# Incorporating genome-based phylogeny and trait similarity into diversity assessments helps to resolve a global collection of human gut metagenomes

Nicholas D. Youngblut*[,1], Jacobo de la Cuesta-Zuluaga[1], Ruth E. Ley[1]

[1]Department of Microbiome Science, Max Planck Institute for Developmental Biology, Max Planck Ring 5, 72076 Tübingen, Germany

* Corresponding author: Nicholas Youngblut

**Running title:** Phylogeny and traits explain metagenome diversity

**Key words:** human gut, metagenome, phylogeny, traits, diversity

## Abstract

Tree-based diversity measures incorporate phylogenetic or phenotypic relatedness into comparisons of microbial communities. This improves the identification of explanatory factors compared to tree-agnostic diversity measures. However, applying tree-based diversity measures to metagenome data is more challenging than for single-locus sequencing (*e.g.,* 16S rRNA gene). The Genome Taxonomy Database (GTDB) provides a genome-based reference database that can be used for species-level metagenome profiling, and a multi-locus phylogeny of all genomes that can be employed for diversity calculations. Moreover, traits can be inferred from the genomic content of each representative, allowing for trait-based diversity measures. Still, it is unclear how metagenome-based assessments of microbiome diversity benefit from incorporating phylogeny or phenotype into measures of diversity. We assessed this by measuring phylogeny-based, trait-based, and tree-agnostic diversity measures from a large, global collection of human gut metagenomes composed of 33 studies and 3348 samples. We found phylogeny- and trait-based alpha diversity to better differentiate samples by westernization, age, and gender. PCoA ordinations of phylogeny- or trait-based weighted UniFrac explained more variance than tree-agnostic measures, which was largely a result of these measures emphasizing inter-phylum differences between *Bacteroidaceae* (*Bacteroidota*) and *Enterobacteriaceae* (*Proteobacteria*) versus just differences within *Bacteroidaceae* (*Bacteroidota*). The disease state of samples was better explained by tree-based weighted UniFrac, especially the presence of Shiga toxin-producing *E. coli* (STEC) and hypertension. Our findings show that metagenome diversity estimation benefits from incorporating a genome-derived phylogeny or traits.

## Importance

Estimations of microbiome diversity are fundamental to understanding spatiotemporal changes of microbial communities and identifying which factors mediate such changes. Tree-based measures of diversity are widespread for amplicon-based microbiome studies due to their utility relative to tree-agnostic measures; however, tree-based measures are seldomly applied to shotgun metagenomics data. We evaluated the utility of phylogeny-, trait-, and tree-agnostic diversity measures on a large scale human gut metagenome dataset to help guide researchers with the complex task of evaluating microbiome diversity via metagenomics.

## Introduction

Sequencing-based assessments of microbiome diversity are fundamental to the field of microbiome science. For instance, 16S rRNA gene and metagenomic sequence-based estimations of human gut microbiome diversity have shown substantial differences among individuals due to disease state, diet, exercise, hygiene, and antibiotic usage (Sommer and Bäckhed, 2013). The choice of diversity measure can be critical, as exemplified in many studies where only diversity assessments that incorporated microbial phylogenetic relatedness were discriminatory, while tree-agnostic diversity measurements could not distinguish between groupings (Bassett *et al.*, 2015; Obregon-Tito *et al.*, 2015; Vogt *et al.*, 2017; Torres *et al.*, 2018). Without incorporating a phylogeny, all microbes in a community are treated as equally related (*i.e.*, a star phylogeny), so within-genus differences in species composition are weighted the same as compositional differences spanning multiple phyla or domains (Matsen, 2015). Closely related species are often phenotypically similar and occupy comparable niches; therefore, a measure of diversity that incorporates phylogenetic information can indirectly assess functional overlap among microbial communities (Lozupone and Knight, 2008). Such an approach is quite powerful, considering that the majority of microbes remain uncultured, and that the common approach of 16S rRNA gene sequencing can only provide coarse inferences of phenotype due to the lack of taxonomic resolution (Hugerth and Andersson, 2017; Louca, Doebeli and Parfrey, 2018). Still, trait-based assessments of microbiome diversity that focused on a few key phenotypes have been employed with great effect in some circumstances (Ortiz-Álvarez *et al.*, 2018; Guittar, Shade and Litchman, 2019). More generally, phylogeny-based measures of within-sample and between-sample diversity (alpha and beta diversity, respectively) have become commonplace for microbiome studies relying on 16S rRNA sequencing (Lozupone and Knight, 2008; Hamady, Lozupone and Knight, 2010).

As the cost of sequencing has declined, shotgun metagenomics has risen in popularity relative to single-locus sequencing, as metagenomics provides a great wealth of information, including i) accurate species-level taxonomic classification and abundance estimation, ii) information on gene and metabolic pathway content, and iii) the ability to assemble genes and metagenome-assembled genomes (MAGs) (Lu *et al.*, 2017; Parks *et al.*, 2017; Franzosa *et al.*, 2018). Recent work has shown that shallow sequencing depths can provide similar or greater coverage of microbial diversity compared to 16S rRNA sequencing (Hillmann *et al.*, 2018). However, generating a phylogeny from shotgun metagenome data is inherently challenging, since read sequences originate from all genomic locations instead of a single locus (Kembel *et al.*, 2011). Various methods exist for extracting 16S rRNA gene sequences, other single loci, or multi-locus data from metagenome reads (*e.g.,* EMIRGE, MATAM, AMPHORA2, PhyloSIFT, and MetaPhlAn2), but this excludes much of the data, limiting the detection sensitivity for less common taxa (Miller *et al.*, 2011; Segata *et al.*, 2012; Wu and Scott, 2012; Darling *et al.*, 2014; Pericard *et al.*, 2018). Alternatively, assembling MAGs enables the construction of multi-locus phylogenies from all assembled genomes, but a very high sequencing depth is required to assemble even the majority of taxa in a diverse microbial community like in soil or the human gut. Another approach is to map all reads to a database of entire genomes (*e.g.,* GenBank or RefSeq), which increases the amount of reads classified relative to single- or multi-locus approaches, but such databases generally lack careful curation of genome assembly quality, a standardized taxonomy, and multi-locus phylogenies for all reference genomes (Parks *et al.*, 2018).

We recently created a pipeline for generating custom metagenome profiling databases from the Genome Taxonomy Database (GTDB) (Parks *et al.*, 2018; de la Cuesta-Zuluaga, Ley and Youngblut, 2019), which is a comprehensive database of *Bacteria* and *Archaea* genomes, that not only provides a coherent microbial taxonomy based on genome relatedness, but it also includes multi-locus genome phylogenies for the reference species. Therefore, one can map all reads to the GTDB reference genomes in order to infer species-level abundances (*e.g.,* with Kraken2) and then utilize a genome phylogeny of reference species for calculating alpha and beta diversity (Wood, Lu and Langmead, 2019). Importantly, the multi-locus genome phylogeny will almost definitely be more robust and better-resolved than a phylogeny inferred from small, hypervariable regions of the 16S rRNA gene, or even the full-length gene sequence (Maiden *et al.*, 1998). Using species-level reference genomes also enables trait inference directly from the loci present in each genome, which is a task that state-of-the-art classifiers can perform accurately, at least for certain common phenotypes such as cell morphology, anaerobiosis, spore formation, and utilization of certain carbohydrates (Weimann *et al.*, 2016). Trait relatedness can then be represented in a tree format by hierarchical clustering of genomes based on trait presence/absence to produce a dendrogram.

While promising, this approach of species-level metagenome profiling, followed by phylogeny- or trait-based diversity calculation has not been robustly assessed and compared to tree-agnostic approaches that are often used for shotgun metagenome studies. We therefore applied this methodology to a large, global human gut metagenome collection, comprising 33 datasets and 3348 samples. We found that, in comparison to tree-agnostic measures, both phylogeny- and trait-based measures of alpha and beta diversity improved our ability to discriminate metagenome samples based on westernization, disease status, age, and gender.

## Methods

### *Data Retrieval*

We retrieved publicly available human gut metagenomes from the Sequence Read Archive (SRA) between December 2019 and February 2020 (Table S1). Sample metadata was obtained from the curatedMetagenomicData v.1.17.0 Bioconductor package (Pasolli *et al.*, 2017) and included according to the following criteria: i) shotgun metagenomes sequenced using the Illumina HiSeq platform with a median read length >95 bp; ii) with available SRA accession; iii)  labeled as adults or seniors, or with a reported age ≥18 years; iv) without report of antibiotic consumption (*i.e.*, no or NA); v) without report of pregnancy (*i.e.*, no or NA); vi) non-lactating women (*i.e.*, no or NA); vii) without report of gangrene, pneumonia, cellulitis, adenoma, colorectal cancer, arthritis, Behcet's disease, cirrhosis or inflammatory bowel disease. Only forward reads were downloaded and further processed. The final dataset was composed of 3348 samples from 33 studies.

### *Sequence processing and taxonomic profiling*

We used the bbtools “bbduk” command and Skewer v0.2.2  (Jiang *et al.*, 2014) to trim adapters and quality-filter raw sequences. The “bbmap” command from bbtools was used to remove human reads mapping to the human genome hg19 assembly. We created quality reports for each step using fastqc v0.11.7 (https://github.com/s-andrews/FastQC) and multiQC v.1.5a (Ewels *et al.*, 2016). Filtered reads were subsampled to 1 million reads per sample and used to obtain taxonomic profiles using Kraken2 (Wood, Lu and Langmead, 2019) and Bracken v2.2 (Lu *et al.*, 2017). Custom databases of Bacteria and Archaea were created using Struo v0.1.6 (de la Cuesta-Zuluaga, Ley and Youngblut, 2019) and based on the Genome Taxonomy Database (GTDB), Release 89.0 (“GTDB-r89”; available at http://ftp.tue.mpg.de/ebio/projects/struo/) (Parks *et al.*, 2018).

### *Genome phylogeny*

The GTDB-r89 “Arc122” and “Bac120” multi-locus phylogenies were obtained from the GTDB ftp server (https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/). The ape R package was used to merge the trees and prune them to the 23,360 species in the Struo-generated GTDB-r89 metagenome profiling database.

### *Trait inference*

We generated a Python v3 implementation of traitar (Weimann *et al.*, 2016) and used it to predict traits for all genomes (https://github.com/nick-youngblut/traitar3), with majority-rules (phypat+PGL model) used for classifying trait presence/absence. We used the vegan R package (Oksanen *et al.*, 2012) to apply the Jaccard dissimilarity metric to the binary traits-per-genome matrix in order to create a distance matrix of trait relatedness among genomes. This distance matrix was clustered via the UPGMA algorithm to create a dendrogram, which was used for tree-based alpha and beta diversity metrics.

### *Congruence of the genome phylogeny and trait similarity*

Global congruence of the genome phylogeny and trait similarity dendrogram was assessed via phytools::cospeciation with 100 permutations for the null model. Local congruence (*i.e.,* per-clade) was assessed via Procrustes superimposition (vegan::procrustes) comparing the genome phylogeny patristic distance matrix versus the Jaccard distance matrix used to generate the trait similarity dendrogram. Due to memory limitation issues with the standard approach for converting a phylogeny to a patristic distance matrix in R (*i.e.,* the “cophenetic” function in the ape R package can only process trees with fewer than ~13,000 tips), we instead ran the procrustes analysis on 1000 randomly pruned subtrees of 1000 tips each and used the mean residuals across all permutations for each taxon.

### *Alpha diversity*

We calculated tree-agnostic (species richness and Shannon index) and tree-based (Faith’s Phylogenetic Diversity: Faith’s PD) alpha diversity measures. Species richness and Shannon index. All tree-agnostic measures were calculated with the vegan R package (vegan::diversity), and Faith’s PD was calculated with PhyloMeasures::pd.query. The lme4 and lmerTest R packages were used to fit linear mixed effects models with dataset as random effect and other variables as fixed effects; *F*-tests and *P*-values were determined via the Satterthwaite's method (ANOVA Type II sum of squares). We adjusted *P*-values for multiple comparisons using the Benjamini-Hochberg method.

### *Beta diversity analyses*

Tree-agnostic weighted and unweighted intersample distances (Bray-Curtis dissimilarity and Jaccard index) were calculated using the vegan R package (vegan::vegdist), while tree-based metrics (weighted and unweighted UniFrac) were calculated with rbiom::unifrac. Principal coordinates analysis (PCoA) was applied to each distance matrix via stats::cmdscale. We used the vegan::envfit function to assess correlations of species abundances to each PCoA. PERMANOVA was performed with vegan::adonis2 (999 permutations; marginal effects of terms assessed). We assessed PCoA ordination similarity via Procrustes superimposition (999 permutations).

### *General data analysis*

General data processing was performed with the tidyverse package in R. The ggplot2 package was used for generating all plots. All code used for this work is available on GitHub at https://github.com/leylabmpi/global_metagenome_diversity.

### *Data availability*

The genome phylogeny, trait tables for each species-genome representative, and trait dendrograms are available at http://ftp.tue.mpg.de/ebio/projects/struo/GTDB_release89/.

## Results

### *Dataset summary*

Our combined human gut metagenome dataset consisted of 33 studies and a total of 3348 samples from 3011 individuals after filtering by required metadata fields and an adequate number of reads following quality control (101 ± 163 s.d. samples per study; Figure S1A). The percent of metagenome reads classified  to our custom GTDB-r89 Kraken2 database was high (mean of 80%), and generally lowest for non-westernized populations (Figure S1B).

### *Broad-scale incongruences between trait and phylogenetic similarity*

To assess alpha and beta diversity based on phenotypic similarity, we inferred the presence/absence of 67 traits for each reference genome in our Kraken2 database (Figure 1A). We quantified the degree of congruence between phylogeny- and trait-based relatedness of all species (taxonomy defined by the GTDB), in order to assess whether each would reveal different patterns of alpha and beta diversity. Congruence was measured via Procrustes superimposition, in which larger incongruences between phylogenetic and trait similarity among taxa will produce larger Procrustes residuals. We found that the congruence between trait and phylogenetic similarity differed greatly across phyla (Figure 1B). The bacterial phyla *Dependentiae*, *Fusobacteriota*, and *Verrucomicrobiota_A* were the most congruent between trait and phylogenetic similarity, while most of the archaeal phyla, including the *Crenarcheota*, *Thermoplasmatota*, and *Nanoarchaeota* were the most incongruent. Notably, *Crenarcheota* were also found in a recent study to be especially variable in phenotypes, as defined by overlap in clusters of orthologous groups (COG) functional categories (Royalty and Steen, 2019). *Firmicutes* and *Proteobacteria* showed the greatest variance in congruence, with many highly incongruent outlier species in both phyla. An inspection at the family level revealed that the *Firmicutes* outliers belonged to *Enterobacteriaceae*, while the *Mycoplasmoidaceae* and *Metamycoplasmataceae* families were the largest outliers in *Proteobacteria* (Figure S2). *Euryarchaeota* trait-phylogeny congruence was relatively high for an archaeal phylum; however, the *Methanosphaera* genus comprised many highly incongruent outliers. Large differences between phylogeny and phenotype in these families may be due to high phenotypic plasticity relative to core genome evolutionary rates. Overall, our findings show that trait and phylogenetic similarities are only partially congruent and would thus likely describe different aspects of microbiome diversity when applied to tree-based diversity measures.

### *More variance is explained by alpha diversity measures incorporating phylogenetic or trait relatedness*

We calculated alpha diversity for all 3348 metagenome samples with four measures: the number of observed taxa, the Shannon Index, and Faith's Phylogenetic Diversity (Faith's PD) with either the genome phylogeny ("PD_phy") or a dendrogram depicting trait relatedness ("PD_trt"). We note that all metagenomes were subsampled to 1 million reads prior to metagenome profiling and thus alpha diversity estimates should not be biased by sampling depth. Both PD_phy and PD_trt clearly separated metagenome samples based on westernization status, while such a separation was less discernible when using the Shannon Index or number of observed species (Figure 2A). When assessing samples with westernization status, age, and gender metadata ($n$ = 1843), we also found that PD_phy and PD_trt more clearly differentiate groups along each variable (Figure 2B). Indeed, linear mixed effects models produced substantially higher and lower $F$-values and $P$-values, respectively, for PD_phy and PD_trt in regards to westernization status, age, and gender (Figure 2C). $F$-values were also slightly higher for BMI when filtering the dataset to just samples with all required metadata ($n$ = 918; Figure 2C). PD_phy $F$-values were consistently higher for age and especially for westernization compared to PD_trt. Indeed, the number of phyla per sample was substantially higher for non-westernized individuals versus westernized (Figure S3A), while no substantial difference was seen for the number of genera (Figure S3B). This finding indicates that coarse taxonomic groups differ substantially by westernization status, which would be emphasized via a phylogenetic measure of diversity. While the boxplots hinted at a substantially greater differentiation between westernized and non-westernized males versus when comparing females (Figure S3B), we did not find a significant interaction between gender and westernization for any diversity measure ($P > 0.1$).

To resolve how the choice of diversity measure influenced per-clade estimations of diversity, we applied our mixed effects model analysis on alpha diversity calculated for each individual family (Figure S4). For all diversity measures, the *Bacteroidales* family F082 was most strongly associated with westernization status, and the strength of association was very consistent among measures. In contrast, most of the other families associated with westernization differed in their strength among the diversity measures (Figure S4A). For instance, the association of *Treponemataceae* was substantially weaker for the Shannon Index versus either tree-based measure. This inconsistency among diversity measures was also observed for associations between family-level diversity and gender or age. *Akkermansiaceae* had the strongest association with gender, but only for Faith's PD based on trait similarity (Figure S4B), suggesting functional differentiation at fine taxonomic levels. Notably, *Methanobacteriaceae* alpha diversity was most strongly associated with age, along with *Butyricicoccaceae*, but the association strength was much lower when measuring diversity via PD_phy versus PD_trt or the Shannon index (Figure S4C). These examples show that fine taxonomic level diversity estimations can differ substantially depending on which aspects of diversity are emphasized: phylogenetic relatedness, trait relatedness, or neither.

### *More variance is explained by beta diversity measures incorporating phylogenetic or trait relatedness*

We calculated beta diversity on all metagenome samples with 6 metrics: Bray Curtis, Jaccard, and UniFrac in all four combinations of unweighted and weighted with either a genome phylogeny or trait-similarity dendrogram. Principal coordinate analysis (PCoA) revealed that substantially more variance was explained by the top principle coordinates (PCs) for both phylogeny-based weighted UniFrac ("w-unifrac_phy") and trait-based weighted UniFrac ("w-unifrac_trt") (Figure 3). This was especially apparent for w-unifrac_phy, with 50% variance explained by PC1 alone, while only 15.4 and 9.3% variance was explained by PC1 for Bray-Curtis and Jaccard, respectively (Figure 3B). In contrast to weighted UniFrac, both unweighted UniFrac measures showed similar amounts of variance explained relative to Bray-Curtis and Jaccard. When summing across the top 5 PCs (Figure 3C), w-unifrac_phy explained 79.1% of the variance, which is more than twice that of Bray-Curtis (38.2%) and more than three times as much as Jaccard (23.8%). The summed percent variance explained by w-unifrac_trt was also substantially higher (53.3%) than Bray-Curtis and Jaccard.

We investigated why w-unifrac_phy and w-unifrac_trt explained so much more variance by correlating species abundances with each of the top three PCs (Figure 4A). The analysis revealed that the top w-unifrac_phy and w-unifrac_trt PCs most strongly differentiates samples based on the abundances of species belonging to *Lachnospiraceae* (*Firmicutes_A*), *Bacteroidaceae* (*Bacteroidota*), and *Enterobacteriaceae* (*Proteobacteria*). In contrast, Bray-Curtis and Jaccard most strongly discern samples differing in species just within *Bacteroidaceae* (*Bacteroidota)*. Specifically, the top PCs for Bray-Curtis and Jaccard correlate strongly with the *Bacteroidaceae* genera: *Bacteroidetes*, *Bacteroidetes_B*, and *Prevotella* (Figure S5). Unlike the weighted UniFrac measures, both unweighted UniFrac measures lacked a strong correlation with *Enterobacteriaceae*, but they did uniquely discern *Oscillospiraceae* (*Firmicutes_A*) and *Ruminococcaceae* (*Firmicutes_A*).

To help illustrate these clade-level differences among the beta diversity measures, we mapped the abundances of these focal clades onto each PCoA ordination. As denoted by our correlation analysis, *Bacteroidaceae* was highly abundant at both ends of PC1 for Bray-Curtis and Jaccard, while its abundance was lowest at the center of the PC (Figure 4B). Conversely, *Bacteroidaceae* was only highly abundant on one side of PC1 for both w-unifrac_phy and w-unifrac_trt. In contrast to *Bacteroidaceae*, *Enterobacteriaceae* was only detectable in 350 samples, with only 28 samples having >1% abundance (Figure 4C). w-unifrac_trt best partitioned the samples with high versus low levels of *Enterobacteriaceae* (Figure 4A & 4C), while w-unifrac_phy also partitioned these samples well, especially along PC2. Plotting *Lachnospiraceae*, *Oscillospiraceae*, and *Ruminococcaceae* abundances on the PCoA ordinations did confirm the correlation analysis, in which *Lachnospiraceae* abundance correlates rather well with PC1 and PC2 of all ordinations, while the *Oscillospiraceae* and *Ruminococcaceae* abundances best correlate the the top PCs for both unweighted UniFrac measures (Figure S6).

We also correlated alpha diversity with the PCoA PCs but found substantially weaker associations ($R^2 < 0.21$ for all measures). Still, gradients of diversity are somewhat apparent across the ordinations, regardless of the diversity measure (Figure S7).

To determine how well each beta diversity measure partitions individuals by age, gender, BMI, westernization, and disease states, we performed PERMANOVA with each measure on all samples with the requisite metadata ($n$ = 1413). Although all model variables were significant due to the large sample size ($P < 0.001$), the effect sizes varied considerably for disease state and westernization (Figure 5). Most notably, w-unifrac_phy had an $R^2$ for disease state that was about twice that of Bray-Curtis or Jaccard (0.082 versus 0.041 and 0.025, respectively). Plotting the location of each metagenome sample from each disease category on PC1 illustrated how Bray-Curtis and Jaccard largely relegate most samples with each disease state to the same half of PC1, while “healthy” samples span the entire PC (Figure 5B). In contrast, the UniFrac measures, especially the weighted versions, partition the various disease states into different regions along the entire length of the PC.

To directly quantify the differences in how each beta diversity measure partitioned samples in each disease category, we performed pairwise Procrustean superposition analyses between each beta diversity measure. Large Procrustes residuals for a disease state indicate that the relative positions of samples in that grouping differ greatly between the two PCoA ordinations. Procrustes residuals were highest for Shiga toxin-producing *E. coli* (STEC) and hypertension disease states when comparing the UniFrac measures to Bray-Curtis or Jaccard (Figure S8). STEC was also moderately divergent between the trait-based and phylogeny-based UniFrac measures (both weighted and unweighted). This discrepancy between diversity measures reflects the incongruence between phylogeny- and trait-based relatedness among *Enterobacteriaceae* species (Figure S2).

## Discussion

Shotgun metagenomics will continue to increase in popularity as the cost of sequencing declines and methods for processing and interpreting metagenomic data continue to develop. A major challenge is to fully harness the heterogeneous sequence data generated by metagenomics, which is vastly more complex than 16S rRNA gene sequences or other single-locus datasets. Measuring community diversity from such heterogeneous data is not straight-forward, and it is often unclear what measures of diversity are most appropriate for metagenome data. Here, we have assessed a method of microbiome diversity measurement by using metrics that incorporate a multi-locus phylogeny or a large set of traits inferred from reference genomes to species-level abundance profiles mapped against species-level genome representatives from the GTDB. Our method is not computationally demanding, generalizable to a wide range of microbiome studies, and flexible in regards to which tree-based measures and which traits are used.

We have shown that our tree-based diversity measures explained more variance, both in regards to overall inter-sample diversity and diversity among individuals differing in westernization, age, gender, and disease status (Figures 2 & 3). While BMI seemed to be slightly better explained by phylogeny- and trait-based measures, the difference was too small to be conclusive. Interestingly, westernization was substantially better explained by the phylogeny-based alpha diversity measure relative to all other measures, while this pattern was not observed for beta diversity. These results suggest that while overall phylogenetic diversity is greater for non-westernized individuals, there is enough broad-scale phylogenetic overlap between individuals to appear highly similar in a direct comparison.

We additionally showed that phylogeny and trait-based diversity measures were more explanatory than tree-agnostic measures due to how each underscored different aspects of community diversity (Figure 4). Bray-Curtis and Jaccard emphasized compositional differences within the *Bacteroidaceae*, which is a prevalent and relatively abundant clade in the human gut. Instead, both the phylogeny and trait-based measures accentuated differences between *Enterobacteriaceae and Bacteroidaceae*, which not only belong to different phyla, but also the former is much less prevalent than the latter. This emphasis on *Enterobacteriaceae* by the tree-based diversity measures likely explains why the disease state that differed most between PCoA ordinations was Shiga toxin-producing *E. coli* (Figure S8). The same may be true for the presence of hypertension, which was the second-most different between PCoA ordinations, as the *Enterobacteriaceae* genus *Klebsiella* has been found to overgrow in hypertensive individuals (Li *et al.*, 2017). Of course, this increased emphasis on *Enterobacteriaceae* by the tree-based measures is just the most prominent, and as we observed for our family-level assessment of alpha diversity, many clades can differ in their apparent diversities depending on the measure used.

In almost all circumstances, phylogeny-based diversity was more explanatory than when incorporating trait relatedness. Our assessment of congruence between phylogenetic and trait-based similarity showed why these diversity measures would differ. For instance, the lower explanatory power of trait-based diversity in regards to disease state can be attributed to the incongruence between trait and phylogeny for many species in the *Enterobacteriaceae* family (Figures 5 & S2) or possibly to the choice of traits included (Figure 1). While we did use a large number of traits relative to other recent trait-based studies of microbial community spatiotemporal diversity (Ortiz-Álvarez *et al.*, 2018; Guittar, Shade and Litchman, 2019), they are likely just a minor subset of all relevant phenotypes. Traits could be defined more abstractly as COG functional categories, KEGG pathways, or other broad classifications of gene function, which may generalize better to novel microbial genetic diversity compared to using a trait classifier trained on a subset of all known microbial species (Royalty and Steen, 2019). However, broad and generalized demarcations of function may obscure particular traits that are most strongly varying across the spatiotemporal gradient of interest. One could choose particular functional categories, like the gut-brain modules defined in a recent study for understanding microbial functional interaction with mental health (Valles-Colomer *et al.*, 2019), although the expert knowledge required to make such targeted selections is often lacking for many systems.

We must note that our method of predicting traits based on the presence of loci presumably produced false negatives for poorly studied clades in which novel genetic mechanisms generate the same phenotypes. Given that the gut microbiome is dominated by a few relatively well-studied clades (Lloyd-Price *et al.*, 2017), the impact of false negatives was likely small for our trait-based weighted UniFrac measure but may have been higher for unweighted UniFrac. Still, both phylogeny- and trait-based unweighted UniFrac were less explanatory than their weighted counterparts, suggesting that inaccuracies in our trait classification approach were negligible. Advances in machine learning models for predicting gene annotations, protein structure and interactions, and metabolic pathways will improve classification of specific microbial phenotypes, especially when generalizing to novel genetic diversity (Celesti *et al.*, 2018; Bileschi *et al.*, 2019).

While our findings demonstrate the potential benefit of incorporating phylogeny or function based on genome representatives of each reference species, much is still unknown about how best to implement these approaches across highly varied microbiome studies. Function-based diversity measures may prove to be highly advantageous for studies of microbial community succession, as some studies have demonstrated (Ortiz-Álvarez *et al.*, 2018; Guittar, Shade and Litchman, 2019). Microbiomes with high numbers of uncultured species such as seafloor sediments may benefit from using a more generalized measure of traits like COG functional categories (Orsi, 2018). We recommend a focus on phylogeny-based diversity measures for shotgun metagenomics data in cases where the most informative traits are unknown, since phylogenetic information will be relevant for most if not all systems, and it will allow for direct cross-study comparisons of microbial diversity.

## Author contributions

Author contributions: N.D.Y. and J.dlC. designed the research; N.D.Y. and J.dlC. performed research; N.D.Y. analyzed data; and N.D.Y., J.dlC., and R.E.L. wrote the paper.

## Acknowledgements

We thank Albane Ruaud and Taichi Suzuki for providing important feedback on this work.

## Funding

This work was supported by the Max Planck Society.

Conflict of Interest: none declared.

## References

Bassett, S. A. *et al.* (2015) ‘Changes in composition of caecal microbiota associated with increased colon inflammation in interleukin-10 gene-deficient mice inoculated with Enterococcus species’, *Nutrients*, 7(3), pp. 1798–1816. doi: 10.3390/nu7031798.

Bileschi, M. L. *et al.* (2019) ‘Using Deep Learning to Annotate the Protein Universe’, *bioRxiv*. doi: 10.1101/626507.

Celesti, F. *et al.* (2018) ‘Why Deep Learning Is Changing the Way to Approach NGS Data Processing: A Review’, *IEEE reviews in biomedical engineering*, 11, pp. 68–76. doi: 10.1109/RBME.2018.2825987.

de la Cuesta-Zuluaga, J., Ley, R. E. and Youngblut, N. D. (2019) ‘Struo: a pipeline for building custom databases for common metagenome profilers’, *Bioinformatics* . doi: 10.1093/bioinformatics/btz899.

Darling, A. E. *et al.* (2014) ‘PhyloSift: phylogenetic analysis of genomes and metagenomes’, *PeerJ*, 2, p. e243. doi: 10.7717/peerj.243.

Ewels, P. *et al.* (2016) ‘MultiQC: summarize analysis results for multiple tools and samples in a single report’, *Bioinformatics* , 32(19), pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

Franzosa, E. A. *et al.* (2018) ‘Species-level functional profiling of metagenomes and metatranscriptomes’, *Nature methods*, 15(11), pp. 962–968. doi: 10.1038/s41592-018-0176-y.

Guittar, J., Shade, A. and Litchman, E. (2019) ‘Trait-based community assembly and succession of the infant gut microbiome’, *Nature communications*, 10(1), p. 512. doi: 10.1038/s41467-019-08377-w.

Hamady, M., Lozupone, C. and Knight, R. (2010) ‘Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data’, *The ISME journal*, 4(1), pp. 17–27.

Hillmann, B. *et al.* (2018) ‘Evaluating the Information Content of Shallow Shotgun Metagenomics’, *mSystems*, 3(6). doi: 10.1128/mSystems.00069-18.

Hugerth, L. W. and Andersson, A. F. (2017) ‘Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing’, *Frontiers in microbiology*, 8, p. 1561. doi: 10.3389/fmicb.2017.01561.

Jiang, H. *et al.* (2014) ‘Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads’, *BMC Bioinformatics*. doi: 10.1186/1471-2105-15-182.

Kembel, S. W. *et al.* (2011) ‘The phylogenetic diversity of metagenomes’, *PloS one*, 6(8), p. e23214. doi: 10.1371/journal.pone.0023214.

Li, J. *et al.* (2017) ‘Gut microbiota dysbiosis contributes to the development of hypertension’, *Microbiome*, 5(1), p. 14. doi: 10.1186/s40168-016-0222-x.

Lloyd-Price, J. *et al.* (2017) ‘Strains, functions and dynamics in the expanded Human Microbiome Project’, *Nature*, 550(7674), pp. 61–66. doi: 10.1038/nature23889.

Louca, S., Doebeli, M. and Parfrey, L. W. (2018) ‘Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem’, *Microbiome*, 6(1), p. 41. doi: 10.1186/s40168-018-0420-9.

Lozupone, C. A. and Knight, R. (2008) ‘Species divergence and the measurement of microbial diversity’, *FEMS microbiology reviews*, 32(4), pp. 557–578.

Lu, J. *et al.* (2017) ‘Bracken: estimating species abundance in metagenomics data’, *PeerJ Computer Science*, p. e104. doi: 10.7717/peerj-cs.104.

Maiden, M. C. J. *et al.* (1998) ‘Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms’, *Proceedings of the National Academy of Sciences*, 95(6), pp. 3140–3145. Available at: http://www.pnas.org/content/95/6/3140.abstract.

Matsen, F. A., 4th (2015) ‘Phylogenetics and the human microbiome’, *Systematic biology*, 64(1), pp. e26–41. doi: 10.1093/sysbio/syu053.

Miller, C. S. *et al.* (2011) ‘EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data’, *Genome biology*, 12(5), p. R44. doi: 10.1186/gb-2011-12-5-r44.

Obregon-Tito, A. J. *et al.* (2015) ‘Subsistence strategies in traditional societies distinguish gut microbiomes’, *Nature communications*, 6, p. 6505. doi: 10.1038/ncomms7505.

Oksanen, J. *et al.* (2012) *vegan: Community Ecology Package*. Available at: http://CRAN.R-project.org/package=vegan.

Orsi, W. D. (2018) ‘Ecology and evolution of seafloor and subseafloor microbial communities’, *Nature reviews. Microbiology*, 16(11), pp. 671–683. doi: 10.1038/s41579-018-0046-8.

Ortiz-Álvarez, R. *et al.* (2018) ‘Consistent changes in the taxonomic structure and functional attributes of bacterial communities during primary succession’, *The ISME journal*, 12(7), pp. 1658–1667. doi: 10.1038/s41396-018-0076-2.

Parks, D. H. *et al.* (2017) ‘Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life’, *Nature microbiology*, 2(11), pp. 1533–1542. doi: 10.1038/s41564-017-0012-7.

Parks, D. H. *et al.* (2018) ‘A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life’, *Nature biotechnology*, 36(10), pp. 996–1004. doi: 10.1038/nbt.4229.

Pasolli, E. *et al.* (2017) ‘Accessible, curated metagenomic data through ExperimentHub’, *bioRxiv*. Cold Spring Harbor Laboratory. doi: 10.1101/103085.

Pericard, P. *et al.* (2018) ‘MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes’, *Bioinformatics* , 34(4), pp. 585–591. doi: 10.1093/bioinformatics/btx644.

Royalty, T. M. and Steen, A. D. (2019) ‘Quantitatively Partitioning Microbial Genomic Traits among Taxonomic Ranks across the Microbial Tree of Life’, *mSphere*, 4(4). doi: 10.1128/mSphere.00446-19.

Segata, N. *et al.* (2012) ‘Metagenomic microbial community profiling using unique clade-specific marker genes’, *Nature methods*, 9(8), pp. 811–814. doi: 10.1038/nmeth.2066.

Sommer, F. and Bäckhed, F. (2013) ‘The gut microbiota--masters of host development and physiology’, *Nature reviews. Microbiology*, 11(4), pp. 227–238. doi: 10.1038/nrmicro2974.

Torres, P. J. *et al.* (2018) ‘Gut Microbial Diversity in Women With Polycystic Ovary Syndrome Correlates With Hyperandrogenism’, *The Journal of clinical endocrinology and metabolism*, 103(4), pp. 1502–1511. doi: 10.1210/jc.2017-02153.

Valles-Colomer, M. *et al.* (2019) ‘The neuroactive potential of the human gut microbiota in quality of life and depression’, *Nature microbiology*. doi: 10.1038/s41564-018-0337-x.

Vogt, N. M. *et al.* (2017) ‘Gut microbiome alterations in Alzheimer’s disease’, *Scientific reports*, 7(1), p. 13537. doi: 10.1038/s41598-017-13601-y.

Weimann, A. *et al.* (2016) ‘From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer’, *mSystems*, 1(6). doi: 10.1128/mSystems.00101-16.

Wood, D. E., Lu, J. and Langmead, B. (2019) ‘Improved metagenomic analysis with Kraken 2’, *Genome biology*, 20(1), p. 257. doi: 10.1186/s13059-019-1891-0.

Wu, M. and Scott, A. J. (2012) ‘Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2’, *Bioinformatics* , 28(7), pp. 1033–1034. doi: 10.1093/bioinformatics/bts079.

## Figure legends

**Figure 1**. *Similarity between phylogenetic and trait-based relatedness differs substantially among phyla.* A) Traits inferred from each genome representative of each species, shown as the percent of all genomes in the phylum (left) or the total for all phyla (right). The numbers next to each column in the right plot denote the x-axis values. B) The boxplots show Procrustes residuals for each genome, grouped by phylum. Higher Procrustes residuals indicate more incongruence between phylogenetic and trait-based relatedness. For clarity, only phyla with ≥10 genomes are shown.

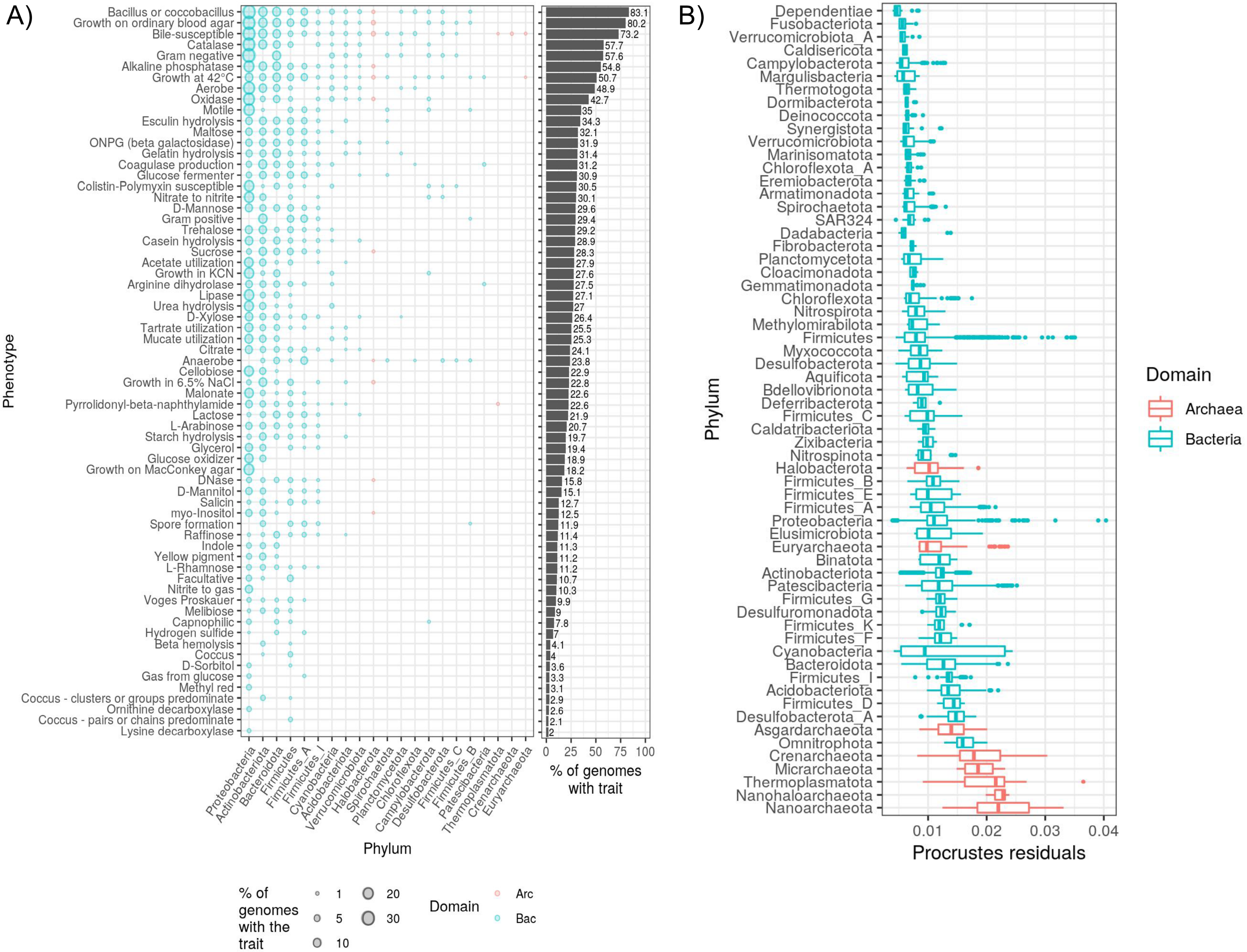
**Figure 2.** *Phylogeny- and trait-based alpha diversity better differentiate samples across key factors.* A) Boxplots of alpha diversity metrics calculated for all samples (*n* = 3348) in all datasets (*n* = 33), grouped by westernization status. “(phy)” denotes that the genome phylogeny was used to calculate Faith’s PD, while “(trt)” means that a dendrogram of trait similarity was

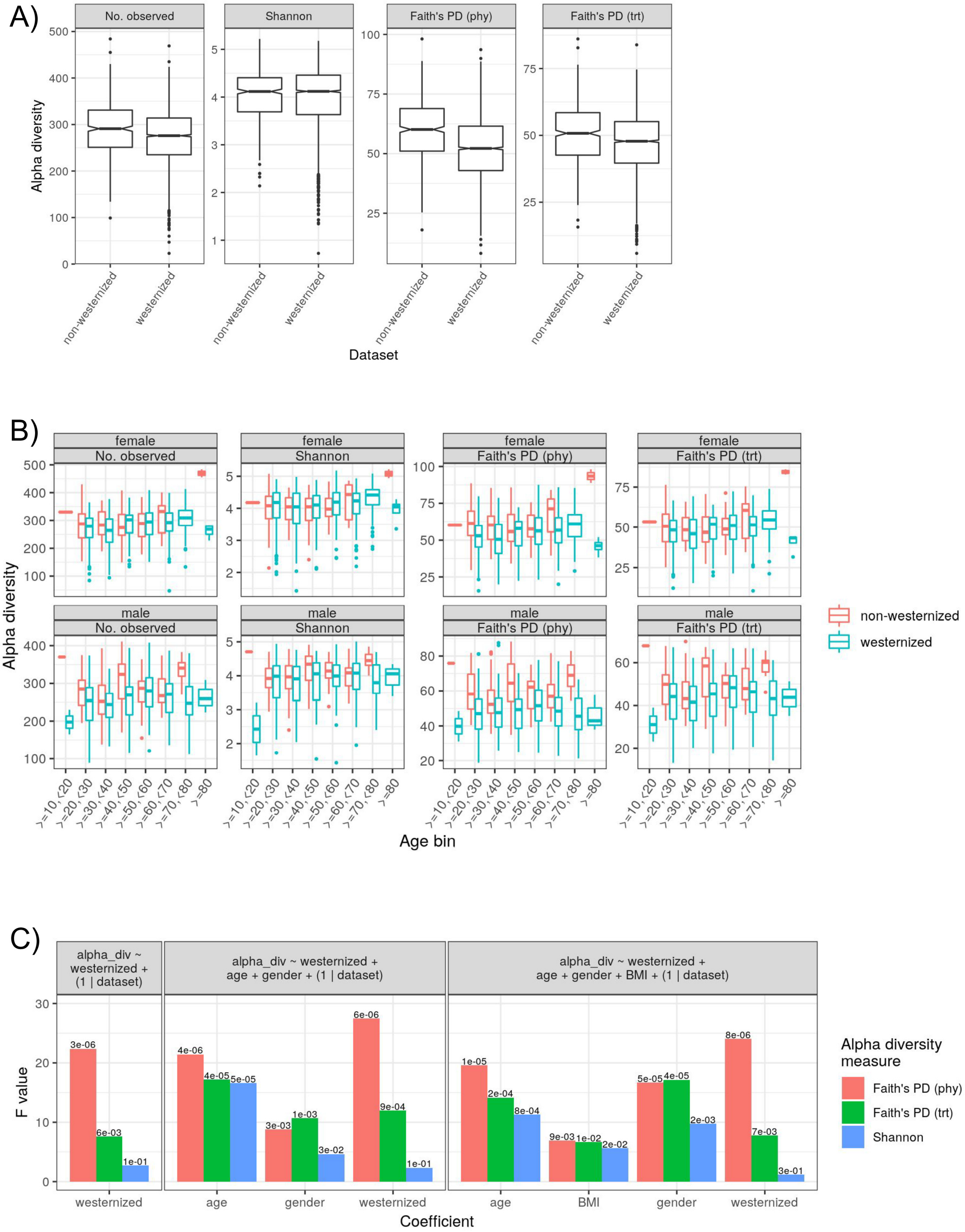
used for the calculation. B) Boxplots of alpha diversity metrics calculated for all samples in which gender and age metadata were available (*n* = 1843) in all datasets (*n* = 17), grouped by westernization of individuals. C) Linear mixed effects model results for assessing the association between alpha diversity and population characteristics while accounting for inter-dataset batch effects. The labels above each bar denote *P*-values. Age was log2-transformed, and BMI Box-Cox transformed. The left facet is on all samples (*n* = 3348) in all datasets (*n* = 33). The middle facet is filtered to samples that have data on gender and age (number of samples = 1843; number of studies = 17). The right facet is filtered to samples that have data on gender, age, and BMI (number of samples = 918; number of studies = 11).

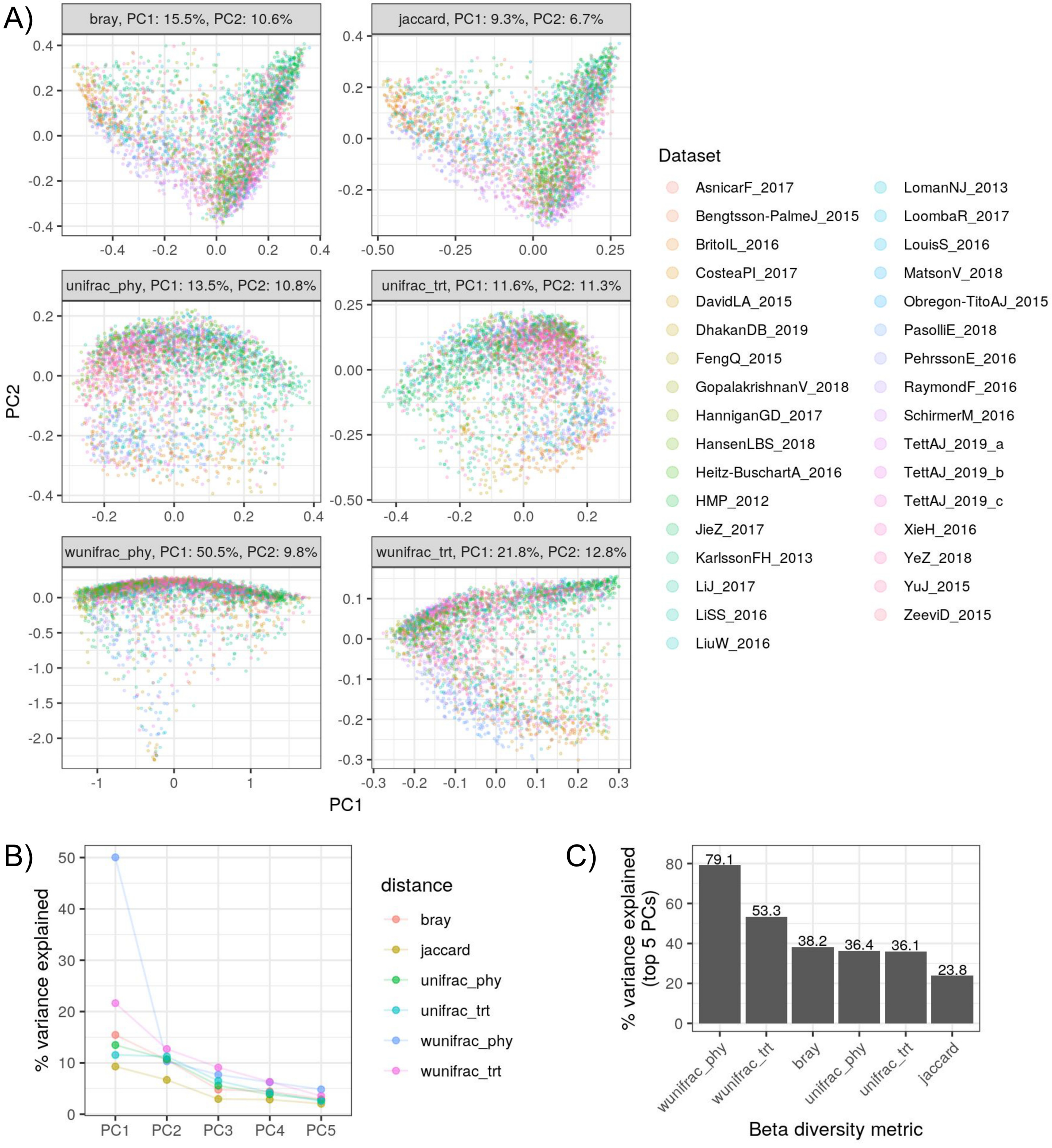
**Figure 3.** *More variance explained when incorporating taxon abundance along with phylogenetic- or trait-based relatedness.* Principal coordinate analysis (PCoA) ordinations for all samples across all datasets (*n* = 3348), colored by dataset and faceted by the beta-diversity metric used ("bray" = Bray-Curtis; "jaccard" = Jaccard; "unifrac_phy" = unweighted UniFrac utilizing the genome phylogeny; "unifrac_trt" = unweighted UniFrac utilizing a dendrogram depicting trait-similarity; "wunifrac_phy" = "unifrac_phy", but using weighted UniFrac; "wunifrac_trt" = "unifrac_trt", but using weighted UniFrac). The percentages in each facet label are the percent variance explained for the first two PCs. B) The percent variance explained by the top five PCs for each ordination shown in A). C) The summed percent variance explained by the top five PCs for each ordination shown in A), with values above each bar denoting the y-axis value.

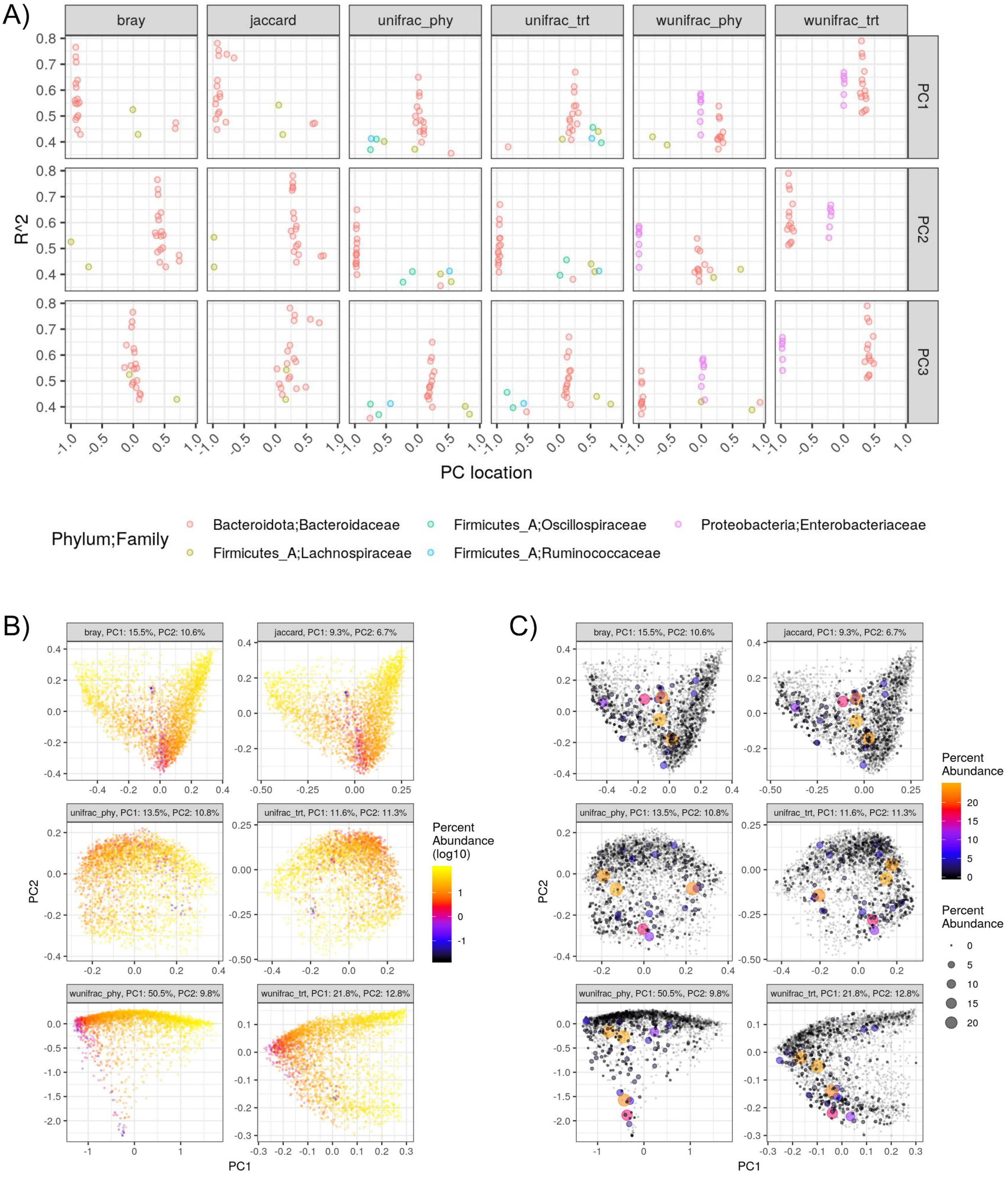
**Figure 4.** *Phylogeny- and trait-based beta-diversity metrics emphasize inter-sample differences in certain taxa that are not emphasized by star-phylogeny measures.* A) Correlations between individual species (points) and the top 3 PCs in the PCoA ordinations shown in Figure 3. The x-axis denotes the direction of the correlation along the PC (*i.e.,* where the taxon abundance is highest), and the y-axis denotes the effect size. For clarity, only species with the top 20 highest effect sizes across all beta diversity metrics are shown. The PCoA ordinations shown in B) and C) are the same as in Figure 3, but samples are colored by the abundance of the *Bacteroidaceae* family (*Bacteroidota* phylum) and *Enterobacteriaceae* (*Proteobacteria* phylum), respectively. Note that abundance is not log10-transformed in C), and point size also represents abundance in order to emphasize the few samples with relatively high *Enterobacteriaceae* abundances, and all grey points indicate samples completely lacking *Enterobacteriaceae*.

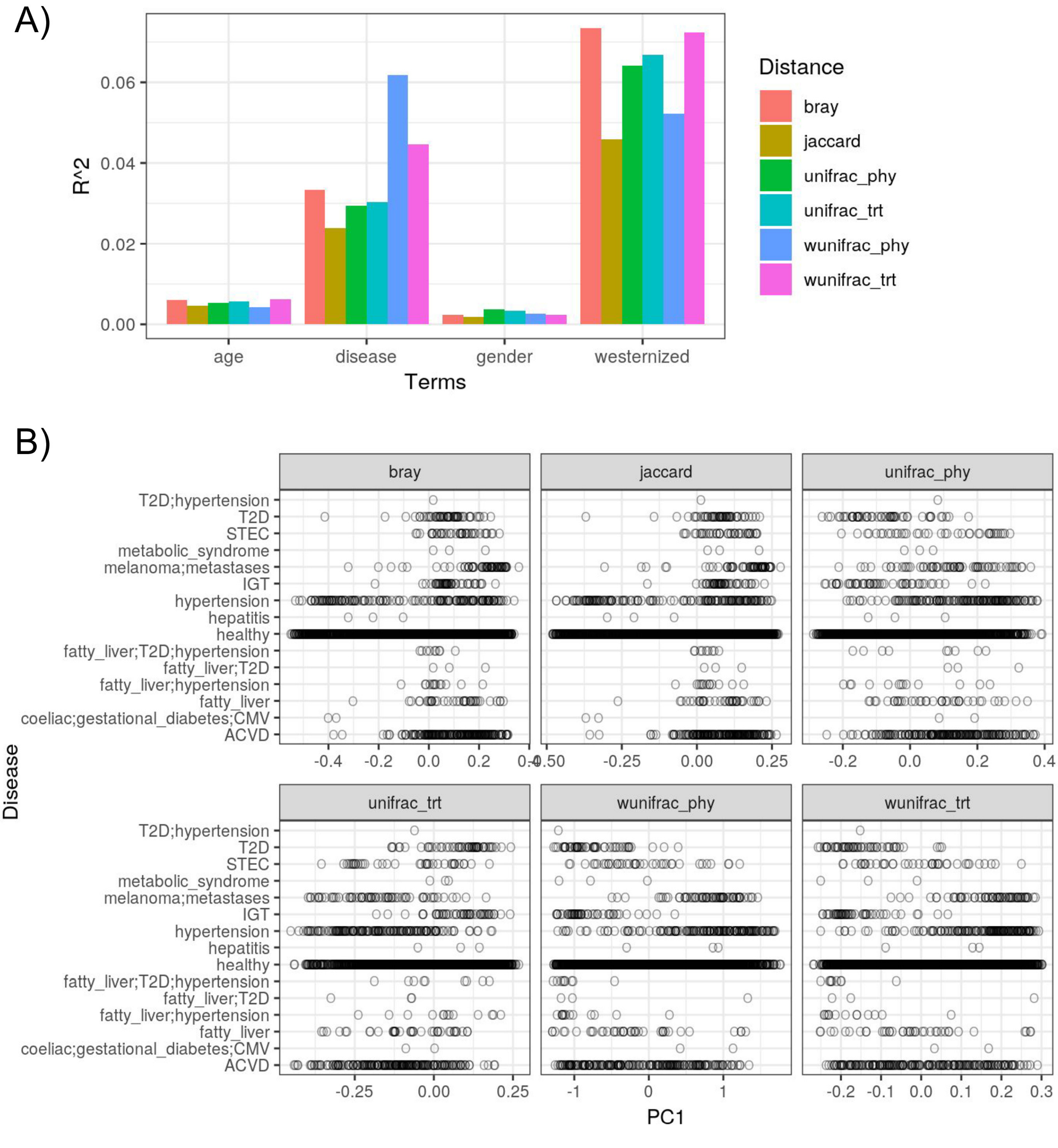
**Figure 5.** *UniFrac-based beta diversity better explains disease status across the metagenome dataset.* A) Variance explained for each covariate in PERMANOVA models (*n* = 1413) applied to each distance matrix as shown in PCoA plots in Figure 3. B) The position of each sample (grouped by disease state) on PC1 for each PCoA of each beta diversity measure as shown in Figure 3. Note that for the tree-agnostic approaches, most disease states fall into the same, constrained range; however, the UniFrac-based approaches (especially weighted UniFrac) generate more separation among disease groups ("STEC" = Shiga toxin-producing *E. coli*; "T2D" = Type 2 diabetes; "ACVD" = atherosclerotic cardiovascular diseases; "CMV" = Cytomegalovirus disease, "IGT" = impaired glucose tolerance). All terms in each PERMANOVA model were significant (number of permutations = 9999; *P* < 0.001).


A)
Phenotype
Bacillus or coccobacillus
Growth on ordinary blood agar
Bile-susceptible
Catalase
Gram negative
Alkaline phosphatase
Growth at 42°C
Aerobe
Oxidase
Motile
Esculin hydrolysis
Maltose
ONPG (beta galactosidase)
Gelatin hydrolysis
Coagulase production
Glucose fermenter
Colistin-Polymyxin susceptible
Nitrate to nitrite
D-Mannose
Gram positive
Trehalose
Casein hydrolysis
Sucrose
Acetate utilization
Growth in KCN
Arginine dihydrolase
Lipase
Urea hydrolysis
D-Xylose
Tartrate utilization
Mucate utilization
Citrate
Anaerobe
Cellobiose
Growth in 6.5% NaCl
Malonate
Pyrrolidonyl-beta-naphthylamide
Lactose
L-Arabinose
Starch hydrolysis
Glycerol
Glucose oxidizer
Growth on MacConkey agar
DNase
D-Mannitol
Salicin
myo-Inositol
Spore formation
Raffinose
Indole
Yellow pigment
L-Rhamnose
Facultative
Nitrite to gas
Voges Proskauer
Melibiose
Capnophilic
Hydrogen sulfide
Beta hemolysis
Coccus
D-Sorbitol
Gas from glucose
Methyl red
Coccus - clusters or groups predominate
Ornithine decarboxylase
Coccus - pairs or chains predominate
Lysine decarboxylase
83.1
80.2
73.2
57.7
57.6
54.8
50.7
48.9
42.7
35
34.3
32.1
31.9
31.4
31.2
30.9
30.5
30.1
29.6
29.4
29.2
28.9
28.3
27.9
27.6
27.5
27.1
27
26.4
25.5
25.3
24.1
23.8
22.9
22.8
22.6
22.6
21.9
20.7
19.7
19.4
18.9
18.2
15.8
15.1
12.7
12.5
11.9
11.4
11.3
11.2
11.2
10.7
10.3
9.9
9
7.8
7
4.1
4
3.6
3.3
3.1
2.9
2.6
2.1
2
Proteobacteria
Actinobacteriota
Bacteroidota
Firmicutes
Firmicutes_A
Firmicutes_I
Cyanobacteria
Acidobacteriota
Verrucomicrobiota
Halobacterota
Spirochaetota
Planctomycetota
Chloroflexota
Campylobacterota
Desulfobacterota
Firmicutes_C
Firmicutes_B
Patescibacteria
Thermoplasmatota
Crenarchaeota
Euryarchaeota
Phylum
0
25
50
75
100
% of genomes with trait
% of genomes with the trait
1
5
10
20
30
Domain
Arc
Bac
B)
Phylum
Dependentiae
Fusobacteriota
Verrucomicrobiota_A
Caldisericota
Campylobacterota
Margulisbacteria
Thermotogota
Dormibacterota
Deinococcota
Synergistota
Verrucomicrobiota
Marinisomatota
Chloroflexota_A
Eremiobacterota
Armatimonadota
Spirochaetota
SAR324
Dadabacteria
Fibrobacterota
Planctomycetota
Cloacimonadota
Gemmatimonadota
Chloroflexota
Nitrospirota
Methylomirabilota
Firmicutes
Myxococcota
Desulfobacterota
Aquificota
Bdellovibrionota
Deferribacterota
Firmicutes_C
Caldatribacteriota
Zixibacteria
Nitrospinota
Halobacterota
Firmicutes_B
Firmicutes_E
Firmicutes_A
Proteobacteria
Elusimicrobiota
Euryarchaeota
Binatota
Actinobacteriota
Patescibacteria
Firmicutes_G
Desulfuromonadota
Firmicutes_K
Firmicutes_F
Cyanobacteria
Bacteroidota
Firmicutes_I
Acidobacteriota
Firmicutes_D
Desulfobacterota_A
Asgardarchaeota
Omnitrophota
Crenarchaeota
Micrarchaeota
Thermoplasmatota
Nanohaloarchaeota
Nanoarchaeota
0.01
0.02
0.03
0.04
Procrustes residuals
Domain
Archaea
Bacteria


A)
No. observed
Shannon
Faith's PD (phy)
Faith's PD (trt)
Alpha diversity
0
100
200
300
400
500
1
2
3
4
5
25
50
75
100
non-westernized
westernized
Dataset
B)
female
male
No. observed
Shannon
Faith's PD (phy)
Faith's PD (trt)
Alpha diversity
>=10,<20
>=20,<30
>=30,<40
>=40,<50
>=50,<60
>=60,<70
>=70,<80
>=80
Age bin
non-westernized
westernized
C)
alpha_div ~ westernized + (1 | dataset)
alpha_div ~ westernized + age + gender + (1 | dataset)
alpha_div ~ westernized + age + gender + BMI + (1 | dataset)
F value
0
10
20
30
3e-06
6e-03
1e-01
4e-06
4e-05
5e-05
3e-03
1e-03
3e-02
6e-06
9e-04
1e-01
1e-05
2e-04
8e-04
9e-03
1e-02
2e-02
5e-05
4e-05
2e-03
8e-06
7e-03
3e-01
westernized
age
gender
westernized
age
BMI
gender
westernized
Coefficient
Alpha diversity measure
Faith's PD (phy)
Faith's PD (trt)
Shannon


A)
bray, PC1: 15.5%, PC2: 10.6%
jaccard, PC1: 9.3%, PC2: 6.7%
unifrac_phy, PC1: 13.5%, PC2: 10.8%
unifrac_trt, PC1: 11.6%, PC2: 11.3%
wunifrac_phy, PC1: 50.5%, PC2: 9.8%
wunifrac_trt, PC1: 21.8%, PC2: 12.8%
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2
0.4
0.4
0.2
0.0
-0.2
-0.50
-0.25
0.00
0.25
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
0.4
0.25
0.00
-0.25
-0.50
-0.4
-0.2
0.0
0.2
0.0
-0.5
-1.0
-1.5
-2.0
-1
0
1
0.1
0.0
-0.1
-0.2
-0.3
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
PC2
PC1
Dataset
AsnicarF_2017
Bengtsson-PalmeJ_2015
BritoIL_2016
CosteaPI_2017
DavidLA_2015
DhakanDB_2019
FengQ_2015
GopalakrishnanV_2018
HanniganGD_2017
HansenLBS_2018
Heitz-BuschartA_2016
HMP_2012
JieZ_2017
KarlssonFH_2013
LiJ_2017
LiSS_2016
LiuW_2016
LomanNJ_2013
LoombaR_2017
LouisS_2016
MatsonV_2018
Obregon-TitoAJ_2015
PasolliE_2018
PehrssonE_2016
RaymondF_2016
SchirmerM_2016
TettAJ_2019_a
TettAJ_2019_b
TettAJ_2019_c
XieH_2016
YeZ_2018
YuJ_2015
ZeeviD_2015
B)
% variance explained
50
40
30
20
10
0
PC1
PC2
PC3
PC4
PC5
distance
bray
jaccard
unifrac_phy
unifrac_trt
wunifrac_phy
wunifrac_trt
C)
% variance explained
(top 5 PCs)
80
60
40
20
0
79.1
53.3
38.2
36.4
36.1
23.8
wunifrac_phy
wunifrac_trt
bray
unifrac_phy
unifrac_trt
jaccard
Beta diversity metric


A)
bray
jaccard
unifrac_phy
unifrac_trt
wunifrac_phy
wunifrac_trt
PC1
PC2
PC3
R^2
0.8
0.7
0.6
0.5
0.4
-1.0
-0.5
0.0
0.5
1.0
PC location
Phylum;Family
Bacteroidota;Bacteroidaceae
Firmicutes_A;Lachnospiraceae
Firmicutes_A;Oscillospiraceae
Firmicutes_A;Ruminococcaceae
Proteobacteria;Enterobacteriaceae
B)
bray, PC1: 15.5%, PC2: 10.6%
jaccard, PC1: 9.3%, PC2: 6.7%
unifrac_phy, PC1: 13.5%, PC2: 10.8%
unifrac_trt, PC1: 11.6%, PC2: 11.3%
wunifrac_phy, PC1: 50.5%, PC2: 9.8%
wunifrac_trt, PC1: 21.8%, PC2: 12.8%
Percent
Abundance
(log10)
1
0
-1
PC1
PC2
C)
bray, PC1: 15.5%, PC2: 10.6%
jaccard, PC1: 9.3%, PC2: 6.7%
unifrac_phy, PC1: 13.5%, PC2: 10.8%
unifrac_trt, PC1: 11.6%, PC2: 11.3%
wunifrac_phy, PC1: 50.5%, PC2: 9.8%
wunifrac_trt, PC1: 21.8%, PC2: 12.8%
Percent
Abundance
20
15
10
5
0
Percent
Abundance
0
5
10
15
20
PC1
PC2


A)
R^2
0.00
0.02
0.04
0.06
age
disease
gender
westernized
Terms
Distance
bray
jaccard
unifrac_phy
unifrac_trt
wunifrac_phy
wunifrac_trt
B)
bray
jaccard
unifrac_phy
unifrac_trt
wunifrac_phy
wunifrac_trt
Disease
T2D;hypertension
T2D
STEC
metabolic_syndrome
melanoma;metastases
IGT
hypertension
hepatitis
healthy
fatty_liver;T2D;hypertension
fatty_liver;T2D
fatty_liver;hypertension
fatty_liver
coeliac;gestational_diabetes;CMV
ACVD
-0.4
-0.2
0.0
0.2
0.4
-0.50
-0.25
0.00
0.25
-0.2
0.0
0.2
0.4
-0.25
0.00
0.25
-1
0
1
-0.2
-0.1
0.0
0.1
0.2
0.3
PC1