TITLE: A phylogenetic transform enhances analysis of compositional microbiota data

AUTHORS: Justin D Silverman*†‡, Alex D 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

CORRESPONDING AUTHOR:
Lawrence A David
CIEMAS, Room 2171
101 Science Drive, Box 3382
Durham, NC 27708
919-668-5388
lawrence.david@duke.edu

ABSTRACT: Surveys of microbial communities (microbiota), typically measured as relative abundance of species, have illustrated the importance of these communities in human health and disease. Yet, statistical artifacts commonly plague the analysis of relative abundance data. Here, we introduce the PhILR transform, which incorporates microbial evolutionary models with the isometric log-ratio transform to allow off-the-shelf statistical tools to be safely applied to microbiota surveys. We demonstrate that analyses of community-level structure can be applied to PhILR transformed data with performance on benchmarks rivaling or surpassing standard tools. Additionally, By decomposing distance in the PhILR transformed space, we identified neighboring clades that may have adapted to distinct human body sites. Decomposing variance revealed that covariation of bacterial clades within human body sites increases with phylogenetic relatedness. Together, these findings illustrate how the PhILR transform combines statistical and phylogenetic models to overcome compositional data challenges and enable evolutionary insights relevant to microbial communities.

IMPACT STATEMENT: The PhILR transform uses an evolutionary model to overcome statistical challenges associated with microbiota surveys.

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 (Figure 1A), but rather on relative data occupying a constrained geometric space and represented in a non-Cartesian coordinate system (Figure 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 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 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 can be built from a sequential binary partition of the original variable space (Figure 1C), creating a new coordinate system with an orthonormal basis (Figure 1D and E) (24). However, a known obstacle to using the ILR transform is the choice of partition such that the resulting coordinates are meaningful (10). To date, microbiota studies have chosen ILR coordinates using ad hoc sequential binary

92 partitions of bacterial groups that are not easily interpreted (25, 26). Alternatively,
93 external covariates have been used to pick groups of bacterial taxa to contrast (27).
94 Here, we introduce the bacterial phylogenetic tree as a natural and informative
95 sequential binary partition when applying the ILR transform to microbiota datasets
96 (Figure 1C). Using phylogenies to construct the ILR transform results in an ILR
97 coordinate system capturing evolutionary relationships between neighboring bacterial
98 groups (clades). Analyses of neighboring clades offer the opportunity for biological
99 insight: clade analyses have linked genetic differentiation to ecological adaptation (28),
100 and the relative levels of sister bacterial genera differentiate human cohorts by diet,
101 geography, and culture (29-31). Datasets analyzed by a phylogenetically aware ILR
102 transform could therefore reveal ecological and evolutionary factors shaping host-
103 associated microbial communities.
104 We term our approach the **Ph**ylogenetic **ILR** (PhILR) transform. Using published
105 environmental and human-associated 16S rRNA datasets as benchmarks, we found
106 that simple Euclidean distances calculated on PhILR transformed data provided a
107 compositionally robust alternative to distance/dissimilarity measures like Bray-Curtis,
108 Jaccard, and Unifrac. In addition, we observed that the accuracy of supervised
109 classification methods on our benchmark datasets was matched or improved with
110 PhILR transformed data relative to applying the same models on untransformed (raw) or
111 log transformed relative abundance data. Decomposing distances between samples
112 along PhILR coordinates identified bacterial clades that may have differentiated to adapt
113 to distinct body sites. Similar decomposition of variance along PhILR coordinates
114 showed that, in all human body sites studied, the degree to which neighboring bacterial

115 clades covary tends to increase with the phylogenetic relatedness between clades.
116 Together, these findings demonstrate that the PhILR transform can be used to enhance
117 existing microbiota analysis pipelines, as well as enable novel phylogenetic analyses of
118 microbial ecosystems.
119
120 **RESULTS**
121 **Constructing the PhILR transform**
122 The PhILR transform has two goals. The first goal is to transform input microbiota
123 data into an unconstrained space with an orthogonal basis while preserving all
124 information contained in the original composition. The second goal is to conduct this
125 transform using phylogenetic information. To achieve these dual goals on a given set of
126 N samples consisting of relative measurements of D taxa (Figure 1B), we transform
127 data into a new space of N samples and $(D-1)$ coordinates termed 'balances' (Figure
128 1C-E) (23, 24). Each balance y_i^* is associated with a single internal node i of a
129 phylogenetic tree with the D taxa as leaves (the asterisk denotes a quantity represented
130 in PhILR space). The balance represents the log-ratio of the geometric mean relative
131 abundance of the two clades of taxa that descend from i (*Methods*). Although individual
132 balances may share overlapping sets of leaves and thus exhibit dependent behavior,
133 the ILR transform rescales and combines leaves to form a coordinate system whose
134 basis vectors are orthonormal and the corresponding coordinates are Cartesian (23,
135 24). The orthogonality of basis vectors allows conventional statistical tools to be used
136 without compositional artifacts. The unit-length of basis vectors makes balances across
137 the tree statistically comparable even when they have differing numbers of descendant

tips or exist at different depths in the tree (10). In addition, the unit-length ensures that the variance of PhILR balances has a consistent scale, unlike the variance of log-ratios originally proposed by Aitchison (5) as a measure of association, in which it can be unclear what constitutes a large or small variance (4).

While the above description represents the core of the PhILR transform, we have also equipped the PhILR transform with two sets of weights that can: 1) address the multitude of zero and near-zero counts present in microbiota data; and, 2) incorporate phylogenetic branch lengths into the transformed space. Because zero counts cause problems when computing logs or performing division, zeros are often replaced in microbiota analyses with small non-zero counts. However, to avoid excess zero replacement that may itself introduce bias, stringent hard filtering thresholds are often employed (*e.g*. removing all taxa that are not seen with at least a minimum number of counts in a subset of samples). Still, hard filtering thresholds may remove a substantial fraction of observed OTUs and do not account for the low precision (or high variability) of near-zero counts (7, 32, 33). We therefore developed a 'taxon weighting' scheme that acts as a type of soft-thresholding, supplementing zero replacement methods with a generalized form of the ILR transform that allows weights to be attached to individual taxa (34). Weights are chosen with a heuristic designed to down weight the influence of taxa with many zero or near-zero counts (*Methods*).

Our second weighting scheme is called branch length weighting. Certain analyses may benefit from incorporating information on evolutionary distances between taxa (35-37). For example, because related bacteria may be more likely to share similar traits (38), it may be desirable to consider communities differing only in the abundance

of closely-related microbes to be more similar than communities differing only in the abundance of distantly-related microbes. Because of the one-to-one correspondence between PhILR balances and internal nodes on the phylogenetic tree, evolutionary information can be incorporated into the PhILR transform by scaling balances using the phylogenetic distance between their direct descendants (*Methods*). We note that we employ both branch length weighting and taxon weighting throughout our following analyses except where noted; still, these weights should be considered optional additions to the core PhILR transform.

Benchmarking community-level analyses in the PhILR coordinate system

To illustrate how the PhILR transform can be used to perform standard community-level analyses of microbiota datasets, we first examined measures of community dissimilarity. Microbiota analyses commonly compute the dissimilarity or distance between pairs of samples and use these computed pairwise distances as input to a variety of statistical tools. We investigated how Euclidean distances calculated on PhILR transformed data compared to common ecological measures of microbiota distance or dissimilarity (UniFrac, Bray-Curtis, and Jaccard) as well as Euclidean distance applied to raw relative abundance data in standard distance-based analysis. We chose three different microbiota surveys as reference datasets: Costello Skin Sites (CSS), a dataset of 357 samples from 12 human skin sites (39, 40); Human Microbiome Project (HMP), a dataset of 4,743 samples from 18 human body sites (*e.g.*, skin, vaginal, oral, and stool) (41); and, Global Patterns (GP), a dataset of 26 samples from

9 human or environmental sites (1) (Supplementary file 1 and Figure 2-figure supplement 1).

Distance-based analyses using Euclidean distances computed on PhILR transformed data exhibited performance rivaling common ecological distance or dissimilarity measures. Principal coordinate analyses (PCoA) qualitatively demonstrated separation of body sites using both Euclidean distances on PhILR transformed data (Figure 2A) and with several standard distance measures calculated on raw relative abundance data (Figure 2-figure supplement 2). To quantitatively compare distance measures, we tested how well habitat information explained variability among distance matrices as measured by the R^2 statistic from PERMANOVA (42). By this metric, the Euclidean distance in the PhILR coordinate system significantly outperformed the five competing distance metrics in all but 1 case (in comparison to Weighted UniFrac when applied to the HMP dataset; Figure 2B).

Next, we tested the performance of predictive statistical models in the PhILR coordinate system. We examined four standard supervised classification techniques: logistic regression (LR), support vector machines (SVM), k-nearest neighbors (kNN), and random forests (RF) (40). 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 (Figure 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 accuracy 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.

Identifying neighboring clades that differ by body site preference

While our benchmarking experiments demonstrated how PhILR transformed data performed in community-level analyses, we also wanted to explore potential biological insights afforded by the PhILR coordinate system. We therefore investigated how distances decomposed along PhILR balances using 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 associated with specialization to new resources or lifestyle preferences (28), 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 (Figure 3A and Figure 3-figure supplement 1). 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 (Figure 3B). Levels of the genus *Bacteroides* relative to the genus *Prevotella* differentiated stool microbiota from other communities on the body (Figure 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 (Figure 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 (Figure 3E and F). These results show that the decomposition of distances between groups of samples along PhILR balances can be used to highlight ancestral balances that distinguish body site microbiota, as well as to identify more recent balances that may separate species that have adapted to inhabit different body sites.

Balance variance and microbiota assembly

As a natural extension of our analysis of how distance decomposes along PhILR balances, we next investigated how balance variance decomposed in the PhILR coordinate system. Balance variance is a measure of association between neighboring bacterial clades. When the variance of a balance between two clades approaches zero, the mean abundance of taxa in each of the two clades will be linearly related and thus covary across microbial habitats (43). By contrast, when a balance exhibits high variance, related bacterial clades exhibit unlinked or exclusionary patterns across samples. Unlike standard measures of association (*e.g.,* Pearson correlation) balance variance is robust to compositional effects (10).

Our preliminary investigation demonstrated a striking pattern in which balance variance decreased for balances closer to the tips of the phylogeny and increased for balances nearer to the root. To determine if this observed pattern was not the result of technical artifact, we took the following three steps. First, we omitted branch length weights from the transform as we anticipated that branch lengths may vary non-randomly as a function of depth in the phylogeny. Second, we anticipated that balances near the tips of the phylogeny would be associated with fewer read counts and thus would be more biased by our chosen heuristics for taxon weighting and zero replacement. We therefore omitted taxon weighting, employed more stringent filtering thresholds, and conditioned our calculation of balance variance on non-zero counts rather than using zero-replacement (*Methods*). Third, we combined regression and a permutation scheme to test the null hypothesis that the degree to which neighboring clades covary is independent of the phylogenetic distance between them (*Methods*). By permuting tip labels on the tree, our test generates a restricted subset of random sequential binary partitions that still maintains the count variability (and potential biases due to our zero handling methods) of the observed data.

With our modified PhILR analysis in place, we observed significantly decreasing balance variances near the tips of the phylogenetic tree for all body sites in the HMP dataset ($p<0.01$, permutation test with FDR correction; Figure 4A-F and Figure 4-figure supplements 1-2). Low variance balances predominated near the leaves of the tree. Examples of such balances involved *B. fragilis* species in stool (Figure 4H), *Rothia mucilaginosa* species in the buccal mucosa (Figure 4J), and *Lactobacillus* species in the mid-vagina (Figure 4L). By contrast, higher variance balances tended to be more basal

on the tree. Three examples of high variance balances corresponded with clades at the order (Figure 4G), family (Figure 4I), and genus (Figure 4K) levels. We also observed that the relationship between balance variance and phylogenetic depth varied at different taxonomic scales. LOESS regression revealed that trends between variance and phylogenetic depth were stronger above the species level than below it (*Methods*; Figure 4D-F and Figure 4-figure supplement 2). Overall, the observed pattern of decreasing balance variance near the tips of the phylogenetic tree suggested that closely related bacteria tend to covary in human body sites.

DISCUSSION

There exists a symbiosis between our understanding of bacterial evolution and the ecology of host-associated microbial communities (44). Microbiota studies have shown that mammals and bacteria cospeciated over millions of years (45, 46), and human gut microbes have revealed the forces driving horizontal gene transfer between bacteria (47). Evolutionary tools and theory have been used to explain how cooperation benefits members of gut microbial communities (48), and raise concerns that rising rates of chronic disease are linked to microbiota disruption (49). Here, we aimed to continue building links between microbiota evolution and ecology by designing a data transform that uses phylogenetic models to overcome the challenges associated with compositional data while enabling novel evolutionary analyses.

We found that the resulting PhILR coordinate system, at least with respect to the performance metrics chosen, led to significantly improved performance for a variety of community-level analyses now used in microbiota analysis. While these results add

297 credence to our proposed approach, we underscore that we do not find it essential that
298 PhILR demonstrates superior benchmark performance to motivate its use in microbiota
299 analysis. We believe that the need for compositionally robust tools has already been
300 well established (3-9) and intended these benchmarks to showcase the flexibility and
301 utility of working with PhILR transformed data. We also note that for some analyses, a
302 phylogeny-based ILR transform will not outperform an ILR transform built from another
303 sequential binary partition. In fact, in the absence of branch length weights, any random
304 ILR partition would yield equivalent results on our benchmark tasks. Instead, what
305 distinguishes the PhILR transform from other ILR transforms is the interpretability of the
306 transformed coordinates. Balances in PhILR space correspond to speciation events,
307 which can be a source for biological insight.

308 For example, performing regression on PhILR transformed data, enabled us to
309 decompose the distance between bacterial communities onto individual locations on the
310 phylogeny, highlighting balances near the tips of the tree that distinguished human body
311 sites. These balances may reflect functional specialization, as ecological partitioning
312 among recently differentiated bacterial clades could be caused by genetic adaptation to
313 new environments or lifestyles (28). Indeed, among oral body sites, we observed
314 consistent site specificity of neighboring bacterial clades within the genera *Actinomyces*
315 (Figure 3F) and *Streptococcus* (Figure 3E). Species within the *Actinomyces* genera
316 have been previously observed to partition by oral sites (50, 51). Even more
317 heterogeneity has been observed within the *Streptococcus* genus, where species have
318 been identified that distinguish teeth, plaque, mucosal, tongue, saliva, and other oral
319 sites (50, 51). This partitioning likely reflects variation in the anatomy and resource

availability across regions of the mouth (50), as well as the kinds of surfaces bacterial strains can adhere to (51).

We also observed evidence for potential within-genus adaptation to body sites that has not been previously reported. Within the genus *Corynebacterium*, we found ratios of taxa varied among oral plaques and select skin sites (Figure 3D). Although the genus is now appreciated as favoring moist skin environments, the roles played by individual Corynebacteria within skin microbiota remain incompletely understood (52). 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 (53, 54). 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 can provide biological insights arose in our analysis of how human microbiota variance decomposes along individual balances. We observed that balances between more phylogenetically related clades were significantly more likely to covary than expected by chance. This pattern could reflect evolutionary and ecological forces structuring microbial communities in the human body. Related bacterial taxa have been hypothesized to have similar lifestyle characteristics (38, 55), and may thus covary in human body sites that favor their shared traits (56, 57). An alternative explanation for the balance variation patterns we observed is that sequencing errors and read clustering artifacts are likely to produce OTUs with similar reference sequences and distributions across samples. While we cannot conclusively rule out this alternative hypothesis, we note that it would not explain

why signal for taxa co-variation is strongest for balances at higher taxonomic levels and appears to plateau for balances near or below the species level. A biological explanation for the plateauing signal would be that lifestyle characteristics enabling bacteria to persist in human body sites are conserved among strains roughly corresponding to the same species. Follow-up studies are needed to more conclusively understand how balance variance patterns across the phylogenies can be interpreted from an evolutionary standpoint.

Though the methods presented here provide a coherent geometric framework for performing microbiota analysis free from compositional artifacts, future refinements are possible. Specifically, we highlight issues relating to our choice of weights, the handling of zero values, and information loss during count normalization. Both the taxa weights and the branch lengths weights we introduce here may be viewed as preliminary heuristics; future work will likely yield additional weighting schemes, as well as knowledge for when a given weighting scheme should be matched to an analysis task. In the case of supervised machine learning, weighting selection could be optimized as part of the training process. Additionally, if it is important that the transformed data has meaningful numerical coordinates, such that one desires to interpret the exact numerical value of a given balance in a sample, we suggest that neither branch length weights nor taxa weights be used as these weights can complicate this type of interpretation. Concerning our handling of zero values, this model design choice confronts an outstanding challenge for microbiota and compositional data analysis (9, 58). Part of this challenge's difficulty is whether a zero value represents a value below the detection limit (rounded zero) or a truly absent taxon (essential zero). Here, we

employ zero replacement, which implies an assumption that all zero values represent rounded zeros. New mixture models that explicitly allow for both essential and rounded zeros (59), as well as more advanced methods of zero replacement (58, 60), may enable us to handle zeros in a more sophisticated manner. Lastly, in regards to informational loss caused by normalization, it is known that the number of counts measured for a given taxon influences the precision with which we may estimate its relative abundance in a sample (7, 32, 33). While our taxa weights are intended to address this idea, a fully probabilistic model of counts would likely provide more accurate error bounds for inference. We believe it would be possible to build such a model in a Bayesian framework by viewing the observed counts as multinomial draws from a point in the PhILR transformed space, as has been done for other log-ratio based spaces (61).

Beyond refining the PhILR transform itself, future effort may also be directed towards interpreting the transform's results at the single taxon level. Microbiota studies frequently focus on individual taxa for tasks such as identifying specific bacteria that are causal or biomarkers of disease. Log-ratio approaches can provide a compositionally robust approach to identifying biomarkers based on changes in the relative abundance of individual taxa. Due to the one-to-one correspondence between CLR coordinates and individual taxa, the CLR transform has been used previously to build compositionally robust models in terms of individual taxa (11, 19, 21). However, CLR transformed data suffers from the drawback of a singular covariance matrix, which can make the development of new models based on the CLR transform difficult (10). ILR transformed data do not suffer this drawback (10) and moreover, can be analyzed at the single taxon

level. To do so, the inverse ILR transform can be applied to model results generated in an ILR coordinate system, yielding analyses in terms of changes in the relative abundance of individual taxa (10). The use of the inverse ILR transform in this manner is well established (10, 62, 63) and the inverse transform is provided in the *Methods* (34).

Despite these avenues for improvement, modification, or extension 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 foresee the PhILR transform being used as a default transformation prior to many microbiota analyses, particularly if a phylogenetic perspective is desired. For example, the PhILR transform could be used in lieu of the conventional log transform, which is often the default choice in microbiota analysis but not robust to compositional effects. Substituting PhILR into existing bioinformatics pipelines should often be seamless and we emphasize that all statistical tools applied to PhILR transformed data in this study were used 'off-the-shelf' and without modification. Importantly, such a substitution contrasts with the alternative approach for accounting for compositional microbiota data, which is to modify existing statistical techniques. Such modification is often challenging because many statistics were derived assuming an unconstrained space with an orthonormal basis, not a constrained and over-determined compositional space. Therefore, while select techniques have already been adapted (*e.g.* distance measures that incorporate phylogenetic information (35) 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 the compositional nature of microbiota datasets. Finally, beyond microbiota surveys, we also recognize that compositional metagenomics datasets are generated when studying the ecology of viral communities (64) or clonal population structure in cancer (65-67). 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.

METHODS

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 (5) and is given by

$$\boldsymbol{x} = \mathcal{C}[(c_1, \cdots, c_D)] = \left(\frac{c_1}{\sum_j c_j}, \ldots, \frac{c_D}{\sum_j c_j}\right)$$

where $\boldsymbol{x}$ represents a vector of relative abundances for the D taxa in the sample. We can represent a binary phylogenetic tree of the D taxa using a sign matrix Θ as introduced by Pawlowsky-Glahn and Egozcue (68) and shown in Figure 5. 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 which of the two clades descending from i that tip is a part of and 0 if that tip is not a descendent of i. The assignment of $+1$ versus -1 determines which clade is represented in the numerator versus the denominator of the corresponding log-ratio (as described below). Exchanging this assignment for a given balance switches which clade is represented in

the numerator versus the denominator of the log-ratio. Following Egozcue and Pawlowsky-Glahn (34), we represent the ILR coordinate (balance) 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

$$y_i^* = \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} \log \frac{g_p(\boldsymbol{y}_i^+)}{g_p(\boldsymbol{y}_i^-)}. \tag{1}$$

Here, $g_p(\boldsymbol{y}_i^+)$ and $g_p(\boldsymbol{y}_i^-)$ represents the weighted geometric mean of the components of $\boldsymbol{y}$ that represent tips in the $+1$ or -1 clade descendant from node i respectively. This weighted geometric mean is given by

$$g_p(\boldsymbol{y}_i^{\pm}) = \exp\left(\frac{\sum_{(\theta_{ij}=\pm 1)} p_j \log y_j}{\sum_{(\theta_{ij}=\pm 1)} p_j}\right) \tag{2}$$

where p_j is the weight assigned to taxa j. The term $\sqrt{n_i^+ n_i^- / n_i^+ + n_i^-}$ in equation 1 is the scaling term that ensures that the ILR basis element has unit length and the terms $n_i^{\pm}$ are given by

$$n_i^{\pm} = \sum_{\theta_{ij}=\pm 1} p_j. \tag{3}$$

Note that when $\boldsymbol{p} = (1, \ldots, 1), \boldsymbol{y} = \boldsymbol{x}$, equation 1 represent the ILR transform as originally published (23), equation 2 represents the standard formula for geometric mean of a vector $\boldsymbol{y}$, and equation 3 represents the number of tips that descend from the $+1$ or -1 clade descendant from node i. However, when $\boldsymbol{p} \neq (1, \ldots, 1)$, these three equations represent a more generalized form of the ILR transform that allows weights to be assigned to taxa in the transformed space (34).

Following Egozcue and Pawlowsky-Glahn (34), we also note that the form of the generalized ILR (which we will denote $\mathrm{ilr_p}$) transform can be rewritten in terms of a

generalized CLR transform (which we will denote $\mathrm{clr_p}$). This formulation in terms of the generalized CLR transform can be more efficient to compute and allows the inverse of the transform to be easily described. We can define the generalized CLR transform as

$$\mathrm{clr_p}(\boldsymbol{y}) = \left(log \frac{y_1}{g_p(\boldsymbol{y})}, \cdots, log \frac{y_D}{g_p(\boldsymbol{y})} \right).$$

The generalized ILR transform can then be written as

$$\boldsymbol{y}^* = \mathrm{ilr_p}(\boldsymbol{y}) = \mathrm{clr_p}(\boldsymbol{y})\ \mathrm{diag}(\boldsymbol{p})\ \Psi^T$$

with the ijth element of the matrix Ψ given by

$$\psi_{ij} = \begin{cases} +\frac{1}{n_i^+}\sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} & \text{if } \theta_{ij} = +1 \\ -\frac{1}{n_i^-}\sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} & \text{if } \theta_{ij} = -1 \\ 0 & \text{if } \theta_{ij} = 0\,. \end{cases}$$

With these components defined the inverse of generalized ILR transform can be written as $\mathcal{C}[\boldsymbol{y}] = \mathrm{ilr_p^{-1}}(\boldsymbol{y}^*) = \mathcal{C}[\exp(\boldsymbol{y}^*\Psi)]$ and $\boldsymbol{x} = \mathcal{C}\left[\mathrm{ilr_p^{-1}}(\boldsymbol{y}^*)\boldsymbol{p}\right]$.

Soft thresholding through weighting taxa

We make use of this generalized ILR transform to down weight the influence of taxa with many zero and near-zero counts since these are less reliable and therefore more variable (32). Our choice of taxa weights is a heuristic that combines two terms multiplicatively: a measure of the central tendency of counts, such as the 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. We add this vector norm term to weight taxa by their site-specificity. 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). Additionally, while both the Euclidean norm and the Aitchison norm improved preliminary benchmark performance compared to using the geometric mean alone, in one case (classification using support vector machine on the global patterns dataset), the Euclidean norm greatly outperformed the Aitchison norm (Supplementary file 1). Therefore, our chosen taxa weighting scheme uses the geometric mean times the Euclidean norm:

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

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. As intended, this scheme tended to assign smaller weights to taxa in our benchmarks with more zero and near-zero counts (Figure 2-figure supplement 1). Despite their heuristic nature, we found that our chosen weights provide performance improvements over alternative weights (or the lack thereof) as measured by our benchmark tasks (Supplementary File 1).

Our taxa weighting scheme supplements the use of pseudo-counts and represents a soft-threshold on low abundance taxa. More generally, these taxa weights represent a form of prior information regarding the importance of each taxon. We note that if prior biological information suggests allowing specific taxa to influence the PhILR transform more (or less) strongly, such a weighting could be achieved for taxon j by increasing (or decreasing) $\mathrm{p_j}$.

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 weighting'. 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 $f(d_i^+, d_i^-) = 1$, the coordinates are not weighted by branch lengths. The form of this transform was chosen so that the weights $d_i^{\pm}$, only influence the corresponding coordinate ($y_i^{*,blw}$).

We also investigated the effect of using $f(d_i^+, d_i^-) = 1$, $f(d_i^+, d_i^-) = d_i^+ + d_i^-$, and based on the results of Chen et al. (42), $f(d_i^+, d_i^-) = \sqrt{d_i^+ + d_i^-}$ on benchmark performance. 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 tasks (Supplementary file 1). Based on these results, except for our analysis of balance variance versus phylogenetic depth (see below), the square root of the summed distances was used throughout our analyses.

Implementation

511 The PhILR transform, as well as the incorporation of branch length and taxa
512 weightings has been implemented in the R programing language as the package *philr*
513 available at https://bioconductor.org/packages/philr/.
514
515 ***Datasets and Preprocessing***
516 All data preprocessing was done in the R programming language using the
517 *phyloseq* package for analysis of microbiome census data (69) as well as the *ape* (70)
518 and *phangorn* (71) packages for analysis of phylogenetic trees.
519 *Data Acquisition*
520 We chose to use previously published OTU tables, taxonomic classifications, and
521 phylogenies as the starting point for our analyses. The Human Microbiome Project
522 (HMP) dataset was obtained from the QIIME Community Profiling Pipeline applied to
523 high-quality reads from the v3-5 region, available at http://hmpdacc.org/HMQCP/. The
524 Global Patterns dataset was originally published in Caporaso*, et al.* (1) and is provided
525 with the *phyloseq* R package (69). The Costello Skin Sites dataset (CSS) is a subset of
526 the dataset collected by Costello et al. (39) featuring only the samples from skin sites.
527 This skin subset was introduced as a benchmark for supervised machine learning by
528 Knights et al. (40) and can be obtained from http://knightslab.org/data.
529 *OTU Table Preprocessing*
530 To accord with general practice, we performed a minimal level of OTU table
531 filtering for all datasets used in benchmarks and analyses. Due to differences in
532 sequencing depth, sequencing methodology, and the number and diversity of samples
533 between datasets, filtering thresholds were set independently for each dataset. For the

HMP dataset, we initially removed samples with fewer than 1000 counts to mimic prior analyses (41). We additionally removed OTUs that were not seen with more than 3 counts in at least 1% of samples. Preprocessing of the Global Patterns OTU table followed the methods outlined in McMurdie and Holmes (69). Specifically, OTUs 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 OTUs with a coefficient of variation ≤ 3.0 were removed. The CSS dataset had lower sequencing depth than the other two datasets; we chose to filter OTUs that were not seen with greater than 10 counts across the skin samples. The PhILR transform, and more generally our benchmarking results in Figure 2b and 2C, were robust to varying our filtering strategies (Supplementary File 2).

Preprocessing Phylogenies

For each dataset, the phylogeny was pruned to include only those taxa remaining after OTU table preprocessing. Except for the Global Patterns dataset, which was already rooted, we chose to root phylogenies by manually specifying an outgroup. For the HMP dataset the phylum Euryarchaeota was chosen as an outgroup. For the CSS dataset, the tree was rooted with OTU 12871 (from phylum Plantomycetes) as the outgroup. For all three phylogenies, any multichotomies were resolved using 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.

Zero replacement and Normalization

555 A pseudocount of 1 was added prior to PhILR transformation to avoid taking log-
556 ratios with zero counts. We found that our benchmarking results were robust to
557 changing the value of this pseudocount from 1 to 2, 3, or 10 (Supplementary file 1).
558 *Grouping Sampling Sites*
559 To simplify subsequent analyses, HMP samples from the left and right
560 retroauricular crease and samples from the left and right antecubital fossa were
561 grouped together, respectively, as preliminary PERMANOVA analysis suggested that
562 these sites were indistinguishable (data not shown).
563
564 ***Benchmarking***
565 *Distance/Dissimilarity Based Analysis*
566 Distance between samples in PhILR transformed space was calculated using
567 Euclidean distance. All other distance measures were calculated using *phyloseq* on the
568 preprocessed data without adding a pseudocount. Principle coordinate analysis was
569 performed for visualization using *phyloseq.* PERMANOVA was performed using the
570 function *adonis* from the R package *vegan* (v2.3.4). The R^2 value from the fitted model
571 was taken as a performance metric. Standard errors were calculated using bootstrap
572 resampling with 100 samples each. Differences between the performance of Euclidean
573 distance in PhILR transformed space and that of each other distance or dissimilarity
574 measure on a given task was tested using two-sided t-tests and multiple hypothesis
575 testing was accounted for using FDR correction.
576 *Supervised Classification*

The performance, as measured by classification accuracy, 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'

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

Identifying balances that distinguish sites

To identify a sparse set of balances that distinguish sampling sites while accounting for the dependencies between nested balances, 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 body site separation (lambda=0.1198). This resulted in 35 balances with non-zero regression coefficients. Phylogenetic tree visualization was done using the R package *ggtree* (72).

Variance and Depth To reduce the likelihood that our analysis of balance variance and phylogenetic depth was affected by statistical artifact, we modified our PhILR transform in several ways. First, we omitted branch length weights (*i.e.,* we set $f(d_i^+, d_i^-) = 1$) as these may vary non-randomly as a function of phylogenetic depth. Second, we also anticipated that any zero replacement method would likely lead to lower variance measurements, which could have greater effects on balances closer to the tips of the tree. We therefore omitted taxa weights and zero replacement; we instead used stricter hard filtering thresholds and calculated balance values based on non-zero counts. In practice, we used the following filtering thresholds for each body site, taxa present in less than 20% of samples from that site were excluded and subsequently samples that had less than 50 total counts were excluded. To calculate balance values based on non-zero counts we retained balances that met the following

criteria: the term $g_p(\boldsymbol{y}_i^+)/g_p(\boldsymbol{y}_i^-)$ had non-zero counts from some taxa within the subcomposition $\boldsymbol{y}_i^+$ (formed by the taxa that descend from the $+1$ clade of node i) and some other taxa within the subcomposition $\boldsymbol{y}_i^-$ (formed by the taxa that descend from the -1 clade of node i) in at least 40 samples from that body site. We believe these two modifications to PhILR resulted in a more conservative analysis for of balance variance versus phylogenetic depth but are likely not optimal in other situations.

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 phylogenetic distance to its descendant tips (d). For a given body site the following model was fit:

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

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 permutations of the tip labels of the phylogenetic tree. For each permutation of the labels, the resultant tree was used to transform the data and β was estimated. We chose this permutation scheme to ensure that the increasing variance we saw with increasing proximity of a balance to 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, for each body site, we found the null distribution for β was symmetric about $\beta = 0$ which further supports that balance variance depends on phylogenetic depth through a biological mechanism and not

through a statistical artifact (Figure 4-figure supplement 3). 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.

The data and code needed to reproduce our analysis of balance variance versus phylogenetic depth is provided in Figure 4 – source data and Figure 4 – source code respectively.

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 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 were estimated by first decorating a phylogenetic tree with taxonomy information using *tax2tree* (v1.0) (73). 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.

The data and code needed to reproduce the taxonomic assignment and estimation of median phylogenetic depths for each taxonomic rank is included in Figure 4 – source data and Figure 4 – source code respectively.

ACKNOWLEDGEMENTS:

We thank Rachel Silverman, Aspen Reese, Firas Midani, Heather Durand, Jesse Shapiro, Jonathan Friedman, Simon Levin, and Susan Holmes for their helpful comments, Dan Knights for providing us with the CSS dataset, as well as Klaus Schliep and Liam Revell for their insight into manipulation of phylogenetic trees in the R programming language.

COMPETING INTERESTS:

The authors declare that they have no competing financial, professional, or personal interests that might have influenced the performance or presentation of the work described in this manuscript.

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, Müller W, Taylor J, Zwart A, & Helliwell C (2011) Proportions, percentages, ppm: do the molecular biosciences treat compositional data right. *Compositional Data Analysis: Theory and Applications*, eds Pawlowsky-Glahn V & Buccianti A (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.
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, Mateu-Figueras G, & Barcelo-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35(3):279-300.
24. Egozcue JJ & Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Mathematical Geology* 37(7):795-828.
25. 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.
26. Le Cao KA, *et al.* (2016) MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities. *PLoS One* 11(8):e0160169.
27. Morton JT, *et al.* (2017) Balance Trees Reveal Microbial Niche Differentiation. *mSystems* 2(1).
28. Hunt DE, *et al.* (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320(5879):1081-1085.
29. 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.
30. Wu GD, *et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334(6052):105-108.
31. Yatsunenko T, *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature* 486(7402):222-227.
32. 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.
33. McMurdie PJ & Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 10(4):e1003531.
34. Egozcue JJ & Pawlowsky-Glahn V (2016) Changing the Reference Measure in the Simplex and its Weightings Effects. *Austrian Journal of Statistics* 45(4):25-44.
35. Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71(12):8228-8235.
36. 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.
37. 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.
38. Martiny JB, Jones SE, Lennon JT, & Martiny AC (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* 350(6261):aac9323.
39. Costello EK, *et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science* 326(5960):1694-1697.

40. Knights D, Costello EK, & Knight R (2011) Supervised classification of human microbiota. *FEMS Microbiol Rev* 35(2):343-359.
41. Consortium HMP (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207-214.
42. Chen J*, et al.* (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28(16):2106-2113.
43. 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.
44. Matsen FA (2015) Phylogenetics and the human microbiome. *Syst Biol* 64(1):e26-41.
45. Moeller AH*, et al.* (2016) Cospeciation of gut microbiota with hominids. *Science* 353(6297):380-382.
46. Ley RE*, et al.* (2008) Evolution of mammals and their gut microbes. *Science* 320(5883):1647-1651.
47. Smillie CS*, et al.* (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241-244.
48. Rakoff-Nahoum S, Foster KR, & Comstock LE (2016) The evolution of cooperation within the gut microbiota. *Nature* 533(7602):255-259.
49. Blaser MJ & Falkow S (2009) What are the consequences of the disappearing human microbiota? *Nat Rev Microbiol* 7(12):887-894.
50. 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.
51. 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.
52. Grice EA & Segre JA (2011) The skin microbiome. *Nat Rev Microbiol* 9(4):244-253.
53. 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.
54. 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.
55. Zaneveld JR, Lozupone C, Gordon JI, & Knight R (2010) Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res* 38(12):3869-3879.
56. 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.
57. Faust K*, et al.* (2012) Microbial co-occurrence relationships in the Human Microbiome. *PLoS Computational Biology* 8(7).
58. Martın-Fernandez JA, Palarea-Albaladejo J, & Olea RA (2011) Dealing with zeros. *Compositional data analysis: Theory and applications*, eds Pawlowsky-Glahn V & Buccianti A), pp 43-58.

59. Bear J & Billheimer D (2016) A Logistic Normal Mixture Model for Compositional Data Allowing Essential Zeros. *Austrian Journal of Statistics* 45(4):3-23.
60. Martín-Fernández J-A, Hron K, Templ M, Filzmoser P, & Palarea-Albaladejo J (2015) Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling* 15(2):134-158.
61. Billheimer D, Guttorp P, & Fagan WF (2001) Statistical interpretation of species composition. *Journal of the American Statistical Association* 96(456):1205-1214.
62. Boogaart GVd & ebrary Inc. (2013) Analyzing compositional data with R. in *UseR!* (Springer,, New York), pp 1 online resource (xv, 258 p.
63. Pawlowsky-Glahn V & Buccianti A (2011) Compositional data analysis : theory and applications. (Wiley,, Chicester, West Sussex ; Hoboken, N.J.), pp xxi, 378 p.
64. Culley AI, Lang AS, & Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. *Science* 312(5781):1795-1798.
65. 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.
66. Yuan K, Sakoparnig T, Markowetz F, & Beerenwinkel N (2015) BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol* 16:36.
67. Roth A*, et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* 11(4):396-398.
68. Pawlowsky-Glahn V & Egozcue JJ (2011) Exploring Compositional Data with the CoDa-Dendogram. *Austrian Journal of Statistics* 40(1&2):103-113.
69. McMurdie PJ & Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8(4):e61217.
70. Paradis E, Claude J, & Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289-290.
71. Schliep KP (2011) phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4):592-593.
72. 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*.
73. 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.

Figure 1. PhILR uses an evolutionary tree to transform microbiota data into an unconstrained coordinate system. (**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 Cartesian. 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 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*.

Figure 2. Performance of standard statistical models on PhILR transformed microbiota data. Benchmarks were performed using three datasets: Costello Skin Sites (CSS), Global Patterns (GP), Human Microbiome Project (HMP) (a summary of these datasets after preprocessing is shown in Supplementary file 1 and Figure 2-figure supplement 1). (**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 Figure 2-figure supplement 2. (**B**) Sample distance (or dissimilarity) was computed by a range of statistics. PERMANOVA R^2 values, which represent how well sample identity explained the variability in sample pairwise distances, were used as a performance metric. 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. 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.

Figure 2-figure supplement 1. Taxa weighting scheme tends to assign smaller weights to taxa with more zero and near zero counts. The weight of a given taxon is calculated as the geometric mean of its counts across all samples times the Euclidean norm of its relative abundance across all samples in the dataset (*Methods*). Data are plotted on a log scale with a pseudocount of 1 added to ease visualization.

Figure 2-figure supplement 2. Principal coordinate analyses using different measures of community distance or dissimilarity.

Figure 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 Figure 3-figure supplement 1). (**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.

Figure 3-figure supplement 1. Balances found to distinguish human body sites by sparse logistic regression. Taxa are listed in balance *numerator/denominator* format.

Figure 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 (Figure 4-figure supplement 1). (**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; *Methods*). Significant trends are seen across all other body sites (Figure 4-figure supplement 2 and 3). 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 (**G**, **I**, **K**) reflect either unlinked or exclusionary dynamics between the geometric means of sister clades abundances.

Figure 4-figure supplement 1. Balances with high and low variance. Shown are the 10 highest and the 10 lowest variance balances for the Stool body site in the HMP dataset. Panels 1-10: Low balance variances reflect a linear relationship (through the origin) between levels of the two clades that descend from a given balance. Panels 11-20: High balances variances lack this linear relationship.

Figure 4-figure supplement 2. Neighboring clades covary less with increasing phylogenetic depth. Shown here are all HMP body sites. Dotted lines indicate median species boundaries (see *Methods*).

Figure 4-figure supplement 3. The null distribution for β. β is the slope in the linear regression between balance variance and phylogenetic depth in log space (regressions shown in Figure 4-figure supplement 2). To form null distributions for β, tip labels on the phylogeny were shuffled (n=20,000) and balance values re-calculated. Distributions for β are symmetric about 0 for each body site, suggesting

Figure 5. Sign matrix representation of a phylogenetic tree. A binary tree (**Left**) can be represented by a sign matrix (**Right)** denoted Θ.

SUPPLEMENTARY FILES

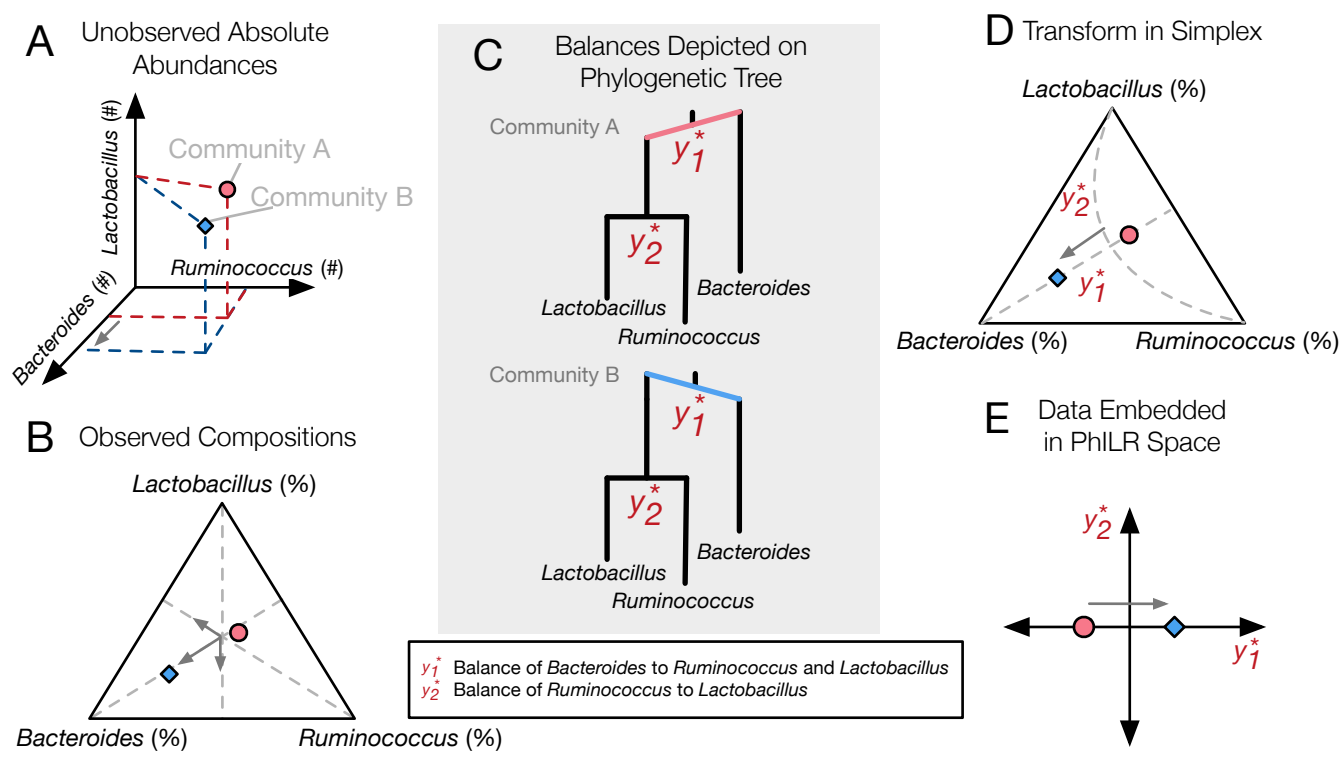
Supplementary file 1. Extended benchmarking results including the effects of changing pseudocount and different choices for taxa and branch length weights. This file also contains summary information for the benchmark datasets after preprocessing.

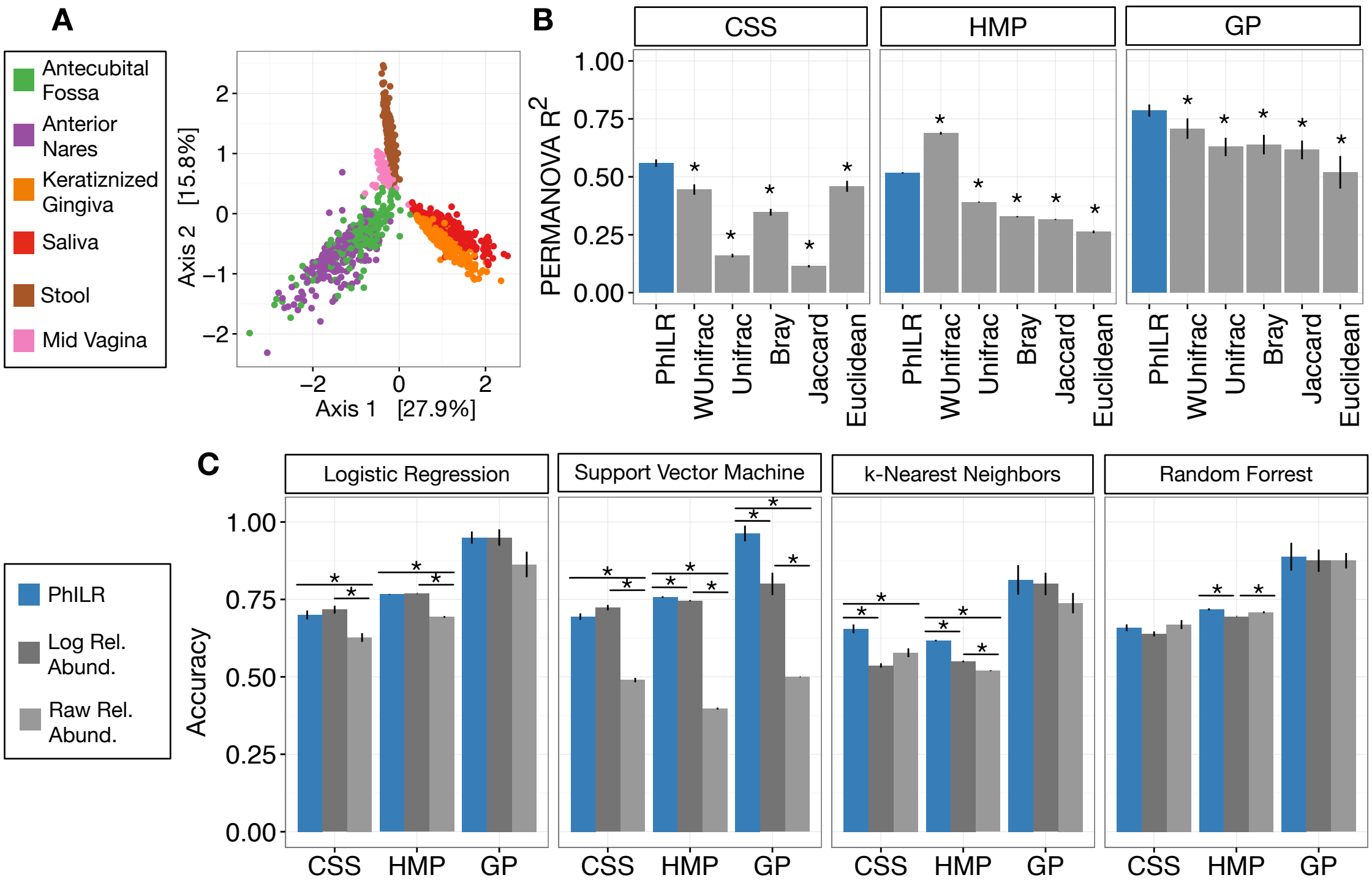
Supplementary file 2. Extended benchmarking results regarding the sensitivity of methods to different OTU filtering schemes.

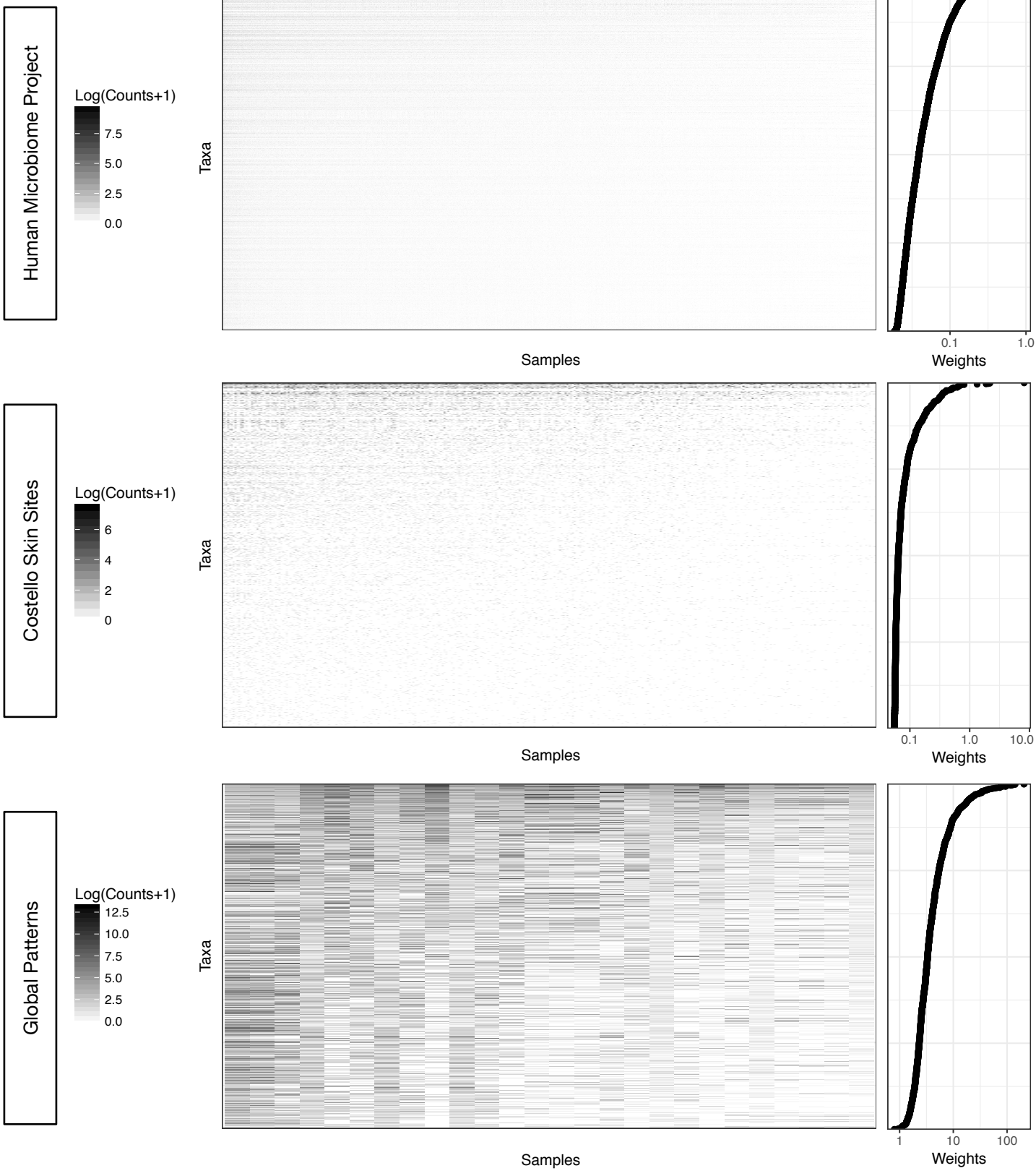
Figure 2 – source data. Source data for Figure 2b and 2c as well as FDR corrected p-values from tests.

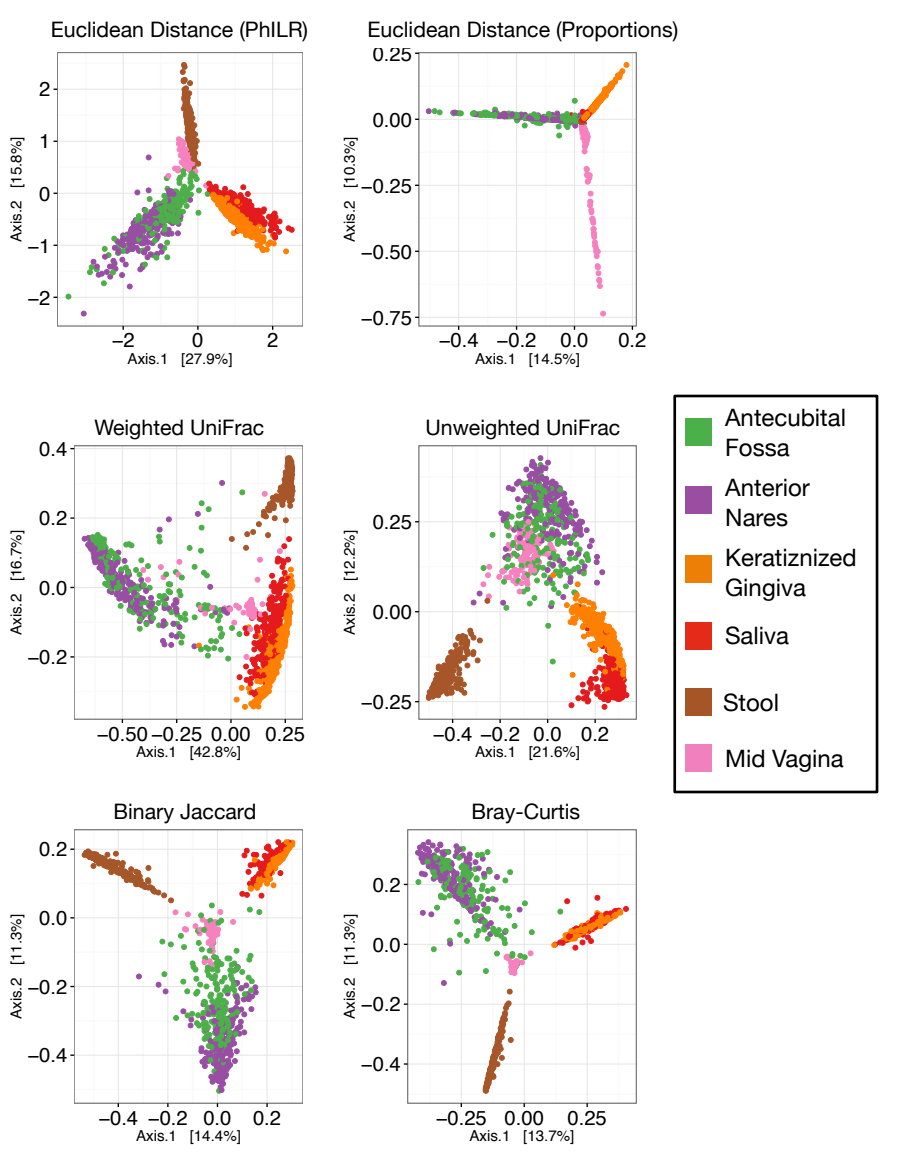
Figure 4 – source code. Source code for Figure 4 and associated supplements.

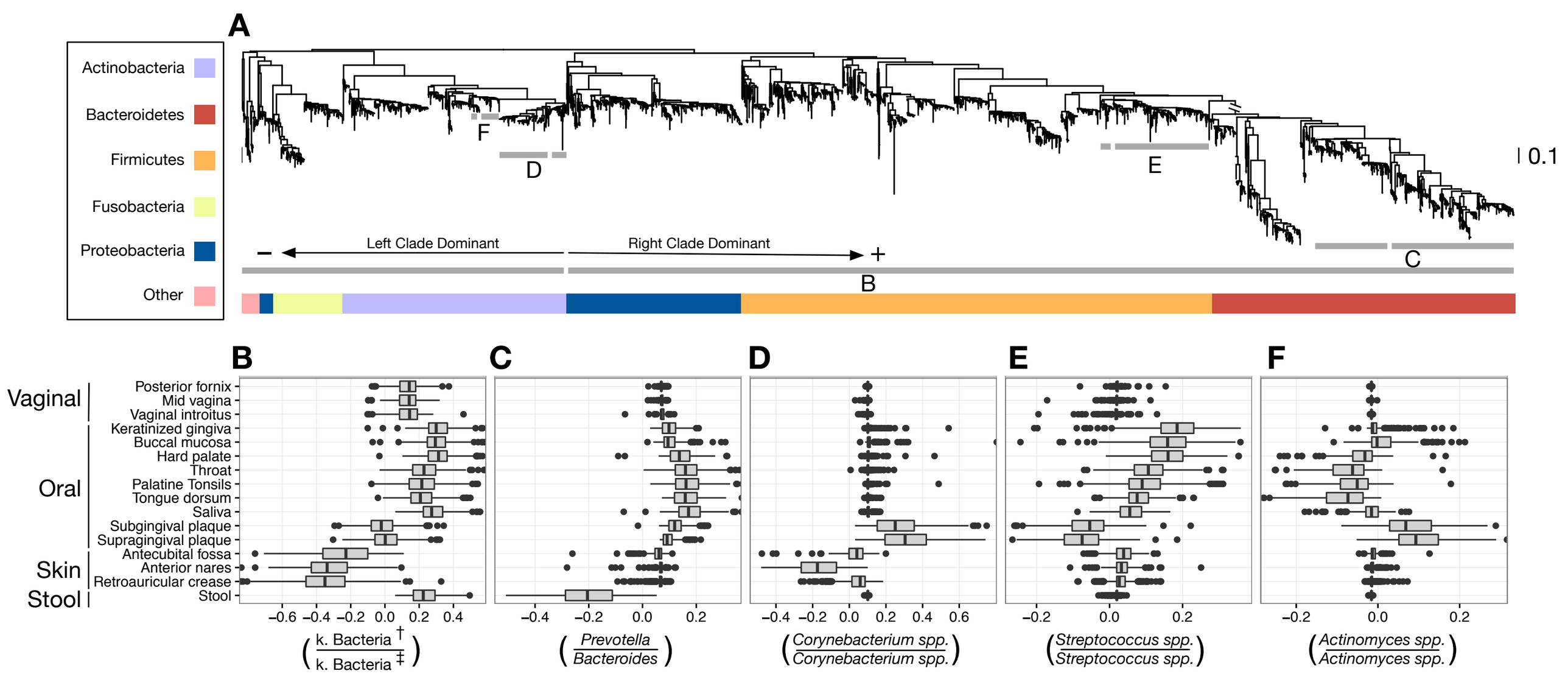
Figure 4 – source data. FDR corrected p-values from permutation tests.

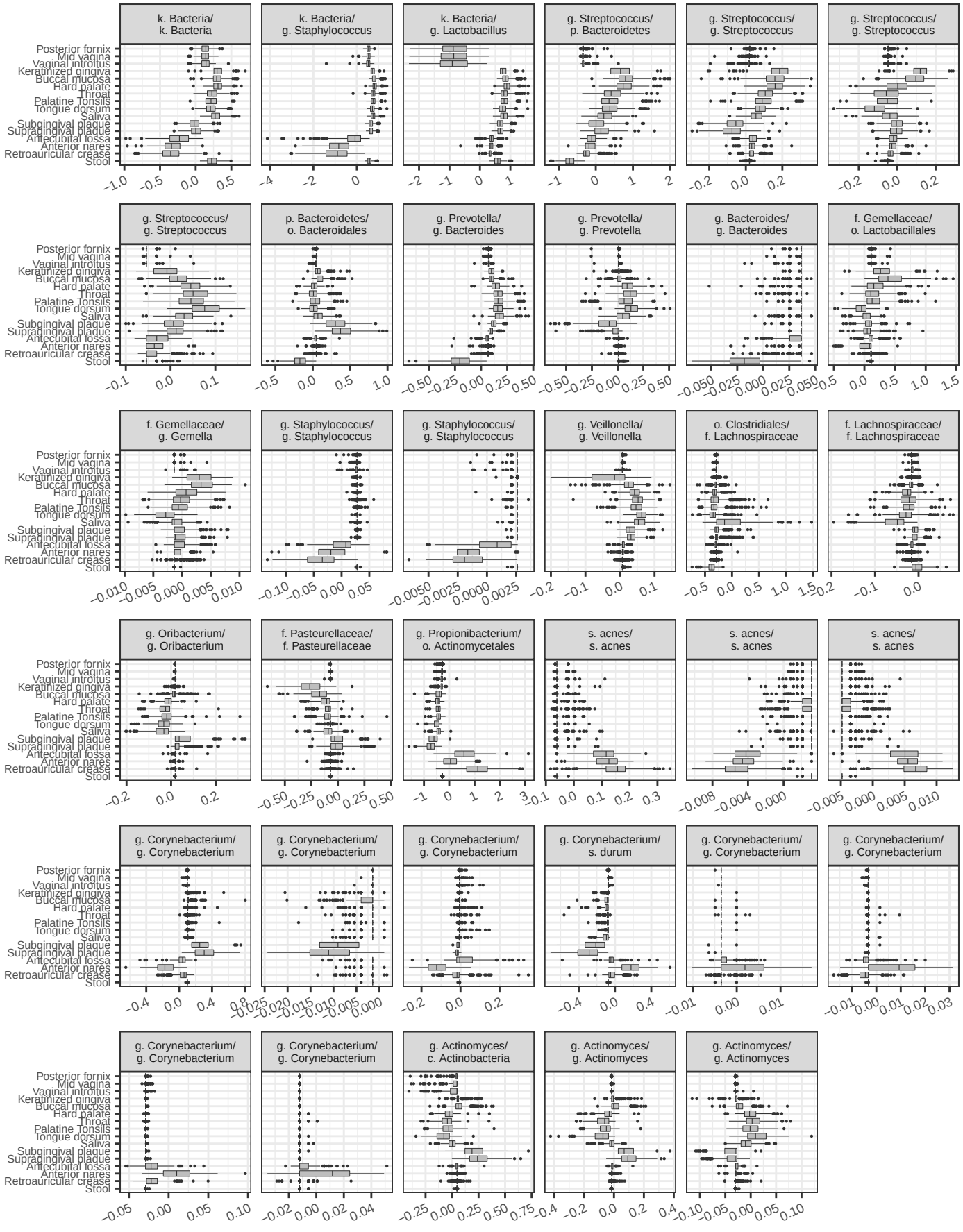

A
Unobserved Absolute Abundances
Lactobacillus (#)
Community A
Community B
Ruminococcus (#)
Bacteroides (#)
B
Observed Compositions
Lactobacillus (%)
Bacteroides (%)
Ruminococcus (%)
C
Balances Depicted on Phylogenetic Tree
Community A
y_1^*
y_2^*
Lactobacillus
Bacteroides
Ruminococcus
Community B
y_1^*
y_2^*
Lactobacillus
Bacteroides
Ruminococcus
y_1^* Balance of Bacteroides to Ruminococcus and Lactobacillus
y_2^* Balance of Ruminococcus to Lactobacillus
D
Transform in Simplex
Lactobacillus (%)
y_2^*
y_1^*
Bacteroides (%)
Ruminococcus (%)
E
Data Embedded in PhILR Space
y_2^*
y_1^*

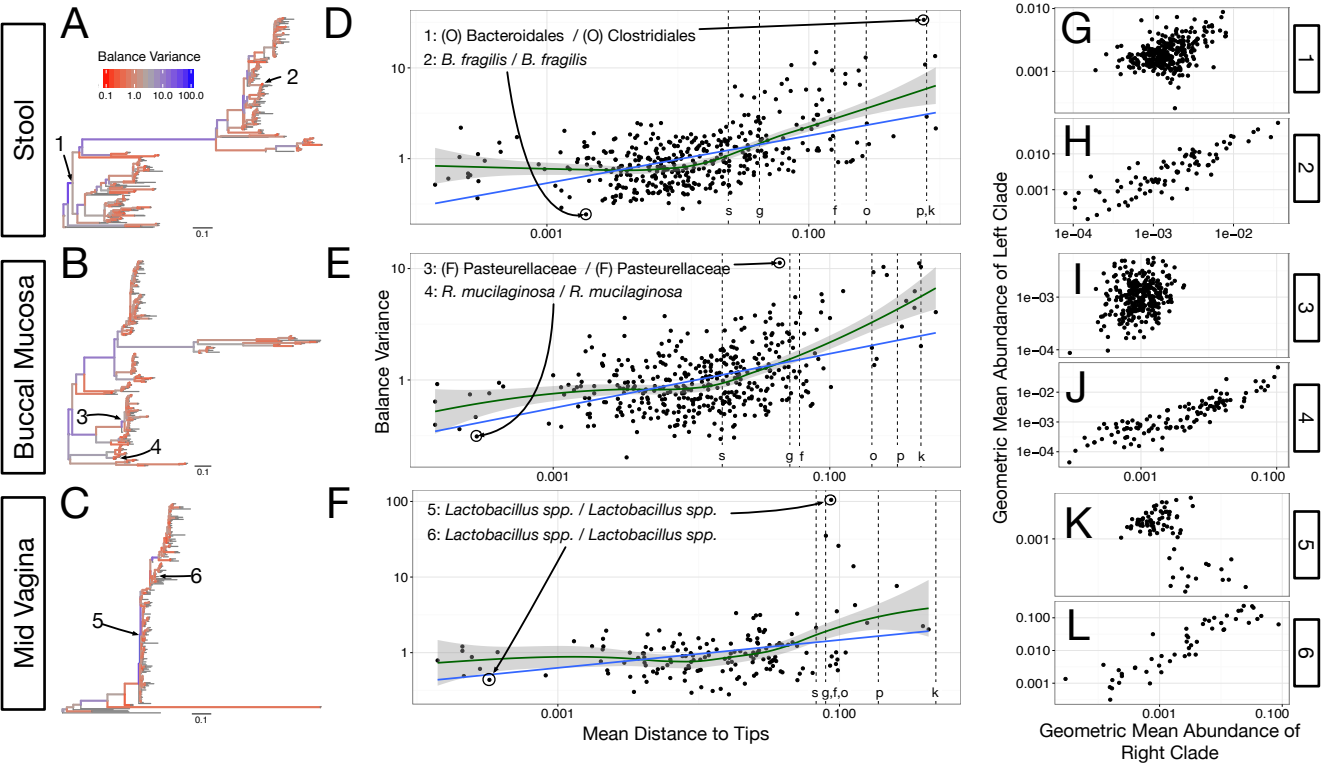

A
Antecubital Fossa
Anterior Nares
Keratiznized Gingiva
Saliva
Stool
Mid Vagina
Axis 2 [15.8%]
Axis 1 [27.9%]
B
CSS
HMP
GP
PERMANOVA R2
1.00
0.75
0.50
0.25
0.00
PhILR
WUnifrac
Unifrac
Bray
Jaccard
Euclidean
C
Logistic Regression
Support Vector Machine
k-Nearest Neighbors
Random Forrest
Accuracy
PhILR
Log Rel. Abund.
Raw Rel. Abund.
CSS
HMP
GP

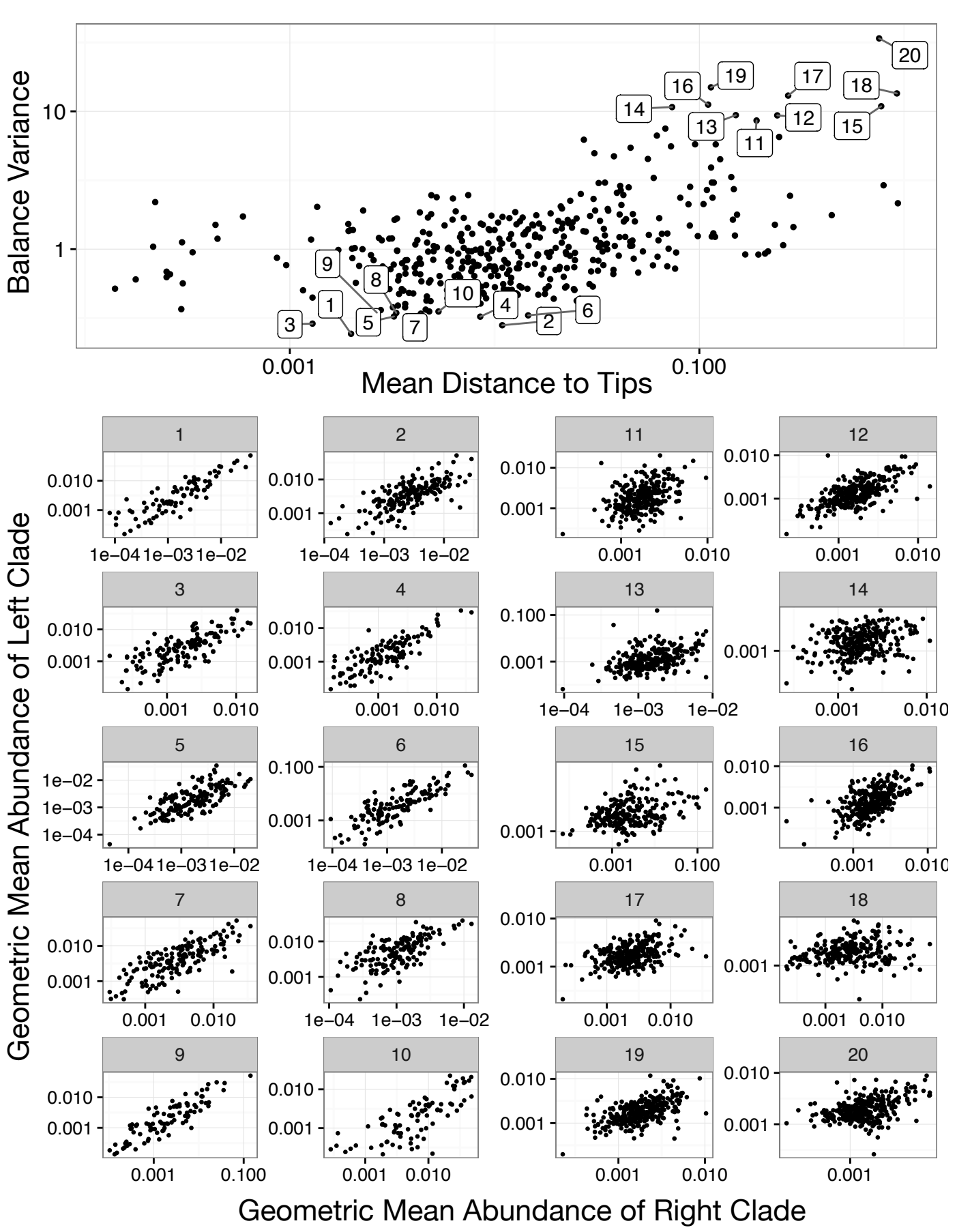

Human Microbiome Project
Log(Counts+1)
7.5
5.0
2.5
0.0
Taxa
Samples
0.1
1.0
Weights
Costello Skin Sites
Log(Counts+1)
6
4
2
0
Taxa
Samples
0.1
1.0
10.0
Weights
Global Patterns
Log(Counts+1)
12.5
10.0
7.5
5.0
2.5
0.0
Taxa
Samples
1
10
100
Weights

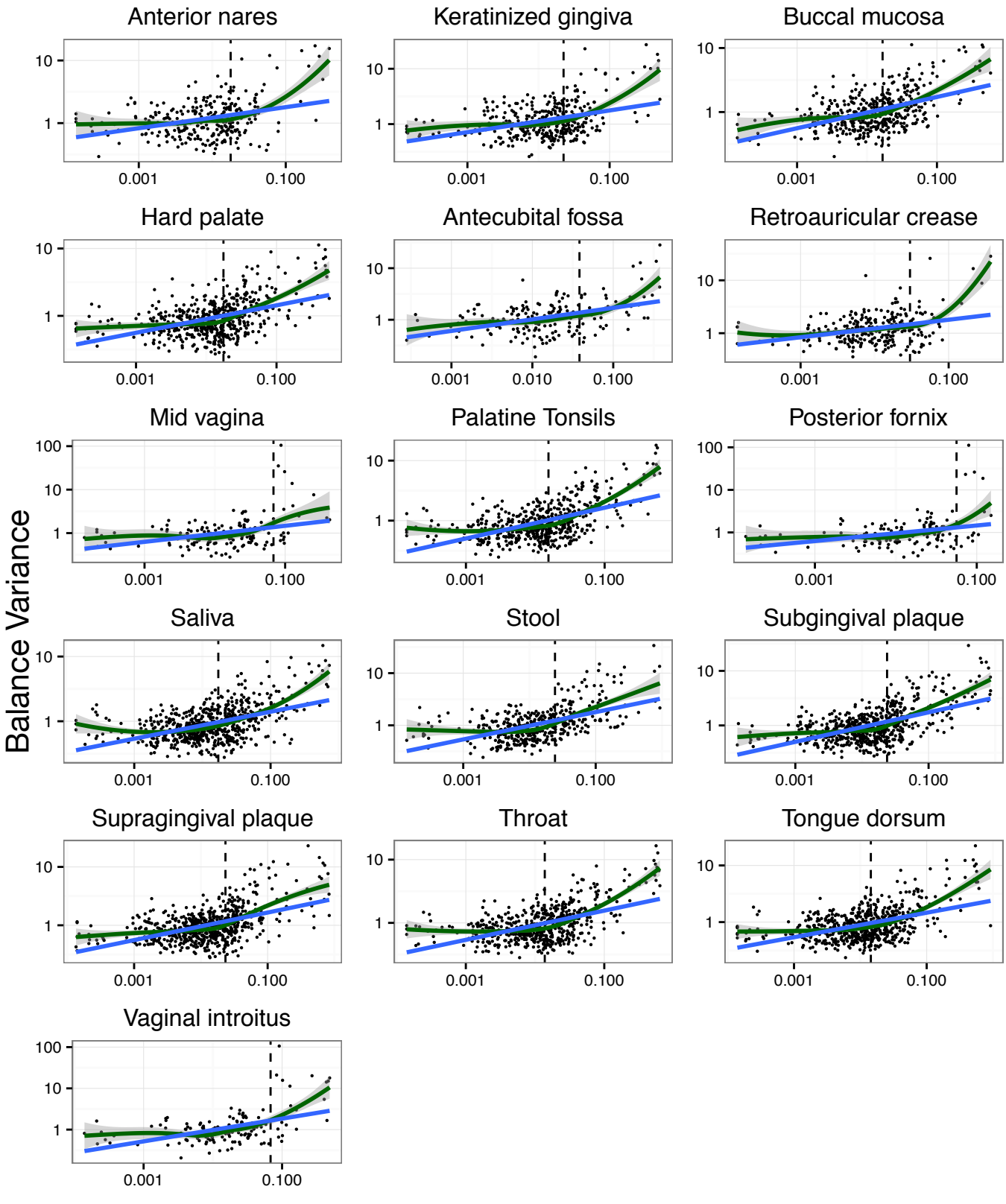

Euclidean Distance (PhILR)
Axis.2 [15.8%]
Axis.1 [27.9%]
Euclidean Distance (Proportions)
Axis.2 [10.3%]
Axis.1 [14.5%]
Weighted UniFrac
Axis.2 [16.7%]
Axis.1 [42.8%]
Unweighted UniFrac
Axis.2 [12.2%]
Axis.1 [21.6%]
Binary Jaccard
Axis.2 [11.3%]
Axis.1 [14.4%]
Bray-Curtis
Axis.2 [11.3%]
Axis.1 [13.7%]
Antecubital Fossa
Anterior Nares
Keratiznized Gingiva
Saliva
Stool
Mid Vagina

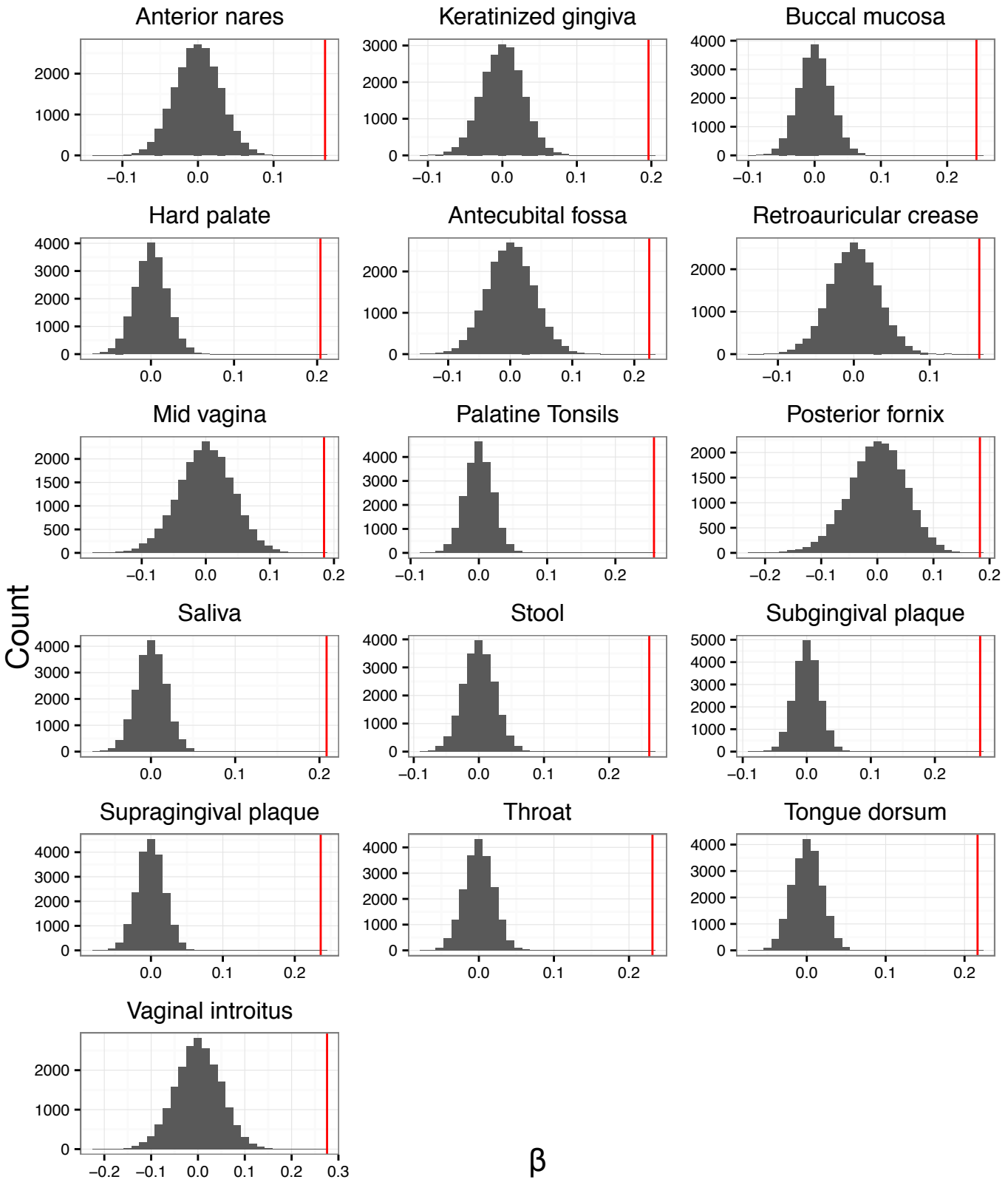

A
Actinobacteria
Bacteroidetes
Firmicutes
Fusobacteria
Proteobacteria
Other
F
D
E
C
B
0.1
–
Left Clade Dominant
Right Clade Dominant
+
B
C
D
E
F
Vaginal
Oral
Skin
Stool
Posterior fornix
Mid vagina
Vaginal introitus
Keratinized gingiva
Buccal mucosa
Hard palate
Throat
Palatine Tonsils
Tongue dorsum
Saliva
Subgingival plaque
Supragingival plaque
Antecubital fossa
Anterior nares
Retroauricular crease
Stool
−0.6 −0.4 −0.2 0.0 0.2 0.4
(k. Bacteria † / k. Bacteria ‡)
−0.4 −0.2 0.0 0.2
(Prevotella / Bacteroides)
−0.4 −0.2 0.0 0.2 0.4 0.6
(Corynebacterium spp. / Corynebacterium spp.)
−0.2 0.0 0.2
(Streptococcus spp. / Streptococcus spp.)
−0.2 0.0 0.2
(Actinomyces spp. / Actinomyces spp.)

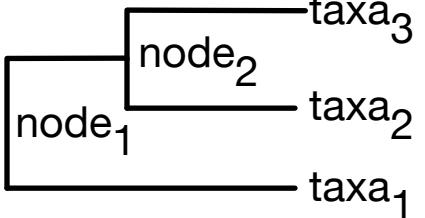

k. Bacteria/
k. Bacteria
k. Bacteria/
g. Staphylococcus
k. Bacteria/
g. Lactobacillus
g. Streptococcus/
p. Bacteroidetes
g. Streptococcus/
g. Streptococcus
g. Streptococcus/
g. Streptococcus
g. Streptococcus/
g. Streptococcus
p. Bacteroidetes/
o. Bacteroidales
g. Prevotella/
g. Bacteroides
g. Prevotella/
g. Prevotella
g. Bacteroides/
g. Bacteroides
f. Gemellaceae/
o. Lactobacillales
f. Gemellaceae/
g. Gemella
g. Staphylococcus/
g. Staphylococcus
g. Staphylococcus/
g. Staphylococcus
g. Veillonella/
g. Veillonella
o. Clostridiales/
f. Lachnospiraceae
f. Lachnospiraceae/
f. Lachnospiraceae
g. Oribacterium/
g. Oribacterium
f. Pasteurellaceae/
f. Pasteurellaceae
g. Propionibacterium/
o. Actinomycetales
s. acnes/
s. acnes
s. acnes/
s. acnes
s. acnes/
s. acnes
g. Corynebacterium/
g. Corynebacterium
g. Corynebacterium/
g. Corynebacterium
g. Corynebacterium/
g. Corynebacterium
g. Corynebacterium/
s. durum
g. Corynebacterium/
g. Corynebacterium
g. Corynebacterium/
g. Corynebacterium
g. Corynebacterium/
g. Corynebacterium
g. Corynebacterium/
g. Corynebacterium
g. Actinomyces/
c. Actinobacteria
g. Actinomyces/
g. Actinomyces
g. Actinomyces/
g. Actinomyces
Posterior fornix
Mid vagina
Vaginal introitus
Keratinized gingiva
Buccal mucosa
Hard palate
Throat
Palatine Tonsils
Tongue dorsum
Saliva
Subgingival plaque
Supragingival plaque
Antecubital fossa
Anterior nares
Retroauricular crease
Stool
Balance Statistic


Stool
Buccal Mucosa
Mid Vagina
A
B
C
Balance Variance
0.1 1.0 10.0 100.0
1
2
3
4
5
6
0.1
D
E
F
1: (O) Bacteroidales / (O) Clostridiales
2: B. fragilis / B. fragilis
3: (F) Pasteurellaceae / (F) Pasteurellaceae
4: R. mucilaginosa / R. mucilaginosa
5: Lactobacillus spp. / Lactobacillus spp.
6: Lactobacillus spp. / Lactobacillus spp.
Balance Variance
Mean Distance to Tips
0.001
0.100
s g f o p,k
s g f o p k
s g,f,o p k
G
H
I
J
K
L
1
2
3
4
5
6
Geometric Mean Abundance of Left Clade
Geometric Mean Abundance of Right Clade
1e−04
1e−03
1e−02
0.010
0.001


Balance Variance
10
1
0.001
0.100
Mean Distance to Tips
20
19
17
16
18
14
12
13
15
11
9
8
10
1
4
6
3
5
7
2
Geometric Mean Abundance of Left Clade
Geometric Mean Abundance of Right Clade
1
2
11
12
3
4
13
14
5
6
15
16
7
8
17
18
9
10
19
20


Anterior nares
Keratinized gingiva
Buccal mucosa
Hard palate
Antecubital fossa
Retroauricular crease
Mid vagina
Palatine Tonsils
Posterior fornix
Saliva
Stool
Subgingival plaque
Supragingival plaque
Throat
Tongue dorsum
Vaginal introitus
Balance Variance
Mean Distance to Tips
0.001
0.010
0.100
1
10
100


Anterior nares
Keratinized gingiva
Buccal mucosa
Hard palate
Antecubital fossa
Retroauricular crease
Mid vagina
Palatine Tonsils
Posterior fornix
Saliva
Stool
Subgingival plaque
Supragingival plaque
Throat
Tongue dorsum
Vaginal introitus
Count
β



	taxa$_1$	taxa$_2$	taxa$_3$
node$_1$	+1	-1	-1
node$_2$	0	+1	-1