. The coefficients of the DBN were fit using the Python package `statsmodels` [118] with a ridge regression penalty.

5.2 Results

In Johnson et al. [95], "we conducted shotgun metagenomic sequencing on each stool sample at a depth of 7,195,302 ± 2,442,901 single-end reads per subject, divided approximately evenly across the time points for each subject, with an average depth of 506,133 ± 323,896 reads per sample after the removal of human DNA." Previously presented tools in chapter 2, have shown that low-depth metagenomic sequencing can recover species-level taxonomic assignments and also allows for the assessment of functional profiles. We removed sequencing adaptors, trimmed and filtered shotgun metagenomic reads according to quality using `shi7`, and assigned taxonomy using `SHOGUN` and a database consisting of the microbial kingdoms *Bacteria* and *Archaea* from the GTDB. We retained 409 microbiome samples for analysis after removal of those with low depth. In Johnson et al. [95], "dietary outliers were identified according to guidelines from ASA24 by comparing macronutrient composition and total energy intake to reference levels to identify low-quality reporting. We retained 566 24-hour food records after the removal of dietary outliers."

Microbiome composition was more variable in some subjects than others across the study period (figure 5.2.A). As has previously been shown [8], the most prevalent microbial functional modules, as annotated by the KEGG, were highly

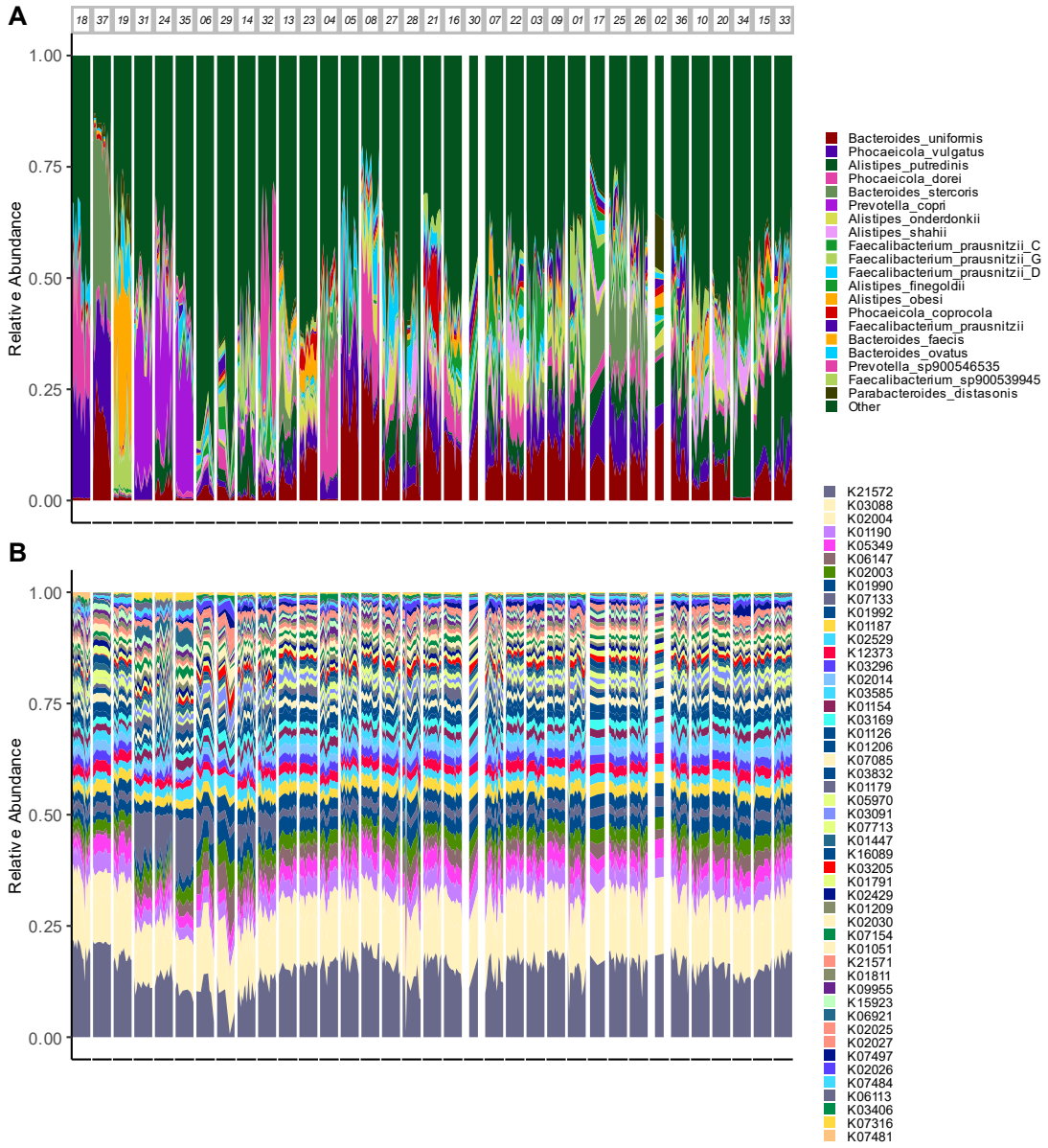

Figure 5.2: **Shallow-shotgun sequencing enables high-resolution taxonomic and functional profiles.** Taxonomic and functional-profiles generated from the tools are variable and high-resolution. The number of taxonomies and genes were limited for visual clarity. (**A**) Taxonomic profiles displayed as a time-series across all subjects. (**B**) Functional profiles displayed as a timeseries across all subjects.

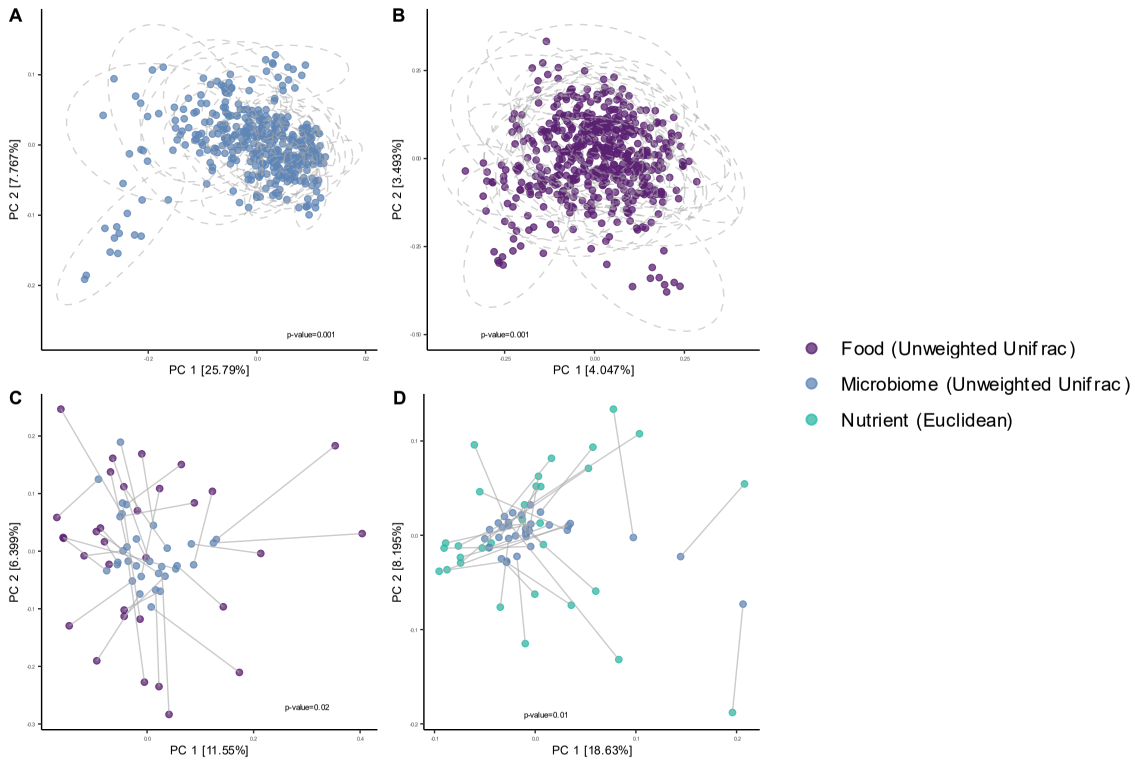

Figure 5.3: **Diet and the microbiome are personalized and vary together.** All beta-diversity profiles were calculated using PCoA and unweighted UniFrac excpet for the nutrient profile in (**D**). Microbiome, diet, and nutrient profiles are all significantly associated with one another. Gray dotted circles depict the 95% confidence interval for subject profiles. (**A**) Figure displaying that taxonomic profiles are personalized. (**B**) Figure displaying that diet profiles are personalized. (**C**) Diet-Microbiome interactions are significantly associated with each other. (**D**) Nutrient-Microbiome profiles are significantly associated with each other.

consistent within and across subjects. Microbiome beta-diversity analysis using unweighted UniFrac distances showed strong grouping by subject (figure 5.3.A; PERMANOVA; $p = 0.001$; 999 permutations). Contrary to the original analysis, the average microbiome beta-diversity did show grouping by gender (figure 5.3.C; PERMANOVA; $p = 0.050$; 999 permutations).

We applied Procrustes analysis to test for microbiome and dietary variation across subjects. Our analysis showed that a subject's average food intake corresponds with that subject's average microbiome composition when analyzed using the unweighted UniFrac-based food distances (figure 5.3.C; Procrustes; Monte Carlo $p = 0.020$) and unweighted UniFrac-based microbiome distance. We found a similar correspondence between the average microbiome and average nutrient intake using 65 macro-and micronutrients (figure 5.3.D; Procrustes of microbiome unweighted UniFrac and standardized nutrient profile Euclidean distances; Monte Carlo $p = 0.010$; 999 permutations).

We evaluated our procedure for inferring a DBN warp against two other models. The first model predicts the current day's taxon abundance as the next day's taxon abundance and is referred to as "Baseline". The second model is the same inference structure as DBN warp, without the temporal warping step known and is referred to as the "DBN" model. There was no significant difference in the mean absolute error of residuals among the Baseline model and DBN warp models (figure 5.4 Tukey HSD) [119]. Both the Baseline and the DBN warp model outperformed the DBN model. While there was no significant difference in means, the error profile of the DBN warp was the lowest, following the trend of earlier DBN microbiome analysis.

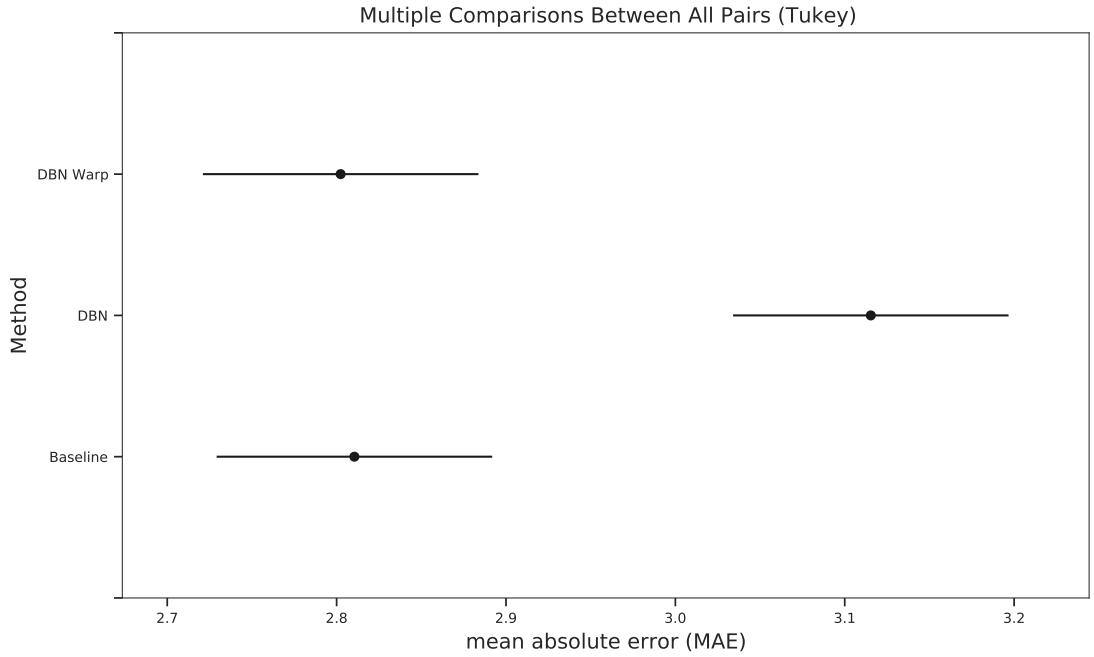

Figure 5.4: **Performance of the DBN-warp model shows no significant improvement above the baseline model.** The DBN warp and Baseline models are significantly different than the DBN model (Tukey HSD DBN-warp - DBN $p\text{-}adj < 0.001$; Tukey HSD Baseline - DBN $p\text{-}adj < 0.001$). We cannot determine if the DBN Warp model is significantly different than the Baseline model (Tukey HSD DBN-warp - Baseline $p\text{-}adj = 0.900$).

5.3 Discussion

5.3.1 Shallow-Shotgun Methods Enables High-Resolution Longitudinal Microbiome Studies

We were able to replicate the original conclusion that food choices were associated with overall microbiome composition between subjects. We verified that both diet and taxonomic profiles are personalized, by having a significant unweighted UniFrac beta-diversity test on both taxonomic and diet profiles stratified by subject. We then showed that diet and microbiome profiles were associated by a Procrustes analysis on the beta-diversity profiles. An original conclusion was that tree-based, phylogenetic methods encode relevant information in profile comparison. Our newer methods were able to utilize the phylogenetic unweighted UniFrac distance on taxonomic profiles, thanks to the taxonomic tree now provided by the GTDB. This circumvented the original need to do custom analysis with the shotgun data and 16S alignments. To our knowledge, this is the first use of unweighted UniFrac on shotgun data. Unweighted UniFrac is beneficial in beta-diversity analysis as it considers the difference in evolutionary divergence in its measurement of distances between taxonomic profiles. The phylogenetic distances allowed us to show a significant association between taxonomic and nutrient profiles, which were originally not obtained when using non-phylogenetic distances.

5.3.2 Diet-Microbiome Network Validation and Interactions

The goal of the DBN analysis was to mine patterns of diet-microbiome interactions. This goal does not require accurate forecasting of held-out microbial

growth rates to be useful, but we cannot recommend their legitimacy without further testing. On average, the DBN temporal warping network did not significantly outperform the baseline model in terms of mean absolute error on the held-out test dataset across all interactions (figure 5.4). This is likely due to factors such as an ill-specified modeling approach for forecasting, insufficient sampling data, and not enough processing time to complete the network inference with all species. Furthermore, earlier studies used multi-omic datasets with the inclusion of host genetics, metabolomics, and RNA-seq data, which could help give us a more complete picture of the complex interactions occuring.

Due to these limitations, the potential associations uncovered by the network are purely speculative and hypothesis-generating. We recommend a targeted in-vitro study in a controlled environment to verify associations if they are of interest. The DBN inference uncovered 184 significant conditional associations out of 2,656,614 hypothesis tests, highlighting potential interactions that may interest future research (figure 5.5). For example, the network model speculates that one strain of *Akkermansia* has a positive growth rate associated with a different strain of *Akkermansia*. This could be a misclassification by the database as they are both from the same species, or perhaps an ecological foundation that the proper environmental conditions promote their growth together. Another interaction highlighted by the network is that of *Barnesiella sp003150885*, which is positively associated with the Supplement MCT MOO. The original study reported no significant associations with the Supplement, so it could be possible that this unnamed species is a newer addition to the database.

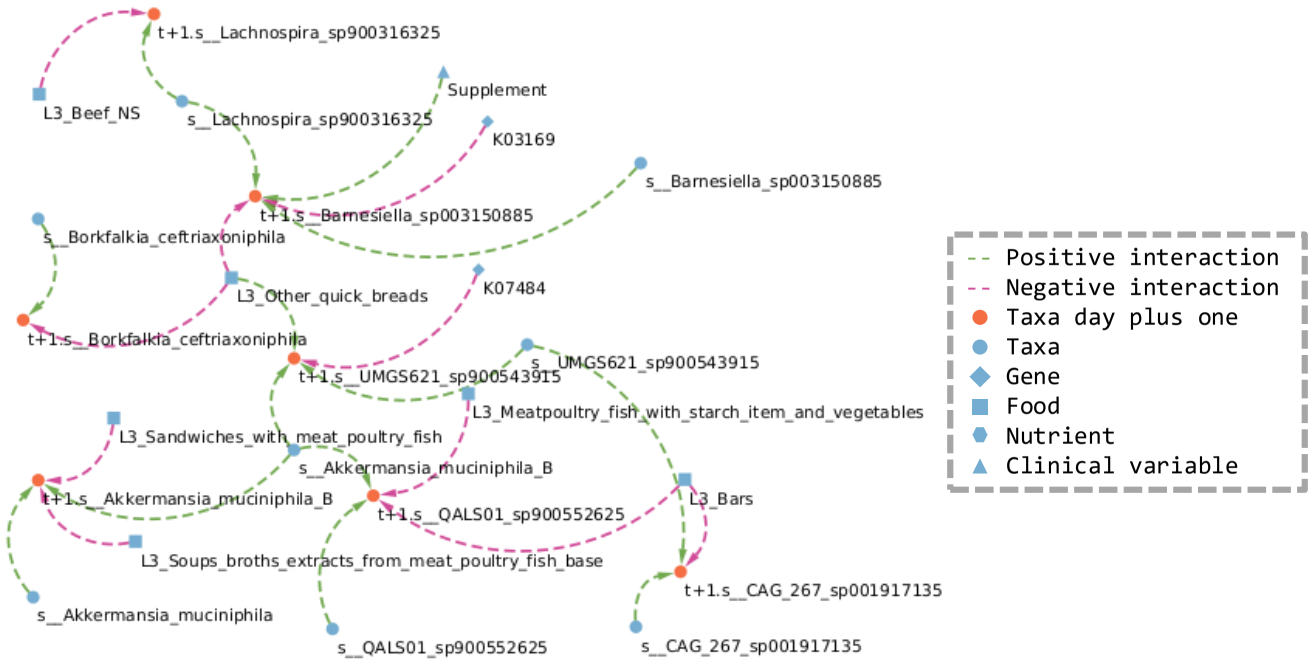

Figure 5.5: **Visualization of the diet-microbe associations hypothesized.** For simplicity of the network, only large connected component of the network is displayed. This component shows just a fraction of the possible associations. Positive conditional dependency edges are shown in red while negative conditional dependencies are shown in green. All previous day feature nodes are blue, while next day are orange. Shapes correspond to different feature types.

5.3.3 Conclusion

The original publication from which these data were derived demonstrated the efficacy of our published shallow-shotgun methods, as described in chapter 2 and chapter 3, to recover meaningful biological signals from human microbiome studies. Furthermore, the use of shallow-shotgun sequencing enabled an entirely new scale of microbiome study by reducing the cost of data generation by nearly a factor of ten. The original Johnson et al. [95] study was the first study to measure daily variation in the human microbiome at the species level, using shotgun metagenomics, and this would not have been possible without our earlier work on shallow-shotgun sequencing in chapter 2 and chapter 3. Our supplementary analysis of the dataset with a modern database with updated methods showed overall similar trends and patterns in diet-microbiome variability as originally discovered, and added a new finding: the association of nutrients with the microbiome profile. These reproduced results with an updated database and methodology showcase the robustness of the study findings originally presented. This should then give researchers confidence when applying our methods to their own metagenomic experiments.

Chapter 6

Concluding Remarks and Future Work

In chapter 2, we optimized the sequencing depth for shotgun microbiome datasets in cost, data efficiency, and computational accuracy. Our shallow-shotgun sequencing method allows for larger sample sizes and greater accuracy than currently used 16S and deep-shotgun sequencing methodologies. We showed that using a lower sequencing depth than what is commonly utilized is an effective way to receive accurate taxonomic and functional profiles of metagenomic experiments. Compared to 16S, this approach was more accurate, had less bias, and was able to identify species outside of the kingdom of Bacteria. Compared to deep-shotgun sequencing, we concluded that if the cost of the experiment is an issue, it is better to sequence more samples with a lower per-sample resolution than to sequence fewer samples with a higher per-sample resolution. In future work, we were concerned about the depth of sequencing and its interaction with the detection of rare species in taxonomic profiles that needed to be assessed. We completed this work in chapter 4. We expect that in the future more sequencing

experiments will be done using shallow- and deep-shotgun methodologies rather than 16S methodologies throughout the field of metagenomics.

In chapter 3, we proposed a computational pipeline `SHOGUN` for accurate taxonomic and functional profiling of shallow- and deep-shotgun sequencing. The software pipeline `SHOGUN` allows for simultaneous taxonomic and functional abundance profiling of metagenomics datasets with Bayesian redistribution of ambiguous mapping. The pipeline is flexible, allowing for user creation of a reference database and selecting the alignment tool that best fits a given user's data and computational resources. The package will enable users to efficiently transform quality-controlled sequences to abundance profiles consistently and accurately, thereby promoting reproducible microbiome research. While we have published the `SHOGUN` pipeline, we also continue the software development cycle with updates and added functionality. With our projection that microbiome experiments will have increased utilization of shotgun sequencing, so will the demand and use of shotgun sequencing processing pipelines. With this increased demand, more reference genomes will be assembled, requiring the software to include processes for database updates. The increased demand may also inspire better methodologies to analyze metagenomic profiles. The software pipeline `SHOGUN`'s flexibility, ease-of-use, updateable codebase, and ability to create a custom database makes it well situated to meet the growing demands of the field.

In chapter 4, we demonstrated the efficacy of a machine learning pipeline in classifying the presence-absence of genomes in metagenomic taxonomic profiling. By utilizing the features of an alignment and a reference genome database, we showed the efficacy of a machine learning pipeline to accurately identify the presence-absence of reference genomes in metagenomic taxonomic profiles. We

experimented with both simulated and real-world strain-level metagenomic sequencing datasets. The machine learning classifier will need to be retrained in future work as reference databases continue to change and grow. This study showed the ability of taxonomic profiles to be accurate in shotgun datasets and should give researchers a confident measurement of reference genome presence in shotgun taxonomic profiles

In chapter 5, we demonstrated the utility of the tools mentioned above on a complete metagenomic sequencing experiment. The taxonomic profiling methods were able to effectively plan and execute a metagenomic intervention study assessing the effect of medium-chain triglycerides (MCTs) on the composition and function of the human microbiome. The shallow-shotgun experiments were able to increase the sample size and allow for longitudinal sampling timelines while still identifying taxonomic and functional profiles for each sample. We inferred a dynamic Bayesian networks to assess the conditional dependence of diet, anthropomorphic variables, MCTs, genes, and taxonomies. A future study is warranted to evaluate the microbiome dynamics uncovered within the network.

6.1 Broader Impacts

The methods described herein will allow for more accurate studies of the microbiome, with the potential ability to increase sample sizes and larger more extensive population studies. The microbial world is extremely complex, and understanding its intricacies have impacts across a broad variety of fields including agriculture, climate change, and human health. Microbial communities are highly complex

and highly variable, necessitating large sample sizes to provide sufficient statistical power for hypothesis testing. Our proposed tools will improve research capacity and scientific accuracy of metagenomic analyses, which will increase the impact of research findings across this broad range of disciplines. Most importantly, these dramatic improvements in data efficiency will help democratize basic metagenomic research by allowing all microbiome researchers to obtain affordable but high-quality taxonomic and functional profiles. This will enable research labs with less funding or without access to a supercomputer to pursue high-impact research that spans discipline boundaries.

References

[1] National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC), 2007.

[2] Carmody K. McCalley, Ben J. Woodcroft, Suzanne B. Hodgkins, Richard A. Wehr, Eun-Hae Kim, Rhiannon Mondav, Patrick M. Crill, Jeffrey P. Chanton, Virginia I. Rich, Gene W. Tyson, and Scott R. Saleska. Methane dynamics regulated by microbial community response to permafrost thaw. *Nature*, 514(7523):478–481, October 2014.

[3] Steven R. Gill, Mihai Pop, Robert T. Deboy, Paul B. Eckburg, Peter J. Turnbaugh, Buck S. Samuel, Jeffrey I. Gordon, David A. Relman, Claire M. Fraser-Liggett, and Karen E. Nelson. Metagenomic analysis of the human distal gut microbiome. *Science (New York, N.Y.)*, 312(5778):1355–1359, June 2006.

[4] Edi Prifti and Jean-Daniel Zucker. The new science of metagenomics and the challenges of its use in both developed and developing countries. *arXiv:1305.2323 [q-bio]*, May 2013. arXiv: 1305.2323.

[5] Clémence Frioux, Dipali Singh, Tamas Korcsmaros, and Falk Hildebrand. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. *Computational and Structural Biotechnology Journal*, 18:1722–1734, January 2020.

[6] Quentin Le Bastard, Tonya Ward, Dimitri Sidiropoulos, Benjamin M. Hillmann, Chan Lan Chun, Michael J. Sadowsky, Dan Knights, and Emmanuel Montassier. Fecal microbiota transplantation reverses antibiotic and chemotherapy-induced gut dysbiosis in mice. *Scientific Reports*, 8(1):6219, April 2018.

[7] David A Muñiz Pedrogo, Jun Chen, Benjamin Hillmann, Patricio Jeraldo, Gabriel Al-Ghalith, Veena Taneja, John M Davis, III, Dan Knights, Heidi Nelson, William A Faubion, Laura Raffals, and Purna C Kashyap. An Increased Abundance of Clostridiaceae Characterizes Arthritis in Inflammatory Bowel Disease and Rheumatoid Arthritis: A Cross-sectional Study. *Inflammatory Bowel Diseases*, 25(5):902–913, April 2019.

[8] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[9] Ran Blekhman, Julia K. Goodrich, Katherine Huang, Qi Sun, Robert Bukowski, Jordana T. Bell, Timothy D. Spector, Alon Keinan, Ruth E. Ley, Dirk Gevers, and Andrew G. Clark. Host genetic variation impacts microbiome composition across human body sites. *Genome Biology*, 16(1):191, September 2015.

[10] Sambhawa Priya and Ran Blekhman. Population dynamics of the human gut microbiome: change is the only constant. *Genome Biology*, 20(1):150, July 2019.

[11] María X. Maldonado-Gómez, Inés Martínez, Francesca Bottacini, Amy O'Callaghan, Marco Ventura, Douwe van Sinderen, Benjamin Hillmann, Pajau Vangay, Dan Knights, Robert W. Hutkins, and Jens Walter. Stable Engraftment of Bifidobacterium longum AH1206 in the Human Gut Depends on Individualized Features of the Resident Microbiome. *Cell Host & Microbe*, 20(4):515–526, October 2016.

[12] Peter J. Turnbaugh, Ruth E. Ley, Michael A. Mahowald, Vincent Magrini, Elaine R. Mardis, and Jeffrey I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–131, December 2006.

[13] Tonya L. Ward, Emmanuel Montassier, Heather Lekatz, Bradley Schmidt, Michael Camilleri, Benjamin M. Hillmann, Madhusudan Grover, Gianrico Farrugia, Dan Knights, and Purna C. Kashyap. 262 Longitudinal Changes in Gut Microbiota in Patients With Irritable Bowel Syndrome. *Gastroenterology*, 150(4):S64, April 2016.

[14] Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I. Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, May 2012.

[15] E. Montassier, E. Batard, T. Gastinne, G. Potel, and M. F. de La Cochetière. Recent changes in bacteremia in patients with cancer: a systematic review of epidemiology and antibiotic resistance. *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology*, 32(7):841–850, July 2013.

[16] Diet rapidly and reproducibly alters the human gut microbiome | Nature.

[17] Francesco Durazzi, Claudia Sala, Gastone Castellani, Gerardo Manfreda, Daniel Remondini, and Alessandra De Cesare. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, 11(1):3030, February 2021. Number: 1 Publisher: Nature Publishing Group.

[18] Gabriel A. Al-Ghalith, Benjamin Hillmann, Kaiwei Ang, Robin Shields-Cutler, and Dan Knights. SHI7 Is a Self-Learning Pipeline for Multipurpose Short-Read DNA Quality Control. *mSystems*, 3(3), June 2018.

[19] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[20] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012.

[21] Duy Tin Truong, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, October 2015.

[22] Alessio Milanese, Daniel R. Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, Renato Alves, Paul I. Costea, Luis Pedro Coelho, Thomas S. B. Schmidt, Alexandre Almeida, Alex L. Mitchell, Robert D. Finn, Jaime Huerta-Cepas, Peer Bork, Georg Zeller, and Shinichi Sunagawa. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications*, 10(1):1014, March 2019. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Microbiome;Software Subject_term_id: microbiome;software.

[23] Derrick E. Wood and Steven L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, March 2014.

[24] Daehwan Kim, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, October 2016.

[25] Eric A. Franzosa, Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, Rob Knight, J. Gregory Caporaso, Nicola Segata, and Curtis Huttenhower. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15(11):962–968, November 2018. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 11 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Metagenomics;Software Subject_term_id: metagenomics;software.

[26] Gavin M. Douglas, Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*, 38(6):685–688, June 2020. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Biomarkers;Computational biology and bioinformatics;Ecology;Microbiology Subject_term_id: biomarkers;computational-biology-and-bioinformatics;ecology;microbiology.

[27] Fredrik H. Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, June 2013.

[28] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. Conference Name: The Bell System Technical Journal.

[29] E. H. Simpson. Measurement of Diversity. *Nature*, 163(4148):688–688, April 1949. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 4148 Primary_atype: Research Publisher: Nature Publishing Group.

[30] C. Lozupone and R. Knight. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, December 2005.

[31] John C. Gower. Principal Coordinates Analysis. In *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470011815.b2a13070.

[32] Dan Knights, Elizabeth K. Costello, and Rob Knight. Supervised classification of human microbiota. *FEMS Microbiology Reviews*, 35(2):343–359, March 2011.

[33] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai, Jordan E. Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J. Brislawn, C. Titus Brown, Benjamin J. Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily Cope, Ricardo Da Silva, Pieter C. Dorrestein, Gavin M. Douglas, Daniel M. Durall, Claire Duvallet, Christian F. Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M. Gauglitz, Deanna L. Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin Huttley, Stefan Janssen, Alan K. Jarmusch, Lingjing Jiang, Benjamin Kaehler, Kyo Bin Kang, Christopher R. Keefe, Paul Keim, Scott T. Kelley, Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan GI Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D. Martin, Daniel McDonald, Lauren J. McIver, Alexey V. Melnik, Jessica L. Metcalf, Sydney C. Morgan, Jamie Morton, Ahmad Turan Naimey, Jose A. Navas-Molina, Louis Felix Nothias, Stephanie B. Orchanian, Talima Pearson, Samuel L.

Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, I. I. Michael S Robeson, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R. Spear, Austin D. Swafford, Luke R. Thompson, Pedro J. Torres, Pauline Trinh, Anupriya Tripathi, Peter J. Turnbaugh, Sabah Ul-Hasan, Justin JJ van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C. Weber, Chase HD Williamson, Amy D. Willis, Zhenjiang Zech Xu, Jesse R. Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J. Gregory Caporaso. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. Technical Report e27295v2, PeerJ Preprints, December 2018.

[34] Benjamin Hillmann, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Qiyun Zhu, Daryl M. Gohl, Kenneth B. Beckman, Rob Knight, and Dan Knights. Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems*, 3(6), October 2018.

[35] Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, Anupriya Tripathi, Sean M. Gibbons, Gail Ackermann, Jose A. Navas-Molina, Stefan Janssen, Evguenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T. Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingjing Jiang, Mohamed F. Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciolek, Nicholas A. Bokulich, Joshua Lefler, Colin J. Brislawn, Gregory Humphrey, Sarah M. Owens, Jarrad Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A. Fuhrman, Aaron Clauset, Rick L. Stevens,

Ashley Shade, Katherine S. Pollard, Kelly D. Goodwin, Janet K. Jansson, Jack A. Gilbert, Rob Knight, and Earth Microbiome Project Consortium. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, 2017.

[36] Connor R. Fitzpatrick, Julia Copeland, Pauline W. Wang, David S. Guttman, Peter M. Kotanen, and Marc T. J. Johnson. Assembly and ecological function of the root microbiome across angiosperm plant species. *Proceedings of the National Academy of Sciences*, page 201717617, January 2018.

[37] Brajesh K. Singh, Richard D. Bardgett, Pete Smith, and Dave S. Reay. Microorganisms and climate change: terrestrial feedbacks and mitigation options. *Nature Reviews Microbiology*, 8(11):779–790, November 2010.

[38] Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, 12(9):635–645, September 2014.

[39] Morgan G I Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron E Burkepile, Rebecca L Vega Thurber, Rob Knight, Robert G Beiko, and Curtis Huttenhower. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9):814–821, August 2013.

[40] Shoko Iwai, Thomas Weinmaier, Brian L. Schmidt, Donna G. Albertson, Neil J. Poloso, Karim Dabbagh, and Todd Z. DeSantis. Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes. *PloS One*, 11(11):e0166104, 2016.

[41] Chengwei Luo, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J. Xavier, and Dirk Gevers. ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology*, 33(10):1045–1052, October 2015. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Data mining;Genome informatics;Metagenomics Subject_term_id: data-mining;genome-informatics;metagenomics.

[42] Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, November 2014.

[43] H. Bjørn Nielsen, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, Laurent Gautier, Anders G. Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, Ida Bonde, Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo B. Quintanilha dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer S. Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf S. Kaas, Sean Kennedy, Karsten Kristiansen, Jens Roat Kultima, Pierre Léonard, Florence Levenez, Ole Lund, Bouziane

Moumen, Denis Le Paslier, Nicolas Pons, Oluf Pedersen, Edi Prifti, Junjie Qin, Jeroen Raes, Søren Sørensen, Julien Tap, Sebastian Tims, David W. Ussery, Takuji Yamada, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, and S. Dusko Ehrlich. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828, August 2014. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetic variation;Genome assembly algorithms;Microbial genetics;Time series Subject_term_id: genetic-variation;genome-assembly-algorithms;microbial-genetics;time-series.

[44] Marcus B. Jones, Sarah K. Highlander, Ericka L. Anderson, Weizhong Li, Mark Dayrit, Niels Klitgord, Martin M. Fabani, Victor Seguritan, Jessica Green, David T. Pride, Shibu Yooseph, William Biggs, Karen E. Nelson, and J. Craig Venter. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences of the United States of America*, 112(45):14024–14029, November 2015.

[45] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(Database issue):D109–114, January 2012. tex.ids: kanehisaKEGGIntegrationInterpretation2012.

[46] Gabriel Al-Ghalith and Benjamin Hillmann. knights-lab/UTree: UTree 2.0 SigNature Edition: SHOGUN release. Technical report, Zenodo, September 2017.

[47] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970.

[48] Gabriel Al-Ghalith and Dan Knights. knights-lab/BURST: BURST v0.99.4a. Technical report, Zenodo, July 2017.

[49] Tatiana Tatusova, Stacy Ciufo, Boris Fedorov, Kathleen O'Neill, and Igor Tolstoy. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Research*, 42(Database issue):D553–559, January 2014.

[50] Benjamin Hillmann, Gabriel A Al-Ghalith, Robin R Shields-Cutler, Qiyun Zhu, Rob Knight, and Dan Knights. SHOGUN: a modular, accurate and scalable framework for microbiome quantification. *Bioinformatics*, 36(13):4088–4090, July 2020.

[51] Daniel McDonald, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6(3):610–618, March 2012.

[52] Nils Homer. DWGSIM: Whole Genome Simulator for Next-Generation Sequencing, July 2017. original-date: 2011-10-19T00:19:14Z.

[53] Lyndsay Radnedge, Peter G. Agron, Karen K. Hill, Paul J. Jackson, Lawrence O. Ticknor, Paul Keim, and Gary L. Andersen. Genome Differences That Distinguish Bacillus anthracis from Bacillus cereus and Bacillus

thuringiensis. *Applied and Environmental Microbiology*, 69(5):2755–2764, May 2003.

[54] Fredrik Karlsson, Valentina Tremaroli, Jens Nielsen, and Fredrik Backhed. Assessing the Human Gut Microbiota in Metabolic Diseases. *Diabetes*, 62(10):3341–3349, October 2013.

[55] Michael Tessler, Johannes S. Neumann, Ebrahim Afshinnekoo, Michael Pineda, Rebecca Hersch, Luiz Felipe M. Velho, Bianca T. Segovia, Fabio A. Lansac-Toha, Michael Lemke, Rob DeSalle, Christopher E. Mason, and Mercer R. Brugler. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*, 7(1):6589, July 2017.

[56] Joshua N. Burton, Ivan Liachko, Maitreya J. Dunham, and Jay Shendure. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda, Md.)*, 4(7):1339–1346, May 2014.

[57] Rachel Poretsky, Luis M. Rodriguez-R, Chengwei Luo, Despina Tsementzi, and Konstantinos T. Konstantinidis. Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLOS ONE*, 9(4):e93827, April 2014.

[58] Felipe da Veiga Leprevost, Björn A. Grüning, Saulo Alves Aflitos, Hannes L. Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, Mingze Bai, Rafael C. Jimenez, Timo Sachsenberg, Julianus Pfeuffer, Roberto Vera Alvarez, Johannes Griss, Alexey I. Nesvizhskii, and Yasset Perez-Riverol. BioContainers: an open-source and

community-driven framework for software standardization. *Bioinformatics (Oxford, England)*, 33(16):2580–2582, August 2017.

[59] Benjamin Hillmann and Dan Knights. knights-lab/SHOGUN: Release 1.0.0. Technical report, Zenodo, September 2017.

[60] George Bernard Dantzig and D. R. Fulkerson. On the Max Flow Min Cut Theorem of Networks. Technical report, RAND Corporation, January 1955.

[61] Gabriel A. Al-Ghalith, Emmanuel Montassier, Henry N. Ward, and Dan Knights. NINJA-OPS: Fast Accurate Marker Gene Alignment Using Concatenated Ribosomes. *PLOS Computational Biology*, 12(1):e1004658, January 2016. Publisher: Public Library of Science.

[62] Rachid Ounit and Stefano Lonardi. Higher Classification Accuracy of Short Metagenomic Reads by Discriminative Spaced k-mers. In Mihai Pop and Hélène Touzet, editors, *Algorithms in Bioinformatics*, number 9289 in Lecture Notes in Computer Science, pages 286–295. Springer Berlin Heidelberg, September 2015.

[63] Jennifer Lu, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3:e104, January 2017.

[64] Alex Bateman, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Andrew Cowley, Alan Da Silva, Maurizio De Giorgi, Tunca Dogan, Francesco Fazzini, Leyla Garcia Castro, Luis Figueira, Penelope Garmiri, George Georghiou, Daniel Gonzalez, Emma Hatton-Ellis, Weizhong Li, Wudong Liu, Rodrigo

Lopez, Jie Luo, Yvonne Lussi, Alistair MacDougall, Andrew Nightingale, Barbara Palka, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Alexandre Renaux, Steven Rosanoff, Rabie Saidi, Tony Sawford, Aleksandra Shypitsyna, Elena Speretta, Edward Turner, Nidhi Tyagi, Vladimir Volynkin, Tony Wardell, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casal-Casas, Edouard de Castro, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Nevila Nouspikel, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh, and Jian Zhang. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, January 2017.

[65] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14):2068–2069, July 2014.

[66] NCBI Staff. Introducing the New Human Genome Assembly: GRCh38, December 2013.

[67] Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804–810, October 2007.

[68] Stinus Lindgreen, Karen L. Adair, and Paul P. Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6:19233, January 2016.

[69] Thomas J. Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5, 2014. Publisher: Frontiers.

[70] Nidhi Shah, Erin K Molloy, Mihai Pop, and Tandy Warnow. TIPP2: metagenomic taxonomic profiling using phylogenetic markers. *Bioinformatics*, (btab023), January 2021.

[71] Yong-Xin Liu, Yuan Qin, Tong Chen, Meiping Lu, Xubo Qian, Xiaoxuan Guo, and Yang Bai. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & Cell*, 12(5):315–330, May 2021.

[72] Ahmed Sobih, Alexandru I. Tomescu, and Veli Makinen. MetaFlow: Metagenomic Profiling Based on Whole-Genome Coverage Analysis with Min-Cost Flows. pages 111–121, April 2016.

[73] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, March 2017.

[74] Victoria Pascal, Marta Pozuelo, Natalia Borruel, Francesc Casellas, David Campos, Alba Santiago, Xavier Martinez, Encarna Varela, Guillaume Sarrabayrouse, Kathleen Machiels, Severine Vermeire, Harry Sokol, Francisco Guarner, and Chaysavanh Manichanh. A microbial signature for Crohn's disease. *Gut*, 66(5):813–822, May 2017. Publisher: BMJ Publishing Group Section: Inflammatory bowel disease.

[75] Siddhartha Mandal, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26:27663, 2015.

[76] Florent Lassalle, Matteo Spagnoletti, Matteo Fumagalli, Liam Shaw, Mark Dyble, Catherine Walker, Mark G. Thomas, Andrea Bamberg Migliano, and Francois Balloux. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular Ecology*, 27(1):182–195, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.14435.

[77] J. R. Bedarf, F. Hildebrand, L. P. Coelho, S. Sunagawa, M. Bahram, F. Goeser, P. Bork, and U. Wüllner. Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. *Genome Medicine*, 9(1):39, April 2017.

[78] Jakob Wirbel, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S. Fleck, Anita Y. Voigt, Albert Palleja, Ruby Ponnudurai, Shinichi Sunagawa, Luis Pedro Coelho, Petra Schrotz-King, Emily Vogtmann, Nina Habermann, Emma Niméus, Andrew M. Thomas, Paolo Manghi, Sara Gandini, Davide Serrano, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Tatsuhiro Shibata, Shinichi Yachida, Takuji Yamada, Levi Waldron, Alessio Naccarati, Nicola Segata, Rashmi Sinha, Cornelia M. Ulrich, Hermann Brenner, Manimozhiyan Arumugam, Peer Bork, and Georg Zeller. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, 25(4):679–689, April 2019. Number: 4 Publisher: Nature Publishing Group.

[79] Volkan Sevim, Juna Lee, Robert Egan, Alicia Clum, Hope Hundley, Janey Lee, R. Craig Everroad, Angela M. Detweiler, Brad M. Bebout, Jennifer Pett-Ridge, Markus Göker, Alison E. Murray, Stephen R. Lindemann, Hans-Peter Klenk, Ronan O'Malley, Matthew Zane, Jan-Fang Cheng, Alex Copeland, Christopher Daum, Esther Singer, and Tanja Woyke. Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Scientific Data*, 6(1):285, November 2019. Number: 1 Publisher: Nature Publishing Group.

[80] James J. Kozich, Sarah L. Westcott, Nielson T. Baxter, Sarah K. Highlander, and Patrick D. Schloss. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology*, 79(17):5112–5120, September 2013.

[81] Denis Bertrand, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M. Senthil Kumar, Chenhao Li, Mirta Dvornicic, Janja Paliska Soldo, Jia Yu Koh, Chengxuan Tong, Oon Tek Ng, Timothy Barkham, Barnaby Young, Kalisvar Marimuthu, Kern Rei Chng, Mile Sikic, and Niranjan Nagarajan. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology*, 37(8):937–944, August 2019. Number: 8 Publisher: Nature Publishing Group.

[82] Esther Singer, Bill Andreopoulos, Robert M. Bowers, Janey Lee, Shweta Deshpande, Jennifer Chiniquy, Doina Ciobanu, Hans-Peter Klenk, Matthew Zane, Christopher Daum, Alicia Clum, Jan-Fang Cheng, Alex Copeland, and Tanja Woyke. Next generation sequencing data of a defined microbial mock community. *Scientific Data*, 3(1):1–8, September 2016.

[83] Samuel M Nicholls, Joshua C Quick, Shuiquan Tang, and Nicholas J Loman. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience*, 8(5), May 2019.

[84] Kelvin Li, Monika Bihan, Shibu Yooseph, and Barbara A. Methé. Analyses of the Microbial Diversity across the Human Microbiome. *PLOS ONE*, 7(6):e32118, June 2012. Publisher: Public Library of Science.

[85] Gavin Turner, Madeline Smith, Andrea L. Hoeschen, Jessica A. Wilson, Jessica Kennedy, Max Abramson, Qing Cao, Najla El Jurdi, Margaret L. MacMillan, Daniel J. Weisdorf, Bruce R. Blazar, Alexander Khoruts, Robin R. Shields-Cutler, Dan Knights, Shernan G.

Holtan, and Armin Rashidi. Shotgun sequencing of the faecal microbiome to predict response to steroids in patients with lower gastrointestinal acute graft-versus-host disease: An exploratory analysis. *British Journal of Haematology*, 192(3):e69–e73, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjh.17238.

[86] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[87] Trang T Le, Weixuan Fu, and Jason H Moore. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1):250–256, January 2020.

[88] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[89] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[90] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, January 2020.

Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computer science;Medical research;Software Subject_term_id: computer-science;medical-research;software.

[91] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432, March 2015. Publisher: Public Library of Science.

[92] Donovan H. Parks, Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, 38(9):1079–1086, September 2020. Number: 9 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Archaeal genomics;Bacterial genomics;Bioinformatics;Taxonomy Subject_term_id: archaea-genomics;bacterial-genomics;bioinformatics;taxonomy.

[93] Donovan H. Parks, Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996–1004, November 2018. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Bacteria;Phylogenetics;Taxonomy Subject_term_id: bacteria;phylogenetics;taxonomy.

[94] Pierre-Alain Chaumeil, Aaron J. Mussig, Philip Hugenholtz, and Donovan H. Parks. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics (Oxford, England)*, November 2019.

[95] Abigail J. Johnson, Pajau Vangay, Gabriel A. Al-Ghalith, Benjamin M. Hillmann, Tonya L. Ward, Robin R. Shields-Cutler, Austin D. Kim, Anna Konstantinovna Shmagel, Arzang N. Syed, Jens Walter, Ravi Menon, Katie Koecher, and Dan Knights. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host & Microbe*, 25(6):789–802.e5, June 2019.

[96] Missing Microbes: How the Overuse of Antibiotics Is Fueling Our Modern Plagues.

[97] Johan S. Bakken, Thomas Borody, Lawrence J. Brandt, Joel V. Brill, Daniel C. Demarco, Marc Alaric Franzos, Colleen Kelly, Alexander Khoruts, Thomas Louie, Lawrence P. Martinelli, Thomas A. Moore, George Russell, and Christina Surawicz. Treating Clostridium difficile Infection With Fecal Microbiota Transplantation. *Clinical Gastroenterology and Hepatology*, 9(12):1044–1049, December 2011. Publisher: Elsevier.

[98] Magdy El-Salhy and Tarek Mazzawi. Fecal microbiota transplantation for managing irritable bowel syndrome. *Expert Review of Gastroenterology & Hepatology*, 12(5):439–445, May 2018.

[99] C. Blanchaert, B. Strubbe, and H. Peeters. Fecal microbiota transplantation in ulcerative colitis. *Acta Gastro-Enterologica Belgica*, 82(4):519–528, December 2019.

[100] Janko Tackmann, João Frederico Matias Rodrigues, and Christian von Mering. Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data. *Cell Systems*, 9(3):286–296.e8, September 2019.

[101] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, August 2012.

[102] Daniel Ruiz-Perez, Jose Lugo-Martinez, Natalia Bourguignon, Kalai Mathee, Betiana Lerner, Ziv Bar-Joseph, and Giri Narasimhan. Dynamic Bayesian Networks for Integrating Multi-omics Time Series Microbiome Data. *mSystems*, 6(2):e01105–20. Publisher: American Society for Microbiology.

[103] Jose Lugo-Martinez, Daniel Ruiz-Perez, Giri Narasimhan, and Ziv Bar-Joseph. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, 7(1):54, April 2019.

[104] Michael J. McGeachie, Joanne E. Sordillo, Travis Gibson, George M. Weinstock, Yang-Yu Liu, Diane R. Gold, Scott T. Weiss, and Augusto Litonjua. Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. *Scientific Reports*, 6:20359, February 2016.

[105] Monica Steffi Matchado, Michael Lauber, Sandra Reitmeier, Tim Kacprowski, Jan Baumbach, Dirk Haller, and Markus List. Network analysis methods for studying microbial communities: A mini review. *Computational and Structural Biotechnology Journal*, 19:2687–2698, May 2021.

[106] Rasnik K. Singh, Hsin-Wen Chang, Di Yan, Kristina M. Lee, Derya Ucmak, Kirsten Wong, Michael Abrouk, Benjamin Farahnik, Mio Nakamura,

Tian Hao Zhu, Tina Bhutani, and Wilson Liao. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*, 15:73, April 2017.

[107] Tanja V. Maier, Marianna Lucio, Lang Ho Lee, Nathan C. VerBerkmoes, Colin J. Brislawn, Jörg Bernhardt, Regina Lamendella, Jason E. McDermott, Nathalie Bergeron, Silke S. Heinzmann, James T. Morton, Antonio González, Gail Ackermann, Rob Knight, Katharina Riedel, Ronald M. Krauss, Philippe Schmitt-Kopplin, and Janet K. Jansson. Impact of Dietary Resistant Starch on the Human Gut Microbiome, Metaproteome, and Metabolome. *mBio*, 8(5):e01343–17, October 2017.

[108] Jethro S. Johnson, Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1):1–11, November 2019.

[109] Marco Scutari. Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimized Implementations in the bnlearn R Package. *Journal of Statistical Software*, 77:1–20, March 2017.

[110] Marco Scutari. Learning Bayesian Networks with the bnlearn R Package. *arXiv:0908.3817 [stat]*, July 2010. arXiv: 0908.3817.

[111] Carlos P. Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12):5825–5829, December 2021.

[112] Pajau Vangay, Abigail J. Johnson, Tonya L. Ward, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Benjamin M. Hillmann, Sarah K. Lucas, Lalit K. Beura, Emily A. Thompson, Lisa M. Till, Rodolfo Batres, Bwei Paw, Shannon L. Pergament, Pimpanitta Saenyakul, Mary Xiong, Austin D. Kim, Grant Kim, David Masopust, Eric C. Martens, Chaisiri Angkurawaranon, Rose McGready, Purna C. Kashyap, Kathleen A. Culhane-Pera, and Dan Knights. US Immigration Westernizes the Human Gut Microbiome. *Cell*, 175(4):962–972.e10, November 2018.

[113] Yikyung Park, Kevin W Dodd, Victor Kipnis, Frances E Thompson, Nancy Potischman, Dale A Schoeller, David J Baer, Douglas Midthune, Richard P Troiano, Heather Bowles, and Amy F Subar. Comparison of self-reported dietary intakes from the Automated Self-Administered 24-h recall, 4-d food records, and food-frequency questionnaires against recovery biomarkers. *The American Journal of Clinical Nutrition*, 107(1):80–93, January 2018.

[114] Zheng Sun, Shi Huang, Meng Zhang, Qiyun Zhu, Niina Haiminen, Anna Paola Carrieri, Yoshiki Vázquez-Baeza, Laxmi Parida, Ho-Cheol Kim, Rob Knight, and Yang-Yu Liu. Challenges in benchmarking metagenomic profilers. *Nature Methods*, 18(6):618–626, June 2021.

[115] Gregory B. Gloor, Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology*, 26(5):322–329, May 2016.

[116] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, and Peer Bork.

eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, January 2019.

[117] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. page 64.

[118] seaborn.barplot — seaborn 0.10.0 documentation.

[119] Tukey: Exploratory data analysis - Google Scholar.