bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TITLE: A phylogenetic transform enhances analysis of compositional microbiota data

AUTHORS: Justin D Silverman*[†‡§], Alex Washburne[¶#], Sayan Mukherjee[*‖], and Lawrence A David[*‡§]

AUTHOR AFFILIATIONS:

* Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708

[†] Medical Scientist Training Program, Duke University, Durham, NC 27708

[‡] Center for Genomic and Computational Biology, Duke University, Durham, NC 27708

[§] Department of Molecular Genetics and Microbiology, Duke University, Durham, NC 27708

[¶] Nicholas School of the Environment, Duke University, Durham, NC 27708

[#] Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, CO 80309

[‖] Departments of Statistical Science, Mathematics, and Computer Science, Duke University, Durham, NC 27708

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT: High-throughput DNA sequencing technologies have revolutionized the study of microbial communities (microbiota) and have revealed their importance in both human health and disease. However, due to technical limitations, data from microbiota surveys reflect the relative abundance of bacterial taxa and not their absolute levels. It is well known that applying common statistical methods, such as correlation or hypothesis testing, to relative abundance data can lead to spurious results. Here, we introduce the PhILR transform, a data transform that utilizes microbial phylogenetic information. This transform enables off-the-shelf statistical tools to be applied to microbiota surveys free from artifacts usually associated with analysis of relative abundance data. Using environmental and human-associated microbial community datasets as benchmarks, we find that the PhILR transform significantly improves the performance of distance-based and machine learning-based statistics, boosting the accuracy of widely used algorithms on reference benchmarks by 90%. Because the PhILR transform relies on bacterial phylogenies, statistics applied in the PhILR coordinate system are also framed within an evolutionary perspective. Regression on PhILR transformed human microbiota data identified evolutionarily neighboring bacterial clades that may have differentiated to adapt to distinct body sites. Variance statistics showed that the degree of covariation of bacterial clades across human body sites tended to increase with phylogenetic relatedness between clades. These findings support the hypothesis that environmental selection, not competition between bacteria, plays a dominant role in structuring human-associated microbial communities.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Microbiota research today embodies the data-rich nature of modern biology. Advances in high-throughput DNA sequencing allow for rapid and affordable surveys of thousands of bacterial taxa across hundreds of samples (1). The exploding availability of sequencing data has poised microbiota research to advance our understanding of fields as diverse as ecology, evolution, medicine, and agriculture (2). Considerable effort now focuses on interrogating microbiota datasets to identify relationships between bacterial taxa, as well as between microbes and their environment.

Increasingly, it is appreciated that the relative nature of microbial abundance data in microbiota studies can lead to spurious statistical analyses (3-9). With next generation sequencing, the number of reads per sample can vary independently of microbial load (6, 9). In order to make measurements comparable across samples, most studies therefore analyze the relative abundance of bacterial taxa. Analyses are thus not carried out on absolute abundances of community members (**Fig. 1A**), but rather on relative data occupying a constrained, non-orthogonal, geometric space (**Fig. 1B**). Such relative abundance datasets are often termed compositional. The use of most standard statistical tools (*e.g.,* correlation, regression, or classification) within a compositional space leads to spurious results (10). For example, three-quarters of the significant bacterial interactions inferred by Pearson correlation on a compositional human microbiota dataset were likely false (4), and over two-thirds of differentially abundant taxa inferred by a t-test on a simulated compositional human microbiota dataset were spurious (11). To account for compositional effects in microbial datasets, bioinformatics

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

efforts have re-derived common statistical methods including correlation statistics (4, 12), hypothesis testing (13), and variable selection (14, 15).

An alternative approach is to transform compositional microbiota data to a space where existing statistical methods may be applied without introducing spurious conclusions. This approach is attractive because of its efficiency: the vast toolbox of existing statistical models can be applied without re-derivation. Normalization methods, for example, have been proposed to modify count data by assuming reads follow certain statistical distributions (*e.g.,* negative binomial) (16, 17). Alternatively, the field of Compositional Data Analysis (CoDA) has focused on formalizing methods for transforming compositional data from a constrained non-orthogonal space into a simpler geometry without having to assume data adhere to a distribution model (18). Previous microbiota analyses have already leveraged CoDA theory and used the centered log-ratio transform to reconstruct microbial association networks and interactions (19, 20) and to analyze differential abundances (21, 22). However, the centered log-ratio transform has a crucial limitation: it yields a coordinate system featuring a singular covariance matrix and is thus unsuitable for many common statistical models (10). This drawback can be sidestepped using another CoDA transform, known as the Isometric Log-Ratio (ILR) transformation (23). The ILR transform uses a sequential binary partition of the original variable space (**Fig. 1C**) to create a new coordinate system with orthonormal bases (**Fig. 1D,E**) (23). However, a known obstacle to using the ILR transform is the choice of partition such that the resulting coordinates are meaningful

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(10). To date, microbiota studies have chosen ILR coordinates using random sequential binary partitions of bacterial groups (24, 25).

Here, we introduce the bacterial phylogenetic tree as a natural and informative sequential binary partition when applying the ILR transform to microbiota datasets (**Fig. 1C**). Using phylogenies to construct the ILR transform results in an ILR coordinate system capturing evolutionary relationships between neighboring bacterial groups (clades). Analyses of neighboring clades offer the opportunity for biological insight: clade analyses have linked genetic adaptation to ecological differentiation (26), and the relative levels of sister bacterial genera differentiate human cohorts by diet, geography, and culture (27-29). Datasets analyzed by a phylogenetically aware ILR transform could therefore reveal ecological and evolutionary factors shaping host-associated microbial communities.

We term our approach the **Ph**ylogenetic **ILR** (PhILR) transform. Using environmental and human-associated 16S rRNA studies as benchmarks, we show that the accuracy of distance-based and machine learning models often increases and never decreases after applying the PhILR transform, relative to applying the same models on untransformed (raw) or log transformed relative abundance data. Moreover, because the PhILR transform incorporates phylogenetic information, statistics applied in the PhILR coordinate system naturally identify bacterial clades that may have differentiated to adapt to distinct body sites. The PhILR coordinate system can also be used to show that, in all human body sites studied, the degree to which neighboring bacterial clades covary tends to increase with the phylogenetic relatedness between clades. This result

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

supports theories that environmental forces, and not competition between bacteria, primarily structure the assembly of human microbiota.

RESULTS

Constructing the PhILR transform

The PhILR transform has two goals. The first goal is to transform input microbiota data into an unconstrained orthogonal space while preserving all information contained in the original composition. The second goal is to conduct this transform using phylogenetic information. To achieve these dual goals on a given set of N samples consisting of relative measurements of D taxa (**Fig. 1B**), we transform data into a new space of N samples and $(D-1)$ coordinates termed 'balances' (**Fig. 1C-E**). Each balance y_i^* is associated with a single internal node i of a phylogenetic tree with the D taxa as leaves. The balance represents the log-ratio of the geometric mean relative abundance of the two clades of taxa that descend from i (see *Methods* and *SI Text*). Balances are by definition orthogonal, which ensures standard statistical tools may be applied to the transformed data without compositional artifacts; however, this orthogonality does not imply statistical independence (*SI Text*). Each balance is also standardized so that balances across the tree are statistically comparable (10), even when balances have differing numbers of descendant tips or exist at different depths in the tree. This standardization also ensures that the variance of PhILR balances has a consistent scale, unlike the variance of standard log-ratios where it is often unclear what constitutes a large or small variance (4).

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Benchmarking statistics in the PhILR coordinate system

To assess how the PhILR transform affects statistical inference on microbiota datasets, we first examined measures of community dissimilarity. Microbiota analyses commonly compute the dissimilarity or distance between pairs of samples and identify groups of samples with differing community structure. We benchmarked how the PhILR transform affected the task of grouping samples using three microbiota surveys as references: Costello Skin Sites (CSS), a dataset of 357 samples from 12 human skin sites (30); Human Microbiome Project (HMP), a dataset of 4,743 samples from 18 human body sites (*e.g.*, skin, vaginal, oral, and stool) (31); and, Global Patterns (GP), a dataset of 26 samples from 9 human or environmental sites (1) (**Fig. S1**). We computed distances between samples in the PhILR coordinate system using Euclidean distances. We compared this measure to common measures of microbiota distance or dissimilarity (Unifrac, Bray-Curtis and Jaccard) as well as a simple measure (Euclidean) applied to raw relative abundance data (32).

The PhILR transform significantly improved distance-based analyses of microbiota samples. Principal coordinate analyses (PCoA) qualitatively demonstrated separation of body sites using both Euclidean distances on PhILR transformed data (**Fig. 2A**) and with a number of standard distance measures calculated on raw relative abundance data (**Fig. S2**). To quantitatively compare distance measures, we tested how well habitat information explained variability among distance matrices using PERMANOVA (33). The Euclidean distance in the PhILR coordinate system

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significantly outperformed the five competing distance metrics across all benchmarks, except in comparison to Weighted Unifrac when applied to the HMP dataset (**Fig. 2B**). These results indicate that the Euclidean distance, when measured in the PhILR coordinate system, generally exhibited superior performance to more sophisticated distance measures used on raw relative abundances.

Next, we tested the performance of predictive statistical models in the PhILR coordinate system. We examined four standard supervised machine learning techniques: logistic regression (LR), support vector machines (SVM), k-nearest neighbors (kNN), and random forests (RF) (34). We applied these methods to the same three reference datasets used in our comparison of distance metrics. As a baseline, the machine learning methods were applied to raw relative abundance datasets and raw relative abundance data that had been log-transformed.

The PhILR transform significantly improved supervised classification accuracy in 7 of the 12 benchmark tasks compared to raw relative abundances (**Fig. 2C**). Accuracy improved by more than 90% in two benchmarks (SVM on HMP and GP), relative to results on the raw data. Log transformation of the data also improved classifier performance significantly on 6 of the 12 benchmarks but also significantly underperformed on 1 benchmark compared to raw relative abundances. In addition, the PhILR transform significantly improved classification accuracy in 5 of the 12 benchmarks relative to the log transform. Overall, the PhILR transform often outperformed the raw and log transformed relative abundances with respect to classification accuracy and was never significantly worse.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identifying neighboring clades that differ by body site preference

We next used a sparse logistic regression model to examine which balances distinguished human body site microbiota in the HMP dataset. Such balances could be used to identify neighboring bacterial clades whose relative abundances capture community-level differences between body site microbiota. Microbial genetic differentiation may be driven by adaptation to new resources or lifestyle preferences (26), meaning that distinguishing balances near the tips of the bacterial tree may correspond to clades adapting to human body site environments.

We identified dozens of highly discriminatory balances, which were spread across the bacterial phylogeny (**Fig. 3A** and **Fig. S3, S4**). Some discriminatory balances were found deep in the tree. Abundances of the Firmicutes, Bacteroidetes, and Proteobacteria relative to the Actinobacteria, Fusobacteria, and members of other phyla, separated skin body sites from oral and stool sites (**Fig. 3B**). Levels of the genus *Bacteroides* relative to the genus *Prevotella* differentiated stool microbiota from other communities on the body (**Fig. 3C**). Notably, values of select balances below the genus level also varied by body site. Relative levels of sister *Corynebacterium* species separated human skin sites from gingival sites (**Fig. 3D**). Species-level balances even differentiated sites in nearby habitats; levels of sister *Streptococcus* species or sister *Actinomyces* species vary depending on specific oral sites (**Fig. 3E,F**). These results show that the PhILR transform can be used to highlight ancestral balances that

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distinguish body site microbiota, as well as to identify more recent balances that may separate species that have adapted to inhabit different body sites.

195

Balance variance and microbiota assembly

Observing that discriminatory balances could be found across the phylogenetic tree suggested investigating theories of microbial community assembly within the PhILR coordinate system. Closely related microbes may directly compete for nutrients and
200 thus exclude one another from a given site (35). By contrast, related taxa may also have similar lifestyle characteristics and thus covary in environments favoring their shared traits (36). Patterns of phylogenetic clustering (36) or predicted metabolic interactions (37, 38) have previously been used to distinguish the relative importance of competition and environmental selection in structuring microbial communities (36).

205 The variance of a balance in the PhILR coordinate system provides an alternative phylogenetic method to measure how bacterial taxa covary across environments. In contrast to standard measures of association (*e.g.*, Pearson correlation), balance variance is robust to compositional artifacts (10). When the variance of a balance between two clades approaches zero, the mean abundance of
210 taxa in each of the two clades will be linearly related and thus exhibit shared dynamics across microbial habitats (39). By contrast, when a balance exhibits high variance, related bacterial clades exhibit unlinked or exclusionary dynamics across samples. A pattern of lower balance variance near the tips of the phylogenetic tree would suggest that closely related taxa tend to covary and support the hypothesis that environmental

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

forces structure sampled microbial communities; by contrast, higher balance variance near the tips of the phylogeny would suggest related taxa do not covary and support a competitive model underlying community structure.

For all body sites in the HMP dataset, we observed significantly decreasing balance variances near the tips of the phylogenetic tree ($p<0.01$, permutation test with FDR correction; *Methods*; **Fig. 4A-F** and **Fig. S5, S6**). Low variance balances predominated near the leaves of the tree. Examples of such balances involved *B. fragilis* species in stool (**Fig. 4H**), *Rothia mucilaginosa* species in the buccal mucosa (**Fig. 4J**), and *Lactobacillus* species in the mid-vagina (**Fig. 4L**). By contrast, higher variance balances tended to be more basal on the tree. Two examples of high variance balances corresponded with clades at the order (**Fig. 4G**) and family (**Fig. 4I**) levels. In the case of select Lactobacilli in the vagina, neighboring clades appeared to exclude one another (**Fig. 4K**). We performed LOESS regression to investigate how the relationship between balance variance and phylogenetic depth varied locally at different taxonomic scales. This regression revealed that trends between variance and phylogenetic depth were stronger above the species level than below this level (*Methods*; **Fig. 4D-F** and **Fig. S6**). Overall, the observed pattern of decreasing balance variance near the tips of the phylogenetic tree demonstrated that closely related bacteria tend to covary in human body sites, supporting the hypothesis that environmental forces structure human-associated microbial communities more than competitive forces. However, the weaker relationship between balance variance and

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phylogenetic depth below the species level suggests that environmental forces induce similar selective pressure on bacterial strains within the same species.

DISCUSSION

The relative nature of microbiota survey data can result in spurious statistical analyses. Here, we addressed this problem by developing a technique to transform conventional microbiota data into a new space free from compositional effects. The resulting data space improves the accuracy of common statistical methods when applied to microbiota data. The PhILR transform also embeds phylogenetic information into statistics computed in its coordinate system. In doing so, the PhILR transform provides a natural means for discovering taxonomic and evolutionary factors structuring microbial communities.

Our benchmarking experiments show that relative to untransformed data, the PhILR transform improves the performance of both distance-based and supervised machine learning algorithms applied to microbiota data. We note that performance gains achieved by the PhILR transform on community distance benchmarks are surpassed in only one instance by the phylogenetic distance Weighted Unifrac. Unifrac down weights the influence of closely related taxa when computing the distance between communities, and a similar effect can be achieved when Euclidean distances are calculated with PhILR transformed data (*Methods*). Because related bacteria often share similar traits (40), this weighting likely biases the grouping of microbiota so that communities with similar functional profiles are nearby in the transformed space. Our

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

benchmarking suggests that in practice, leveraging phylogenetic information and accounting for compositional constraints can improve statistical analysis of microbiota surveys.

The PhILR transform's use of phylogenetic information also helps formalize the practice of distinguishing microbiota using taxonomic ratios. For example, the enteric Firmicutes to Bacteroidetes ratio has repeatedly been compared between obese and lean individuals (41-44). In part, such ratios are relied on because they simplify microbiota with hundreds of component species into single variables. The PhILR transform provides a statistical framework that guides the process of finding pairs of bacterial clades that differentiate groups of samples. Important balances identified correspond to ratios already known to distinguish microbiota in practice; *e.g.* relative abundances of Actinobacteria to other bacterial phyla, which we find separate skin samples from other human body sites (**Fig. 3B**), have previously been used to identify skin microbiota (45). Interestingly, the PhILR transform also identifies new uses for well-known ratios. The balance between the genera *Bacteroides* and *Prevotella*, which has been previously linked to inter-individual stool variation (46, 47), emerged as one of the best discriminants separating human stool samples from other body sites (**Fig. 3C**). This finding was likely sensitive to the use of a Western subject cohort in the Human Microbiome Project; a cohort drawn from non-industrialized settings would likely have exhibited higher levels of enteric *Prevotella* (27). Nevertheless, we anticipate the PhILR coordinate system to be a useful tool for identifying clades of bacteria that vary by habitat.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Although discriminatory balances between habitats could be constructed between unrelated clades, the PhILR transform's reliance on phylogenetically defined balances also carries the benefit of linking subsequent statistical analyses to evolutionary models. A symbiosis exists between our understanding of bacterial evolution and the ecology of host-associated microbial communities (48). Microbiota studies have shown that mammals and bacteria cospeciated over millions of years (49, 50), and human gut microbes have revealed the forces driving horizontal gene transfer between bacteria (51). Evolutionary tools and theory have been used to explain how cooperation benefits members of gut microbial communities (52), and raise concerns that rising rates of chronic disease are linked to microbiota disruption (53). The PhILR transform provides a convenient framework for carrying out statistical analyses in a coordinate system that is evolutionarily informed.

Regression on PhILR transformed data, for example, highlighted balances near the tips of the bacterial phylogeny that distinguished human body sites. These balances may reflect functional specialization, as ecological partitioning among recently differentiated bacterial clades could be caused by genetic adaptation to new environments or lifestyles (26). Indeed, among oral body sites, we observed consistent site specificity of neighboring bacterial clades within the genera *Actinomyces* (**Fig. 3F**) and *Streptococcus* (**Fig. 3E**). Species within the *Actinomyces* genera have been previously observed to partition separately between the teeth, gingival plaque, buccal mucosa and tongue in healthy subjects (54, 55). Even more heterogeneity has been observed within the *Streptococcus* genus, where species have been identified that

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distinguish teeth, plaque, mucosal, tongue, saliva, and other oral sites (54, 55). This partitioning likely reflects variation in the anatomy and resource availability across regions of the mouth (54), as well as the kinds of surfaces bacterial strains can adhere to (55).

We also observed evidence for potential within-genus adaptation to body sites that has not been previously reported. In particular, within the genus *Corynebacterium*, we found ratios of taxa varied among oral plaques and select skin sites (**Fig. 3D**). Although the genus is now appreciated as favoring moist skin environments, the roles played by individual Corynebacteria within skin microbiota remain incompletely understood (45). Precisely linking individual *Corynebacterium* species or strains to body sites is beyond the scope of this study due to the limited taxonomic resolution of 16S rRNA datasets (56, 57). Nevertheless, we believe the PhILR coordinate system may be used in the future to identify groups of related bacterial taxa that have undergone recent functional adaptation.

Another example of how the PhILR transform may be used to provide evolutionary insight arises in our analysis of balance variance and phylogenetic depth. The relative importance of environmental and competitive forces in shaping human microbiota remains an outstanding question for microbial ecology (35). Reports of paired strains within the same genus or species that inhibit growth of one another (58-62) have suggested that competitive forces are dominant. By contrast, we observed decreasing balance variance near the tips of the phylogenetic tree (**Fig. 4A-F** and **Fig. S6**), supporting the hypothesis that microbiota in different body sites are shaped

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

primarily by environmental forces. Such forces could include moisture, oxygen level, or resource availability (45, 63, 64). Our findings complement previous studies that used metabolic interactions to show that in the human gastrointestinal and oral microbiomes, species tend to co-occur with other species with which they strongly compete (37, 38). In addition, our conclusions are supported by recent fecal transplant experiments in humans showing that the presence of conspecific bacteria increases the likelihood that a bacterial strain engrafts in the human gut (65).

We also found that the relationship between balance variance and phylogenetic depth varies with taxonomic scale, appearing stronger at balances corresponding to higher taxonomic scales and weaker at balances near or below the species level. Although it is often believed that microbial phenotypes are linked to phylogenetic distance (40, 66), the precise taxonomic levels at which this relationship degrades remains debated (67). Our phylogenetic analysis suggests that lifestyle characteristics enabling bacteria to persist in human body sites are conserved among strains roughly corresponding to the same species.

Ultimately, though the methods presented here provide a coherent geometric framework for performing microbiota analysis in a compositionally robust manner, future refinements and modifications are possible. As it is often unclear as to when a zero value represents a value below the detection limit (rounded zero) or a truly absent taxa (essential zero), the handling of zero values remains an outstanding challenge for microbiota analysis and compositional data analysis. Here, we have used two methods for handling zeros depending on the biological question of interest (*Methods, SI Text*);

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

however, new mixture models that explicitly allow for both essential and rounded zeros (68) appear promising for microbiota data analysis. Additionally, we chose to use phylogenies to create the sequential binary partitions needed for the ILR transform. This algorithm design choice provided our analyses with evolutionary context, but such context may not be needed for every analysis. Alternative balances between non-phylogenetically neighboring groups of taxa can be constructed and used with the ILR transform, provided the overall partition of the taxa is binary. Lastly, if analytical insights are desired on the level of individual taxa, and not ratios of clades, analysis can be performed in the transformed ILR space and the results then converted back into compositional space using the inverse of the ILR transform (10, 69). This provides an alternative approach to adapting statistical methods for use with compositional microbiota data.

Yet, despite these avenues for improvement or modification, we believe the PhILR transform already enables existing statistical methods to be applied to metagenomic datasets, free from compositional artifacts and framed according to an evolutionary perspective. We emphasize that all statistical tools applied to PhILR transformed data in this study were used 'off-the-shelf' and without modification. This approach contrasts with the more standard practice of adapting current statistical techniques to the limitations of microbiota survey data. Such adaptation is often challenging because many statistics were derived assuming unconstrained orthogonal coordinate systems, not constrained and over-determined compositional spaces. Therefore, while select techniques have already been adapted (*e.g.* distance measures

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that incorporate phylogenetic information (70) and feature selection methods that handle compositional input (14, 15)), it is likely that certain statistical goals, such as non-linear community forecasting or control system modeling, may prove too complex for adapting to microbiota datasets. Beyond microbiota surveys, we also recognize that compositional metagenomics datasets are generated when studying the ecology of viral communities (71) or clonal population structure in cancer (72-74). We expect the PhILR transform to aid other arenas of biological research where variables are measured by relative abundance and related by an evolutionary tree.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

Overview of the PhILR Transform

The PhILR transform is an Isometric Log-Ratio transform (23) defined by using a binary phylogenetic tree as a sequential binary partition. The PhILR transform also involves an optional scaling step to integrate phylogenetic distances into the transformed space (branch length weighting) and two methods for handling zero values (taxa weighting and conditioning on non-zero counts). We describe these methods and their motivation in more detail below. A more detailed description of their derivation and the underlying theory of the transform is provided in the *SI Text*.

The ILR Transform

A typical microbiome sample consists of measured counts c_j for taxa $j \in \{1, \ldots, D\}$. A standard operation is to take count data and transform it to relative abundances. This operation is referred to as closure in compositional data analysis and is given by

$$\boldsymbol{x} = (x_1, \ldots, x_D) = \left(\frac{c_1}{\sum_j c_j}, \ldots, \frac{c_D}{\sum_j c_j} \right)$$

where x_j represents the relative abundance of taxa j in the sample. We can represent a binary phylogenetic tree of the D taxa using a sign matrix Θ. Each row of the sign matrix indexes an internal node i of the tree and each column indexes a tip of the tree. A given element in the sign matrix is ± 1 depending on whether that tip is in the left or right subtree descending from i and 0 if that tip is not a descendent of i (**Fig. S7**). Following Egozcue and Pawlowsky-Glahn (69), we represent the ILR coordinate (balance)

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associated with node i in terms of the shifted composition $\boldsymbol{y} = \boldsymbol{x}/\boldsymbol{p} = (x_1/p_1, \ldots, x_D/p_D)$ as

$$y_i^* = \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} \log \frac{g_p(\boldsymbol{y}_i^+)}{g_p(\boldsymbol{y}_i^-)}, \quad n_i^{\pm} = \sum_{\theta_{ij}=\pm 1} p_j, \quad g_p(\boldsymbol{y}_i^{\pm}) = \exp\left(\frac{\sum_{(\theta_{ij}=\pm 1)} p_j \log y_j}{\sum_{(\theta_{ij}=\pm 1)} p_j}\right)$$

where $g_p(\boldsymbol{y}_i^{\pm})$ represents the weighted geometric mean of the components of $\boldsymbol{y}$ that are descendants of the left or right subtree of node i respectively and p_j is given by the weights assigned to taxa j. When $\boldsymbol{p} = (1, \ldots, 1)$, $\boldsymbol{y} = \boldsymbol{x}$ and the above equation represents the ILR transform as originally published (75). However, when $\boldsymbol{p} \neq (1, \ldots, 1)$, the above equation represents a more generalized form of the ILR transform (69) that allows the effects of very low abundance taxa to be down weighted. *(See the SI Text for more background on the form of this transformation)*

Addressing sparsity through weighting taxa

We make use of this generalized ILR transform and weights $\boldsymbol{p}$ to address the challenge of zero and near-zero counts. In most analyses involving count data the challenge of modeling zero or near-zero counts must be addressed. After replacing zeros with small non-zero counts, a standard procedure to address this challenge is variance stabilization, down-weighting the influence of small counts since these are less reliable and therefore more variable (76). Microbiota datasets are often sparse, with more than than 90% zero counts. Through the use of the generalized ILR we may down weight the less reliable rare and very low abundance taxa by adjusting the weights $\boldsymbol{p}$. This reduces the sensitivity of statistics in the PhILR coordinate system to very low

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

abundance taxa. As total read counts contain variance information (77), we utilize total read counts to choose these weights. We refer to this method as 'taxa weighting'.

Our choice of taxa weights combines two terms multiplicatively, a mean or median of the raw counts for a taxon across the N samples in a dataset and the norm of the vector of relative abundances of a taxon across the N samples in a dataset. This second term ensures that highly site-specific taxa are not unduly down weighted (*SI Text*). Preliminary studies showed that the geometric mean of the counts (with a pseudocount added to avoid skew from zero values) outperformed both the arithmetic mean and median as a measure of central tendency for the counts (data not shown). Both the Euclidean norm and the Aitchison norm improved performance (as measured by classification accuracy or PERMANOVA R^2, see below) in our benchmark tasks as compared to using the geometric mean alone (**Fig. S8**). However in one case (classification using support vector machine on the global patterns dataset) the Euclidean norm greatly outperformed the Aitchison norm and was therefore chosen for our analysis here. Thus, the taxa weights we used are given by

$$p_j = \|\boldsymbol{x}_j\| \sqrt[N]{(c_{j1}+1)\cdot \ldots \cdot (c_{jN}+1)}.$$

Note that we add the subscript j to the right hand side of the above equation to emphasize that this is calculated with respect to a single taxon across the N samples in a dataset. These taxa weights supplement the use of pseudo-counts with variance information from total count data and, with the exception of our analysis of balance variance as a function of phylogenetic depth (see below), are used throughout the analyses presented here.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

445 ***Incorporating branch lengths***

Beyond utilizing the connectivity of the phylogenetic tree to dictate the partitioning scheme for ILR balances, branch length information can be embedded into the transformed space by linearly scaling ILR balances (y_i^*) by the distance between neighboring clades. We call this scaling by phylogenetic distance 'branch length 450 weighting'. This has the effect of scaling distances in the PhILR coordinate system by the relatedness of bacteria present in a community, which is a feature shared by other phylogenetic methods that utilize phylogenetically informed distances (70, 78, 79). However, importantly, whereas those methods are based on reducing the data to a set of distances, the PhILR transform provides an explicit coordinate system of balances 455 where each balance identifies a distinct location on the phylogenetic tree and has evolutionary meaning. Specifically for each coordinate y_i^*, corresponding to node i we use the transform

$$y_i^{*,blw} = y_i^* \cdot f(d_i^+, d_i^-)$$

where $d_i^{\pm}$ represent the branch lengths of the two direct children of node i. When 460 $f(d_i^+, d_i^-) = 1$, the coordinates are not weighted by branch lengths. With the exception of our analysis of balance variance as a function of phylogenetic depth (see below), here we have used $f(d_i^+, d_i^-) = \sqrt{d_i^+ + d_i^-}$ for our branch length weights. When coupled with the taxa weights specified above, the square root of the summed distances had the highest rank in 9 of the 12 supervised classification tasks and 2 of the 3 distance based 465 tasks (compared to either $f(d_i^+, d_i^-) = d_i^+ + d_i^-$ or $f(d_i^+, d_i^-) = 1$; **Fig. S8**). This choice of

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

branch length weights most similarly resembles the generalized UniFrac distance with $d = 0.5$ (33).

Implementation

The PhILR transform, as well as the incorporation of branch length and taxa weightings has been implemented in the R programing language as the package *philr* available at https://github.com/jsilve24/philr. The implementation is limited in time and space complexity by a single matrix multiplication step involving a $D \times (D-1)$ contrast matrix (see *SI Text*) and the $N \times D$ sample dataset and the runtime is therefore expected to be $\mathcal{O}(ND^2)$.

Datasets and Preprocessing

All data preprocessing was done in the R programming language using the *phyloseq* package for analysis of microbiome census data (80) as well as the the *ape* (81) and *phangorn* (82) packages for analysis of phylogenetic trees. For each of the three 16s rRNA datasets (consisting of an Operational Taxonomic Unit (OTU) table, taxonomic classifications, and phylogenetic tree) the preprocessing pipeline was as follows: 1) If unrooted, manually root the phylogenetic tree by specifying an outgroup, 2) resolve multichotomies, if present, with the function multi2di from the *ape* package which replaces multichotomies with a series of dichotomies with one (or several) branch(es) of length zero, 3) filter low abundance taxa and prune the tree accordingly, 4) filter low abundance samples, 5) add a pseudocount of 1 prior to PhILR transformation to avoid

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

taking log-ratios with zero counts. The PhILR transform is robust to changing the value of this pseudocount (**Fig. S9**).

Human Microbiome Project (HMP)

This dataset was obtained from the QIIME Community Profiling Pipeline applied to high-quality reads from the v3-5 region, available at http://hmpdacc.org/HMQCP/. The phylogenetic tree was rooted with the phylum Euryarchaeota as an outgroup and multichotomies were resolved. Samples with fewer than 1000 counts were removed so that our analysis included the same samples as prior analyses (31). Taxa that were not seen with more than 3 counts in at least 1% of samples were removed. Samples from the left and right retroauricular crease and samples from the left and right antecubital fossa were grouped together, respectively, as preliminary PERMANOVA analysis suggested that these sites were indistinguishable (data not shown).

Global Patterns

The Global Patterns dataset was originally published in Caporaso*, et al.* (1). This dataset is provided with the *phyloseq* package and our preprocessing followed the methods outlined in McMurdie and Holmes (80). Specifically, taxa that were not seen with more than 3 counts in at least 20% of samples were removed, the sequencing depth of each sample was standardized to the abundance of the median sampling depth, and finally taxa with a coefficient of variation ≤ 3.0 were removed (80). The tree was rooted with Archaea as the outgroup and no multichotomies were present.

Costello Skin Sites (CSS)

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The original dataset was collected by Costello et al. (30). A subset of the dataset containing only the skin samples was introduced as a benchmark for supervised machine learning by Knights et al. (34). This dataset was obtained from http://knightslab.org/data. The CSS dataset consists of counts from the v2 region of bacterial 16s rRNA genes. The tree was rooted at OTU 12871 (from phylum Plantomycetes) and multichotomies were resolved. This dataset had lower sequencing depth than the other two benchmarks. To retain a reasonable number of taxa while still removing potential spurious reads, we chose to filter taxa that were not seen with greater than 10 counts across the skin samples.

Benchmarking

Distance/Dissimilarity Based Analysis

Distance between samples in PhILR transformed space was calculated using Euclidean distance. All other distance measures were calculated using *phyloseq* on the preprocessed data without adding a pseudocount. Principle coordinate analysis was performed for visualization using *phyloseq*. PERMANOVA was performed using the function *adonis* from the R package *vegan* (v2.3.4). Standard errors were calculated using bootstrap resampling with 100 samples each. Differences between the performance of Euclidean distance in PhILR transformed space and that of each other measure on a given task was tested using two-sided t-tests and multiple hypothesis testing was accounted for using FDR correction.

Supervised Classification

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The performance of PhILR transformed data was compared against data preprocessed using one of two standard strategies for normalizing sequencing depth: the preprocessed data was transformed to relative abundances (*e.g.*, each sample was normalized to a constant sum of 1; *raw*); or, a pseudocount of 1 was added, the data was transformed to relative abundances, and finally the relative abundances were log-transformed (*log)*.

All supervised learning was implemented in Python using the following libraries: *Scikit-learn* (v0.17.1), *numpy* (v1.11.0) and *pandas* (v0.17.1). Four classifiers were used: penalized logistic regression, support vector classification with RBF kernel, random forest classification, and k-nearest-neighbors classification. Each classification task was evaluated using the mean and variance of the test accuracy over 10 randomized test/train (30/70) splits which preserved the percentage of samples from each class at each split. For each classifier, for each split, the following parameters were set using cross-validation on the training set. Logistic regression and Support Vector Classification: the *'C'* parameter was allowed to vary between 10^{-3} to 10^{3} and multi-class classification was handled with a one-vs-all loss. In addition, for logistic regression the penalty was allowed to be either l_1 or l_2. K-nearest-neighbors classification: the 'weights' argument was set to 'distance'. Random forest classification: each forest contained 30 trees and the 'max_features' argument was allowed to vary between 0.1 and 1. All other parameters were set to default values. Due to the small size of the Global Patterns dataset, the supervised classification task was simplified to distinguishing human vs. non-human samples. Differences between each methods'

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

accuracy in a given task was tested using two-sided t-tests and multiple hypothesis testing was accounted for using FDR correction.

555

Identifying balances that distinguish sites

To identify a sparse set of balances that distinguish sampling sites, we fit a multinomial regression model with a grouped l_1 penalty using the R package *glmnet (v2.0.5*). The penalization term lambda was set by visually inspecting model outputs for clear bodysite
560 separation (lambda=0.1198). This resulted in 35 balances with non-zero regression coefficients. Phylogenetic tree visualization was done using the R package *ggtree* (83).

Variance and Depth

The orthonormality of PhILR balances and their association with internal nodes of
565 the phylogenetic tree enabled us to investigate how the association between neighboring clades varied with phylogenetic depth. The original variance of log-ratios proposed by Aitchison as a measure of association (5) are not comparable on the same scale (it is unclear what constitutes a large or small variance) (4). However, the variance of ILR coordinates can be compared on the same scale because of the unit length of
570 ILR basis elements (10).

We computed the variance of balances that did not include taxa weights (*i.e.,* $\boldsymbol{p} = (1, \ldots, 1)$). We also omitted branch length weights (*i.e.,* $f(d_i^+, d_i^-) = 1$) when computing balance variances to avoid directly weighting variances by phylogenetic distances. We omitted taxa weights because of concern that zero values would vary as a function of

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

575 phylogenetic depth and could therefore systematically bias our analysis of balance variance as a function of phylogenetic depth. That is, balances closer to the root of the tree have more descendant tips that could be non-zero compared to balances closer to the tips of the tree. We therefore calculated balance values (y_i^*) on non-zero counts. In practice, we retained balances that met the following criteria: the term $g_p(\boldsymbol{y}_i^+)/g_p(\boldsymbol{y}_i^-)$ 580 had non-zero counts from some part(s) within the subcomposition $\boldsymbol{y}_i^+$ and some other part(s) within the subcomposition $\boldsymbol{y}_i^-$ in at least 40 samples from that body site. To further focus our analysis of each HMP body site on non-zero counts, prior to calculation of balance values, taxa present in less than 20% of samples from that site were excluded and subsequently samples that had less than 50 total counts were 585 excluded.

In order to investigate the overall relationship between balance variance and phylogenetic depth we used linear regression. A balance's depth in the tree was calculated as its mean distance to its descendant tips (d). For a given body site the following model was fit:

590 $$\log var(y^*) = \beta \log d + \alpha$$

where d represents mean distance from a balance to its descendant tips. We then set out to test the null hypothesis that $\beta = 0$, or that the variance of the log-ratio between two clades was invariant to the distance of the two clades from their most recent common ancestor. For each site, a null distribution for β was constructed by 595 permutations of the tip labels of phylogenetic tree. We chose this permutation scheme to ensure that the increasing variance we saw with increasing proximity of a balance to

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the root was not because deeper balances had more descendant tips, an artifact of variance scaling with mean abundance, or due to bias introduced due to our handling of zeros. Furthermore, the null distribution for β is symmetric about $\beta = 0$ which further supports that balance variance depends on phylogenetic depth through an ecological mechanism and not through a statistical artifact (**Fig. S10**). Two tailed p-values were calculated for β based on 20000 samples from each site's respective null distribution. FDR correction was applied to account for multiple hypothesis testing between body sites.

To visualize local trends in the relationship between balance variance and phylogenetic depth, a LOESS regression was fit independently for each body site. This was done using the function *geom_smooth* from the R package ggplot2 (v2.1.0) with default parameters.

Integrating Taxonomic Information

Taxonomy was assigned to OTUs in the HMP dataset using the *assign_taxonomy.py* script from *Qiime* (v1.9.1) to call *uclust* (v1.2.22) with default parameters using the representative OTU sequences obtained as described above. Taxonomic identifiers were assigned to the two descendant clades of a given balance separately using a simple voting scheme and combined into a single name for that balance. The voting scheme occurs as follows: (1) for a given clade, the entire taxonomy table was subset to only contain the OTUs that were present in that clade (2) starting at the finest taxonomic rank the subset taxonomy table was checked to see if

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

any species identifier represented ≥95% of the table entries at that taxonomic rank, if so that identifier was taken as the taxonomic label for the clade (3) if no consensus identifier was found, the table was checked at the next most-specific taxonomic rank.

Median phylogenetic depths for each taxonomic rank was estimated by first decorating a phylogenetic tree with taxonomy information using *tax2tree* (v1.0) (84). For a given taxonomic rank the mean distance to tips was calculated for each internal node possessing a label that ended in that rank. The median of these distances was used to display an estimate of the phylogenetic depth of that given rank. This calculation of median phylogenetic depth of different taxonomic ranks was done separately for each body site.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements:

We thank Jesse Shapiro, Aspen Reese, Firas Midani, Heather Durand, Jonathan Friedman, Susan Holmes, and Simon Levin for their helpful comments, Dan Knights for providing us with the CSS dataset, and Klaus Schliep and Liam Revell for their insight into manipulation of phylogenetic trees in the R programming language. JS was supported in part by the Duke University Medical Scientist Training Program (GM007171). LAD was supported by the Global Probiotics Council, a Searle Scholars Award, and an Alfred P. Sloan Research Fellowship.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1. Caporaso JG*, et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *P Natl Acad Sci USA* 108 Suppl 1:4516-4522.
2. Waldor MK*, et al.* (2015) Where next for microbiome research? *PLoS Biol* 13(1):e1002050.
3. Jackson DA (1997) Compositional data in community ecology: The paradigm or peril of proportions? *Ecology* 78(3):929-940.
4. Friedman J & Alm EJ (2012) Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 8(9):e1002687.
5. Aitchison J (1986) *The statistical analysis of compositional data* (Chapman and Hall, London ; New York).
6. Lovell D*, et al.* (2011) Proportions, percentages, ppm: do the molecular biosciences treat compositional data right. *Compositional Data Analysis: Theory and Applications*, (John Wiley & Sons, Ltd.), pp 193-207.
7. Gloor GB, Macklaim JM, Vu M, & Fernandes AD (2016) Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian Journal of Statistics* 45(4):73.
8. Li HZ (2015) Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application, Vol 2* 2(1):73-94.
9. Tsilimigras MC & Fodor AA (2016) Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol* 26(5):330-335.
10. Pawlowsky-Glahn V, Egozcue JJ, & Tolosana-Delgado R (2015) *Modeling and analysis of compositional data* (John Wiley & Sons, Ltd).
11. Mandal S*, et al.* (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* 26:27663.
12. Fang H, Huang C, Zhao H, & Deng M (2015) CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* 31(19):3172-3180.
13. La Rosa PS*, et al.* (2012) Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one* 7(12):e52078-e52078.
14. Chen J & Li H (2013) Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. *Ann Appl Stat* 7(1):418-442.
15. Lin W, Shi PX, Feng R, & Li H (2014) Variable selection in regression with compositional covariates. *Biometrika* 101(4):785-797.
16. Paulson JN, Stine OC, Bravo HC, & Pop M (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 10(12):1200-1202.
17. Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
18. Bacon-Shone J (2011) A Short History of Compositional Data Analysis. *Compositional Data Analysis*, eds Pawlowsky-Glahn V & Buccianti A (John Wiley & Sons, Ltd), pp 1-11.
19. Kurtz ZD*, et al.* (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* 11(5):e1004226.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20. Lee SC*, et al.* (2014) Helminth colonization is associated with increased diversity of the gut microbiota. *PLoS Negl Trop Dis* 8(5):e2880.
21. Fernandes AD*, et al.* (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2:15.
22. Gloor GB, Wu JR, Pawlowsky-Glahn V, & Egozcue JJ (2016) It's all relative: analyzing microbiome data as compositions. *Ann Epidemiol* 26(5):322-329.
23. Egozcue JJ & Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Mathematical Geology* 37(7):795-828.
24. Finucane MM, Sharpton TJ, Laurent TJ, & Pollard KS (2014) A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. *PLoS One* 9(1):e84689.
25. Le Cao KA*, et al.* (2016) MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities. *PLoS One* 11(8):e0160169.
26. Hunt DE*, et al.* (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320(5879):1081-1085.
27. De Filippo C*, et al.* (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *P Natl Acad Sci USA* 107(33):14691-14696.
28. Wu GD*, et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334(6052):105-108.
29. Yatsunenko T*, et al.* (2012) Human gut microbiome viewed across age and geography. *Nature* 486(7402):222-227.
30. Costello EK*, et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science* 326(5960):1694-1697.
31. Consortium HMP (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207-214.
32. Kuczynski J*, et al.* (2010) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods* 7(10):813-819.
33. Chen J*, et al.* (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28(16):2106-2113.
34. Knights D, Costello EK, & Knight R (2011) Supervised classification of human microbiota. *FEMS Microbiol Rev* 35(2):343-359.
35. Cavender-Bares J, Kozak KH, Fine PV, & Kembel SW (2009) The merging of community ecology and phylogenetic biology. *Ecol Lett* 12(7):693-715.
36. Horner-Devine MC & Bohannan BJ (2006) Phylogenetic clustering and overdispersion in bacterial communities. *Ecology* 87(7 Suppl):S100-108.
37. Levy R & Borenstein E (2013) Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *P Natl Acad Sci USA* 110(31):12804-12809.
38. Levy R & Borenstein E (2014) Metagenomic systems biology and metabolic modeling of the human microbiome: from species composition to community assembly rules. *Gut Microbes* 5(2):265-270.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

39. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, & Bahler J (2015) Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol* 11(3):e1004075.
40. Martiny JB, Jones SE, Lennon JT, & Martiny AC (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* 350(6261):aac9323.
41. Ley RE*, et al.* (2005) Obesity alters gut microbial ecology. *P Natl Acad Sci USA* 102(31):11070-11075.
42. Mariat D*, et al.* (2009) The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol* 9:123.
43. Ley RE, Turnbaugh PJ, Klein S, & Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* 444(7122):1022-1023.
44. Sze MA & Schloss PD (2016) Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *MBio* 7(4).
45. Grice EA & Segre JA (2011) The skin microbiome. *Nat Rev Microbiol* 9(4):244-253.
46. Arumugam M*, et al.* (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174-180.
47. Koren O*, et al.* (2013) A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol* 9(1):e1002863.
48. Matsen FAt (2015) Phylogenetics and the human microbiome. *Syst Biol* 64(1):e26-41.
49. Moeller AH*, et al.* (2016) Cospeciation of gut microbiota with hominids. *Science* 353(6297):380-382.
50. Ley RE*, et al.* (2008) Evolution of mammals and their gut microbes. *Science* 320(5883):1647-1651.
51. Smillie CS*, et al.* (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241-244.
52. Rakoff-Nahoum S, Foster KR, & Comstock LE (2016) The evolution of cooperation within the gut microbiota. *Nature* 533(7602):255-259.
53. Blaser MJ & Falkow S (2009) What are the consequences of the disappearing human microbiota? *Nat Rev Microbiol* 7(12):887-894.
54. Aas JA, Paster BJ, Stokes LN, Olsen I, & Dewhirst FE (2005) Defining the normal bacterial flora of the oral cavity. *J Clin Microbiol* 43(11):5721-5732.
55. Mager DL, Ximenez-Fyvie LA, Haffajee AD, & Socransky SS (2003) Distribution of selected bacterial species on intraoral surfaces. *J Clin Periodontol* 30(7):644-654.
56. Janda JM & Abbott SL (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* 45(9):2761-2764.
57. Vetrovsky T & Baldrian P (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* 8(2):e57923.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

58. Iwase T*, et al.* (2010) Staphylococcus epidermidis Esp inhibits Staphylococcus aureus biofilm formation and nasal colonization. *Nature* 465(7296):346-349.
59. Gilmore MS*, et al.* (2015) Pheromone killing of multidrug-resistant Enterococcus faecalis V583 by native commensal strains. *P Natl Acad Sci USA* 112(23):7273-7278.
60. Gebhart D*, et al.* (2012) Novel high-molecular-weight, R-type bacteriocins of Clostridium difficile. *J Bacteriol* 194(22):6240-6247.
61. Buffie CG*, et al.* (2015) Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* 517(7533):205-208.
62. Kommineni S*, et al.* (2015) Bacteriocin production augments niche competition by enterococci in the mammalian gastrointestinal tract. *Nature* 526(7575):719-722.
63. David LA*, et al.* (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505(7484):559-563.
64. Morris RL & Schmidt TM (2013) Shallow breathing: bacterial life at low O(2). *Nat Rev Microbiol* 11(3):205-212.
65. Li SS*, et al.* (2016) Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* 352(6285):586-589.
66. Morrissey EM*, et al.* (2016) Phylogenetic organization of bacterial activity. *ISME J* 10(9):2336-2340.
67. Philippot L*, et al.* (2010) The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Microbiol* 8(7):523-529.
68. Bear J & Billheimer D (2016) A Logistic Normal Mixture Model for Compositional Data Allowing Essential Zeros. *Austrian Journal of Statistics* 45(4):3-23.
69. Egozcue JJ & Pawlowsky-Glahn V (2016) Changing the Reference Measure in the Simlex and its Weightings Effects. *Austrian Journal of Statistics* 45(4):25-44.
70. Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71(12):8228-8235.
71. Culley AI, Lang AS, & Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. *Science* 312(5781):1795-1798.
72. Britanova OV*, et al.* (2014) Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J Immunol* 192(6):2689-2698.
73. Yuan K, Sakoparnig T, Markowetz F, & Beerenwinkel N (2015) BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol* 16:36.
74. Roth A*, et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* 11(4):396-398.
75. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, & Barcelo-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35(3):279-300.
76. Good IJ (1956) On the Estimation of Small Frequencies in Contingency-Tables. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 18(1):113-124.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

77. McMurdie PJ & Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 10(4):e1003531.
78. Fukuyama J, McMurdie PJ, Dethlefsen L, Relman DA, & Holmes S (2012) Comparisons of distance methods for combining covariates and abundances in microbiome studies. *Pacific Symposium on Biocomputing.*, (NIH Public Access), p 213.
79. Purdom E (2011) Analysis of a Data Matrix and a Graph: Metagenomic Data and the Phylogenetic Tree. *Annals of Applied Statistics* 5(4):2326-2358.
80. McMurdie PJ & Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8(4):e61217.
81. Paradis E, Claude J, & Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289-290.
82. Schliep KP (2011) phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4):592-593.
83. Yu G, Smith DK, Zhu H, Guan Y, & Lam TTY (2016) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*.
84. McDonald D*, et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6(3):610-618.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

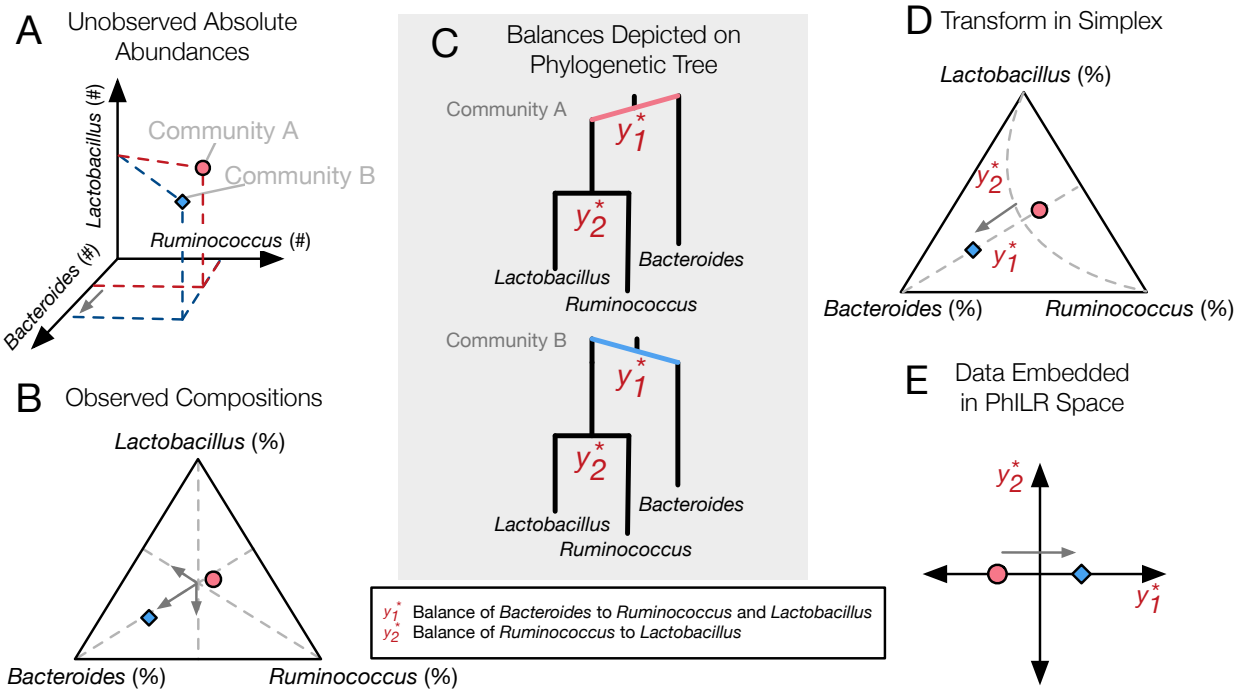

Fig. 1 PhILR uses an evolutionary tree to embed microbiota data into an orthogonal, phylogenetically informed space. (**A**) Two hypothetical bacterial communities share identical absolute numbers of *Lactobacillus*, and *Ruminococcus* bacteria; they differ only in the absolute abundance of *Bacteroides* which is higher in community A (red circle) compared to community B (blue diamond). (**B**) A ternary plot depicts proportional data typically analyzed in a sequencing-based microbiota survey. Note that viewed in terms of proportions the space is constrained and the axes are not orthogonal. As a result, all three genera have changed in relative abundance between the two communities. (**C**) Schematic of the PhILR transform based on a phylogenetic sequential binary partition. The PhILR coordinates can be viewed as 'balances' between the weights (relative abundances) of the two subclades of a given internal node. In community B, the greater abundance of *Bacteroides* tips the balance y_1^* to the right. (**D**) The PhILR transform can be viewed as a new coordinate system (grey dashed lines) in the proportional data space. (**E**) The data transformed to the orthogonal PhILR space. Note that in contrast to the raw proportional data (**B**), the PhILR space only shows a change in the variable associated with *Bacteroides*.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

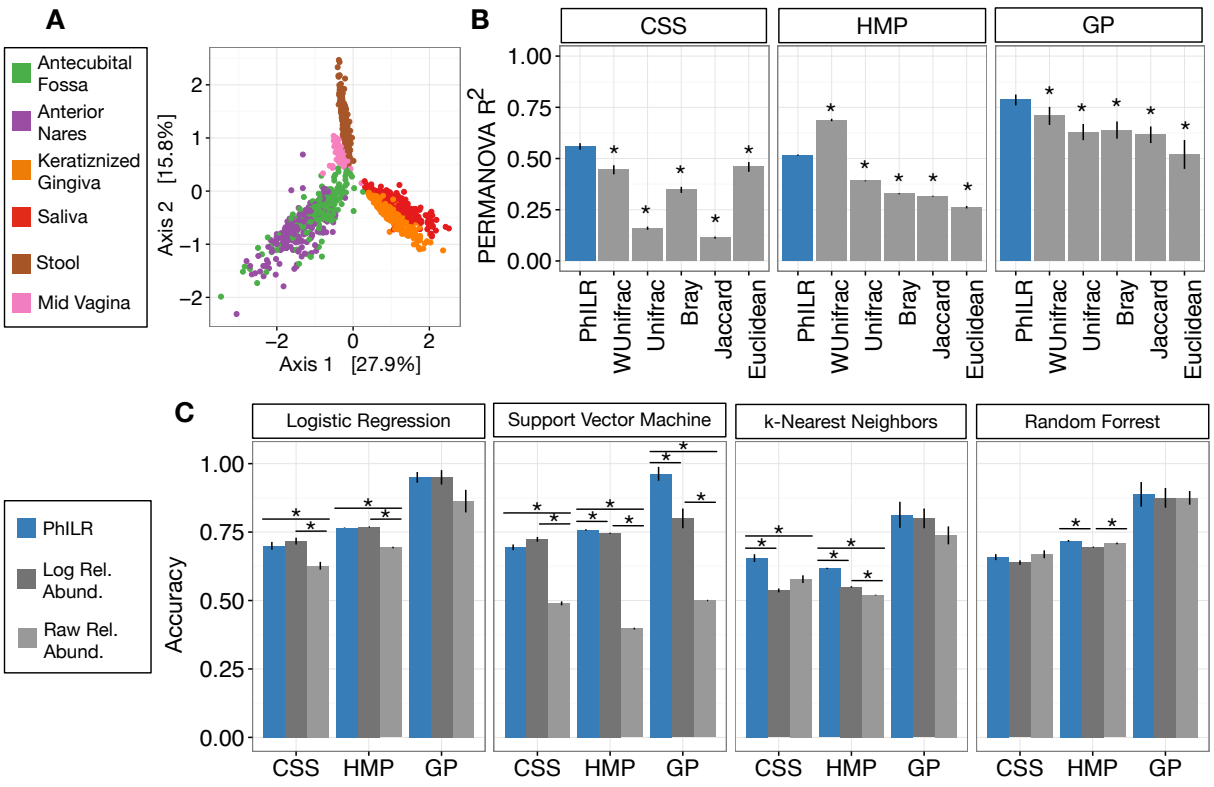

Fig. 2 PhILR transform improves performance of standard statistical models on microbiota data. Benchmarks were performed using three datasets: Costello Skin Sites(CSS), Global Patterns (GP), Human Microbiome Project (HMP). (**A**) Sample distance visualized using principal coordinate analysis (PCoA) of Euclidean distances computed in PhILR coordinate system. A comparison to PCoAs calculated with other distance measures is shown in **Fig. S2**. (**B**) Sample distance (or dissimilarity) was computed by a range of statistics. R^2 values from PERMANOVA were used to measure how well sample location explained the variability in distances between samples. Distances in the PhILR transformed space were calculated using Euclidean distance. Distances between samples on raw relative abundance data were computed using Weighted and Unweighted UniFrac (WUnifrac and Unifrac, respectively), Bray-Curtis, Binary Jaccard, and Euclidean distance. Error bars represent standard error measurements from 100 bootstrap replicates and (*) denotes a p-value of ≤0.01 after FDR correction of pairwise tests against PhILR. (**C**) Accuracy of supervised classification methods tested on benchmark datasets. The PhILR transform significantly improved supervised learning algorithm accuracy in 7 out of 12 supervised classification benchmarks compared to raw or log-transformed raw data. Error bars represent standard error measurements from 10 test/train splits and (*) denotes a p-value of ≤0.01 after FDR correction of all pairwise tests.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

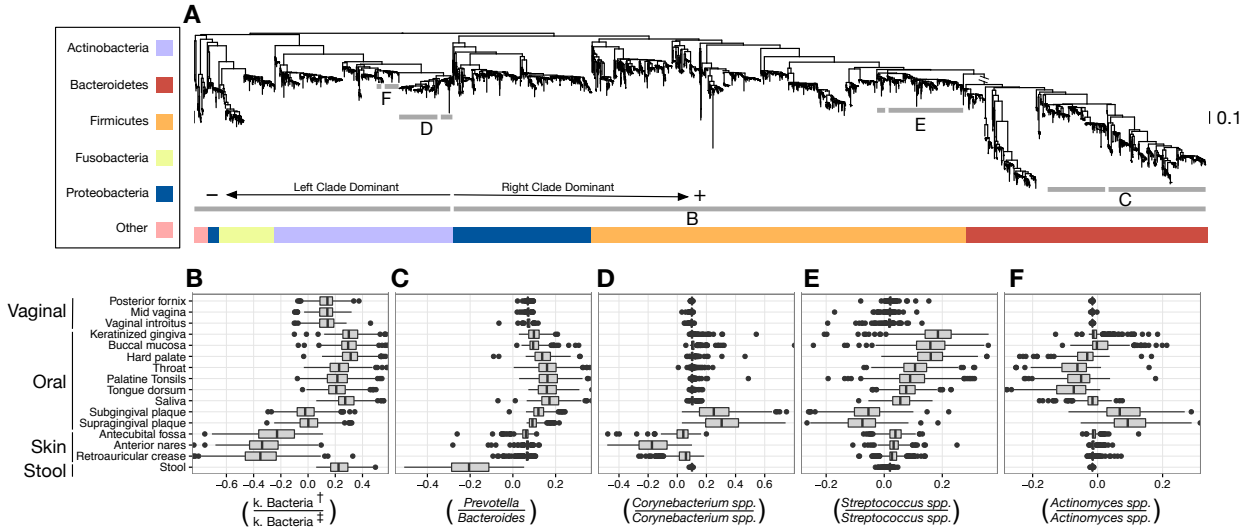

Fig. 3 Balances distinguishing human microbiota by body site. Sparse logistic regression was used to identify balances that best separated the different sampling sites (full list of balances provided in **Fig. S3-4**). (**A**) Each balance is represented on the tree as a broken grey bar. The left portion of the bar identifies the clade in the denominator of the log-ratio, and the right portion identifies the clade in the numerator of the log-ratio. The branch leading from the Firmicutes to the Bacteroidetes has been rescaled to facilitate visualization. (**B**-**F**) The distribution of balance values across body sites. Vertical lines indicate median values, boxes represent interquartile ranges (IQR) and whiskers extend to 1.5 IQR on either side of the median. Balances between: (**B**) the phyla Actinobacteria and Fusobacteria versus the phyla Bacteroidetes, Firmicutes, and Proteobacteria distinguish stool and oral sites from skin sites; (**C**) *Prevotella spp.* and *Bacteroides spp.* distinguish stool from oral sites; (**D**) *Corynebacterium spp.* distinguish skin and oral sites; (**E**) *Streptococcus spp*. distinguish oral sites; and, (**F**) *Actinomyces spp.* distinguish oral plaques from other oral sites. (†) Includes Bacteroidetes, Firmicutes, Alpha-, Beta-, and Gamma-proteobacteria. (‡) Includes Actinobacteria, Fusobacteria, Epsilon-proteobacteria, Spirochaetes, and Verrucomicrobia.

bioRxiv preprint doi: https://doi.org/10.1101/072413; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

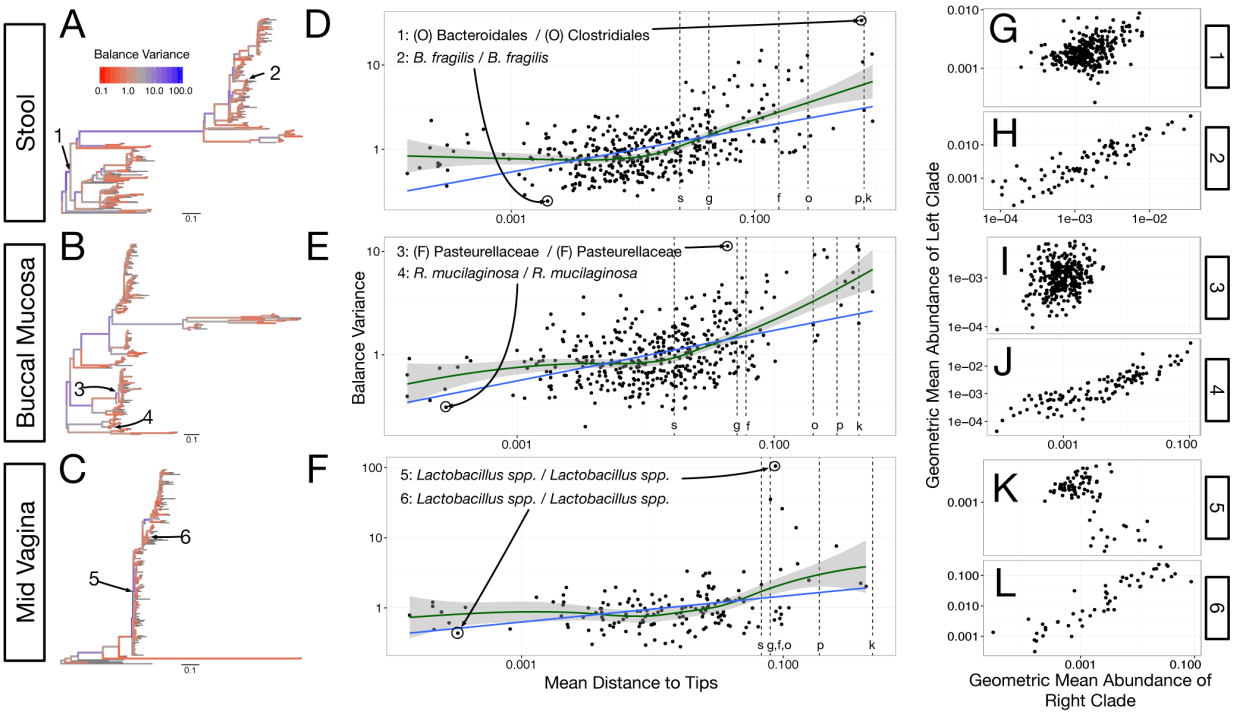

Fig. 4 Neighboring clades covary less with increasing phylogenetic depth. The variance of balance values captures the degree to which neighboring clades covary, with smaller balance variances representing sister clades that covary more strongly. (**A**-**C**) Balance variances were computed among samples from stool (**A**), buccal mucosa (**B**), and the mid-vagina (**C**). Red branches indicate small balance variance and blue branches indicate high balance variance. Balances 1-6 are individually tracked in panels (**D**-**L**). (**D**-**F**) Balance variances within each body site increased linearly with increasing phylogenetic depth on a log-scale (blue line; $p<0.01$, permutation test with FDR correction). Significant trends are seen across all other body sites (**Fig. S6**). Non-parametric LOESS regression (green line and corresponding 95% confidence interval) reveals an inflection point in the relation between phylogenetic depth and balance variance. This inflection point appears below the estimated species level ('s' dotted line; the median depth beyond which balances no longer involve leaves sharing the same species assignment; *Methods*). (**G**-**L**) Examples of balances with high and low variance from panels (**A**-**F**). Low balance variances (**H, J, L**) reflect a linear relationship between the geometric means of sister clades abundances. High balance variances reflect either unlinked (**G, I**) or exclusionary (**K**) dynamics between the geometric means of sister clades abundances.