A Phylogenetic Approach to Overcoming Compositional Problems in Microbiome Data

Justin D. Silverman[1,2,7], Alex D. Washburne[3,4], Sayan Mukherjee[1,6], Lawrence A. David[1,5,7]

[1] Program in Computational Biology and Bioinformatics, Duke University, [2] Medical Scientist Training Program, Duke University, [3] Nicholas School of the Environment, Duke University, [4] Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, [5] Department of Molecular Genetics and Microbiology, Duke University, [6] Department of Statistical Science, Duke University, [7] Center for Genomic and Computational Biology, Duke University,

1. 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. The PhILR transform[1] incorporates microbial evolutionary models with the isometric log-ratio transform[2] to allow off-the-shelf statistical tools to be safely applied to microbiota surveys. Here we describe our methodology (**Fig. 1**) and demonstrate its use by identifying neighboring bacterial clades that distinguish human body sites (**Fig. 2**). The PhILR transform is implemented in the R programing language as the package *philr* available on Bioconductor.

2. Motivation

Compositional Effects

DNA sequencing techniques cause total sequencing read counts to be arbitrary (*i.e.* not reflective of biological signal, such as bacterial load).[3] As a consequence, the resulting count data conveys information regarding the relative amounts of different bacteria only. Such relative data is often termed compositional. The use of most standard statistical tools (*e.g.,* correlation, regression, or classification) within a compositional space leads to spurious results.[3-5]

Phylogenetic Structure

Closely related bacteria may share traits and thus respond in similar ways to a subtle perturbation, like a shift in diet. Despite increasing appreciation that bacterial traits are strongly predicted by evolutionary distance[6], phylogenetic structure often remains unaccounted for in microbiome analyses.

Data Transformation

Rather than adapting statistical models to work with compositional data, we have chosen a data transformation approach which allows a wide variety of standard statistical models to be employed without modification. Specifically we have chosen the Isometric Logratio (ILR) transform[2] as it, along with other logratio transforms, satisfies scale invariance, permutation invariance, and subcompositional coherence. However, unlike the more well known additive log-ratio (ALR) and centered logratio (CLR) transforms, the ILR transform is an isometry from $\boldsymbol{S}^D \leftrightarrow \boldsymbol{R}^{D-1}$ and corresponds to an orthonormal basis in the simplex making it a good choice for both statistical modeling and distance based analysis.

3. Phylogenetic Partitioning for the ILR Transform

A key challenge in adapting the ILR transform to human microbiome data is ensuring that data in the resulting space remain interpretable. To address this challenge, we note that the ILR transform can be build from a sequential binary partition (SBP) of the original data space.[7] The PhILR transform uses the bacterial phylogenetic tree as an SBP to define an ILR transform for the microbiome. This gives each PhILR coordinate the interpretation as the balance between the relative abundance of the two clades of taxa that descend from the corresponding internal node of the phylogenetic tree (**Fig. 1C**).

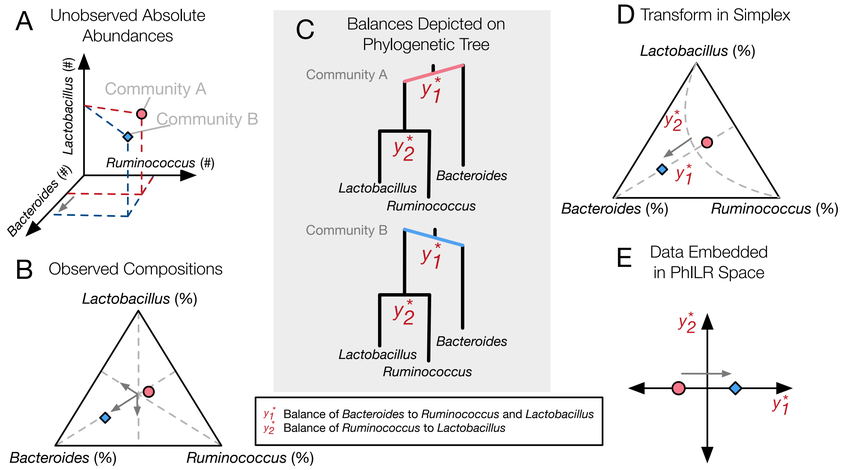

Fig. 1: **PhILR uses an evolutionary tree to transform microbiota data into an unconstrained coordinate system.** (**A**) Two hypothetical bacterial communities share identical absolute numbers of *Lactobacillus*, and *Ruminococcus* bacteria; they differ only in the absolute abundance of *Bacteroides* 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*.

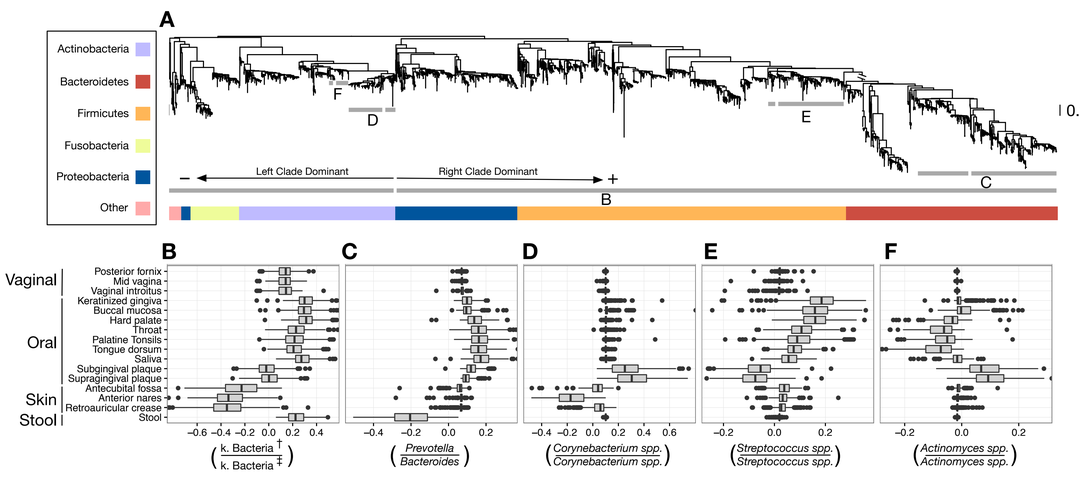

Fig. 2: **Balances distinguishing human microbiota by body site**. Sparse logistic regression was used to identify balances that best separated the different sampling sites in the Human Microbiome Project dataset.[8] (**A**) Each balance is represented on the tree as a broken grey bar. The left portion 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.

4. The Unweighted PhILR Transform

For a microbiome sample represented as a vector of relative abundance measurements $\boldsymbol{y} = (y_1, \ldots, y_D)$, we denote the PhILR transformed data as $\boldsymbol{y}^* = (y_1^*, \ldots, y_{D-1}^*)$. We represent the PhILR coordinate (balance) associated with internal node i (where $i \in \{1, \ldots, D-1\}$) of the phylogenetic tree as

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

Here, $g_p(\boldsymbol{y}_i^+)$ and $g_p(\boldsymbol{y}_i^-)$ represents the geometric mean relative abundance of the taxa that are part of either of the two clades that descend from internal node i. The terms n_i^+ and n_i^- represent the number of taxa that descend from each of the two clades descendant from internal node i.

References

1. Silverman, J. D., Washburne, A. D., Mukherjee, S., David, L. A., *eLife*. 2017.
2. Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barceló-Vidal, C., *Math. Geol.* 2003, 35.
3. Tsilimigras, M. C., Fodor, A. A., *Ann. Epidemiol.* 2016.
4. Friedman, J. & Alm, E. J., *PLoS Comput. Biol.* 2012, 8.
5. Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R., *Modeling and Analysis of Compositional Data*. Wiley. 2015.
6. Martiny, A. J. B. H., Jones, S. E., Lennon, J. T. & Adam, C., *Science*. 2015.
7. Egozcue, J. J., Pawlowsky-Glahn, V., *Math. Geol.* 2005, 37.
8. HMP Consortium, *Nature*. 2012, 486.