EDGE PRINCIPAL COMPONENTS AND SQUASH CLUSTERING: USING THE SPECIAL STRUCTURE OF PHYLOGENETIC PLACEMENT DATA FOR SAMPLE COMPARISON

FREDERICK A. MATSEN AND STEVEN N. EVANS

Abstract. It is becoming increasingly common to analyze collections of sequence reads by first assigning each read to a location on a phylogenetic tree. In parallel, quantitative methods are being developed to compare samples of reads using the information provided by such phylogenetic placements: one example is the *phylogenetic Kantorovich-Rubinstein (KR) metric* which calculates a distance between pairs of samples using the evolutionary distances between the assigned positions of the reads on the phylogenetic tree. The KR distance generalizes the *weighted UniFrac* metric. Classical, general-purpose ordination and clustering methods can be applied to KR distances, but we argue that more interesting and interpretable results are produced by two new methods that leverage the special structure of phylogenetic placement data. *Edge principal components analysis* enables the detection of important differences between samples containing closely related taxa and allows the visualization of the principal component axes in terms of edges of the phylogenetic tree. *Squash clustering* produces informative internal edge lengths for clustering trees by incorporating distances between averages of samples, rather than the averages of distances between samples used in general-purpose procedures such as UPGMA. We present these methods and illustrate their use with data from the microbiome of the human vagina.

1. Introduction

Microbial sequence data for a given locus are naturally endowed with a somewhat hidden special structure: the phylogenetic relationships between the organisms represented by each sequence. That structure can be inferred by building the phylogenetic tree of the sampled sequences from scratch or by using *phylogenetic placement* techniques to assign to each sampled sequence a location on a previously constructed reference phylogenetic tree.

In 2005, Lozupone and Knight proposed a way to incorporate this hierarchical structure when computing distances between samples. Their method, *unweighted UniFrac* [5], was followed by *weighted UniFrac* in 2007 [6]. A key feature of both distances is that differences in community structure due to closely related organisms are weighted less heavily than differences arising from distantly related organisms. The UniFrac methodology has been widely adopted, and the papers describing the UniFrac variants have hundreds of citations as of the beginning of 2011.

Once distances have been computed between samples using UniFrac, these distances are typically fed into general-purpose ordination and clustering methods,

Date: March 7, 2011.

such as classical principal components analysis and UPGMA. Although it is appropriate to apply such techniques to distance matrices of this sort, the classical methods do not use the fact that the underlying distances were calculated in a specific manner. Consequently, in an application of principal components analysis, it is difficult to describe what the axes represent. Similarly, in hierarchical clustering, it is unclear what is driving a certain agglomeration step; although it can be explained in terms of an arithmetic operation, a certain amount of interpretability in terms of the original microbial data is lost.

In this paper, we propose ordination and clustering procedures specifically designed for the comparison of microbial sequence samples that do take advantage of the underlying phylogenetic structure. The input for these methods are collections of assignments of sequencing reads to locations on a pre-existing reference phylogenetic tree: that is, phylogenetic placements. Our *edge principal components analysis* (edge PCA) algorithm applies the standard principal components construction to a "data matrix" generated from the differences between proportions of phylogenetic placements on either side of each internal edge of the reference phylogenetic tree. Our *squash clustering* algorithm is hierarchical clustering with a novel definition of distances between clusters that incorporates information concerning how the data sit on a phylogenetic tree.

The primary advantage of these methods is that of transparency — namely, that the results of these analyses can be readily visualized and understood. For example, with edge PCA the principal component axes can be pictured directly in terms of the reference phylogenetic tree, thereby attaching a clearer interpretation to the position of a data point along that axis. Edge PCA is also capable of picking up minor — but consistent — differences in collections of placements between samples: a feature that will be important in our example application. The squash hierarchical clustering method has the advantage that each vertex of the clustering tree induces a certain natural distribution of mass on the phylogenetic tree and the length of an edge in the clustering tree has a simple interpretation as the distance between the mass distributions associated with the two incident vertices.

2. Results

2.1. **General setting for methods.** Phylogenetic placement is a way to analyze second-generation sequencing applied to DNA extracted in bulk from an environmental sample of microbes. It is simply the assignment of sequencing reads to a "reference" phylogenetic tree constructed from previously-characterized DNA sequences; recent algorithms have focused on doing so according to the phylogenetic maximum-likelihood criterion [1, 7]. By fixing a reference tree rather than attempting to build a phylogenetic tree for the sample from scratch, recent algorithms of this type are able to place tens of thousands of query sequences per hour per processor on a reference tree of one thousand taxa (e.g. species), with performance scaling linearly in the number of reference taxa, the number of query sequences, and the length of the query sequences.

A collection of reads placed on a phylogenetic tree can be thought of as a distribution of a unit amount of mass across the tree. In the simplest setting, for a collection of N placements on a tree each read is given mass $1/N$ and that mass is assigned to the most likely position for that read on the tree. Another option is to

distribute the $1/N$ mass for a given read across the tree in proportion the posterior probability of assignment of that read to various positions [7].

This mass distribution can then be used to produce distances between collections of phylogenetic placements. Given two samples for a given locus, each sample is placed individually on the phylogenetic tree, and then each sample is thought of as a distribution of mass on the tree. The Kantorovich-Rubinstein or "earth-mover's" distance can then be used to quantify the difference between those two samples. This distance is defined rigorously in [4], but the idea is simply explained. Imagine that the phylogenetic tree is a road network, and each sample is represented by the distribution of a unit of mass into piles of dirt along this road network. The distance between two samples is then defined to be the minimal amount of "work" required to move the dirt in the first configuration to that in the second configuration (in this context the amount of work to move an infinitesimal mass δ a distance x is defined to be $\delta \cdot x$). Thus, similar collections of phylogenetic placements result in similar dirt pile configurations that don't require much mass movement from one to the other, while quite different collections of placements require that significant amounts of mass must move long distances across the tree. This distance is classical, having roots in 18th century mathematics, and is a generalization of the *weighted UniFrac* distance [4, 6]. One can perform PCA and clustering on distance matrices derived from these distances, or, as presented next, develop algorithms that work directly on the underlying data.

2.2. **Edge principal components analysis.** Suppose that each of S samples is encoded by a mass distribution on a reference tree with E internal edges. We distinguish an arbitrary vertex of the tree as the *root* and map each mass distribution to an E-dimensional vector by recording for each internal edge the difference between the total mass on the root side of the edge and the total mass on the non-root side of the edge. This results in an $S \times E$ "data matrix".

Edge principal components analysis (edge PCA) applies the usual principal components procedure of constructing the $E \times E$ covariance matrix of this data matrix and then calculating its eigenvalues and their corresponding eigenvectors.

Each eigenvector can be displayed on the tree, because the coordinates of the eigenvector correspond to internal edges of the tree. A large entry in an eigenvector corresponding to one of the bigger eigenvalues identifies an edge for which there is substantial heterogeneity among the associated set of mass differences (see Methods). In our visualization tool, each eigenvector is represented by a single colored and thickened reference tree: the thickness of an edge is proportional to the magnitude of the corresponding entry of the eigenvector and the color specifies the sign of that entry (Figures 1 and 2). Moreover, we can project each sample onto an eigenvector to visualize how the sample is spread out with respect to that "axis" (Figure 3 (a))

2.3. **Squash clustering.** Squash clustering is hierarchical clustering with a novel way of calculating distances between clusters. Rather than taking averages of distances as is done in average-linkage clustering (also known as UPGMA), in squash clustering we take distances between averages of samples. That is, given a collection of mass distributions on the phylogenetic tree that each correspond to a cluster that has been built at some stage of the procedure, when the procedure merges two

clusters we simply take a weighted average of the two corresponding mass distributions to get the mass distribution that corresponds to the new, larger cluster (see Methods). The "squash" terminology describes this averaging procedure: the original mass distributions for a given cluster are stacked on top of one another and then "squashed" down to produce a new object with unit total mass. That is, if we merge two clusters that correspond to sets of m and n original mass distributions and represented by averaged mass distributions μ and ν, then the new cluster is represented by the mass distribution

$$\frac{m}{m+n}\mu + \frac{n}{m+n}\nu.$$

Equivalently, if the points in the two clusters were originally represented by the mass distributions $\mu_1, \ldots, \mu_m$ and $\nu_1, \ldots, \nu_n$, respectively, then the two clusters are now represented by the mass distributions

$$\mu = \frac{\mu_1 + \cdots + \mu_m}{m}$$

and

$$\nu = \frac{\nu_1 + \cdots + \nu_n}{n},$$

respectively, and the new cluster obtained by merging them is represented by the mass distribution

$$\frac{(\mu_1 + \cdots + \mu_m) + (\nu_1 + \cdots + \nu_n)}{m+n}.$$

Apart from that difference, the sequence of steps in the algorithm is identical to that which would occur in the usual agglomerative hierarchical clustering procedure applied with the Kantorovich-Rubinstein distances between the initial mass distributions (i.e. those representing the individual samples) as input. Each step of agglomerative hierarchical clustering is associated with a pairwise distance matrix and the algorithm proceeds by merging the pair of clusters that have the smallest distance between them. As described above, we take the distance between two clusters A and B to be the earth-mover's distance from the average of the mass distributions in A to the corresponding average for those in B. The series of merges in the clustering algorithm determines the topology of the rooted *clustering tree* that the algorithm produces. Leaves of the tree correspond to individual samples. Every internal vertex is associated with a cluster (the collection of samples below that vertex) and thus with a distribution of mass on the phylogenetic tree. The length an edge between two arbitrary adjacent vertices on the tree can be computed by using the earth-movers distance between the distributions of mass corresponding to those vertices. This edge length calculation gives the resulting trees an appearance that differs from that of UPGMA trees (Figure 4).

2.4. **Example application: the vaginal microbiome.** In this section we apply our clustering and ordination methods to pyrosequencing data from the vaginal microbiome. For this study, swabs were taken from 242 women from the Public Health, Seattle and King County Sexually Transmitted Diseases Clinic (STD clinic) between September 2006 and June 2010 (of which 222 samples resulted in enough material to analyze) [12]. DNA was extracted and the 16s gene was amplified in the V3-V4 hypervariable region using broad-range primers and sequenced using a 454 sequencer with FLX chemistry. Sequences were pre-processed using the R / Bioconductor package *microbiome*. A custom maximum likelihood reference tree consisting

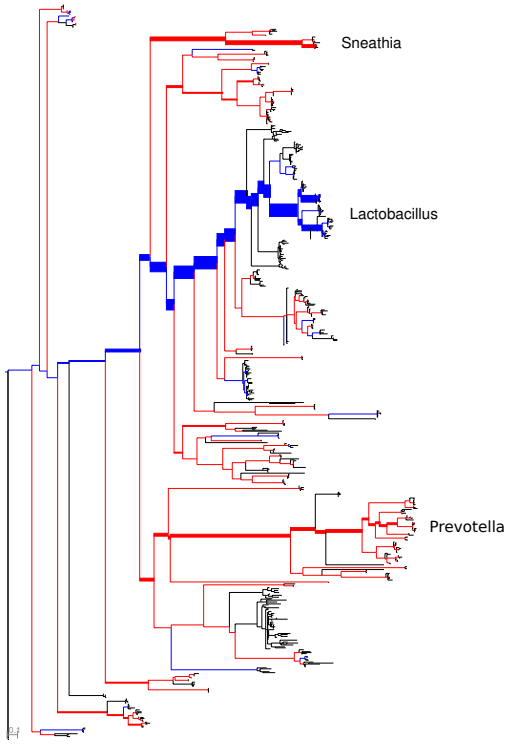


FIGURE 1. The first principal component for the vaginal data set, with eigenvalue ≈ 11.57. The edges across which maximal between-sample heterogeneity is found are those leading to the *Lactobacillus* clade and those leading to the Sneathia and Prevotella clade. This axis appears to represent the bacterial vaginosis axis, as *Sneathia* and *Prevotella* are associated with bacterial vaginosis, while *Lactobacillus* is associated with its absence.

of sequences from RDP [3] and our local collection was built using *RAxML* 7.2.7 [13] using GTR+4Γ. Sequences were placed into this tree using pplacer [7] with the default parameter choices. The methods presented here are implemented as part of the pplacer package, available at `http://matsen.fhcrc.org/pplacer/`.

The principal components for the vaginal samples independently recover previous knowledge about the contribution of certain microbial species to distinct types of vaginal microbial environment. In a medical setting, the diagnosis of bacterial vaginosis is often done by looking for rod-shaped bacteria that have a positive Gram stain; these are typically *Lactobacillus*. The edge principal component algorithm indicates the importance of this genus: the first principal component for the vaginal data set picks out the presence of *Lactobacillus* versus *Sneathia* and *Prevotella* (Figure 1). The second principal component reveals that important differences between samples exist at the species level. Indeed, it highlights the substantial amount of heterogeneity between the amount of two *Lactobacillus* species observed: *L. iners* and *L. crispatus* (Figure 2).

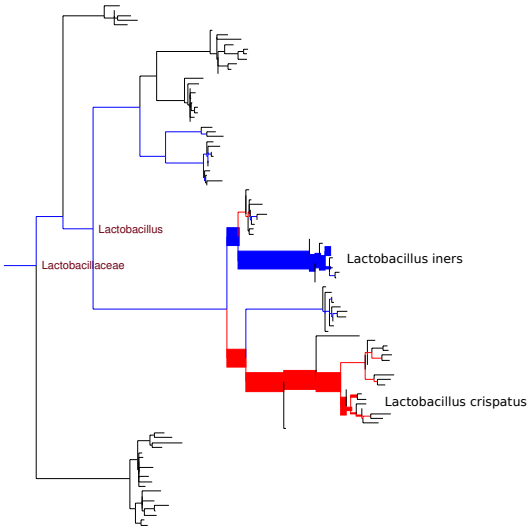
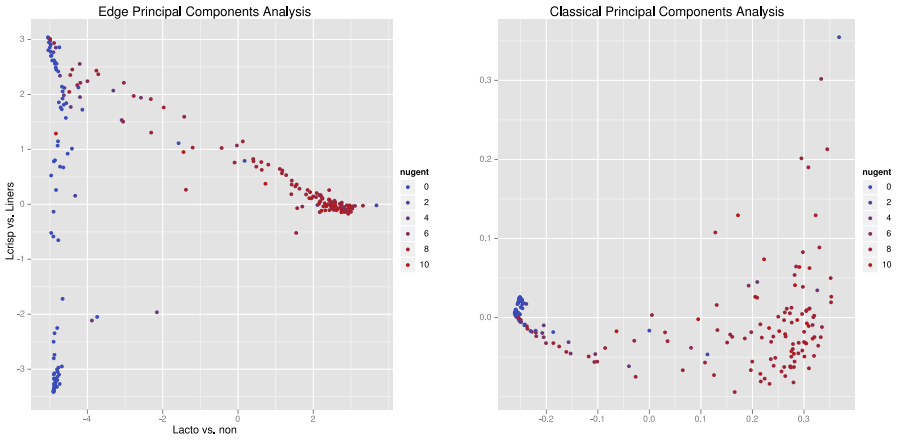


FIGURE 2. The second principal component for the vaginal data set, with eigenvalue ≈ 3.17 (low-weight regions of the tree excluded from the figure). The edges across which maximal between-sample heterogeneity is found are those between two different *Lactobacillus* clades: *L. iners* and *L. crispatus*. Thus, the second important "axis" appears to correspond to the presence or absence of these two species.

The samples form an interesting pattern when plotted on these axes with their diagnostic score (Figure 3 (a)). As described above, samples on the left side have *Lactobacillus* and lack *Sneathia* and *Prevotella*, while those on the right side have the opposite. Samples on the bottom have lots of *L. crispatus* and a small amount of *L. iners*, while those on the top have the opposite. A continuum of samples exists from the lower left to the upper left (mixes of the two prominent *Lactobacillus* species) and from the upper left to the right (from *L. iners*-dominant to *Sneathia* and *Prevotella*), but there is no continuum from lower left to the right (from *L. crispatus*-dominant to *Sneathia* and *Prevotella*). Reviewing the data from [12] confirms that this pattern can be observed from taxonomically classified reads.

The features detected by the edge PCA algorithm correspond to ones that are pertinent to clinical laboratory diagnosis. As part of the bacterial vaginosis study, these samples were also processed according to traditional diagnostic criteria. Vaginal samples were classified in the clinical laboratory according to the Nugent score, which quantifies the presence of various morphotypes (i.e. shapes of bacteria) under a microscope after gram staining. The Nugent score is considered to be a rigorous standard for the diagnosis of bacterial vaginosis (BV); it ranges from 0 (healthy) to 10 (severe BV). Swabs were also evaluated for pH.

In general, *Lactobacillus* is associated with a low Nugent score and thus a negative BV diagnosis. More specifically, *L. crispatus* dominated samples are not found to have a high Nugent score (indicating BV), while *L. iners* sometimes are, and



(a) Edge principal components (b) Classical principal components

FIGURE 3. A comparison of edge principal components analysis (edge PCA) and classical PCA on the vaginal data set. The axes for the edge principal components plot are described in Figures 1 (x-axis) and 2 (y-axis). The Nugent score is a diagnostic score for bacterial vaginosis, with high score indicating bacterial vaginosis.

those on the right side always are. Coloring the samples according to pH (Figure S2) shows a similar pattern. These plots indicate the possibility of a medically relevant difference between these two *Lactobacillus* species. We emphasize that the PCA was **not** informed of the Nugent score, pH, or the taxonomic classifications.

Applying classical PCA to the pairwise distance matrix does reproduce some of the same features (Figure 3 (b)). However, there is no immediate interpretation of the meaning of the axes output by the classical algorithm. Furthermore, the important difference between the two *Lactobacillus* species is lost.

Squash clustering was applied to the collection of vaginal samples in our cohort. Because meaningful internal edge lengths can be assigned to the squash clustering tree, it is not ultrametric, whereas the UPGMA tree is (Figure 4). The two tight clusters at the bottom of (a) and (b) contain the *Lactobacillus*-dominated vaginal samples seen on the left side of (Figure 3 (a)) and correspond to *L. iners* (upper tight cluster) and *L. crispatus* (lower tight cluster). A more detailed leaf-labeled comparison between the two trees is available in the supplementary material (Figure S1).

Although we argue that these clusters do have more meaningful internal edge lengths than the UPGMA tree, the clusters found by squash clustering for this example data set do not have a significantly different pH *phenotypic consistency* than those found by UPGMA. For a subtree c of the clustering tree, define S_c to be its leaf set; for any real-valued phenotypic character of a sample we define the phenotypic consistency of c to be the variance of the phenotypic character for the samples in S_c. To investigate how phenotypic consistency corresponds to statistically supported subtrees of the cluster, the data was bootstrapped at the level of reads 100 times, such that every bootstrap replicate for a given sample was

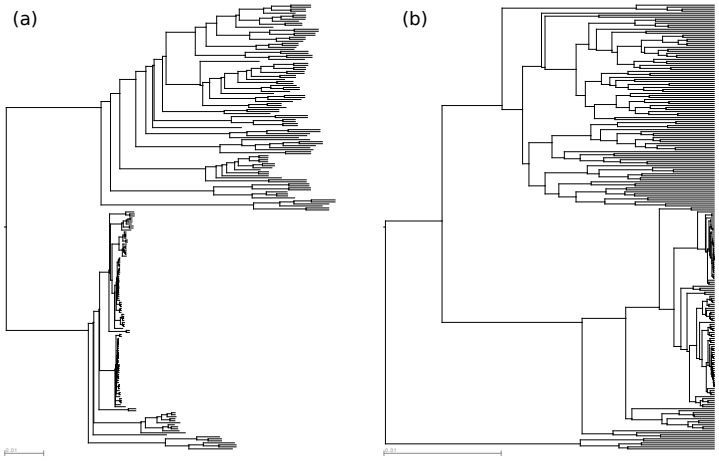


FIGURE 4. The results of (a) squash clustering and (b) UPGMA as applied to the vaginal data set. The trees are unlabeled, and th labels do not appear in the same order on the two trees; for a comparison of labeled trees see Figure S1.

a uniform sample of the reads for that sample with replacement. Then, for every subtree c, the bootstrap support and the pH phenotypic consistency for the samples in that subtree can be represented by a point in a scatterplot (Figure 5). If one method found clusters with better pH phenotypic consistency than the other, then one method's points would appear lower on average than the other.

3. Discussion

3.1. **Generalization and limitations.** The methods described here, although implemented for comparison of microbial communities, can in fact be used in more general settings. Edge PCA can be used whenever each sample can be represented by a collection of mass distributed over a common tree structure. Squash clustering can be applied in any case where there is a well-defined notion of the distance between two samples and a well-defined procedure for averaging two samples to produce another object of the same type.

There are some limitations to the sort of comparisons that can be performed using these methods simply because the underlying data is a collection of phylogenetic placements on a tree. For example, if a clade of the reference tree is missing, then differences in diversity within that clade will not be accounted for in the comparison. Such issues will be present whenever a reference tree is being used, whether using phylogenetic placements directly or mapping reads to the tree using BLAST as a preliminary step in a UniFrac analysis. This disadvantage can be balanced with the advantage of not having to define OTU's, which can be sensitive to clustering parameters [14].

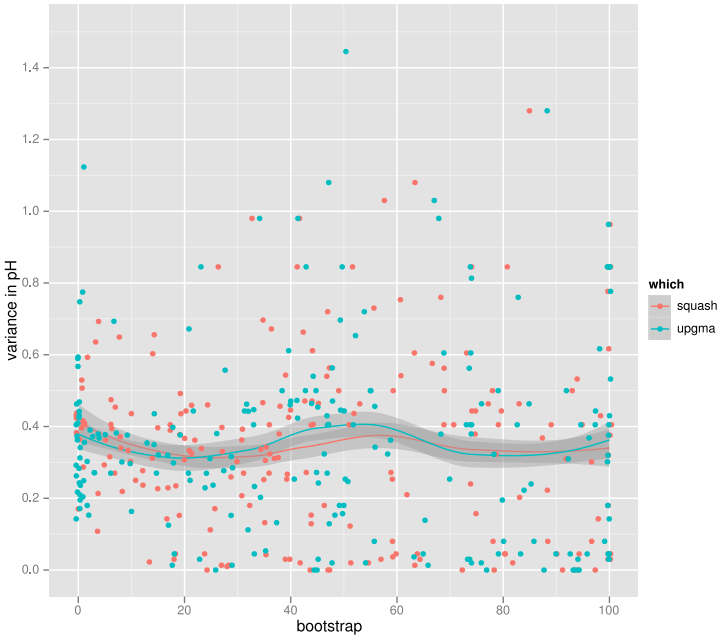


FIGURE 5. Squash clustering does not appear to have significantly better correlation with phenotype than UPGMA. Each point represents a subtree in the cluster, where the x-axis shows the bootstrap value (100 replicates) for that subtree, and the y axis shows the variance in pH for all of the samples in that cluster. Lines show smoothing performed with a generalized additive model using splines.

The methods presented here also depend on the number of phylogenetic placements being correlated with the number of organisms of that type found in the sample. This is not always true. Loci such as 16s are often sequenced by first amplifying using a polymerase chain reaction with a broad-spectrum primer; this primer may have different efficiencies for different organisms, or may miss certain organisms altogether. In addition, genetic material extraction efficiency varies by organism [9]. Nevertheless, the results that come from using our methods do appear to correspond with non-genetic methods such as morphological comparison (Figure 3) and pH (Figure S2).

3.2. **Relation to previous work.** The work presented here has a similar intent as double principal components (DPCoA) analysis as applied to distributions of phylotypes on a phylogenetic tree [2, 11]. The idea of a DPCoA analysis is to perform a principal components analysis on the phylotype abundance table in a way that down-weights differences between species that are close to one another on the phylogenetic tree. As such, it shares some similarity to doing multidimensional scaling or principal components on the pairwise distance matrix generated by a KR/UniFrac analysis. It differs from the methods presented here because it only

uses the tree in the form of a pairwise distance matrix; consequently it cannot leverage the edge-by-edge structure of the tree as is done here.

There are also some connections with the statistical comparison features of MEGAN [8] in that we use the structure of the tree as part of a comparative framework. Our method and the MEGAN method both highlight regions of the tree for which important differences exist between samples. However, the details are quite different: the "directed homogeneity" test of [8] employs statistical criteria to decide if two samples have significant differences either across an edge or between daughters of a given edge. The edge PCA algorithm, on the other hand, does not attempt to make hypothesis-testing statistical statement and it uses all of the samples available simultaneously.

3.3. **Future work.** The basic step of the methods presented here— transforming phylogenetic placement samples into vectors indexed by the edges of the tree— is general and can be applied in a number of contexts. In this paper, we followed this transformation with an application of principal components analysis, but many other options are possible. We plan to apply classical supervised learning techniques to similarly transformed data.

4. Methods

A probability measure on the reference phylogenetic tree is obtained from a collection of sequence reads as follows. A given read can be assigned to the phylogenetic tree in its maximum likelihood or maximum posterior probability location using the phylogenetic likelihood criterion to obtain a "point placement." A point placement can be thought of as a probability measure with all of the mass concentrated at the best attachment location. Alternatively, one can express uncertainty in the optimal location by spreading the probability mass according to posterior probability (assuming some priors) or "likelihood weight ratio"; see [7] for details. In either case, each read is thought of as a probability measure on the reference phylogenetic tree. A probability measure for a collection of reads can be obtained by averaging the measures for each read individually (that is, by constructing the probability measure that is the convex combination or mixture of the probability measures for each read in which each such measure is given an equal weight).

4.1. **Edge principal components analysis.** Begin with a phylogenetic tree T and probability measures $P_1, \ldots, P_S$ on T, each of which comes from an assignment of the reads in one of S samples to the phylogenetic tree, as described above. If T is not already rooted at some vertex, pick an arbitrary vertex to be the root. Removing a given internal edge e from the tree splits T into two components: $T_+(e)$ containing the root and $T_-(e)$ without. For a probability measure P on T, define the corresponding *edge mass difference*

$$\delta_P(e) = P(T_+(e)) - P(T_-(e)).$$

Suppose that T has E internal edges. The *edge mass difference matrix* Δ is the $S \times E$ matrix that has the vectors of edge mass differences for the successive samples as its rows.

Edge principal components analysis is then performed by first deriving the $E \times E$ covariance matrix Σ from the matrix Δ of "observations" followed by computing the E eigenvectors of Σ ordered by decreasing size of eigenvalue.

Each resulting eigenvector is a signed weighting on the internal edges of the tree, and these weightings may be used to highlight those edges of the tree for which there is substantial between-sample heterogeneity in the masses assigned to the two components of the tree defined by the edge. Indeed, recall the variational characterization of the eigenvectors $v_1, \ldots, v_E$ of an $E \times E$ non-negative definite matrix M listed in order of decreasing eigenvalue:

$$\begin{aligned} v_1 &= \arg\max_{||v||=1} \langle v, Mv \rangle \\ v_2 &= \arg\max_{||v||=1, v \perp v_1} \langle v, Mv \rangle \\ &\ldots \\ v_E &= \arg\max_{||v||=1, v \perp \{v_1, \ldots, v_{E-1}\}} \langle v, Mv \rangle, \end{aligned}$$

where $\|v\|$ is the usual Euclidean length of the vector v, $\langle v, w \rangle$ is the usual Euclidean inner product of the vectors v and w, and $v \perp \{v_1, \ldots, v_k\}$ indicates that v is perpendicular to each of the vectors $v_1, \ldots, v_k$. Thus, an edge that receives a weight with large magnitude from an eigenvector corresponding to one of the bigger eigenvalues of the covariance matrix Σ may be viewed as an edge for which there are substantial dissimilarities between samples in the amount of mass placed in the components on either side of the edge.

When looking at the weight assigned to a single edge in isolation, only the magnitude of the weight matters and not the sign, because if v is an eigenvector for a particular eigenvalue, then so is $-v$. However, sign matters when comparing the weights assigned to two or more edges: if the edge mass differences for two edges are negatively associated, then the corresponding entry of the covariance matrix will be negative, and the corresponding two entries of the eigenvector for a large eigenvalue will tend to have different signs.

Changing the chosen root from vertex x to vertex y does not affect the eigenvalues or the magnitudes of the entries in the corresponding eigenvectors, and it only changes the signs of the entries for the edges between x and y. We may see this as follows. Note first that if an edge e is between x and y, then re-rooting flips the sign of $\delta_P(e)$. Define K be the diagonal $E \times E$ matrix such that $K_{e,e} = -1$ for edges e on the path between x and y, and 1 otherwise. Note that $K = K^{-1}$. The covariance matrix Σ' for the re-rooted tree and that for the original tree will then be related by a similarity transformation: $\Sigma' = K\Sigma K$. Thus, the eigenvalues for Σ are the same as those for Σ', and v_k is an eigenvector of Σ if and only if Kv_k is an eigenvector of Σ'.

4.2. **Squash clustering.** Given probability measures P and Q on the rooted tree T, the Zolatarev-like Z_p generalization of the KR distance is defined for $p \geq 1$ as

$$Z_p(P, Q) = \left[\int_T |P(\tau(y)) - Q(\tau(y))|^p \ \lambda(dy) \right]^{\frac{1}{p}}, \tag{1}$$

where λ is the natural length measure on the tree and $\tau(y)$ is the subtree on the other side of y to the root. The classical KR distance is (1) with $p = 1$; for the examples in this paper, $p = 1$ was used exclusively. It is shown in [4] that choosing a different root does not change the distance. It is also noted in [4] that if P and

Q only assign mass to leaves of the tree and y is in the interior of edge e then

$$|P(\tau(y)) - Q(\tau(y))| = \frac{1}{2}\left(|P(T_+(e)) - Q(T_+(e))| + |P(T_-(e)) - Q(T_-(e))|\right)$$

furnishing a connection with edge PCA.

At each stage of the squash clustering algorithm we have a pairwise distance matrix with rows and columns indexed by the clusters that have already been made by the algorithm. Initially, the clusters are just the individual samples and the entries in the pairwise distance matrix are computed using the formula (1).

Agglomerative hierarchical clustering in general proceeds by iterating the following sequence of steps until there is a single cluster and a corresponding 1×1 pairwise distance matrix.

(1) Find the smallest off-diagonal element in the current pairwise distance matrix. Say it is the distance between clusters i and j.
(2) Merge the i and j clusters, making a cluster k.
(3) Remove the ith and jth rows and columns from the distance matrix.
(4) Calculate the distance from the cluster k to the other clusters.
(5) Insert the distances from k into the distance matrix.

Classical hierarchical clustering methods calculate the distance in step number 4 as some function of the distance matrix. In particular, average-linkage clustering or UPGMA calculates the distance between two clusters as the average between pairs of items in the clusters. Thus, if clusters i and j containing respective numbers of items a and b are merged to form a cluster k with $a+b$ items, then the average-linkage distance between another cluster ℓ with c items and the new cluster k is (writing $d(\cdot,\cdot)$ for the distance between individual items)

$$\begin{aligned}
\text{distance}(k,\ell) &= \frac{1}{(a+b)c}\sum_{y\in k, z\in \ell} d(y,z) \\
&= \frac{a}{a+b}\frac{1}{ac}\sum_{w\in i, z\in \ell} d(w,z) + \frac{b}{a+b}\frac{1}{bc}\sum_{x\in j, z\in \ell} d(x,z) \\
&= \frac{a}{a+b}\text{distance}(i,\ell) + \frac{b}{a+b}\text{distance}(j,\ell),
\end{aligned}$$

and so the entries of the updated UPGMA distance matrix are just suitably weighted averages of the entries of the previous distance matrix.

At each stage of squash clustering, on the other hand, a cluster is associated with a probability measure on the tree T. When two clusters i and j containing respective numbers of items a and b and associated with respective probability measures P and Q are merged to form a cluster k, then the new cluster k is associated with the probability measure $\frac{a}{a+b}P + \frac{b}{a+b}Q$ and the distance from k to some other cluster ℓ associated with the probability measure R is

$$Z_p\left(\frac{a}{a+b}P + \frac{b}{a+b}Q, R\right) \tag{2}$$

which is analogous to the above formula for the UPGMA averaging procedure. As remarked above, the "squash" interpretation of (2) comes from recalling that the probability measures associated with the two clusters are each simple averages of all of the measures for the items in the clusters. That is, if S_z is the probability

measure associated with original item z, then

$$P = \frac{1}{a} \sum_{x \in i} S_x$$

and

$$Q = \frac{1}{b} \sum_{y \in j} S_y,$$

and the probability measure associated with the new cluster k is

$$\frac{a}{a+b} P + \frac{b}{a+b} Q = \frac{1}{a+b} \sum_{z \in k} S_z.$$

A natural question to ask is whether the distance between a probability measure R and the weighted average of two probability measures P and Q is equal to the similarly weighted average of the distance between R and P and the distance between R and Q. The answer is "no": starting from (1) we have from the Minkowski inequality that for $0 < t < 1$:

$$\begin{aligned}
&Z_p(tP + (1-t)Q, R) \\
&\quad = \left[\int_T \left| t \left(P(\tau(y)) - R(\tau(y)) \right) + (1-r) \left(Q(\tau(y)) - R(\tau(y)) \right) \right|^p \lambda(dy) \right]^{\frac{1}{p}} \\
&\quad \leq t \left[\int_T \left| P(\tau(y)) - R(\tau(y)) \right|^p \lambda(dy) \right]^{\frac{1}{p}} + (1-t) \left[\int_T \left| Q(\tau(y)) - R(\tau(y)) \right|^p \lambda(dy) \right]^{\frac{1}{p}} \\
&\quad = t Z_p(P, R) + (1-t) Z_p(Q, R),
\end{aligned}$$

and the two sides are the same if and only if $P = Q$. Intuitively, then, the UPGMA and squash clustering steps are very similar for the initial steps when the objects being clustered are themselves similar. This can be seen in Figure S1, where regions of the tree representing closely related samples (short edge length) show good agreement between the two methods (thin edge width).

5. Acknowledgments

This work would not have been possible without an ongoing collaboration with David Fredricks, Noah Hoffman, Martin Morgan, and Sujatha Srinivasan at the Fred Hutchinson Cancer Research Center. The authors are grateful for early feedback on this work from Jonathan Eisen and Aaron Darling. Graphics made with ggplot2 [15] and archaeopteryx (http://www.phylosoft.org/archaeopteryx/).

References

[1] Simon A Berger and Alexandros Stamatakis. Evolutionary Placement of Short Sequence Reads Performance & Accuracy of Evolutionary Placement Algorithms for Short Sequence Reads under maximum-likelihood Simon A. Berger, Alexandros Stamatakis. *Systematic Biology*, pages 1–58, 2011.

[2] Elisabeth M Bik, Paul B Eckburg, Steven R Gill, Karen E Nelson, Elizabeth a Purdom, Fritz Francois, Guillermo Perez-Perez, Martin J Blaser, and David a Relman. Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3):732–7, January 2006.

[3] J R Cole, Q Wang, E Cardenas, J Fish, B Chai, R J Farris, A S Kulam-Syed-Mohideen, D M McGarrell, T Marsh, G M Garrity, and J M Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(Database issue):D141–D145, 2009.

[4] S.N. Evans and F.A. Matsen. The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *Arxiv preprint arXiv:1005.1699*, 2010.

[5] Catherine Lozupone and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228, 2005.

[6] Catherine a Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, 73(5):1576–85, March 2007.

[7] Frederick A Matsen, Robin B Kodner, and E Virginia Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1):538, October 2010.

[8] Suparna Mitra, Bernhard Klar, and Daniel H Huson. Visual and statistical comparison of metagenomes. *Bioinformatics*, 25(15):1849–55, August 2009.

[9] Jenna L. Morgan, Aaron E. Darling, and Jonathan A. Eisen. Metagenomic Sequencing of an In Vitro-Simulated Microbial Community. *PLoS ONE*, 5(4), April 2010.

[10] Tom M W Nye, Pietro Liò, and Walter R Gilks. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics (Oxford, England)*, 22(1):117–9, January 2006.

[11] Elizabeth Purdom. Analyzing data with graphs: Metagenomic data and the phylogenetic tree. pages 1–22, 2008.

[12] Sujatha Srinivasan, Noah G. Hoffman, Martin T. Morgan, Frederick A. Matsen, Tina L. Fiedler, Robert W. Hall, Roger Bumgarner, Jeanne M. Marrazzo, and David N. Fredricks. Deep Sequencing with High Resolution Phylogenetic Analysis of Bacterial Community Profiles in Women with Bacterial Vaginosis, 2011. Submitted to *PLoS Biology*.

[13] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[14] James White, Saket Navlakha, Niranjan Nagarajan, Mohammad Reza Ghodsi, Carl Kingsford, and Mihai Pop. Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. *BMC Bioinformatics*, 11(1), 2010.

[15] Hadley Wickham. *ggplot2: elegant graphics for data analysis*, volume 35 of *use R!* Springer, 2009.

6. Supplementary Figures

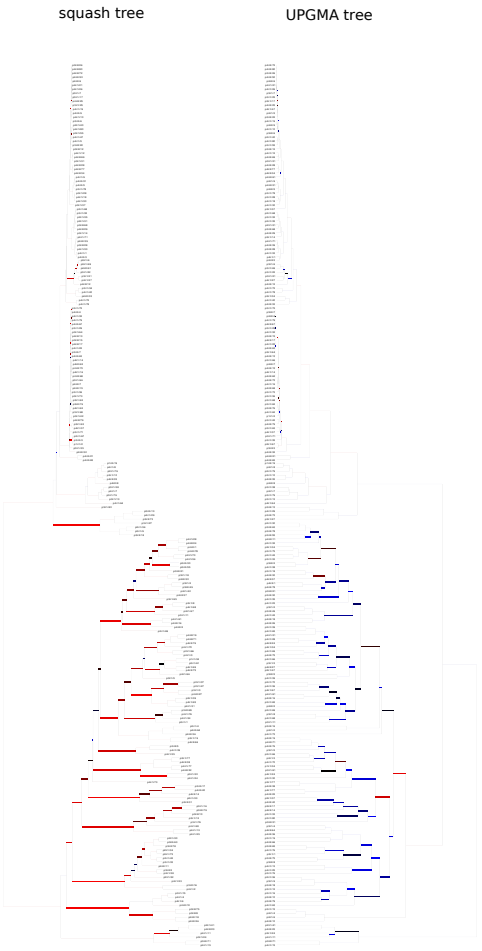


FIGURE S1. A comparison of the clustering results for the vaginal data set using the software of [10]. The software uses the Hungarian (a.k.a. Munkres) algorithm to find an optimal one-to-one matching between edges of the trees minimizing differences in a topological score between pairs of matched branches as follows. Given two trees T and S on the same samples, let $\Sigma(T)$ and $\Sigma(S)$ be the bipartitions of the samples induced by cutting the edges of T and S. For two bipartitions i and j, one can associate a "agreement score" $s(i,j)$ describing the proportion of shared elements between the sides of the bipartitions. The algorithm finds a one-to-one matching between $\Sigma(T)$ and $\Sigma(S)$ that minimizes the total agreement score between matched bipartitions. Now each tree can be drawn in a way which shows the agreement scores, such that a thick branch represents an edge which has a low agreement score with its partner in the matching. The program then arranges the trees such that matched edges are close to one another on the tree.

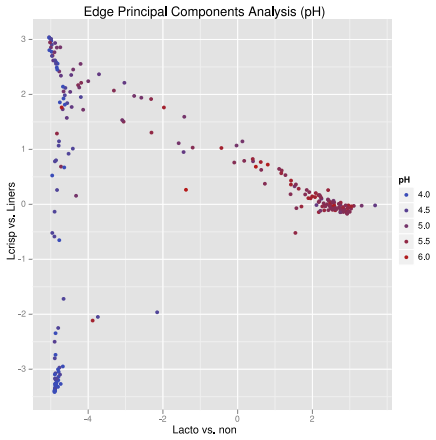


FIGURE S2. The samples plotted with respect to the first two principal components and colored according to pH.