Title

Exploration of whole-genome data of vertebrate gut microbiome reveals dynamics of strain phylosymbiosis

Author

Chiu, Jeffrey Huey-Chuan

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Exploration of whole-genome data of vertebrate gut microbiome reveals dynamics of strain phylosymbiosis

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Biology

by

Jeffrey Huey-Chuan Chiu

Committee in charge:

Professor Rob Knight, Chair
Professor Eric Allen, Co-chair
Professor Aspen Reese

2021

The Thesis of Jeffrey Huey-Chuan Chiu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I want to thank my parents for the non-trivial task of raising me. In a world where most have so little, I had the privilege to grow up in a nourishing, open, and supportive environment. I also want to acknowledge my family and friends that have pushed me on during the ups and help me find strength during the downs. Special shout out to Bob Kucer, who encouraged my early scientific interest, and Isaac Castillo, who helped me troubleshoot various mind-boggling computational obstacles. I will forever be indebted to this community I call home.

I would like to acknowledge my PI and chair of my committee, Dr. Rob Knight, for giving me a chance to work with this terrific group of scientists at the Knight lab. This work would not have been possible without his honest, critical, and clear-eyed guidance. I would also like to acknowledge my co-chair, Dr. Eric Allen, who has been a source of inspiration and pillar of support. Last but not least, I want to acknowledge Dr. Aspen Reese, whose detailed and useful feedback elevated the quality of my work as well as directed me to the research behind evolutionary ecology.

I also want to acknowledge my colleagues' support. Thank you Dr. Se jin Song for hatching the basic idea underlying my research topic, directing me to the data available for this study, and being a source of guidance. Thank you Victor Cantu for the many conceptual introductions to topics within microbiome science, your help with editing this thesis, and, most importantly, your friendship. Thank you Dr. Justin Shaffer for the critiques on best research practices, and technical assistance regarding R and bowtie2. Thank you, Dr. Holly Lutz and Dr.

Smruthi Karthikeyan for your uplifting positivity and many useful scientific discussions. Lastly, thank you Jon Sanders for sharing your scientific and technical insights regarding evolutionary studies. Other notable colleagues that have helped me overcome some obstacles or taught me something new include Jeff De Reus, George Armstrong, Dr. Qiyun Zhu, Cameron Martino, and Dr. Shi Huang.

This acknowledgment section would be incomplete if I did not give credit to a couple of online resources that have helped me tremendously with the technical aspect of this project. Shout out to Pat Schloss, whose Riffomonas Youtube channel is a goldmine for beginners trying to learn R, and Rob Edwards, whose Youtube channel is a great gateway to understand the NGS technologies and microbiome science.

Material from this thesis is currently being prepared for submission for publication of the material. Chiu, Jeffrey H.; Song, Se J.; Cantu, Victor; Shaffer, Justin; Lutz, Holly L.; Knight, Rob. The thesis author was the primary author of this material.

VITA

2019	Bachelor of Science, University of California San Diego
2016-2019	Instructional Assistant, University of California San Diego
2021	Master of Science, University of California San Diego
2019-2021	Teaching Assistant, University of California San Diego
2016-2021	Research Assistant, University of California San Diego

FIELDS OF STUDY

Major Field: Biology

Minor: Computer Science

Studies in Biology
Dr. Rob Knight

ABSTRACT OF THE THESIS

Exploration of whole-genome data of vertebrate gut microbiome reveals dynamics of strain phylosymbiosis

by

Jeffrey Huey-Chuan Chiu

Master of Science in Biology

University of California San Diego, 2021

Professor Rob Knight, Chair
Professor Eric Allen, Co-Chair

The gut microbiota is a complex community of microbial species inhabiting the digestive tract. Each microbial species is further composed of microbes with slightly different genetic variants also known as strains. While most evolutionary studies of the gut microbiome occur at the community level or focus on narrow clades of vertebrates, few studies have examined the evolution of wildlife and their gut microbiome at the strain level across the animal kingdom. In this exploratory study, we examine a wildlife gut metagenomic dataset to investigate the evolutionary dynamics of bacterial species and their respective host. In particular, this is the first examination of whether there is significant congruence between the phylogeny of bacterial strains and that of their respective hosts, which we refer to as strain phylosymbiosis, across the animal kingdom. Our analysis of the most abundant bacteria in our dataset revealed *Akkermansia muciniphila* and *Bacteroides vulgatus* exhibited strong signals of strain phylosymbiosis.

INTRODUCTION

The advent of next-generation sequencing (NGS) has significantly expanded our ability to investigate the microbial world (1). Such technologies have revolutionized the field of microbiology from clinical applications to environmental studies by allowing us to efficiently sequence unculturable microbes and analyze entire microbial communities (2-3). With such advances, researchers have discovered the vast, once unknown, number and diversity of microbes in our gut, learned how our behavior (what we eat, antibiotic use, FMT, how we give birth, etc.) affects the make-up or composition of our gut microbiome and uncovered the links between various chronic diseases and microbes (4-6).

We are continuously learning about the intricate role microbes play in host fitness. In humans, microbes affect not just our metabolic and digestive systems, but also our nervous and immune systems (7-9). In the wild, although we have known about the various symbiotic relationships between microbial species and their insect hosts as far back as 1965, only in recent decades has it been possible to uncover the relationships between microbial species and vertebrates, which usually carries much larger, and more diverse microbial community than that of insects (10, 11). For example, researchers have found sanguivorous, or blood-feeding, organisms such as sea lamprey, leeches, mosquitoes, and blood-feeding bats all possessing digestive tracts dominated by *Aeromonas spp.* (12). This shared association in sanguivores from distant host lineages highlights *Aeromonas spp*.'s important role in the digestion of blood for the energy requirement of its host. As another example, one study examining the microbiome of the American alligator has found the alligator's core gut microbiome to be uniquely enriched with Fusobacteria while low in Bacteroidetes and Proteobacteria, forming a core microbiome distinct from that of mammalian, avian, and other reptilian guts (13). The authors speculate that

Fusobacteria interacts closely with the American alligator's immune system as well as serves important roles in the animal's nutrient acquisition and organ development. With increasing evidence of microbial communities playing important roles in host fitness beyond humans, biologists have begun exploring the evolutionary dynamics of and partnership between microbial communities and their respective host species (14).

Early studies of the wildlife gut microbiome have hypothesized that consistent vertical transmission (passing of microbes from mother to child within a host species) of the microbial community could result in co-diversification of the host and its microbial community (14). Not to be confused with co-diversification (also sometimes referred to as co-evolution) of host and specific symbionts which implies parallel evolutionary changes reflected in each partner's genetics, co-diversification of host and microbial community does not involve reciprocal evolutionary changes, since the microbial community does not possess a shared genome. Instead, the diversification of the microbial community is mainly driven by ecological processes of microbial dispersal (ways in which microbes and hosts come in contact) and microbial selection (ways in which microbes and hosts select for each other) that determine the community assemblage (15). Over time, consistent co-diversification could lead to signals of phylosymbiosis, which is the overall congruence between the host phylogeny measured in evolutionary relatedness and microbial community phylogeny measured in ecological relatedness (16). Put more simply, phylosymbiosis examines if the makeup of microbial communities is more similar between more closely related host species (e.g. tiger & lion) than between two more distantly related host species (e.g. tiger & cuttlefish).

To dissect the evolutionary dynamics of phylosymbiosis, many biologists have traditionally interrogated 16S amplicon sequences of wildlife microbiomes. Amplicon

sequencing involves the amplification and sequencing of a usually universal, yet slightly variable genomic region for the identification of bacteria, and therefore, community-level characterization of the microbiome. A recent review by Mallot & Amato condensing various studies on host specificity of microbiome across vertebrates has found that more "primitive" classes of organisms such as Insect, Anthozoa (Sea anemones, corals, etc.), and Actinopterygii (bony fishes) with low microbial richness (<100 ASVs) exhibit stronger signals of phylosymbiosis, while more "complex" classes of organisms such as Amphibia, Reptilia, Aves with higher microbial richness (>100 ASVs) exhibit weaker signals of phylosymbiosis (15). This general trend could be explained by it being harder for a group of bacteria co-evolving with a host to influence the composition of a richer community than a less rich community. However, the notable exception to this trend is the microbiome within the Mammalia class which generally exhibits strong signals of phylosymbiosis while having rich microbial communities. They propose that Mammalian traits such as milk feeding, viviparous birth, and parental care, all of which facilitate vertical transmission of core microbial communities, are key factors in elevating signals of phylosymbiosis in Mammals (15). Overall, the host specificity of the gut microbiome appears to be weaker in taxonomically richer gut microbial communities but stronger in vertically transmitted microbial communities.

Researchers have also investigated the co-speciation (co-diversification could lead to parallel speciation) of certain bacteria lineages and their associated hosts (17). A study by Moeller et al. has explored the co-speciation of specific bacterial strains within the great ape microbiome by examining the amplicon sequence of the gyrase B gene. Unlike the 16S rRNA gene, the Gyrase B gene is more variable and enables robust resolution of closely related bacterial species and strains in the host gut microbiome. By comparing phylogenies and

calculating divergence time of wild apes and their associated bacterial species, the authors found strong evidence for co-speciation of Bacteroidaceae and Bifidobacteriacea family of bacteria with hominids that started millions of years ago.

Previous studies of phylosymbiosis and co-diversification have inspired us to wonder if we can detect signals of phylosymbiosis at the strain level not just in great apes but across the animal kingdom. Could we find a bacterial species where the phylogeny of its strain and phylogeny of each strain's respective host correlate? A significant positive correlation would suggest strains from a bacterial species are more similar between more closely related host species (e.g. tiger & lion) than between two more distantly related host species (e.g. tiger & cuttlefish). We refer to this evolutionary dynamic as strain phylosymbiosis. Unlike phylosymbiosis that compares the ecological relatedness of the microbial community and the evolutionary relatedness of its respective host, strain phylosymbiosis compares the genetic relatedness between strains and the evolutionary relatedness of each strain's respective host.

Furthermore, since shotgun metagenomic (or whole-genome shotgun) data are becoming increasingly prevalent, we sought to investigate strain phylosymbiosis using shotgun metagenomic data. As opposed to various amplicon data that contain sequences of a particularly useful marker such as 16S rRNA or Gyrase B, shotgun metagenomic data contains the entire collection of genetic information within an environmental sample. Therefore, shotgun metagenomic data allows for strain-level resolution of gut microbiome samples much like Gyrase B amplicon data. However, compared to Gyrase B amplicon dataset, shotgun metagenomic dataset is much more prevalent and, therefore, available, which provides a major advantage for research looking for scientific insights about strain-level evolutionary dynamics. With the development of tools that leverage the potential of shotgun metagenomic reads to

profile strains quickly and accurately, culture-independent strain-profiling with shotgun metagenomic datasets have enabled the investigation of strain-level evolutionary dynamics of wildlife gut microbes across the animal kingdom (18).

Thus, in this exploratory study, we leveraged a large shotgun metagenomic dataset aggregated from five Qiita studies (Qiita study ID: 2338, 11166, 13114, 11212, 13881) to explore the diversity of the vertebrate gut microbiome as well as to investigate the presence of strain-level phylosymbiosis in vertebrate (19-21). Our dataset contains samples from 288 vertebrate species spanning 6 host classes: Mammalia, Aves, Reptilia, Amphibia, Actinopterygii, Hyperoartia. To get a general community-level understanding of our dataset, we first examined the diversity and composition within our samples. Then to examine signals of strain phylosymbiosis of available bacteria in the dataset, we applied Mantel Spearman statistics to examine congruence between each strain tree, found with a marker-based strain profiler StrainPhlAn 3.0, and each corresponding host tree. Due to the advent of tools with different strain profiling approaches, the precise definition of strain has become quite ambiguous and fluid. It is therefore important to define strain considered in this study. Determined by our choice of strain profiler, strains examined in this study are the dominant genotype per bacterial species found in a sample. As one of the first works exploring the congruent phylogenies of bacterial strains and hosts using a large shotgun metagenomic dataset, our results provide evidence that *Akkermansia muciniphila* and *Bacteroides vulgatus* exhibit strain phylosymbiosis across distantly related hosts from across the animal kingdom.

METHOD

Dataset selection

To examine the diversification of strain across vertebrates, we searched for shotgun metagenomic datasets containing wildlife gut samples within the Qiita database (https://qiita.ucsd.edu/) (22). At the time of selection, the most datasets available in Qiita were either 16S amplicon sequences or related to human studies. Nevertheless, we were first able to identify 4 viable studies (Qiita study ID: 2338, 11166, 13114, 11212) to incorporate. By far the smallest study, study 2338 includes 6 samples of wild bats. Study 11166 primarily consists of 90 bird and bat samples, around half of which are wild (51%). Study 11212 consists of 95 wild primate samples. Lastly, study 13114 contains 182 mostly wild samples (67%) with the majority from the Mammalia class and the rest distributed among Aves, Reptilia, and Amphibia. Notably, out of the 182 samples from study 13114, 120 samples come from one mammalian species *Myodes glareolus*, commonly known as the bank vole. To improve the diversity of hosts within the final dataset, we sought out and incorporated shotgun metagenomic data from a recent study from Youngblue et al., which is now available as study 13881 on Qiita (20). Study 13881 contains 289 mostly wild samples (67%) from Mammalia, Aves, Reptilia, and Amphibia.

Metadata construction

Metadata was manually built by aggregating the available metadata information submitted to Qiita. Some categories such as sample name, sample type, taxonomic information of hosts, host common name, captivity information, and country of collection were already present in the available metadata so these categories were directly combined.

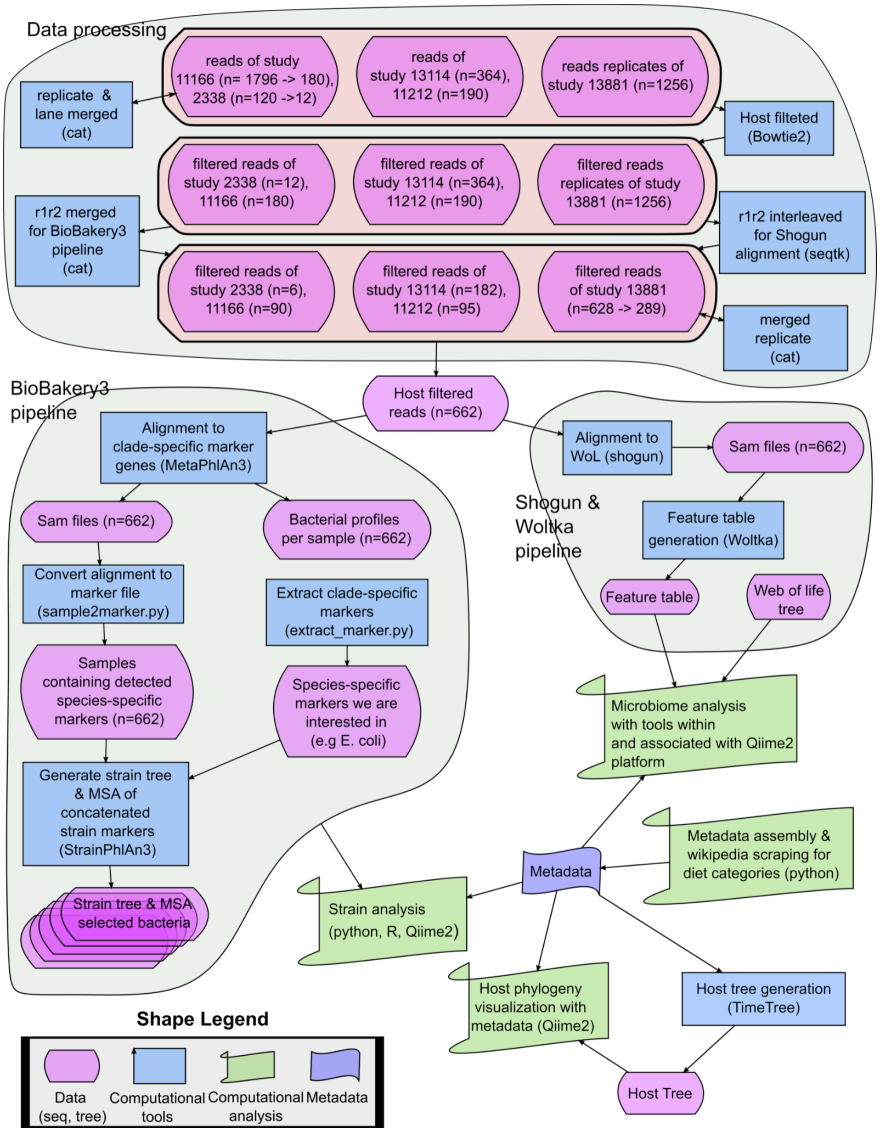

Figure 1. Overview of major steps carried out in this study.

Other metadata categories were present in some studies but not others. For example, information about host flight was missing in study 13881, host modality was present in 13114 only, and information about trophic guilds such as carnivore, herbivore, or omnivore was not present in study 2338 and study 11166.

Because of diet's strong influence on the gut microbial community, it is important to have a detailed category that details the dietary information for each host (23). So far, there is no tool

or database publicly available to obtain diets of various vertebrates, so to complete this metadata category, a custom script based in Python (v. 3.6.11) was built to extract dietary information on all hosts by scraping information from Wikipedia. The script counts keywords on a host's Wikipedia page to deduce the trophic guild of the host. For example, keywords such as carnivore, carnivorous, and meat-eating were associated with the dietary mode carnivore, whereas keywords such as herbivore, herbivorous, and plant-eating were associated with the dietary mode herbivore. The script was able to directly find the dietary mode for 33.3% of our hosts. If dietary mode cannot be determined directly, sentences containing keywords such as eat, consume, and feed were extracted for manual curation. Then dietary modes for the remaining hosts were determined from the relevant dietary information extracted from Wikipedia and manually curated with information available online. References used to make judgments about dietary mode have been recorded in a datasheet. We categorized host diet into 12 trophic guilds, which are Sanguivore (blood feeder), Scavenger, Carnivore, Insectivore, Omnivore, Herbivore, Folivore, Frugivore, Gummivore (specialist in tree sap), Granivore (specialist in seeds), Nectivore, and Filter Feeders. This metadata category was titled diet_category_12. We tried to be as detailed and inclusive in our approach as possible. If a host organism is known to feed on fruit and leaves, we tag it as both a frugivore and folivore. If a host shifts between an omnivorous and carnivorous diet during different seasons, it is categorized as both an omnivore and a carnivore. A more general category (diet_category_3) which contained only 3 trophic guilds (carnivore, omnivore, herbivore) was made by generalizing the trophic guilds in the more diet_category_12. For example, Insectivore would be generalized to Carnivore and Folivore would be generalized to Herbivore.

To get an understanding of the sequencing depth, we also incorporated read counts of files before and after the host filtering into the metadata. In addition, we also incorporated the host name whose genome was used to filter the associated file and the scientific names of host species submitted to the TimeTree database to construct our host tree (24).

The final metadata consists of 662 rows (excluding header) for the 662 samples and 28 columns (sample_name, studyID, SampleID, file_filtered_r1r2_combined, reads_per_unfiltered_file, reads_per_file, species_id, host_phylum, host_class, host_family, host_genus, host_kingdom, host_order, host_species, reference_for_filtering, TimeTree_returned, host_common_name, host_flight, host_modality, host_diet, country, habitat, sample_type, diet_category_12, diet_category_3, captive_wild).

Host tree construction

The TimeTree database was relied upon to construct the host tree of our dataset (24). Scientific names of our host species within our dataset were submitted to the TimeTree (http://www.timetree.org/) to obtain a complete newick tree containing 285 of 288 available host species. 3 species (*Aspius aspius*, *Cervus canadensis*, and *Geospiza acutirostris*) were not present in the TimeTree database and were not represented in our complete host tree.

Sequence preprocessing

The raw FASTQ reads associated with each Qiita study were acquired directly from barnacle, which is the supercomputer hosting files submitted to the Qiita database. These files had already been demultiplexed using Trimmamatics and adapter trimmed with minimap2 (v. 92021.1) on Qiita before being downloaded for additional processing (25). Files that were

generated from multiple sequencing runs of the same sample (samples from study 2338, 11166, 13881 were sequenced in multiple runs and lanes) were merged into one sample file. All files were then filtered by each respective host's reference genome obtained from the NCBI genome database (26). To prepare the files for strain evolution analysis, the forward and reverse reads were concatenated. To prepare the files for microbiome diversity analysis, the forward and reverse reads were interleaved using seqtk (v. 1.3) (https://github.com/lh3/seqtk). Figure 1 is an illustrative representation of the major steps taken in this project.

Host filtering

Samples were host filtered using either a host genome that was a direct match or the closest phylogenetic relative with a host genome using bowtie2 (v.2.4.4) (27). A custom bash script was built to expedite gathering host genomes for our samples. The script utilized the assembly_summary_genbank.txt file downloaded from NCBI's FTP server (https://ftp.ncbi.nih.gov/genomes/ASSEMBLY_REPORTS), which stores FTP links to host genomes. The script first checked if a direct host genome assembly was available for download. Out of the 288 host species represented in our dataset, 128 had direct host genomes assemblies. Because host genomes for the remaining 160 hosts were not available on NCBI, the genome assembly of the closest phylogenetic relative was found on the NCBI genome database and used for filtering. The closest relative was found by searching for viable hosts at higher taxonomic levels and then plotting candidates on a phylogenetic tree using TimeTree if multiple options existed at the same taxonomic level. For example, Transcaspian wild ass (*Equus hemionus kulan*) did not have a direct host genome available on NCBI, so the three relatives belonging to the genus Equus with genomes available were considered: *Equus caballus, Equus przewalskii,*

Equus asinus. These relative host species along with our host species of interest were submitted to TimeTree to construct a phylogenetic tree that was used to determine the closest phylogenetic relative to our species of interest. Once the closest phylogenetic relative was determined, in this case, *Equus asinus*, its host genome was used to filter samples from our species of interest, *Equus hemionus kulan*. We did not set a specific cutoff at which we did not pursue host genome for filtering, such that if there were no relative species at a specific taxonomic level with host genomes, we would look for closest relative species at one level higher until we found a viable

Table 1. Summary on taxonomic level of host genomes available for host filtering. 44% of the species represented by our samples had direct host genomes for filtering, while, of the 66% of species that did not have a direct host genome for filtering, 55% of the relative host genome was found at the genus level.

	Relative host genome					
Direct host genome	Genus level	Subfamily level	Family level	Superfamily level	Infraorder level	Suborder level
128	89	26	35	5	3	2

relative reference genome for host filtering. More than half (~55%) of the relative host genome used for host filtering was found on the genus level, while the most distantly related host used for host filtering was found on the suborder level (Table 1). For cases where multiple host genomes were available for one species, we selected the host genome based on the following three criteria by order of decreasing importance: 1) the most complete assembly (Chromosomes > Scaffolds > Contigs), 2) the most recent genome assembly, 3) the genome assembly with the largest size.

Microbiome Compositional analysis

Microbiome analysis of our dataset relied upon SHOGUN (v. 1.0.8), Woltka (v. 0.1.2), and QIIME2 (v. 2020.1) (28-30). Shogun was used for sequence alignment of our interleaved FASTA files against the Web of Life (WoL) database (30, 31). The WoL database includes 10,575 evenly sampled bacteria and archaea genomes as well as a reference phylogeny built with 381 single-copy marker genes (31). The alignments are then used to generate feature tables at the phylum, species, and Operational Genomic Unit levels (OGU-level). Proposed as a new feature providing the finest resolution possible for a shotgun metagenomic dataset, OGU refers to the taxonomically independent reference genomes with which shotgun metagenomic reads are mapped to. The OGU-level and species-level feature tables were randomly subset to contain one sample per species. For the alpha and beta diversity analysis, the OGU-level feature table was rarified to 50,000 reads per sample, which retained 2.15% of features in 65.28% of the available samples. The feature tables were then used for diversity analysis with QIIME 2's Python API. The subsetted species-level feature table was not rarified for the compositionally-aware PCA analysis by DEICODE (v. 0.2.4) (32).

QIIME2's heatmap function was used to generate a heatmap of phyla represented in our dataset from the phylum-level feature table. Robust Atchison PCA analysis was performed on the species feature table. The PERMANOVA and PERMDISP function available through scikit-bio (v. 0.5.6) was used for the multivariate analysis. Alpha and beta diversity significance were calculated by QIIME2's diversity alpha group significance and beta group significance plugin which runs the Kruskal-Wallis test and PERMANOVA, respectively. The heatmap and diversity figures were generated with the Python package Dokdo (v. 1.11.0), which enables

visualization of QIIME2 figures. The Kruskal-Wallis test was used to calculate significance. Statistical significance was defined as having a p-value < 0.05 for all analyses.

BioBakery3 pipeline overview

MetaPhlAn 3 (v. 3.0.11) aligned our merged FASTQ files against the ChocoPhlAn 3.0 database, which contains 1.1M unique clade-specific marker genes, to produce alignment files as well as species profiles for each sample (33). The species profiles for each sample were then merged to make a species-level feature table. This species-level feature table was used for bacterial abundance analysis using Python (v. 3.6.11) packages.

StrainPhlAn 3.0, a marker-based strain profiling tool within MetaPhlAn 3.0, was used to investigate strains from a particular bacterial species in this dataset (34). In brief, StrainPhlAn 3.0 concatenated the clade-specific markers of a bacterial species from ChocoPhlAn3 into a species-specific marker sequence with which metagenomic reads of a sample could be aligned against to estimate the consensus sequence of detected species-specific markers (34). The consensus sequences of each detected species-specific marker are then concatenated to form a strain-specific consensus marker sequence. Then, for a bacterial species of interest, StrainPhlAn 3.0 generated a multiple sequence alignment (MSA) of strain-specific consensus marker sequences, each representing the most dominant strain of bacteria found in a sample (34). It is important to note that the markers that are chosen to be concatenated into the strain-specific consensus markers sequence are influenced by the marker threshold. For example, if the marker threshold is set at 80%, then markers found in less than 80% of the strains are discarded. Finally, the MSA generated by StrainPhlAn 3.0 is used to build a maximum-likelihood phylogenetic tree with the GTRGAMMA model using RAxML (v 8.2.12) (35). Since consensus marker sequence

represents the dominant genotype of bacterial strains in a sample, we will refer to strain found by StrainPhlAn 3.0 as representative strain for the rest of the method section.

It is important to note that StrainPhlAn 3.0 has two important parameters that affect the output of StrainPhlAn MSA considerably; one parameter (i.e --marker_in_n_samples), which was touched upon in the last paragraph, specifies the marker inclusion criteria by the minimum percentage of representative strain each marker has to be found in (default: 80%), whereas the other parameter (i.e. --marker_in_n_samples) specifies the strain inclusion criteria by the minimum number of markers each representative strain needs to have (default: 20). To simplify the following explanation, we will refer to the marker inclusion criteria as marker threshold and strain inclusion criteria as the strain threshold. To explore the effect marker threshold on the outcome of our strain phylosymbiosis analysis, MSAs at five different marker thresholds (i.e 20%, 35%, 50%, 65%, 80%) were generated for each bacterial species examined. However, to maximize the number of representative strains without neglecting alignment quality, the strain threshold was determined using a custom approach that seeks to maximize the alignment score. This custom approach is described in detail in the next section.

Custom approach to choosing strain threshold at a specific marker threshold

At a given marker threshold, setting a lower minimum strain threshold (i.e. minimum markers present in a given strain to include a strain in MSA) will generate an MSA with more strains (Fig. 2A). Since the MSA will be used to generate a phylogenetic tree detailing the relationship between our strain, it will lead to a larger phylogenetic tree. However, this tree would be populated with strains whose phylogenetic relationship to others is more uncertain since they harbor fewer markers used for tree building (Fig. 2A). On the other hand, setting a

higher minimum strain threshold will lead to MSA of strains with very high marker counts, therefore a more phylogenetically confident tree. However, this approach likely unnecessarily excludes strains from downstream analysis (Fig. 2B). The extremes of the two simple scenarios above illustrate the difficulty in determining the strain threshold (or marker count per strain) without some sort of objective measurement we could use to maximize or minimize.

Thus, we introduce the concept of alignment score, which is the sum of nucleotides present in markers that lie within the marker threshold (Fig. 3). We believe MSA with a higher alignment score would improve MSA's performance in tree building or ordination analysis since there are more overlapping nucleotides across strains for comparison. In our custom StrainPhlAn 3.0 approach, we sought to find the minimum strain threshold that can maximize the alignment score for a set of strains iteratively (Fig. 3). We believe this approach optimizes the MSA for tree building as it maximizes the overlapping alignments of nucleotide positions given a certain marker threshold while avoiding the two problems with the previous two approaches; it does not maximize sample count by including samples with low marker count nor does it unnecessarily exclude samples with marker count that does not satisfy the often arbitrarily set threshold (Fig. 3 with sample calculation).

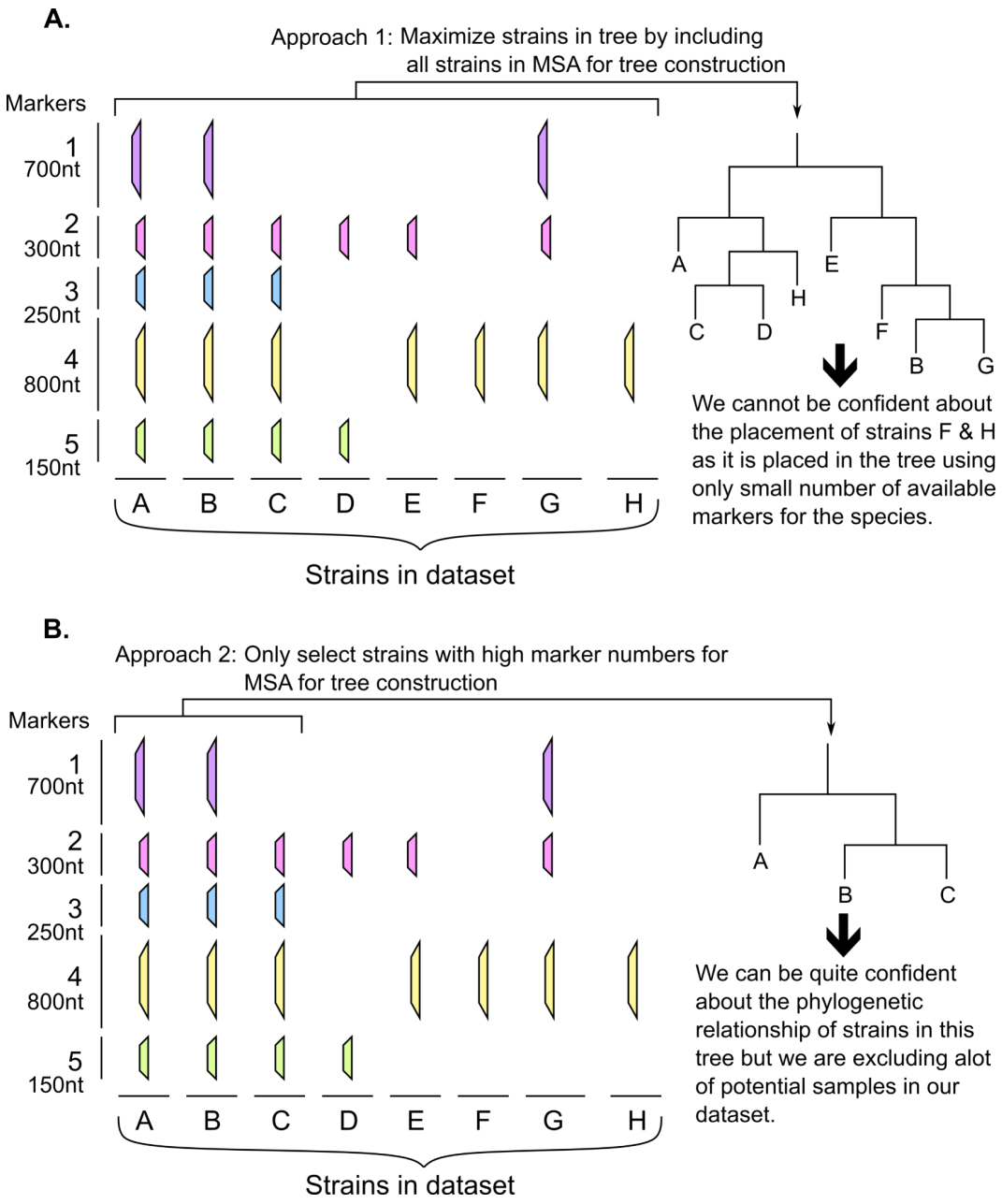

Figure 2. Approaches considered to strain selection for multiple sequence alignment. A) Approach 1 seeks to maximize strain count but produces a bloated tree. B) Approach 2 seeks to maximize the quality of strains chosen but neglect available strain diversity.

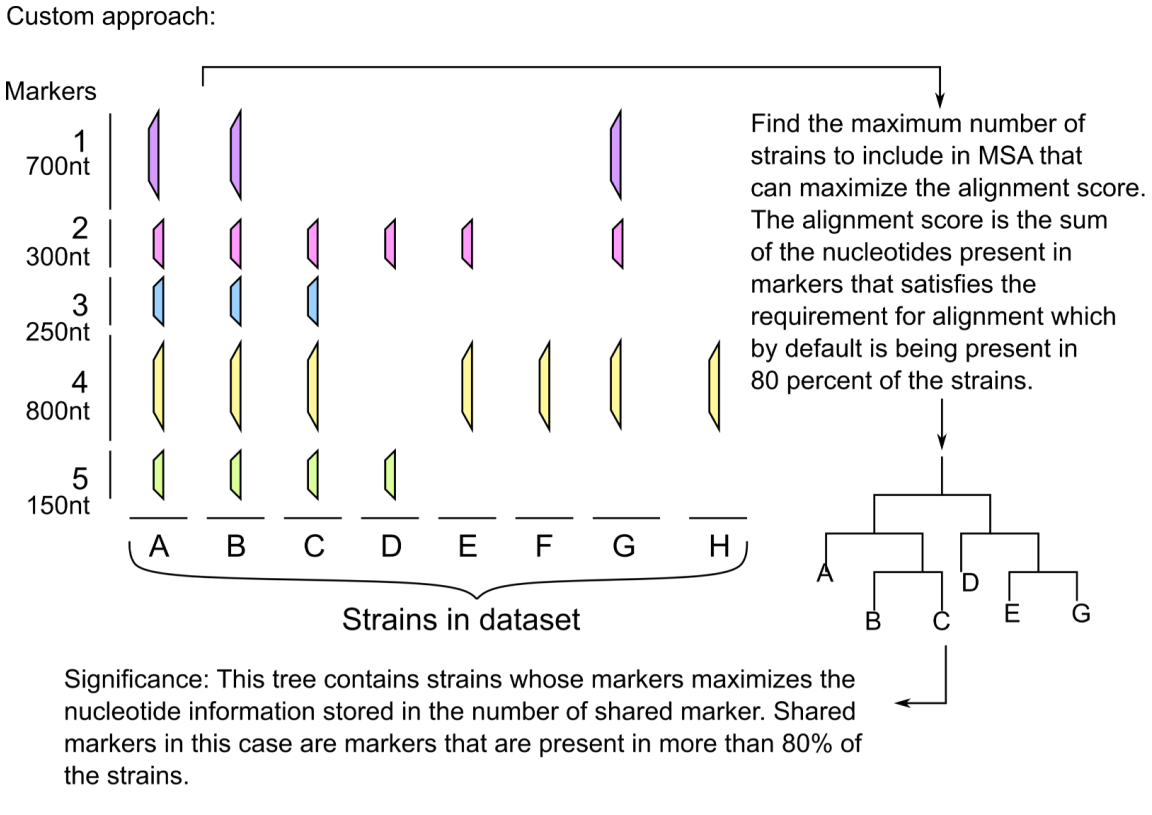

Figure 3. Illustration of custom approach to strain selection. The optimal number of strains to include for each species is determined by maximizing the alignment score, which is the sum of nucleotides in shared markers.

Bacterial selection for StrainPhlAn 3.0 processing

MetaPhlAn3 profiled around 86% (~570/662) of our metagenomic samples at the species level. 92 samples contained reads with unknown taxonomical identity. There appears to be a higher occurrence of unprofiled samples at a lower read count per sample (Fig. 4). Nevertheless, we targeted the 20 most abundant bacterial species by normalized abundance and by sample

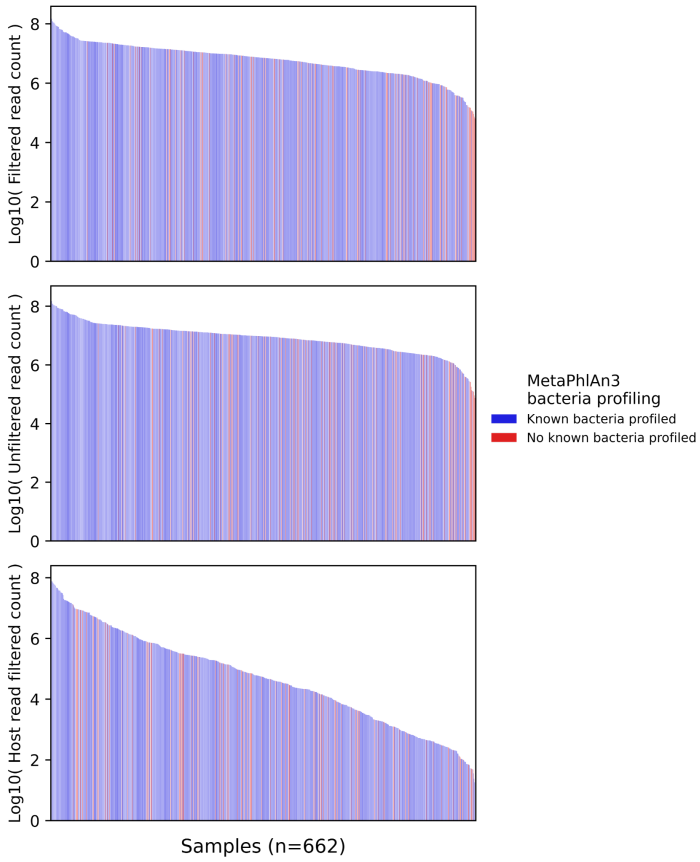

Figure 4. Distribution of read count per sample colored by MetaPhlAn3's success in profiling bacteria in each sample. Panel A & B represents a filtered and unfiltered read count of each sample whereas panel C represents the read count of host reads filtered out from each sample. Samples that failed to be profiled by MetaPhlAn3 seem to be concentrated in low read count samples and not related to the number of host reads in the sample.

presence in our dataset for strain analysis (Fig. 5). However, of these species, four species (i.e. *Bacillaceae bacterium EAG3*, *Lactobacillus apodemi*, *Plesiomonas shigelloides*, *Pseudomonas lundensis*) could not be processed by StrainPhlAn 3.0 because each had fewer than 4 strains

remaining in the MSA even at the least conservative marker threshold (20%) with custom strain threshold. These species were excluded from downstream analysis.

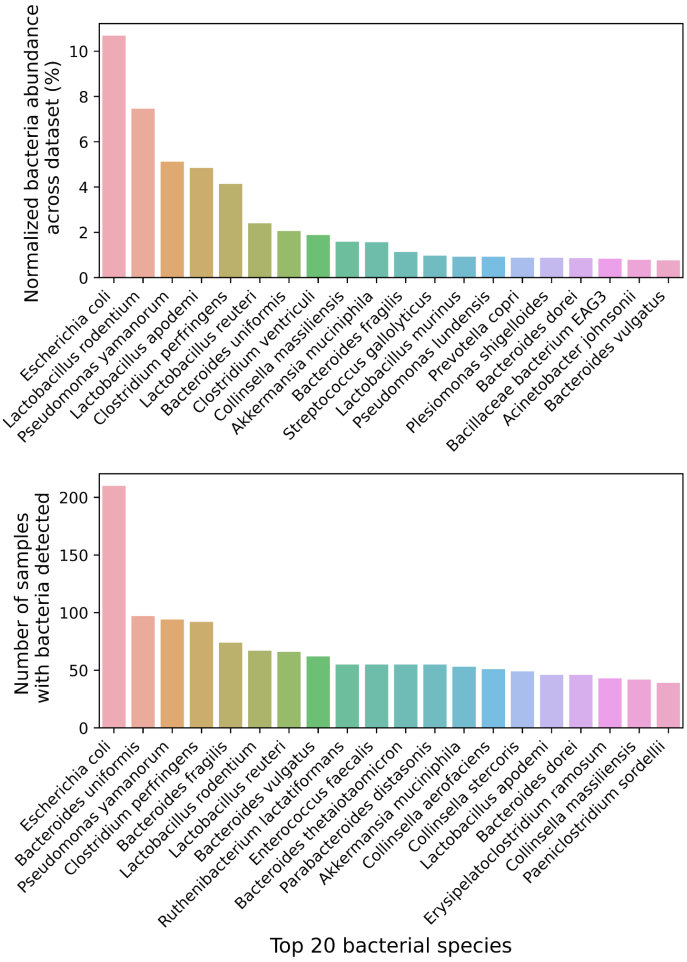

Figure 5. Top 20 abundant bacterial species profiled by MetaPhlAn3 measured by normalized bacterial abundance (A) and sample presence (B). In total, 27 bacterial species were found by both measurements. Strain analysis for all 27 bacterial species was attempted. *E. coli* is the most commonly detected bacterial species by far in both metrics.

The rest of the 23 bacteria species were processed through StrainPhlAn 3.0 at five different marker thresholds of 20%, 35%, 50%, 65%, 80%. For a given marker threshold, the custom approach to StrainPhlAn 3.0 parameter was applied, which relied on Pandas (v. 1.1.4) to

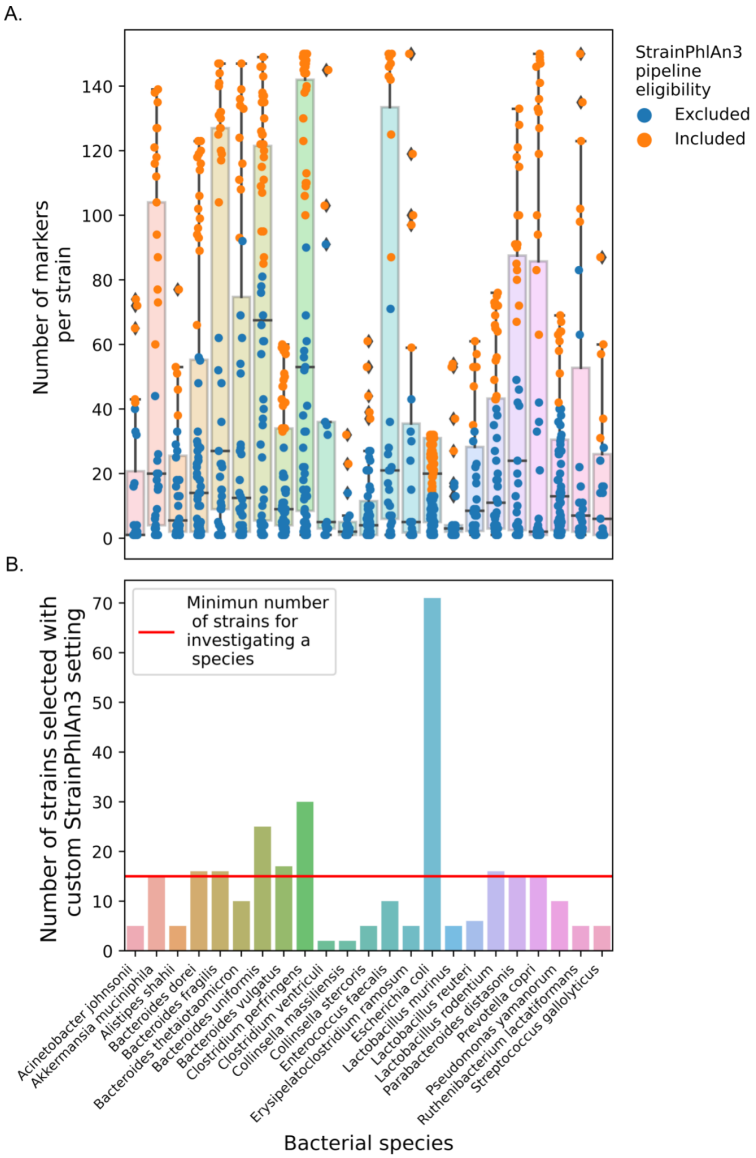

Figure 6. Overview of strains and bacteria selected for strain evolution analysis when setting marker threshold at 80%. A) The distribution of markers with each strain found per bacterial species is represented by the box plot. The box represents the interquartile range, while the whiskers extend to the minimum and maximum value. The box plot is overlaid with a strip plot of marker count of each strain per bacterial species colored by strain inclusion based on our custom approach. The yellow dots represent strains that are included in the MSA using our custom approach. B) 10 bacteria have at least 15 representative strains in the dataset at this marker threshold and will be processed by StrainPhlAn3.

calculate the strain threshold that maximizes the alignment score (36). The number of strains that were retained at the chosen marker threshold and calculated strain threshold was then calculated. If at least 15 strains were retained for a bacterial species, then the bacterial species was chosen

for StrainPhlAn 3.0 processing. Figure 5 illustration of bacterial species selection process when the marker threshold was set at 80%.

Strain phylosymbiosis and diversification analysis

Since our dataset includes samples from the same host species, we often find representative strains from the same host in the MSA as well as the strain tree for a bacteria of interest (i.e. of the *Akkermansia muciniphila* strains found two were from Hoffman's two-fingered sloth in figure 13). Therefore, to conduct the Mantel test, the strain tree is subsetted to one strain per species. If the subsetted strain tree had less than 15 strains, it was excluded from further analysis. Once a strain tree is constructed, the corresponding host tree is found by subsetting the host tree of our dataset to contain only host species represented in the strain tree. The subset strain tree and corresponding host tree are then converted to a patristic distance matrix for Mantel Spearman correlation test. Patristic distance is the sum of all branch lengths between two leaves of a tree. For each strain tree in the analysis, Mantel tests were carried out 100 times, each time with a random selection of one strain per species. The final Spearman correlation and p-value were averages of the Spearman correlations and p-values from 100 iterations of the Mantel Spearman test. The Mantel Spearman correlation test was carried out by the mantel function from the ecopy python package.

Strain trees generated by the GTRGAMMA model with RAxML were annotated by host class, host captivity status, and diet. in ITOL (v. 6.3.2) (37). In addition, for a strain tree of a bacteria species of interest, we calculated the Spearman correlation coefficient of each of its subtree that has at least 3 strains from distinct hosts (e.g. Fig. 15A). The level of Spearman correlations from each subtree regardless of statistical significance are then represented as branch colors on the strain tree with ITOL.

For a species of interest, a pairwise patristic distance of a randomly subsetted strain tree was plotted against the patristic distance of the corresponding host tree with Python (v. 3.6.11). Patristic distance of the host tree represents the estimated divergence time between hosts in terms of millions of years. This scatterplot was used as background visualization to present the overall Mantel Spearman correlation result (e.g. 15B).

For bacterial strains investigated for strain phylosymbiosis, evidence of divergence across host classes was investigated by examining the clustering of strain groups from different host classes. The patristic distance matrix of the strain tree was also used for ordination analysis and permutational multivariate analysis of variance [PERMANOVA] analysis of strains group across host class, host captivity, or host diet (e.g. 15C). Patristic distances were generated with R (v. 4.0.5)'s APE (v. 5.5) package (38). Ordinations were generated with Tidyverse (v. 1.3.1) (39). To assist visualization of the ordinations, ellipses are drawn using the stat_ellipse function in Tidyverse. Statistical multivariate analysis of strain divergence across the host class was processed using R's adonis, betadisper, ANOVA, TukeyHSD function from Vegan (v. 2.5.7) (40). The betadisper and ANOVA functions were used in conjunction to examine the variance of strain groups involved in the comparison. To find out if there are significant differences between the means of two groups, we also applied Tukey's HSD (honest significance difference) test.

RESULTS

Dataset assembly and filtering overview

Merging the samples from the 5 Qiita studies, our dataset consists of 662 samples from 288 species spanning 6 taxonomic classes: Mammalia, Aves (birds), Reptilia, Amphibia, Actinopterygii (bony fishes), Hyperoartia (Lampreys) (Fig. 7). Overall, mammals make up around half of the dataset (51%), whereas Aves accounted for more than a quarter (27%). By the

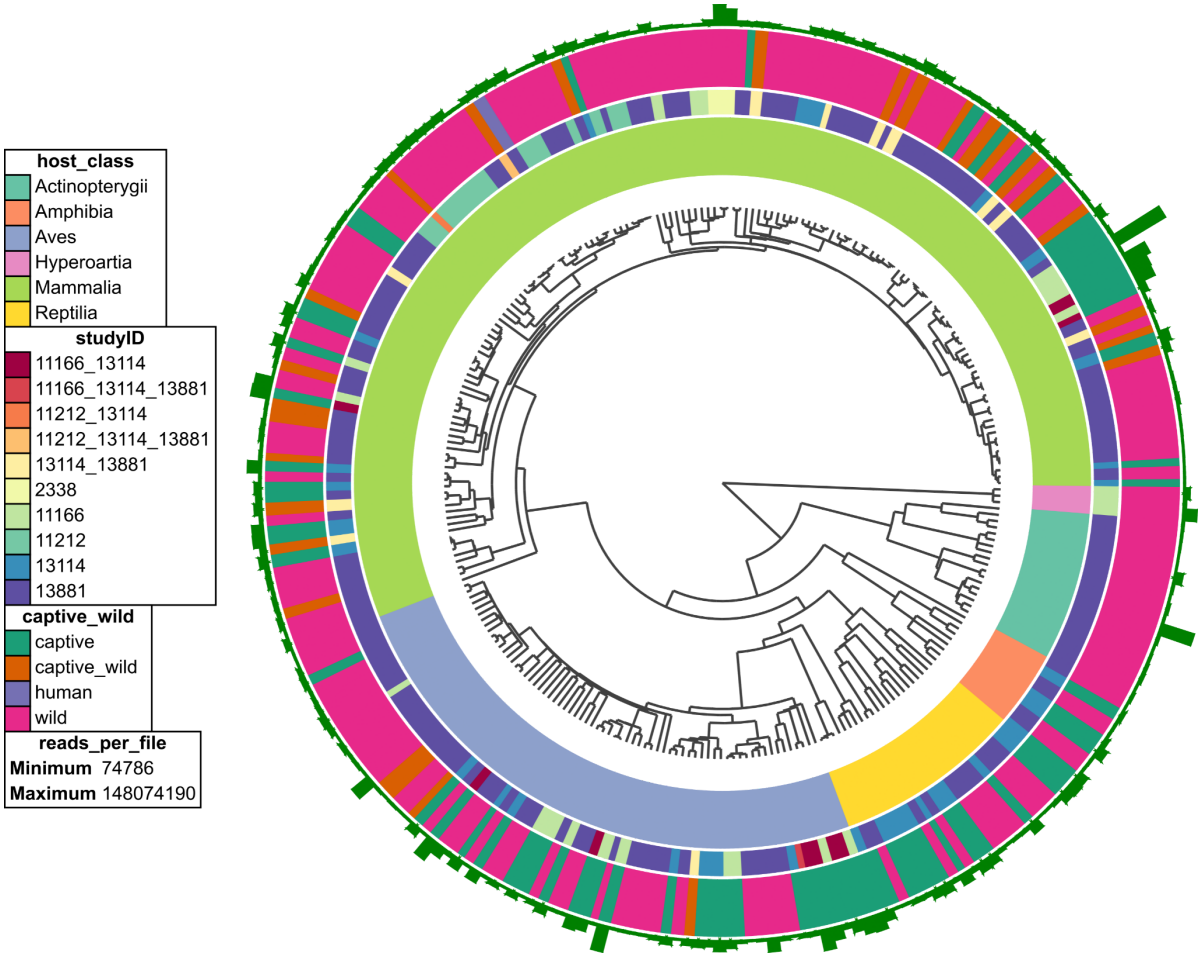

Fig 7. Overview of host species represented in our vertebrate gut dataset. Starting with the innermost ring, host species are labeled by host class, Qiita study ID, captivity information. The outermost ring represents the level of average read from associated gut samples after host filtering.

general dietary group, our dataset is split into 40.2% carnivores, 38.04% herbivores, and 21.7% omnivores (Fig. 7). Lastly, the vast majority (76.1%) of samples come from wild hosts (Fig. 7).

During sample processing, host filtering removed 17.8% of the initial sample reads (Fig. 8). Samples were processed for future projects hoping to generate MAGs from these samples. Interestingly, samples from sea lampreys and vampire bats were found to contain the highest level of host reads (data not shown). Since both animals are known sanguivores (blood feeders), this points to the potential that blood-feeding may lead to higher host shedding in the gut.

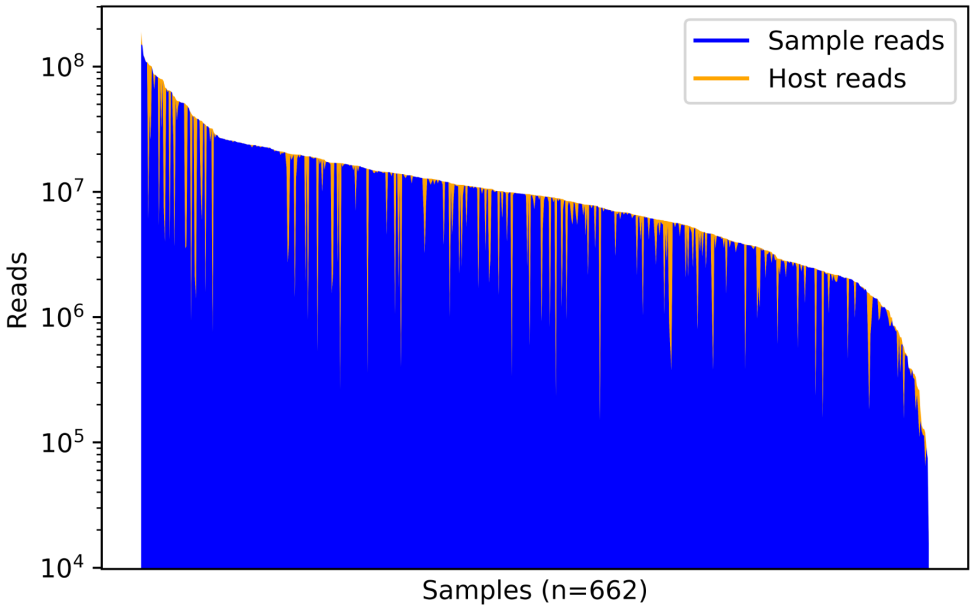

Figure 8. Distribution of host reads removed from samples in the dataset. In total, 17.8% of initial metagenomic reads were removed through host filtering.

Microbiome diversity analysis results

As expected, commonly known enteric bacterial phyla such as Firmicutes, Bacteroidetes are represented in our gut and fecal samples (Fig. 9). Looking at figure 8's heatmap more closely, Proteobacteria appear to be more abundant in herbivore gut than that of carnivore or omnivore. Unfortunately, because of time constraints, I did not conduct differential abundance analysis to

investigate feature enrichment in certain trophic guilds or host classes, which would be an interesting future project.

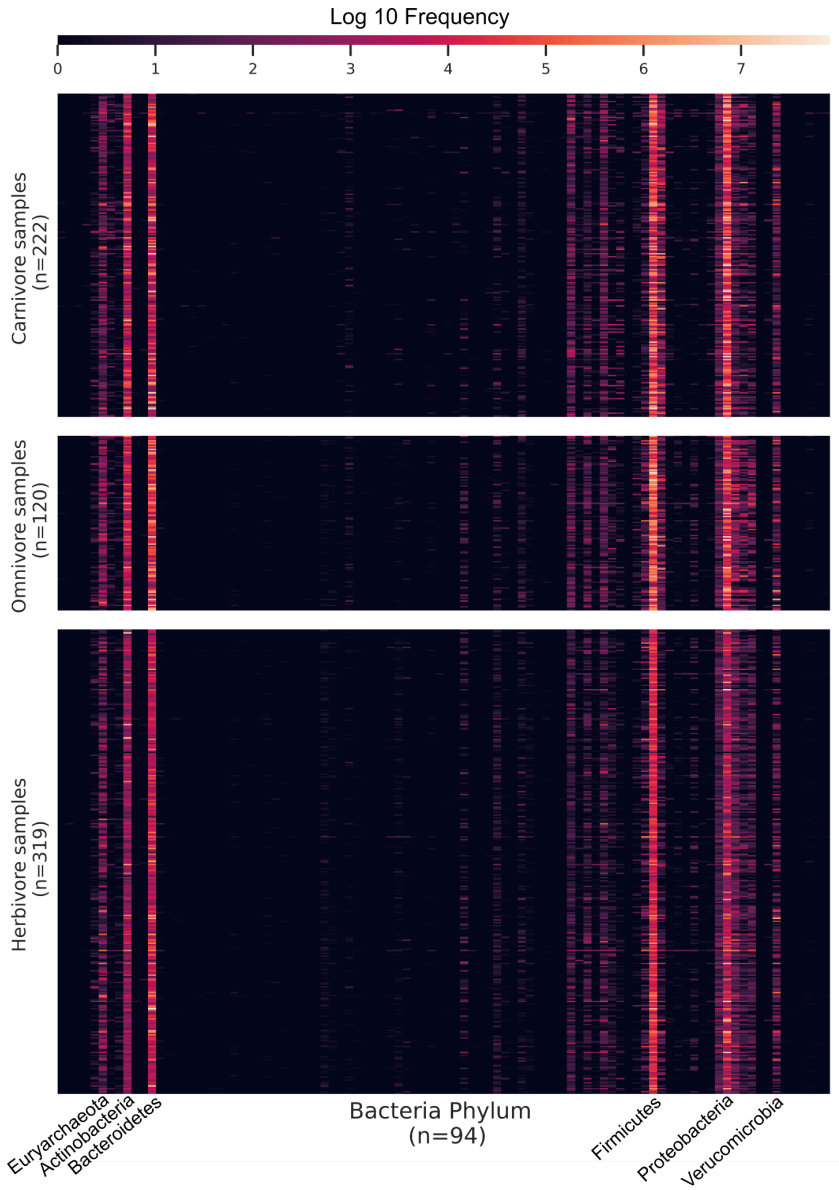

Figure 9. Heatmap of the absolute abundance of bacteria phyla represented in our dataset clustered by three general trophic guilds. Expected enrichment of known enteric bacterial phyla such as Bacteroides, Firmicutes, and Proteobacteria was observed.

To decipher the diversity within our samples, we used three different metrics (observed features, faith's phylogenetic distance, shannon entropy) to evaluate alpha diversity. We found microbiome diversity to be significantly different across host classes ([Kruskal-Wallis (all

groups)] Observed features: H = 12.28, p-value = 0.031 **; Faith's pd: H = 15.42, p-value = 0.009 **; Shannon entropy: H = 22.89, p-value = 0.00035 ***) (Fig. 10). Pairwise comparisons between host classes revealed that mammalian gut microbiome to be generally more diverse than that of avian gut microbiome ([Kruskal-Wallis (pairwise)] Observed features: H = 12.28, p-value = 0.0082 **, q-value = 0.123; Faith's pd: H = 10.12, p-value = 0.0015 **, q-value = 0.0220 *; Shannon entropy: H = 17.28, p-value = 0.000032 ****, q-value = 0.000484 ***) (Fig. 10). We also found microbiome diversity to be significantly different across general dietary categories ([Kruskal-Wallis (all groups)] Observed features: H = 11.65, p-value = 0.0029 **; Faith's pd: H = 18.40 , p-value = 0.0001 ***; Shannon entropy: H = 11.52, p-value = 0.0031 **). Pairwise comparisons between general dietary categories revealed that herbivore gut microbial communities to be generally more diverse than that of carnivores ([Kruskal-Wallis (pairwise)] Observed features: H = 11.14, p-value = 0.00085 **, q-value = 0.0025 **; Faith's pd: H = 18.28, p-value = 0.000019 ****, q-value = 0.000057 ****; Shannon entropy: H = 11.86, p-value = 0.00058 ***, q-value = 0.0017 **) and omnivores ([Kruskal-Wallis (pairwise)] Observed features: H = 5.135, p-value = 0.023 *, q-value = 0.035 *; Faith's pd: H = 6.184, p-value = 0.013 *, q-value = 0.019 *; Shannon entropy: H = 3.465, p-value = 0.063, q-value = 0.094) (Fig. 11).

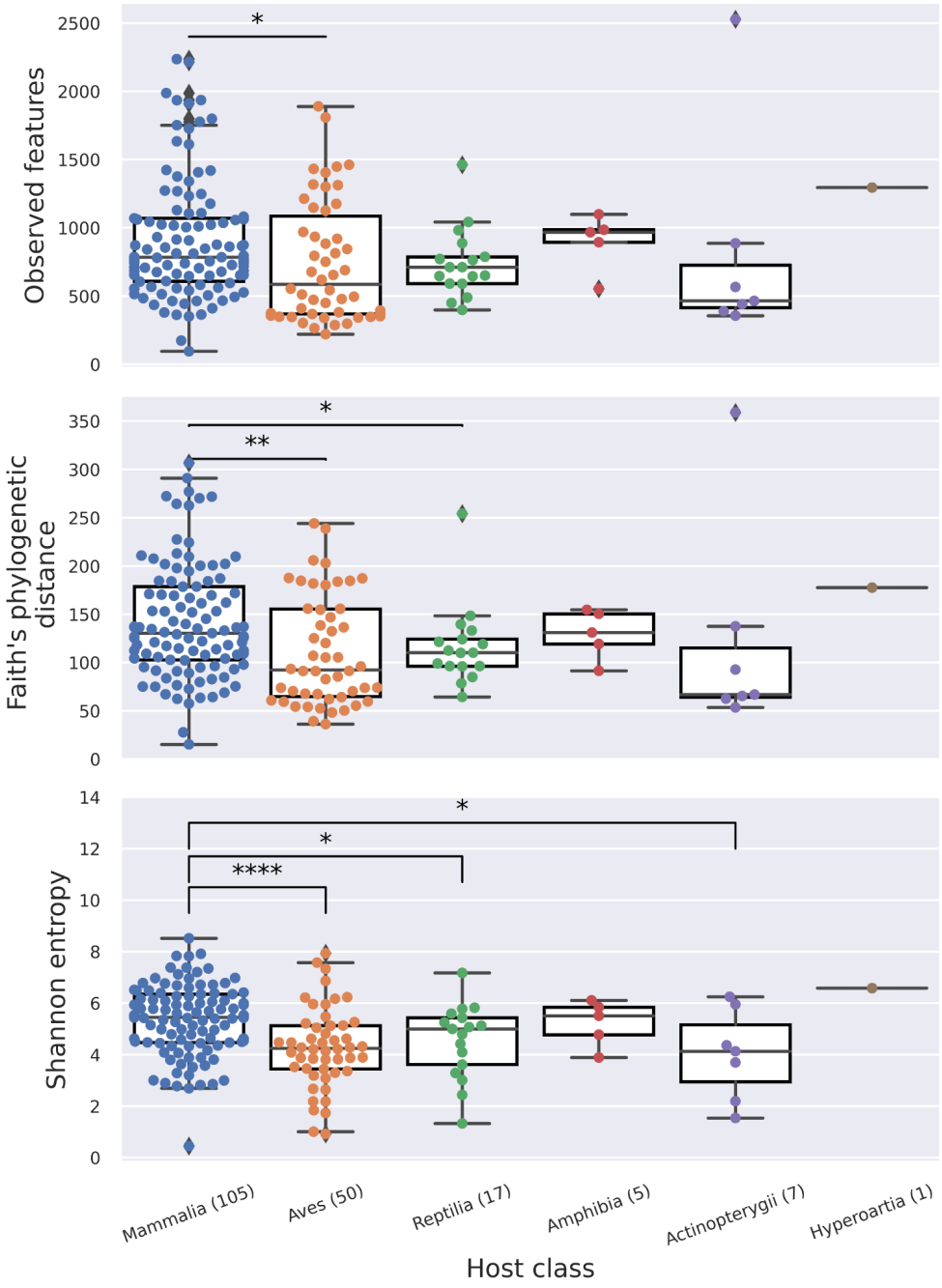

Figure 10. Alpha diversity compared across 6 host classes. Gut samples from Mammals exhibited statistically significant higher diversity than that of Aves. Feature table used is at OGU-level and subset to 1 sample per species rarified to 50,000 reads/sample, which retains 2.15% of features in 65.28% of the samples. * represents 0.05 > p-value > 0.01, ** represents 0.01 > p-value > 0.001, *** represents 0.001 > p-value > 0.0001.

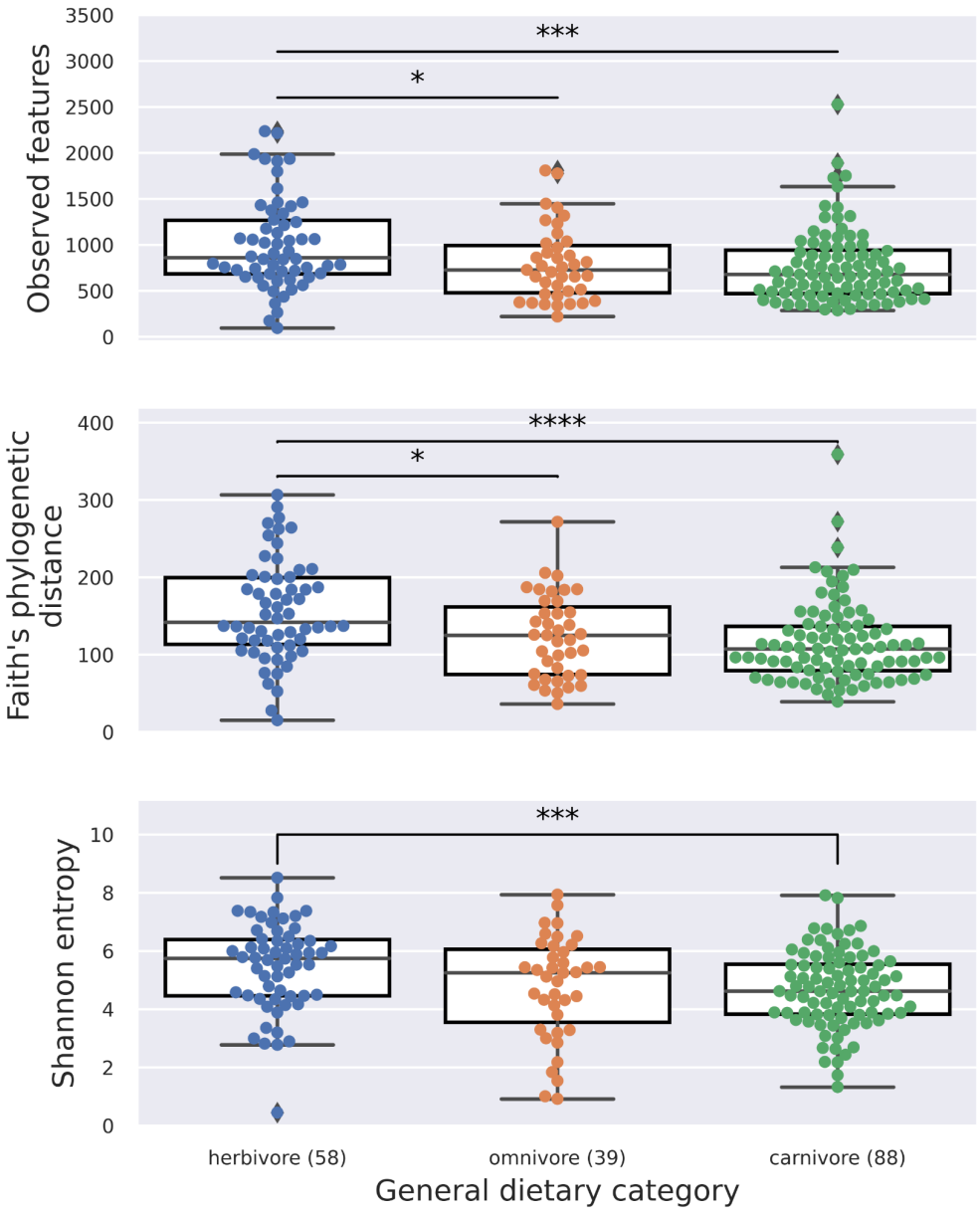

Figure 11. Alpha diversity compared across 3 general trophic guilds. Gut samples from herbivores exhibited statistically significant higher alpha diversity than samples from carnivores and omnivores. Feature table used is at OGU-level and subset to 1 sample per species rarified to 50,000 reads/sample, which retains 2.15% of features in 65.28% of the samples. * represents 0.05 > p-value > 0.01, ** represents 0.01 > p-value > 0.001, *** represents 0.001 > p-value > 0.0001.

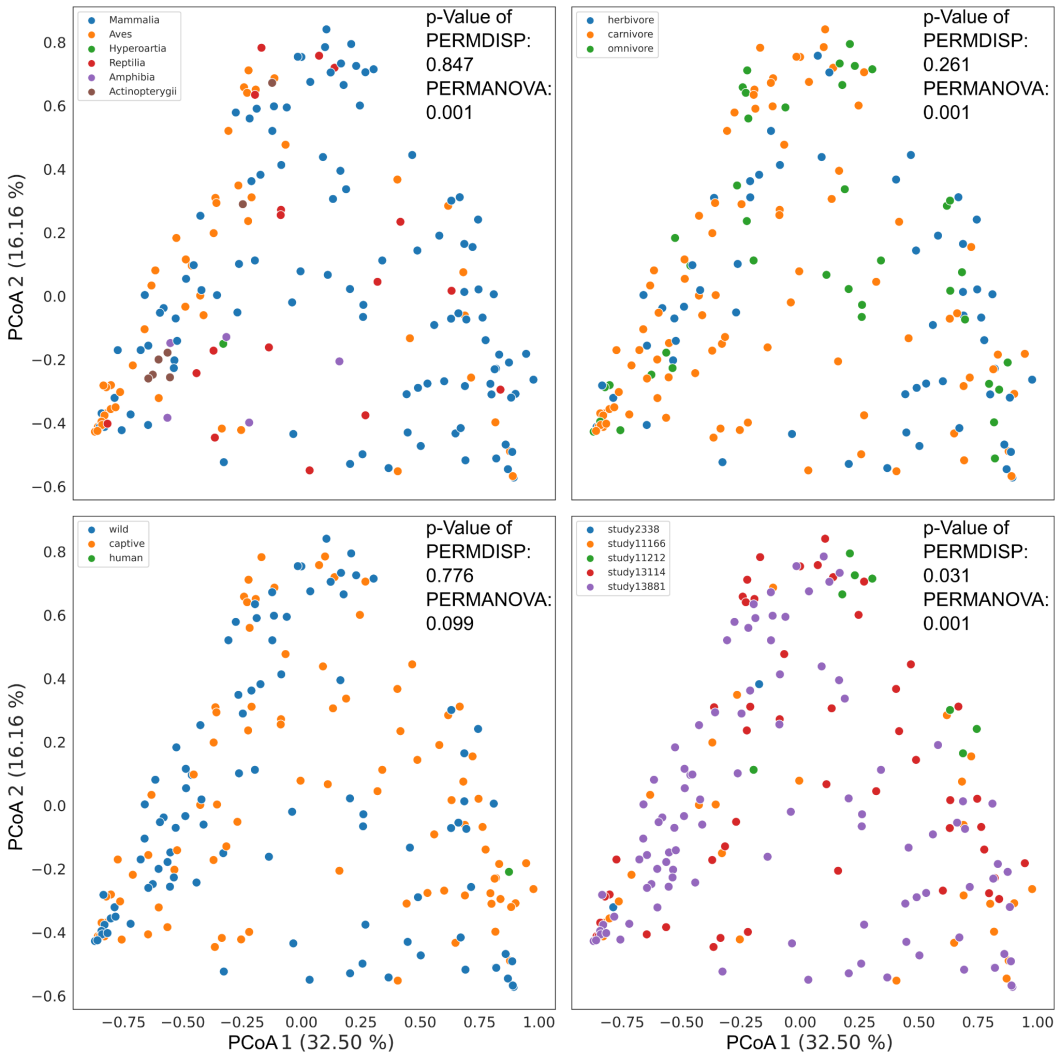

Figure 12. PCoA ordinations of weighted UniFrac distances colored by 4 different categories. From left to right and top to bottom, the microbiome communities are colored by host class, general host dietary category, host captivity, and study ID. The microbial composition differs across host class and general host dietary category. The feature table used is at OGU-level and subset to 1 sample per species rarified to 50,000 reads/sample, which retains 2.15% of features in 65.28% of the samples.

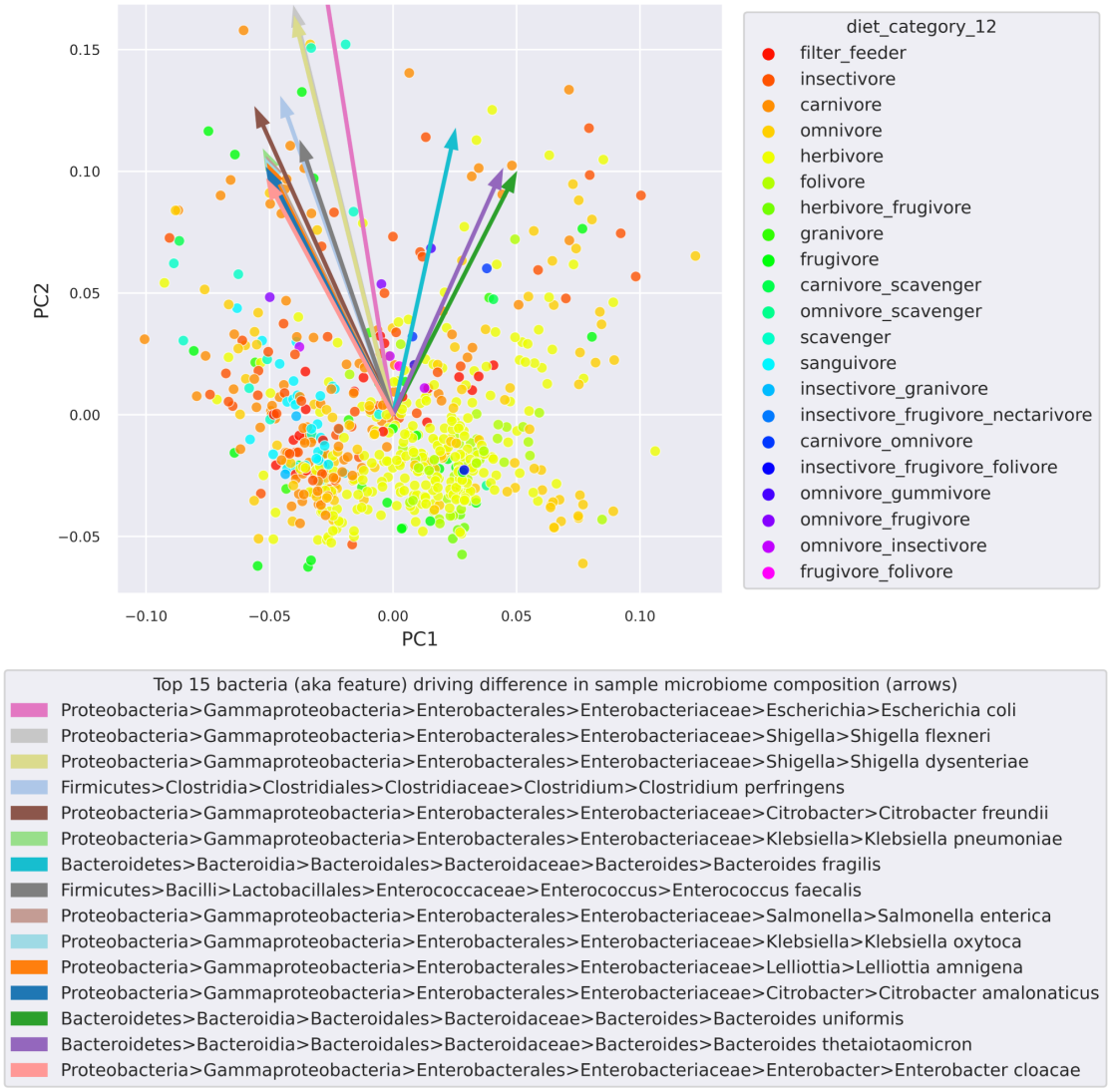

Figure 13. Biplot of Robust PCA of bacteria community colored by detailed trophic guilds. The Species-level feature table was used in this compositionally aware PCA analysis. The top 15 features of the PCA are dominated by bacteria of the Enterobacteriaceae family.

Beta-diversity analysis of UniFrac distances between microbiome samples revealed significant clustering of microbiome communities across host class (number of samples = 185, number of groups = 6: [PERMANOVA] pseudo-F = 3.3, p-value = 0.001 **; [PERMDISP] F-value = 7.71, p-value = 0.83) as well as general dietary categories (number of samples = 185,

number of groups = 3: [PERMANOVA] pseudo-F = 3.880, p-value = 0.001 **; [PERMDISP] F-value = 1.288, p-value = 0.248) (Fig. 12).

Unlike the weighted UniFrac distances, robust Aitchison distance between microbiomes did not reveal any statistically significant clustering of groups via multivariate analysis (Fig. 13). Nevertheless, DEICODE revealed the top 15 features driving the differences among samples, which are dominated by bacterial species of the Enterobacteriaceae family (Fig. 13).

Strain evolution analysis results

Evidence for strain phylosymbiosis was observed via the Mantel Spearman correlation test. Out of the 13 bacterial species investigated for strain phylosymbiosis, 7 bacterial species exhibited significant congruence between the host and strain tree at one or more marker thresholds (Fig. 14 & Table 2). Of these 7 bacterial species, *A. muciniphila* and *B. vulgatus* had significant Mantel Spearman correlation across all marker thresholds at which at least 15 strains were available (Fig. 14 & Table 2). The bacterial species with the highest number of strains, *E. coli* did not exhibit strain phylosymbiosis at any of the 5 marker thresholds (Fig. 14 & Table 2).

Exhibiting strong evidence for strain phylosymbiosis, *A. muciniphila* strains at 50% marker threshold were examined more closely (Fig. 15). The signal of strain phylosymbiosis is stronger near the root of the tree as shown by the yellow branch concentrated near the root of the tree (Fig. 15A). Taken as a whole, the strain tree exhibits significant strain phylosymbiosis via 100 iterations of the Mantel Spearman statistic ([Mantel Spearman results averaged over 100 iterations] r = 0.253, p-value =0.022) (Fig. 15B).

Figure 14. Overview of strain phylosymbiosis results, strain clustering across host class results, and basic biology of bacterial species examined for strain phylosymbiosis. A) Strain phylosymbiosis signals shown via the Mantel results averaged over 100 iterations of the mantel test. The numbers of samples involved in each mantel test are shown along with the significance of each test. Mantel test conducted on bacterial species involving at least 15 strains across 5 marker thresholds (20%, 35%, 50%, 65%, 80%). Green tick marks highlight mantel test results with positive correlation and statistically significant standard p-value < 0.05 value, whereas red cross marks highlight insignificant mantel test results. Mantel statistics were based on a two-sided spearman rank correlation with 999 permutations. B) Green tick marks indicate significant PERMANOVA results and insignificant ANOVA results when examining strains from different host classes. For panel A & B, ~ represents 0.1 > p-value > 0.05, * represents 0.05 > p-value > 0.01, ** represents 0.01 > p-value > 0.001, *** represents 0.001 > p-value > 0.0001. C) Basic biological characteristics of examined bacterial species (41).

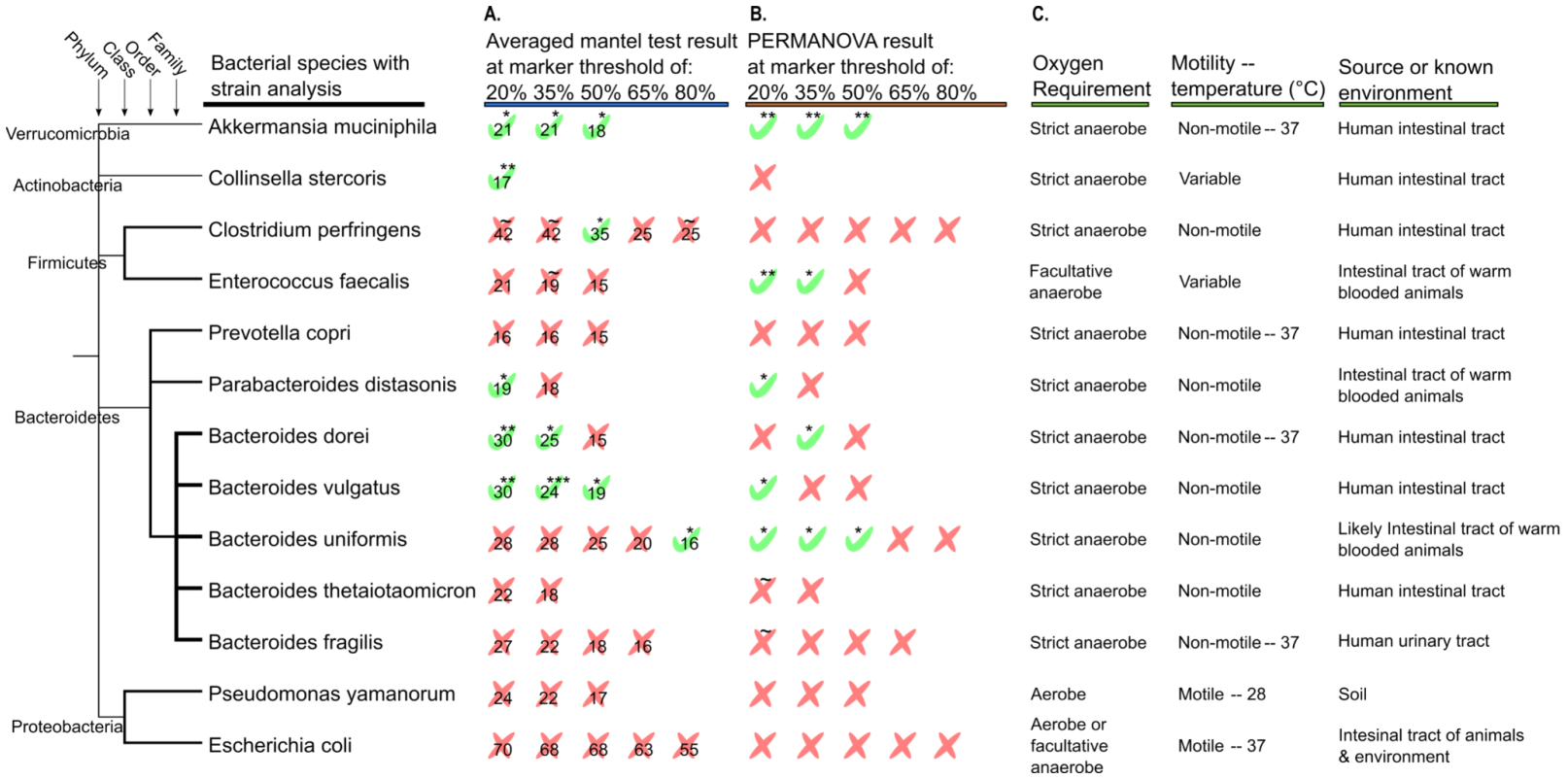
Phylum
Class
Order
Family
Bacterial species with strain analysis
A.
Averaged mantel test result at marker threshold of:
20% 35% 50% 65% 80%
B.
PERMANOVA result at marker threshold of:
20% 35% 50% 65% 80%
C.
Oxygen Requirement
Motility -- temperature (°C)
Source or known environment
Verrucomicrobia
Actinobacteria
Firmicutes
Bacteroidetes
Proteobacteria
Akkermansia muciniphila
Collinsella stercoris
Clostridium perfringens
Enterococcus faecalis
Prevotella copri
Parabacteroides distasonis
Bacteroides dorei
Bacteroides vulgatus
Bacteroides uniformis
Bacteroides thetaiotaomicron
Bacteroides fragilis
Pseudomonas yamanorum
Escherichia coli
Strict anaerobe | Non-motile -- 37 | Human intestinal tract
Strict anaerobe | Variable | Human intestinal tract
Strict anaerobe | Non-motile | Human intestinal tract
Facultative anaerobe | Variable | Intestinal tract of warm blooded animals
Strict anaerobe | Non-motile -- 37 | Human intestinal tract
Strict anaerobe | Non-motile | Intestinal tract of warm blooded animals
Strict anaerobe | Non-motile -- 37 | Human intestinal tract
Strict anaerobe | Non-motile | Human intestinal tract
Strict anaerobe | Non-motile | Likely Intestinal tract of warm blooded animals
Strict anaerobe | Non-motile | Human intestinal tract
Strict anaerobe | Non-motile -- 37 | Human urinary tract
Aerobe | Motile -- 28 | Soil
Aerobe or facultative anaerobe | Motile -- 37 | Intesinal tract of animals & environment

Table 2. Numeric values of mantel spearman statistic result across all five marker thresholds. Mantel test results are averaged over 100 iterations; mantel tests are only conducted on strains trees with at least 15 strains across 5 marker thresholds (20%, 35%, 50%, 65%, 80%). The numbers of samples involved in each mantel test are shown along with the significance of each test. Each iteration of the mantel statistics was based on a two-sided spearman rank correlation with 999 permutations. * represents 0.05 > p-value > 0.01, ** represents 0.01 > p-value > 0.001, *** represents 0.001 > p-value > 0.0001.

	20% marker threshold			35% marker threshold			50% marker threshold			65% marker threshold			80% marker threshold		
Bacteria investigated	# of host & strain pairs	r	p-value	# of host & strain pairs	r	p-value	# of host & strain pairs	r	p-value	# of host & strain pairs	r	p-value	# of host & strain pairs	r	p-value
Akkermansia muciniphila	21	0.266	0.023 *	21	0.261	0.0242 *	18	0.253	0.0215 *						
Bacteroides dorei	30	0.273	0.0047 **	25	0.177	0.0465 *	15	0.001	0.857						
Bacteroides fragilis	27	0.085	0.353	22	0.006	0.958	18	0.054	0.748	16	0.030	0.880			
Bacteroides thetaiotaomicron	22	0.132	0.133	18	0.127	0.249									
Bacteroides uniformis	28	0.044	0.721	28	0.037	0.752	25	0.020	0.854	20	0.160	0.215	16	0.311	0.047 *
Bacteroides vulgatus	30	0.355	0.0049 **	24	0.391	0.00718 **	19	0.327	0.032 *						
Clostridium perfringens	42	0.121	0.078	42	0.117	0.091	35	0.146	0.0351 *	25	0.029	0.793	20	0.157	0.073
Collinsella stercoris	17	0.559	0.0041 **												
Escherichia coli	70	0.034	0.417	68	0.022	0.607	68	0.039	0.371	63	0.025	0.552	55	0.022	0.568
Enterococcus faecalis	21	0.095	0.322	19	0.068	0.536	15	0.028	0.810						
Parabacteroides distasonis	19	0.301	0.0189 *	16	0.140	0.375									
Prevotella copri	16	0.145	0.352	18	0.199	0.141	15	0.227	0.161						
Pseudomonas yamanorum	24	0.131	0.271	22	0.144	0.296	17	0.204	0.180						

Table 3. Numeric values of multivariate results selected bacterial strains across five marker thresholds (20%, 35%, 50%, 65%, 80%). For each bacteria and marker threshold, the columns from left to right contain the R^2 and p-value of overall and pairwise adonis permanova results, the F-value and p-value of ANOVA of beta dispersion results, and TukeyHSD p-value between groups. * represents 0.05 > p-value > 0.01, ** represents 0.01 > p-value > 0.001, *** represents 0.001 > p-value > 0.0001. The table is split into two parts and presented on the following two pages for clearer visuals.

		20% marker threshold					35% marker threshold					50% marker threshold					65% marker threshold					80% marker threshold				
		Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD	Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD	Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD	Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD	Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD
Bacterial strains examined	Groups	R2	P-value	F value	P-value	P-value	R2	P-value	F value	P-value	P-value	R2	P-value	F value	P-value	P-value	R2	P-value	F value	P-value	P-value	R2	P-value	F value	P-value	P-value
Akkermansia muciniphila	all groups	0.45538	0.0021 **	0.9791	0.3914		0.45502	0.0017 **	1.042	0.3695		0.45328	0.0037 **	0.621	0.5485											
	Mammals-Aves	0.0261	0.6968			0.9277527	0.02957	0.6639			0.9367071	0.03008	0.77			0.9997362										
	Aves-Rep	0.58925	0.0224 *			0.5091896	0.59543	0.0214 *			0.4756166	0.60146	0.0224 *			0.5619129										
	Mammals-Rep	0.5356	0.006899 **			0.3598822	0.52753	0.007199 **			0.3372172	0.52099	0.0124 *			0.5362922										
Clostridium stercoris	all groups	0.07195	0.1339	1.3749	0.253																					
	Mammals-Aves	0.07195	0.1339																							
Clostridium perfringens	all groups	0.07451	0.0447 *	4.2252	0.01973 *		0.09368	0.0435 *	4.2252	0.01973 *		0.09368	0.0441 *	5.2853	0.008865 **		0.09804	0.1083	6.4538	0.004309 **		0.01423	0.7598	10.264	0.003373 **	0.0033728
	Amphibia-Aves	0.05757	0.1647			0.1041412	0.05757	0.1625			0.1041412	0.07142	0.1086			0.1195107	0.07021	0.1587			0.0107046					
	Mammals-Amphibia	0.05757	0.1521			0.3110085	0.13017	0.038 *			0.3110085	0.24243	0.0487 *			0.4661641	0.50247	0.08739 .			0.6036722					
	Mammals-Aves	0.0337	0.1214			0.0684211	0.0337	0.1278			0.0684211	0.03587	0.1624			0.0214732	0.03141	0.339			0.0991519					
Enterococcus faecalis	all groups	0.22214	0.002 **	0.8277	0.4508		0.21023	0.0207 *	0.1811	0.836		0.18618	0.1419	3.00E-04	0.9997											
	Mammals-Aves	0.17483	0.0037 **			0.4253138	0.16242	0.0113 *			0.8353051	0.12426	0.1383			0.9998298										
	Aves-Rep	0.16184	0.0163 *			0.9720529	0.15176	0.0488 *			0.9999738	0.16947	0.1685			0.9999652										
	Mammals-Rep	0.17631	0.08329 .			0.7457335	0.16716	0.1675			0.9145083	0.15348	0.3466			0.999726										
Prevotella copri	all groups	0.05027	0.4473	1.4578	0.2429		0.05049	0.4612	1.4069	0.251		0.16354	0.1095	3.3538	0.08574 .											
	Mammals-Aves	0.05027	0.4473																							
Parabacteroides distasonis	all groups	0.16297	0.0455 *	1.0153	0.378		0.09122	0.2833	0.6545	0.5295																
	Mammals-Aves	0.04196	0.3468			0.6296297	0.03435	0.5044			0.5253754															
	Aves-Rep	0.21882	0.0255 *			0.3473596	0.26328	0.05669 .			0.9453831															
	Mammals-Rep	0.15	0.0297 *			0.6662833	0.07165	0.2437			0.8628321															
Bacteroides dorei	all groups	0.09122	0.2774	0.6545	0.5295		0.28404	0.0218 *	1.029	0.3682		0.00681	0.9587	0.0456	0.8329											
	Mammals-Aves	0.03435	0.4942			0.5253754	0.00386	0.8884			0.9835261					0.8328796									
	Aves-Rep	0.26328	0.05129 .			0.9453831	0.64286	0.1268			0.4074815															
	Mammals-Rep	0.07165	0.2532			0.8628321	0.32695	0.0307 *			0.3354781															

		20% marker threshold					35% marker threshold					50% marker threshold					65% marker threshold					80% marker threshold				
		Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD	Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD	Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD	Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD	Adonis PERMANOVA		ANOVA of betadisper		Tukey HSD
Bacterial strains examined	Groups	R2	P-value	F value	P-value	P-value	R2	P-value	F value	P-value	P-value	R2	P-value	F value	P-value	P-value	R2	P-value	F value	P-value	P-value	R2	P-value	F value	P-value	P-value
Bacteroides vulgatus	all groups	0.21473	0.0402 *	0.5643	0.573		0.2944	0.0046 **	3.3196	0.0482 *		0.03334	0.4517	1.3324	0.2602											
	Mammals-Aves	0.00859	0.7294			0.7870096	0.0665	0.1111			0.0741041					0.2602398									
	Aves-Rep	0.77923	0.1009			0.8148428	0.37382	0.1128			0.1738553															
	Mammals-Rep	0.22113	0.05489 .			0.6564305	0.39756	0.0357 *			0.570898															
Bacteroides uniformis	all groups	0.24059	0.0338 *	0.888	0.419		0.21473	0.0404 *	0.5643	0.573		0.22292	0.0407 *	0.6839	0.5109		0.0387	0.2933	0.7402	0.3964		0.0291	0.5927	2.2278	0.1491	
	Mammals-Aves	0.01233	0.5893			0.6323826	0.00859	0.7367			0.7870096	0.01113	0.6343			0.6886413					0.3964077					0.1491357
	Aves-Rep	0.80719	0.09859 .			0.7829039	0.77923	0.09289 .			0.8148428	0.81539	0.1118			0.8504733										
	Mammals-Rep	0.24683	0.05779 .			0.5605165	0.22113	0.05479 .			0.6564305	0.2261	0.06649 .			0.6498132										
Bacteroides thethaiotaomicron	all groups	0.14205	0.09659 .	0.8508	0.4395		0.11591	0.3939	0.5916	0.5644																
	Mammals-Aves	0.0616	0.3519			0.501992	0.05214	0.4383			0.8433234															
	Aves-Rep	0.17445	0.2254			0.9683143	0.10873	0.5766			0.9307594															
	Mammals-Rep	0.11757	0.07979 .			0.639519	0.10414	0.2606			0.5538324															
Bacteroides fragilis	all groups	0.13628	0.08939 .	0.1847	0.8323		0.04964	0.5703	0.0135	0.9866		0.12611	0.2341	0.3497	0.709		0.05879	0.41	0.5501	0.5868						
	Mammals-Aves	0.01752	0.4986			0.8402498	0.00389	0.5525			0.9862469	0.02522	0.3845			0.7535175	0.02623	0.4344			0.7250542					
	Aves-Rep	0.13318	0.2074			0.9965046	0.12153	0.282			0.9937401	0.54025	0.07569 .			0.754016	0.3861	0.2857			0.5967673					
	Mammals-Rep	0.19545	0.0284 *			0.929226	0.05632	0.2168			0.9999741	0.08915	0.1806			0.9649214	0.02628	0.2798			0.8483853					
Pseudomos yamanorum	all groups	0.05408	0.4083	0.5226	0.5965		0.03774	0.7074	3.1706	0.05359 .		-0.01073	0.9845	3.2041	0.05771 .											
	Aves-Hyperoartia	0.0552	0.2709			0.7302038	0.01015	0.7833			0.0802671	-0.04156	0.9839			0.1606406										
	Mammals-Aves	0.01189	0.8547			0.6120764	0.06098	0.1981			0.9933697	0.04102	0.7486			0.8873948										
	Mammals-Hyperoartia	0.05611 .	0.05869			0.987571	-0.02059	0.9731			0.0859553	-0.07416	0.9958			0.0589853										
Escherichia coli	all groups	0.00841	0.8264	1.6844	0.1912		0.01708	0.4233	0.8617	0.426		0.01224	0.5793	1.2674	0.2866		0.00717	0.8793	1.5115	0.2267		0.03844	0.2195	3.1751	0.04805 *	
	Mammals-Aves	0.00586	0.6447			0.4398388	0.01569	0.2142			0.6833361	0.01008	0.3912			0.4239641	0.005	0.7735			0.3438267	0.03605	0.05029 .			0.0759831
	Aves-Rep	0.00558	0.7287			0.4875994	0.0025	0.8824			0.6340113	0.00367	0.7315			0.6862326	0.00418	0.8536			0.6683664	0.00555	0.9232			0.6111476
	Mammals-Rep	0.00496	0.8564			0.2918464	0.00365	0.7819			0.5239211	0.0051	0.7473			0.5212076	0.00452	0.7617			0.4779282	0.00655	0.8246			0.3115724

Overall, *A. muciniphila* strains exhibit genotypic differences across the host class. *A. muciniphila* strains cluster significantly across host classes ([adonis PERMANOVA] R^2 =0.45328, p-value = 0.0048 ** with no difference in variance), but not across host captivity status ([adonis PERMANOVA] R^2 =0.05533, p-value = 0.2309) or general diet category ([adonis PERMANOVA] R^2 = 0.05161, p-value = 0.2237) (Fig. 15C). Pairwise comparisons of *A. muciniphila* strain across host classes using suggests there are genotypic difference between mammalian strains and reptilian strains ([adonis PERMANOVA] R^2 = 0.52099, p-value = 0.0124 * with no difference in variance), avian strains and reptilian strain ([adonis PERMANOVA] R^2 = 0.60146, p-value = 0.0224 * with no difference in variance), but not mammalian strains and avian strains (adonis [PERMANOVA] R^2 = 0.03008, p-value = 0.77). However, pairwise comparisons via Tukey's HSD test reveal no differences in means between strain groups from different host classes (Table 3).

Similarly, *B. vulgatus* also presented a strong case for strain phylosymbiosis and its strain tree at 50% marker threshold was examined more closely (Fig. 16). The general level of strain phylosymbiosis decreases traveling down the strain tree except in the subtree containing humans, fallow deer, and tamandua in which it is elevated (Fig. 16A). The *B. vulgatus* strain tree proves to exhibit significant strain phylosymbiosis via 100 iterations of the Mantel Spearman statistic (Fig. 16B). Similar to that of *A. muciniphila* strains, we see a positive relationship between the evolutionary relatedness of *B. vulgatus* strains and that of their corresponding hosts ([Mantel Spearmam results averaged over 100 iterations] r = 0.337, p-value =0.029) (Fig. 16B).

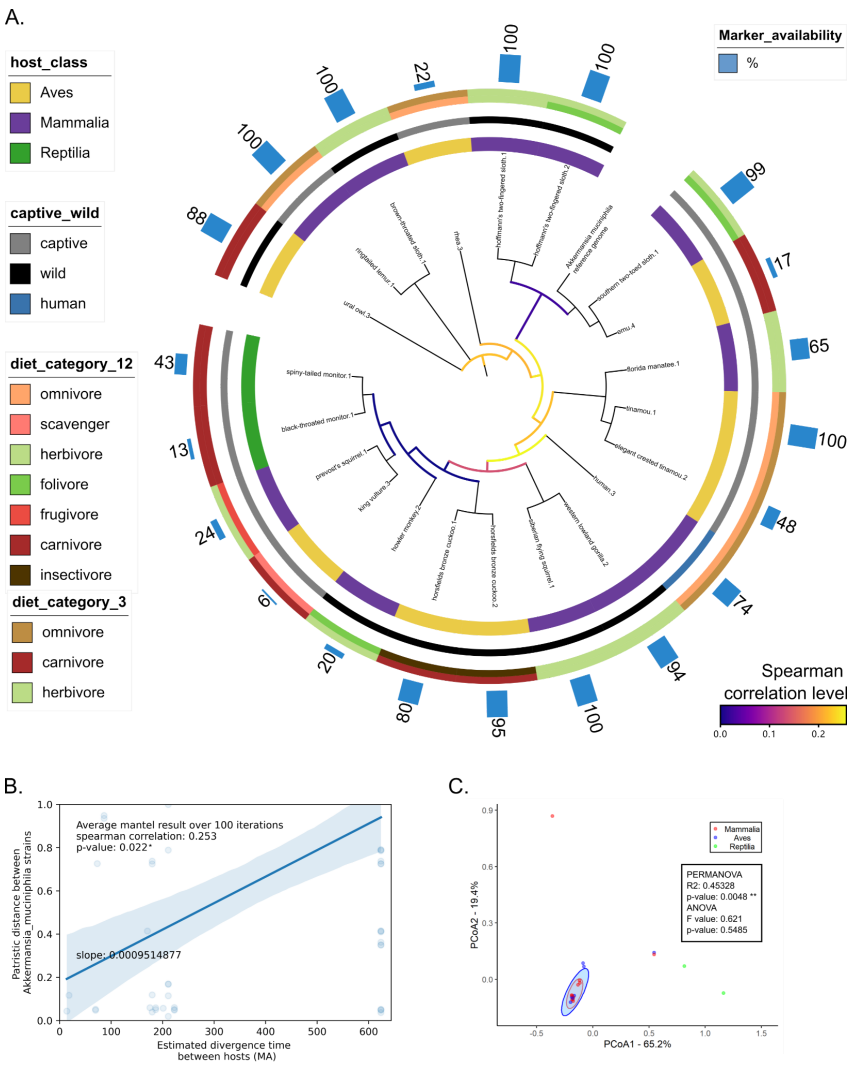

Figure 15. Analysis of *Akkermansia muciniphila* MSA with 50% marker inclusion criteria reveals evidence for strain phylosymbiosis and strain diversification across host classes. A) Strain tree of detected *A. muciniphila* strains with tips labeled by the common name of each strain's host. Strength of strain phylosymbiosis as measured by Mantel Spearman correlation of each subtree with at least 3 leaves is represented by branch color. From the innermost to the outermost ring, strains are colored by hosts' taxonomic class, captivity information, fine and general dietary information. The percent of markers for *A. muciniphila* available for each strain for tree building is represented by the annotated bar chart. The strain tree is built with the GTRGAMMA model and visualized with branch length ignored. B) Strong signal of strain phylosymbiosis detected for *A. muciniphila*. Mantel spearman test reveals a statistically significant positive correlation between host divergence time and patristic distance between *A. muciniphila* strains. C) *A. muciniphila* strains differ across host classes. Multivariate analysis was conducted on the distance matrix based on the patristic distance of the GTRGAMMA maximum-likelihood (ML) tree above. (As a clarification, values under "ANOVA" represent ANOVA results testing if the multivariate dispersions (average distance to centroid calculated with betadisper) are significantly different between groups compared.)

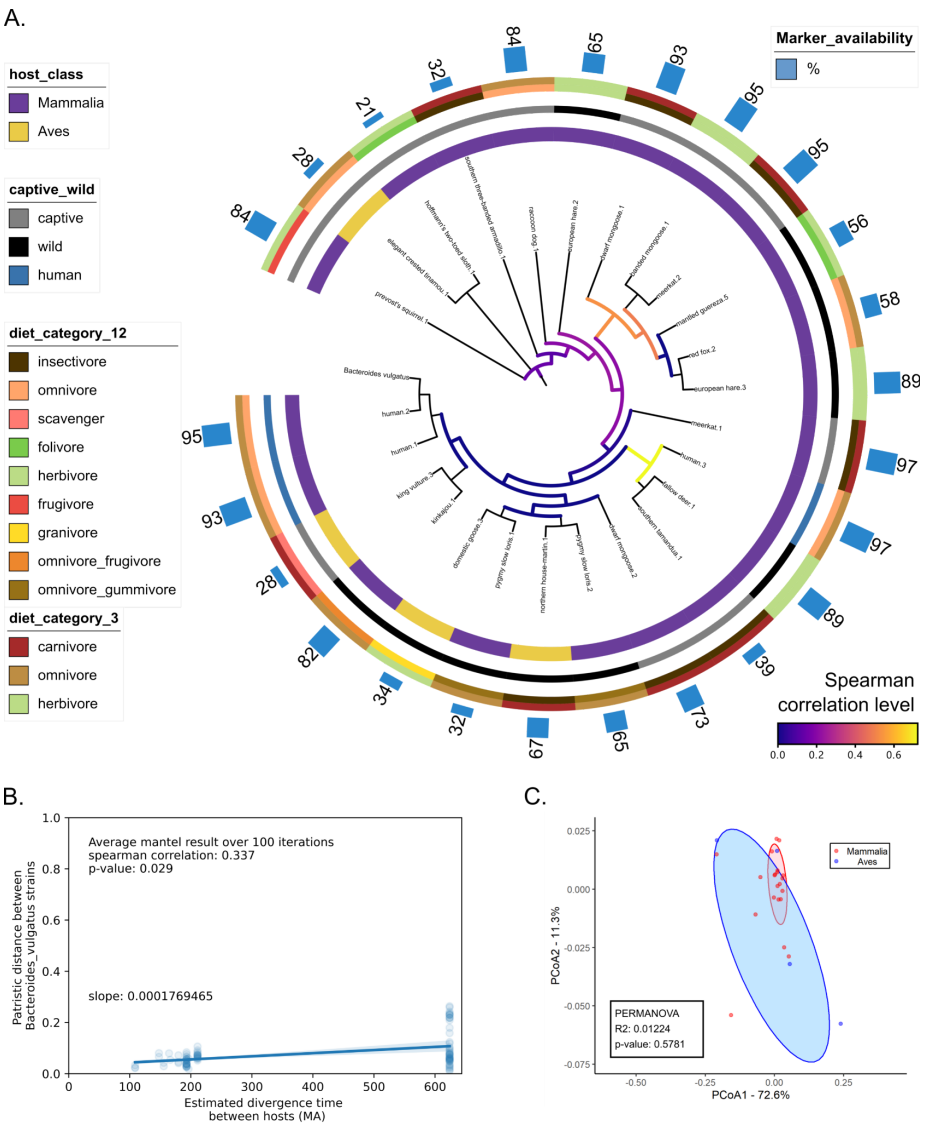

Figure 16. Analysis of *Bacteroides vulgatus* MSA with 50% marker inclusion criteria reveals evidence for strain phylosymbiosis, but no strain diversification across host classes. A) Strain tree of detected strains with tips labeled by the common name of each strain's host. Strength of strain phylosymbiosis as measured by Mantel Spearman correlation of each subtree with at least 3 leaves is represented by branch color. From the innermost to the outermost ring, strains are colored by hosts' taxonomic class, captivity information, fine and general dietary information. The percent of markers for *B. vulgatus* available for each strain for tree building is represented by the annotated bar chart. The strain tree is built with the GTRGAMMA model and visualized with branch length ignored. B) Strong signal of strain phylosymbiosis detected for *B. vulgatus*. Mantel spearman test reveals a statistically significant positive correlation between host divergence time and patristic distance between *B. vulgatus* strains. C) *B. vulgatus* strains did not differ across host classes. Multivariate analysis was conducted on the distance matrix based on the patristic distance of the GTRGAMMA ML tree above. (As a clarification, values under "ANOVA" represent ANOVA results testing if the multivariate dispersions (average distance to centroid calculated with betadisper) are significantly different between groups compared.)

Overall, *B. vulgatus* strains does not cluster significantly across host classes (Mammalia, Aves) ([PERMANOVA] $R^2 = 0.01224$, p-value = 0.5781) or general diet category ([adonis PERMANOVA] $R^2 = 0.04974$, p-value = 0.6723), but does across host captivity status ([adonis PERMANOVA] $R^2 = 0.18761$, p-value = 0.0428 * with no difference in variance) (Fig. 16C).

However, when we examine *B. vulgatus* strains at 20% marker threshold, we do see strains cluster significantly across host classes ([adonis PERMANOVA] $R^2 = 0.21473$, p-value = 0.0402*; [ANOVA of betadisper] F value = 0.5643, p-value = 0.573) (Fig. 17A). Similar to before, pairwise comparisons via Tukey's HSD test reveal no differences in means between strain groups from different host classes (Table 3)

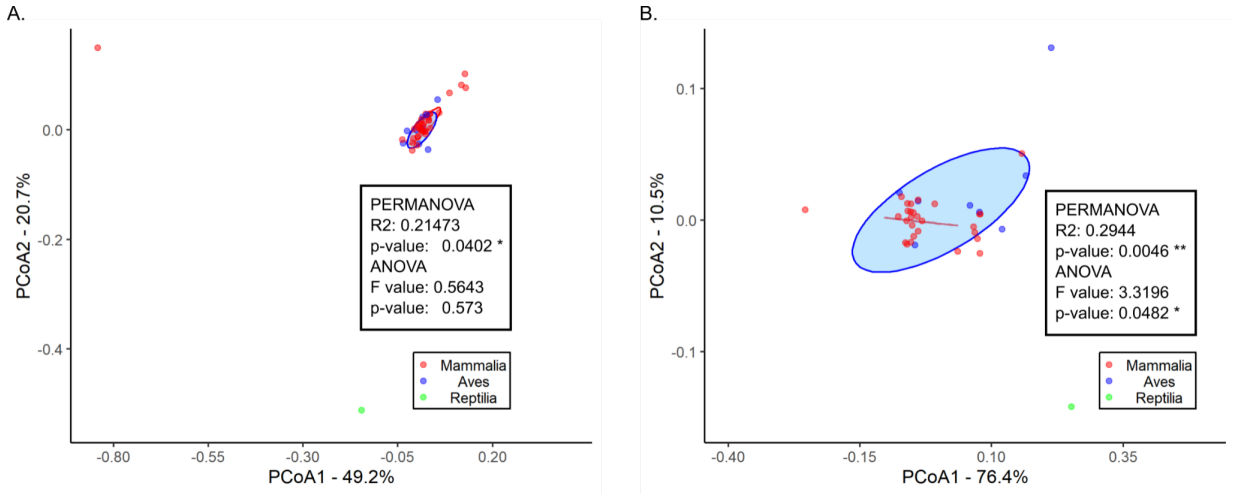

Figure 17. PCoA ordination and multivariate analysis of patristic distance between *B. vulgatus* strains found at 20% and 35% marker threshold. A) *B. vulgatus* strains found with a 20% marker threshold do cluster significantly across host classes. B) *B. vulgatus* strains found with a 35% marker threshold exhibit significant variance among groups. (As a clarification, values under "ANOVA" represent ANOVA results testing if the multivariate dispersions (average distance to centroid calculated with betadisper) are significantly different between groups compared.)

Although *E. faecalis* strains do not exhibit significant levels of strain phylosymbiosis, they do exhibit genotypic difference between strains from different host classes at 20%, 35% marker thresholds examined (20%: [adonis PERMANOVA] $R^2 = 0.22214$, p-value = 0.0002 ***

with no difference in variance; 35% [adonis PERMANOVA] R^2 =0.21023, p-value = 0.0207 * with no difference in variance) (Fig. 18). Across both marker thresholds, pairwise comparisons of strains from different host classes reveal genotypic differences between stains from mammals

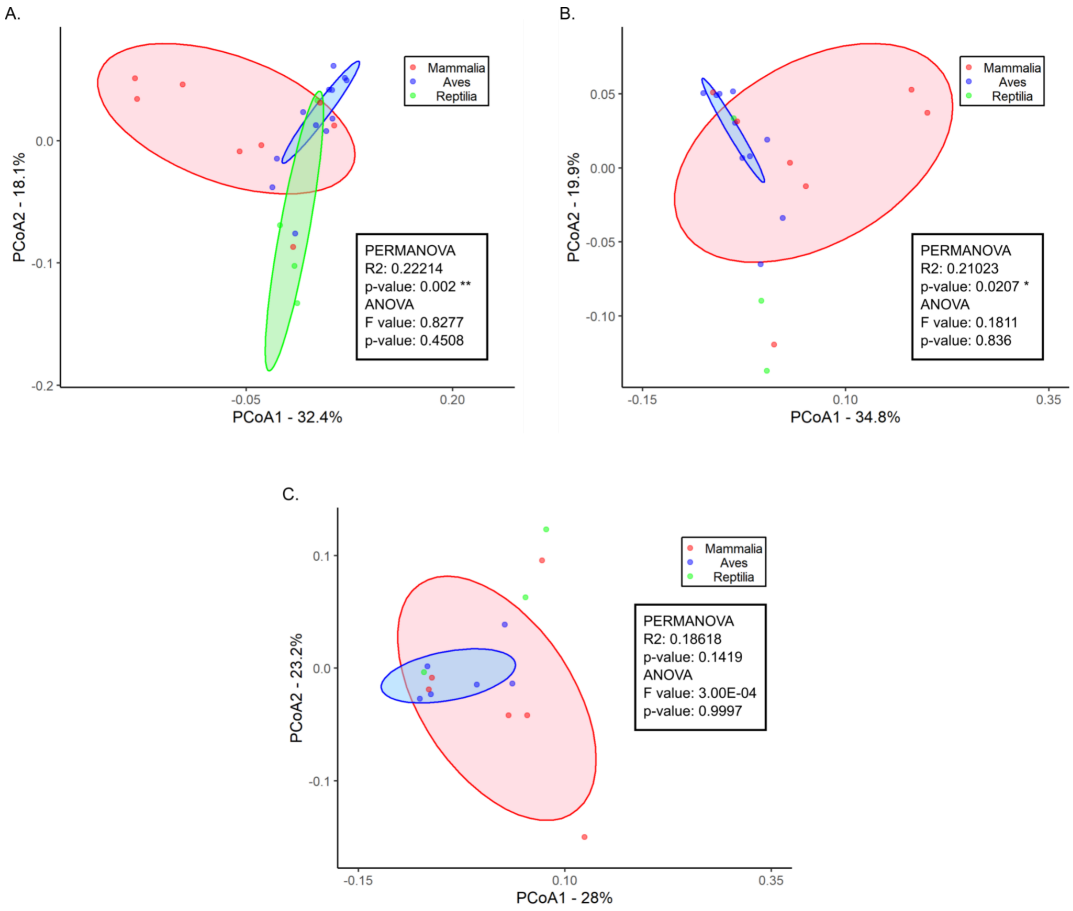

Figure 18. PCoA ordination and multivariate analysis of patristic distance between *E. faecalis* strains found at A) 20%, B) 35%, C) 50% marker thresholds. A & B) *E. faecalis* strains found with 20% and 35% marker threshold do cluster significantly across host classes. C) *E. faecalis* strains found with a 50% marker threshold does not exhibit difference across host classes. (As a clarification, values under "ANOVA" represent ANOVA results testing if the multivariate dispersions (average distance to centroid calculated with betadisper) are significantly different between groups compared.)

and birds (20% marker threshold [adonis PERMANOVA] R^2 = 0.17483, p-value = 0.0037 ** with no difference in variance; 35% marker threshold [adonis PERMANOVA] R^2 = 0.16242, p-value = 0.0113 * with no difference in variance) as well as reptiles and birds (20% marker threshold [PERMANOVA] R^2 = 0.16184, p-value = 0.0163 * with no difference in variance;

35% marker threshold [PERMANOVA] $R^2 = 0.15176$, p-value = 0.0488 * with no difference in variance). However, pairwise comparisons via Tukey's HSD test reveal no differences in means of strain groups from different host classes (Table 3)

Lastly, as the most abundant bacteria from the dataset, *Escherichia coli* was examined for its lack of strain phylosymbiosis (Fig. 19). The general level of strain phylosymbiosis was weak across the stain tree except in certain clades (Fig. 17A). Overall, we see a lack of relationship between the evolutionary relatedness of *E. coli* strains and that of their corresponding hosts ([Mantel Spearman result averaged over 100 iterations] r = -0.038, p-value =0.377) (Fig. 17B). In addition, the strains does not cluster significantly across host classes (Mammalia, Aves, Reptilia) or general diet category ([PERMANOVA] $R^2 = 0.04974$, p-value = 0.6723), but does across host captivity status ([PERMANOVA] $R^2 = 0.18761$, p-value = 0.0428 * with no difference in variance) (Fig. 19C).

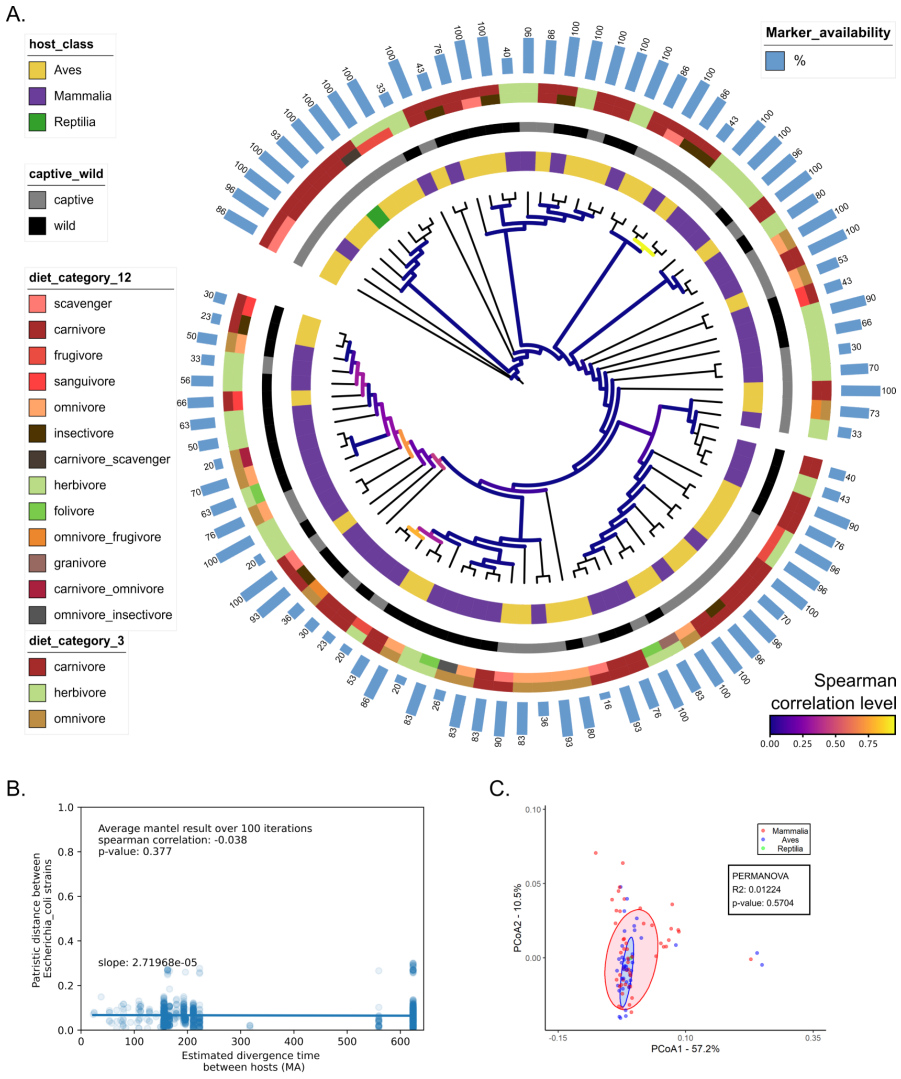

Figure 19. Analysis of *Escherichia coli* MSA with 50% marker inclusion criteria does not reveal evidence for strain phylosymbiosis and strain diversification across host classes. A) Strain tree of detected E. coli strains. Strength of strain phylosymbiosis measured by Mantel Spearman correlation is plotted as branch color. From the innermost to the outermost ring, strains are colored by hosts' taxonomic class, captivity information, fine and general dietary information. The percent of markers for *E. coli* available for each strain for tree building is represented by the annotated bar chart. The strain tree is built with the GTRGAMMA model and visualized with branch length ignored. B) Signal of strain phylosymbiosis was not detected for *E. coli* via Mantel spearman statistics. C) *E. coli* strains do not differ across host classes. Multivariate analysis was conducted on the distance matrix based on the patristic distance of the GTRGAMMA ML tree above. (As a clarification, values under "ANOVA" represent ANOVA results testing if the multivariate dispersions (average distance to centroid calculated with betadisper) are significantly different between groups compared.)

DISCUSSION

Overall, analysis of our vertebrate gut dataset found expected trends of microbiome diversity and bacterial species that exhibited strain phylosymbiosis across the animal kingdom. Diversity of the vertebrate gut microbiome was found to be higher in herbivores than carnivores and omnivores, and higher in mammals than birds (Fig. 10 & Fig. 11). These relationships of diversity found at the OGU-level echoed results from previous studies (20, 21, 42). In addition, we also found host class and general dietary strategy (herbivore, omnivore, carnivore) to affect microbiome composition (Fig. 12). Top bacteria driving differences in microbiome composition as measured by Aitchison distance are dominated by bacteria from the Proteobacteria phylum (Fig. 11). Our microbiome composition analysis provides a general understanding of the microbiome diversity represented in our dataset.

Strain phylosymbiosis analysis was conducted on 13 bacterial species in the dataset (Fig. 14). As mentioned above, only bacterial species with strains found from at least 15 distinct hosts at a given marker threshold are analyzed. All of these species are known to be common inhabitants of the human intestinal tract, except for Pseudomonas yamanorum, which is a psychrotolerant bacteria found in the subarctic soil (Fig. 14C) (41, 42). For most species, we did not observe consistent signals of strain phylosymbiosis across marker thresholds tested. Six bacterial species *E. faecalis*, *P. copri*, *B. thetaiotaomicron*, *B. fragilis*, *P. yamanorum*, *E. coli* did not exhibit signals of strain phylosymbiosis at any marker thresholds, whereas four bacterial species *C. perfringens*, *P. distasonis*, *B. dorei*, *B. uniformis* had signals of strain phylosymbiosis depending on the marker threshold used. Only *A. muciniphila* and *B. vulgatus* exhibited strain phylosymbiosis across all marker thresholds (Fig. 14).

For a species of interest, if the signal of strain phylosymbiosis is present at the host class level, we could potentially find strains to exhibit genotypic differences across host classes. Therefore, we sought to investigate evidence of host class divergence by examining if there was significant clustering of strains from different host classes. Results showed that *A. muciniphila* strains differed across host class at all examined marker thresholds, which mirrors the bacteria's signal of strain phylosymbiosis. Interestingly, when looking at the pairwise comparison between *A. muciniphila* strains from different host classes, we see a significant difference between mammalian strains and reptilian strains, and avian strains and reptilian strains, but no difference between avian strains and mammalian strains. This suggests that, at least for *A. muciniphila*, reptilian strains are significantly different from that of avian or mammalian strains, which may be due to the degree of gut physiological difference between the host class such as reptiles having a shorter intestine than that of mammals (44). Unlike that of *A. muciniphila*, *B. vulgatus* strains did not cluster significantly across host classes at 35% and 50% marker thresholds. Whereas E. faecalis, which had no signal of strain phylosymbiosis, exhibits the differences between strains of mammalian, avian, and reptilian origins.

We find that the presence of strain phylosymbiosis and analysis of strain genotype from different host classes are both sensitive to the marker threshold levels. This is expected as the level of marker threshold would determine the markers available for strain-specific consensus marker sequence construction, therefore, affect the nucleotide identity of the available strains under analysis. Since the strain trees are based on the MSA of strain-specific consensus marker sequence, differences in any consensus sequence would affect the resulting strain tree and, therefore, both the Mantel test results and multivariate results. One important consideration is that the marker threshold determines the criteria of marker inclusion by each marker's

uniqueness. At a higher marker threshold, markers included in the MSA need to be shared by a higher number of strains, whereas, at a lower marker threshold, markers shared by a lower percentage of strains (more unique) will be included in the MSA. Therefore, strains from certain host classes such as reptiles may potentially have more unique markers that can only be included in the MSA for analysis at a lower marker threshold. This would affect the host class represented by bacterial strains available for this analysis. Therefore, although ideally, a high marker threshold would improve analysis as there will be more overlap between marker sequences for comparison, lower marker thresholds also have their advantages in including markers more unique to strains from certain host clades (see Fig. 18 for an example).

It is important to note that Mantel test statistics using Spearman correlation tests for congruence of the relative positions of the leaves and not the relative distances between the leaves in the tree. This is because our Mantel test results presented in this study uses Spearman correlation which correlates the rankings of leaf pairs derived from leaf distances and not the distances themselves. For example, in a particular patristic distance matrix where the "distance A-B" is 100 branch unit, "distance B-C" is 101 branch unit, and "distance C-D" is 300 branch unit, Spearman correlation would convert the distances into the respective ranks for correlation tests while ignoring the relative unit differences (rank 1: distance A-B < rank 2: distance B-C < rank 3: distance C-D). To factor in the relative differences between branch lengths, one could theoretically apply Pearson correlation through the Mantel test statistics. Thus, we did also run Mantel tests based on Pearson correlation, the results of which mostly match results obtained with Spearman correlation. However, since Pearson correlation is a parametric test, which requires our data to satisfy a normal distribution, we decided not to present our Mantel test results based on the Pearson correlation as our input data is not normally distributed. Therefore,

our phylosymbiosis results presented above reflect the overall structural or architectural congruence between phylogenies and not the overall shape of the tree.

It is also important to consider that our strain phylosymbiosis analysis searches for structural congruences between phylogenies while being agnostic about the forces responsible for these congruences. Nevertheless, most evolutionary biologists would point to co-speciation, host-shift speciation, intrahost speciation, and extinction as the main factors affecting the congruence of phylogenies (45). As briefly touched upon before, co-speciation refers to the parallel speciation of host and microbial species (or more generally symbionts) that could occur from long-term co-diversification. Host-shift speciation, on the other hand, refers to the speciation of symbionts to occupy niches available in new host species (45). Intrahost speciation (also referred to as duplication) refers to the one-sided speciation of the symbionts within a host lineage (45). As *Vienne et al.* eloquently illustrated in their review, other than the process of co-speciation, host-shift speciation alone could also lead to congruencies of phylogenies. On the other hand, intrahost speciation and extinction events could mitigate signals of congruencies even for host-symbiont systems that experience co-speciation. To confidently deduce which of these events affected the congruences of phylogenies in the past, one would have to calculate and compare estimated divergence time via molecular clocks such as in Moeller et al. 2016. In the 103 available studies on evolutionary dynamics of host-symbiont published before 2013, the majority (48 studies) found host-shift speciation as the primary evolutionary force while less than a dozen (9 studies) found convincing cases of co-speciation. Based on these studies thus far, *Vienne et al.* reasonably point out that the host-symbiont evolutionary relationship is primarily affected by host-shift speciation events and only rarely driven by co-speciation events.

However, regardless of whether host-symbiont congruence is based on host-shift speciation or co-speciation, we propose that bacterial species with the highest chance of exhibiting strain phylosymbiosis are those species that experience consistent microbial dispersal and favorable microbial selection within the host species. Consistent microbial dispersal within species refers to the reliable, potentially long-term, exposure of microbial symbionts within populations of host species through successive generations. It is important to stress that the source of the exposure should not be from the environment or different species. Microbial dispersal is facilitated by both vertical transmission (transfer from mother to child) and horizontal transmission (transfer between members of the same species except mother to child) (15). Favorable microbial selection refers to microbial species' successful colonization of the host gut, which enables close interaction of host and microbial species. We hypothesize that only when these two conditions are met will microbial strains from different host species accumulate enough respective reciprocal genotypic changes that allow for congruent phylogenetic detection. On the other hand, we believe sufficient environmental transmission or cross-species transmission of microbial strains would overwhelm signals of strain phylosymbiosis, even if the original strain has been diversifying with the host. If our hypothesis is correct, then strain phylosymbiosis analysis across the animal kingdom may be a tool to search for bacterial species that are prone to consistent microbial dispersion and favorable microbial selection within a wide range of vertebrates. These bacterial species could be of special interest to biologists investigating host-microbe interaction or microbial adaptation to the host environment.

Interestingly, *A. muciniphila* was suggested to possess vertical transmissibility in humans as it is found to be present in human milk and being able to metabolize human milk oligosaccharides (46). Additionally, *A. muciniphila* has been found to enhance the intestinal

barrier in humans and induce an adaptive immune response in mice (47, 48). These findings suggest that *A. muciniiphila* interacts closely with hosts and, potentially, does so within many different host species. Another recent study has found vertical transmission of *B. vulgatus*, *B. fragilis*, *P. distasonis,* and *E. coli* in humans (49). A prior study by Moeller et al. has provided compelling evidence of co-speciation of hominids and bacteria of the Bacteroidaceae family including *B. vulgatus* (17). In addition to the technical limitations of our study, the lack of evidence for strain phylosymbiosis in *B. fragilis* and *E. coli* could be a result that these bacterial species experience sufficient levels of environmental transmission. In particular, *E. coli*, as a motile facultative anaerobe, can be commonly found in the environment and could potentially spread to different wildlife species via water sources.

One limitation of our analysis worth noting is the unevenly distributed dataset. Although the number of host species represented in the dataset is quite large (n=288), our dataset is heavily biased towards the mammalian class (51%) (Fig. 7). This impedes our ability to detect strain phylosymbiosis across host classes as well as analyze if there were different degrees of strain phylosymbiosis within different host classes. Given that pairwise PERMANOVA analysis of *A. muciniphila* strains revealed that there are only differences between strains of Mammalia and Reptilia, and strains of Aves and Reptilia, but not stains of Mammalia and Aves, it would be interesting to have more Reptilia samples to reaffirm if strains of reptilian origin are different than that of mammalian and avian origins. An ideal dataset to analyze strain phylosymbiosis would include deeply sequenced WGS samples ($>10^6$ reads per file) representing a large number of host species evenly distributed across host classes of interest.

One apparent limitation of our strain analysis is that 14% of samples in our dataset cannot be profiled by MetaPhlAn3 (Fig. 4). These samples did not have enough alignments against

markers of ChocoPhlAn3 database for species classification. Most of these unprofiled samples are concentrated at lower sequencing depth ($<10^6$ reads per sample) (Fig 4.). Without sufficient alignments, the strains available in the samples could not be included for the strain analysis. However, some unprofiled samples are also present at higher sequencing depth ($>10^6$ reads per sample) (Fig. 4). Therefore, low sampling depth ($<10^6$ reads per sample) should not be the only reason for the presence of unprofiled samples.

We suspect the presence of novel, uncharacterized bacteria in our samples to be a large factor in the lack of species detection. Previous de novo studies have found that, as of March 2021, large proportions of (up to 75%) metagenomic samples from wildlife microbiomes remain uncharacterized (20, 42). Since around three-quarters of our dataset is from wild animals, we should expect a large proportion of diversity to remain uncharacterized in our dataset and, therefore, unprofiled with our reference-based profiler. Therefore, the most abundant bacteria selected for strain phylosymbiosis potentially does not reflect the actual abundance of bacteria represented in our sample.

Furthermore, in addition to undetected species, we suspect that the available markers in ChocoPhlAn3 do not capture the genetic variability of bacterial species found in wildlife samples. Since these markers used for strain detection are thoughtfully selected from characterized genes in Uniprot and the genes in Uniprot is mostly assembled from studies related to human and the human gut microbiome, the strains detected by StrainPhlan 3.0 could be limited to strains that resemble characterized strains found in human (33). This would exclude highly divergent strains of a bacterial species that is only present in certain wildlife gut microbiomes, therefore, limiting our ability to detect potential signals of strain phylosymbiosis in wildlife samples.

All in all, a reference-based approach to analysis involving wildlife gut microbiome likely neglects large, crucial diversity. Given these challenges, we could utilize de novo assembly of metagenome-assembled genomes (MAGs) to alleviate this problem. This could be done in two ways. First, apply MAG-based strain profiling tools such as the newly released inStrain to the study of strain phylosymbiosis (50). With higher sensitivity than StrainPhlAn 3.0, the de novo approach of inStrain should be able to detect more strains especially those uncharacterized, allowing for more complete assessment of strain phylosymbiosis for our species of interest. And, unlike StrainPhlAn 3.0 that exclusively considers the consensus sequence that represents the dominant genotype, inStrain considers both major and minor alleles during genomic comparisons. This allows for better resolution of the natural genetic variability of strains present within a sample, which would improve the accuracy of our analysis. Another way, perhaps more technically challenging, is to utilize MAGs to find additional markers that could detect more divergent wildlife bacterial strains. Since ChocoPhlAn3 is customizable, these markers could then be added to ChocoPhlAn3 to supplement our existing marker-based strain analysis. Both methods should be able to detect more uncharacterized bacterial strains and, therefore, should improve our ability to investigate the evolutionary dynamics of strain phylosymbiosis.

Overall, our reference-based approach exhibits limitations with regards to detecting divergent strains in variable species we would expect in wildlife samples. In addition, most strain profiling tools including StrainPhlAn3 require deep sequencing depth (> 10x coverage) to perform effectively (18). However, these challenges are not inherent to the shotgun metagenomic dataset. As technology allows for sequencing with deeper coverage and new de novo methods emerge such as inStrain that utilizes MAGs to profile strains, shotgun metagenomic data could become the standard medium through which we profile strains in wildlife samples. In contrast,

traditional culture-based comparative genetics of isolate could not be used to investigate unculturable bacterial diversity within wildlife metagenomic samples. However, it remains an essential tool in identifying and tracking certain culturable bacterial strains, especially for those with clinical interest (18). Furthermore, although high-throughput single-cell sequencing offers unprecedented insight into the genetic, transcriptomic, and proteomic composition of individual cells within a cellular community, this emerging technology is more applicable for the analysis of eukaryotic cells than microbial cells from gut samples (18). This is, in large part, due to the heterogeneity of microbial cell walls and complexities associated with environmental samples such as animal stool (18). The Shotgun metagenomic dataset should remain a compelling medium to study bacterial strains found on or within wildlife in the foreseeable future.

In summary, as an exploratory study investigating strain phylosymbiosis across the animal kingdom, we were able to detect signals of strain phylosymbiosis in various bacteria. Within the confines of a marker-based approach, we found *A. muciniphila* and *B. vulgatus* exhibit the strongest signals of strain phylosymbiosis. The correlation between the strain of these bacteria and their respective host, once again, highlights the interconnectedness of the host-microbiota system. Future studies should utilize a de novo approach to the study of strain phylosymbiosis and strive to uncover the intricate biological mechanisms underlying this interesting evolutionary dynamic within our tree of life.

Material from this thesis is currently being prepared for submission for publication of the material. Chiu, Jeffrey H.; Song, Se J.; Cantu, Victor; Shaffer, Justin; Lutz, Holly L.; Knight, Rob. The thesis author was the primary author of this material.

REFERENCES

1. Hall, N. Advanced sequencing technologies and their wider impact in microbiology. Journal of Experimental Biology vol. 210 1518–1525 (2007).

2. Pallen, M. J., Loman, N. J. & Penn, C. W. High-throughput sequencing and clinical microbiology: Progress, opportunities and challenges. Current Opinion in Microbiology vol. 13 625–631 (2010).

3. Logares, R., Haverkamp, T. H. A., Kumar, S., Lanzén, A., Nederbragt, A. J., Quince, C., & Kauserud, H. (2012). Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches. In *Journal of Microbiological Methods* (Vol. 91, Issue 1, pp. 106–113). Elsevier. https://doi.org/10.1016/j.mimet.2012.07.017

4. Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., & Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature*, *568*(7753), 499–504. https://doi.org/10.1038/s41586-019-0965-1

5. Hasan, N. & Yang, H. Factors affecting the composition of the gut microbiota, and its modulation. PeerJ 7, (2019).

6. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., Fitzgerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., … White, O. (2012). Structure, function and diversity of the healthy human microbiome. *Nature 2012 486:7402*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

7. Gilbert, J., Blaser, M. J., Caporaso, J. G., Jansson, J., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*, *24*(4), 392. https://doi.org/10.1038/NM.4517

8. Mayer, E. A., Knight, R., Mazmanian, S. K., Cryan, J. F. & Tillisch, K. Gut microbes and the brain: Paradigm shift in neuroscience. J. Neurosci. 34, 15490–15496 (2014).

9. Fung, T. C., Olson, C. A. & Hsiao, E. Y. Interactions between the microbiota, immune and nervous systems in health and disease. Nature Neuroscience vol. 20 145–155 (2017).

10. Hosokawa, T. & Fukatsu, T. Relevance of microbial symbiosis to insect behavior. Curr. Opin. Insect Sci. 39, 91–100 (2020).

11. Buchner P. 1965. Symbiosis in animals which suck plant juices. In Endosymbiosis of Animals with Plant Microorganisms, pp. 210–432. New York: Interscience

12. Hanning, I., Diaz-Sanchez, S. The functionality of the gastrointestinal microbiome in non-human animals. Microbiome 3, 51 (2015). https://doi.org/10.1186/s40168-015-0113-6

13. SW, K., AS, E. & RM, E. The alligator gut microbiome and implications for archosaur symbioses. Sci. Rep. 3, 2877–2877 (2013).

14. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: Evolution of the vertebrate gut microbiota. Nat. Rev. Microbiol. 6, 776–788 (2008)

15. Mallott, E.K., Amato, K.R. Host specificity of the gut microbiome. Nat Rev Microbiol (2021). https://doi.org/10.1038/s41579-021-00562-3

16. Brooks, A. W., Kohl, K. D., Brucker, R. M., van Opstal, E. J. & Bordenstein, S. R. Correction to: Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History (PLOS Biology, (2016), 14, 11, (e2000225), 10.1371/journal.pbio.2000225). PLoS Biol. 15, 1–29 (2017).

17. Moeller, A. H., Caro-Quintero, A., Mjungu, D., Georgiev, A. V., Lonsdorf, E. V., Muller, M. N., Pusey, A. E., Peeters, M., Hahn, B. H., & Ochman, H. (2016). Cospeciation of gut microbiota with hominids. Science, 353(6297), 380–382. https://doi.org/10.1126/science.aaf3951

18. Yan, Y., Nguyen, L. H., Franzosa, E. A., & Huttenhower, C. (2020). Strain-level epidemiology of microbial communities and the human microbiome. *Genome Medicine 2020 12:1*, *12*(1), 1–16. https://doi.org/10.1186/S13073-020-00765-Y

19. Song, S. J., Sanders, J. G., Delsuc, F., Metcalf, J., Amato, K., Taylor, M. W., Mazel, F., Lutz, H. L., Winker, K., Graves, G. R., Humphrey, G., Gilbert, J. A., Hackett, S. J., White, K. P., Skeen, H. R., Kurtis, S. M., Withrow, J., Braile, T., Miller, M., … Knight, R. (2020). Comparative analyses of vertebrate gut microbiomes reveal convergence between birds and bats. MBio, 11(1), 1–14. https://doi.org/10.1128/mBio.02901-19

20. Youngblut, N. D., de la Cuesta-Zuluaga, J., Reischer, G. H., Dauser, S., Schuster, N., Walzer, C., Stalder, G., Farnleitner, A. H., & Ley, R. E. (2020). Large-Scale Metagenome

Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity. *MSystems*, *5*(6). https://doi.org/10.1128/MSYSTEMS.01045-20

21. Amato, K. R., G. Sanders, J., Song, S. J., Nute, M., Metcalf, J. L., Thompson, L. R., Morton, J. T., Amir, A., J. McKenzie, V., Humphrey, G., Gogul, G., Gaffney, J., L. Baden, A., A.O. Britton, G., P. Cuozzo, F., Di Fiore, A., J. Dominy, N., L. Goldberg, T., Gomez, A., … R. Leigh, S. (2019). Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. *ISME Journal*, *13*(3), 576–587. https://doi.org/10.1038/s41396-018-0175-0

22. Qiita: rapid, web-enabled microbiome meta-analysis. Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D. Swafford, Stephanie B. Orchanian, Jon G. Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J. Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J. Gregory Caporaso, Pieter C. Dorrestein & Rob Knight. Nature Methods, volume 15, pages 796–798 (2018); https://doi.org/10.1038/s41592-018-0141-9.

23. Youngblut, N. D., Reischer, G. H., Walters, W., Schuster, N., Walzer, C., Stalder, G., Ley, R. E., & Farnleitner, A. H. (2019). Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. *Nature Communications*, *10*(1), 1–15. https://doi.org/10.1038/s41467-019-10191-3

24. Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol doi:10.1093/molbev/msx116

25. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34:3094-3100. doi:10.1093/bioinformatics/bty191

26. Genome [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2021 Sep 16]. Available from: https://www.ncbi.nlm.nih.gov/genome/

27. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology 2009 10:3*, *10*(3), 1–10. https://doi.org/10.1186/GB-2009-10-3-R25

28. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A,

Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857. https://doi.org/10.1038/s41587-019-0209-9

29. Zhu, Q., Huang, S., Gonzalez, A., McGrath, I., McDonald, D., Haiminen, N., Armstrong, G., Vázquez-Baeza, Y., Yu, J., Kuczynski, J., Sepich-Poore, G. D., Swafford, A. D., Das, P., Shaffer, J. P., Lejzerowicz, F., Belda-Ferre, P., Havulinna, A. S., Méric, G., Niiranen, T., … Knight, R. (2021). OGUs enable effective, phylogeny-aware analysis of even shallow metagenome community structures. *BioRxiv*, 2021.04.04.438427. https://doi.org/10.1101/2021.04.04.438427

30. Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R. R., Zhu, Q., Knight, R., & Knights, D. (2020). SHOGUN: a modular, accurate and scalable framework for microbiome quantification. *Bioinformatics*, *36*(13), 4088–4090. https://doi.org/10.1093/BIOINFORMATICS/BTAA277

31. Zhu Q*, Mai U*, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, Kosciolek T, Yin JB, Huang S, Salam N, Jiao J, Wu Z, Xu ZZ, Sayyari E, Morton JT, Podell S, Knights D, Li W, Huttenhower C, Segata N, Smarr L, Mirarab S, Knight R. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nature Communications. 2019. 10(1):5477. doi: 10.1038/s41467-019-13443-4.

32. Martino, C., Morton, J. T., Marotz, C. A., Thompson, L. R., Tripathi, A., Knight, R., & Zengler, K. (2019). A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *MSystems*, *4*(1). https://doi.org/10.1128/MSYSTEMS.00016-19

33. Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *ELife*, *10*. https://doi.org/10.7554/ELIFE.65088

34. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Res. 27, 626–638 (2017).

35. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).

36. Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010) (publisher link)

37. Letunic I and Bork P (2021) Nucleic Acids Res doi: 10.1093/nar/gkab301 Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation (full article)

38. Paradis E, Schliep K (2019). "ape 5.5: an environment for modern phylogenetics and evolutionary analyses in R." Bioinformatics, 35, 526-528.

39. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

40. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2020). vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan

41. John G. Holt ... [and others]. Bergey's Manual of Determinative Bacteriology. Baltimore :Williams & Wilkins, 1994.

42. Levin, D., Levin, D., Raab, N., Pinto, Y., Rothschild, D., Zanir, G., & Godneva, A. (2021). *Diversity and functional landscapes in the microbiota of animals in the wild*. *5352*(March), 1–20.

43. Arnau VG, Sánchez LA, Delgado OD. Pseudomonas yamanorum sp. nov., a psychrotolerant bacterium isolated from a subantarctic environment. Int J Syst Evol Microbiol. 2015

44. Hoppe, M. I., Meloro, C., Edwards, M. S., Codron, D., Clauss, M., & Duque-Correa, M. J. (2021). Less need for differentiation? Intestinal length of reptiles as compared to mammals. *PLOS ONE*, *16*(7), e0253182. https://doi.org/10.1371/JOURNAL.PONE.0253182

45. Vienne, D. M. de, Refrégier, G., López-Villavicencio, M., Tellier, A., Hood, M. E., & Giraud, T. (2013). Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytologist*, *198*(2), 347–385. https://doi.org/10.1111/NPH.12150

46. Qixiao Zhai, Saisai Feng, Narbad Arjan & Wei Chen (2019) A next generation probiotic, Akkermansia muciniphila, Critical Reviews in Food Science and Nutrition, 59:19, 3227-3236, DOI: 10.1080/10408398.2018.1517725

47. Naito, Y., Uchiyama, K. & Takagi, T. A next-generation beneficial microbe: Akkermansia muciniphila. J. Clin. Biochem. Nutr. 63, 33 (2018).

48. Ansaldo, E., Slayden, L. C., Ching, K. L., Koch, M. A., Wolf, N. K., Plichta, D. R., Brown, E. M., Graham, D. B., Xavier, R. J., Moon, J. J., & Barton, G. M. (2019). *Akkermansia muciniphila induces intestinal adaptive immune responses during homeostasis Eduard*. *01*(June), 1–7.

49. Li, W., Tapiainen, T., Brinkac, L., Lorenzi, H. A., Moncera, K., Tejesvi, M. V, Salo, J., & Nelson, K. E. (2020). Vertical Transmission of Gut Microbiome and Antimicrobial Resistance Genes in Infants Exposed to Antibiotics at Birth. *The Journal of Infectious Diseases*. https://doi.org/10.1093/infdis/jiaa155

50. Olm, M. R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B. A., Morowitz, M. J., & Banfield, J. F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*, *39*(6), 727–736. https://doi.org/10.1038/s41587-020-00797-0