RESEARCH

Intensity Plots and other options for Visualizing Microbiome Data

Eugenie Jackson[†], Ekaterina Smirnova[†], Anju Menon and S Huzurbazar[*†]

Correspondence: lata@uwyo.edu
epartment of Statistics,
niversity of Wyoming, 1000 E.
niversity Avenue, Box 3332, WY
2072 Laramie, USA
ull list of author information is
vailable at the end of the article
Equal contributor

Abstract

Background: Dimension reduction for high-dimensional sparse 16S rRNA data is necessary for descriptive data analysis. Most visualization options are restricted to 3 dimensions, while more than 3 dimensions are necessary to capture the relationships between taxa or subjects in microbiome data. We develop intensity plots as a way of highlighting the changing contribution of taxa or subjects as the number of principal components of the ordination method is changed.

Methods: For ordination methods implemented in R packages, use of visualization tools from the QIIME pipeline requires the time intensive task of formatting R output to make it readable by QIIME; we present code for such an output transfer. We develop intensity plots for taxa and subjects where the intensity indicates the relative position of the taxon/subject within the data, and we also develop 'inclusion in intensity deciles' plots which provide a quick visualization of taxa/subjects that are close to the 'center' or that contribute to dissimilarity. For 2 OTU tables linked to the same subjects, we develop co-inertia analysis based on principal coordinates analysis. We provide R code for all the visualizations.

Results: Using two data sets, we illustrate the various plots for data from 1 OTU table, as well as for longitudinal data where 2 tables are linked to the same subjects. We illustrate and discuss the differences between covariance versus correlation PCA, and using the latter, show how patterns for subjects and taxa change as the number of PCs, or alternatively, the dimension of the reduction is varied. For longitudinal data, we examine changing intensity patterns over the two time periods for taxa as well as subjects.

Conclusion: PCA based on correlation avoids the confounding of high abundance taxa with highest contributions to overall variability. Exploring intensity plots as a function of increasing number of PCs gives insight into the changing role of taxa as well as subjects. Inclusion in decile plots help clarify similarities and differences between covariate classes as the number of PCs changes.

Keywords: ordination methods; dimension reduction; compositional shifts of taxa

Background

Data visualization is a crucial part of the first steps of any statistical analysis. Such visualization is complicated in many modern data settings, including for microbiome data where sequencing of the variable region of the 16S rRNA gene yields counts on a large number of taxa with counts of zero in a number of samples. For human microbiome data, the taxa are often the species-level taxonomic units, also called operational taxonomic units (OTUs), with the samples as individuals. The resulting data form large sparse matrices that cannot be readily visualized or analyzed, necessitating a first task of dimension reduction by one of many multivariate dimension reduction methods. These ordination methods, widely used in ecology to understand relationships between ecological communities, include principal components analysis (PCA), correspondence analysis (CA), principal coordinate analysis (PCoA) also called multidimensional scaling (MDS) as well as extensions such as non-symmetric correspondence analysis (NSCA); see [1] for details. There are two key steps in implementing ordination methods, data transformation followed by singular value decomposition; ordination methods are characterized by specific data transformations. For example, in PCA, transformations of the observed taxa matrix are used to obtain either a covariance or correlation matrix, while PCoA requires a matrix with distances between taxa or subjects. In the second step, the method-specific transformed data matrix is decomposed to find the *principal components* (PCs) of that particular ordination analysis. Figure 1 gives an overview of the options as used in various microbiome studies. Analyses in the microbiome literature often start with plots of the first two or three PCs or dimensions especially in relation to covariate categories, and from these, patterns are gleaned and conclusions are drawn.

In ‘textbook’ multivariate data sets, adequate reduction to two or three dimensions is readily achieved, but with microbiome data, this is hardly the case. As an example, consider PCA with the correlation or covariance matrix as the transformed matrix; adequacy of the reduced dimension is often judged as the number of PCs needed to explain a large percentage of the original variance or correlation, examples of total inertia. Traditionally, this percentage is closer to 90 (see [2] among others), while a quick perusal of the human and other microbiome literature yields a list of papers where conclusions are drawn from patterns observed from the first two or three PCs that often explain less than 40% of the variation. For example, in [3], two PCs explaining 14.7% of total inertia are used; since these plots *do not explain* 85.3% of total inertia, the conclusions drawn from these visualizations might be incomplete. Along those lines, the following studies present PCA or PCoA analyses to draw conclusions based on two PCs that cumulatively account for 6.64% [4], 27.44% [5], 40.6% [6], 24.73% [7], 22.9% [8], 30.2% [9], 24.9% [10], 29.71% [11], 70.6% [12] and 25.7% [13] of the total inertia. In addition, other studies neglect to mention how much of the total inertia is explained by the ordination presented, eg. [14] and [15]. These studies span a variety of microbiome analyses: human airways (oral, throat, nose, sputum, lung) [4, 5, 6, 15], human gut [7, 8], chicken gut [10], and mice gut [11, 12, 13].

Ordination methods, widely used in microbial ecology to describe relationships among communities, have been implemented in various statistical software packages, most notably using R [16]. The package *ade4* [17, 18, 19] is the most comprehensive in the range of ordination methods implemented. The *adegraphics* [20] package provides improved visualization for the output from *ade4*. The *vegan* package [21], in addition to having many ordination methods implemented, can also be used to compute distance matrices with respect to different metrics, and to display distance and similarity information for the taxa. The *phyloseq* package [22] links major packages developed for phylogenetic analyses using sequenced data and handles internal tasks associated with organizing data in formats required by the individual packages. However, the graphics produced by these packages lack interactivity and other tools which could help in visualizing ordination results for a large number of taxa in more than two or three dimensions while simultaneously examining relationships across covariate classes or subgroups.

Traditional visualization of ordination results in more than 2 dimensions is difficult, but the loss of information from restricting to an inaccurate approximation of the data matrix is equally disconcerting. Keeping this in mind, we explored other options for visualization of more than two or three PCs, and also developed code for some novel visualizations. In this paper, we consider a common scenario where an ordination method is chosen and implemented via the R package *ade4* [17, 18, 19] and the task at hand is to visually explore results from the ordination methods; i.e., examine the patterns of the PCs with respect to taxa and samples, with the latter potentially grouped according to covariate classes. An important point is that we would like to examine a sufficient number of PCs that collectively explain a high proportion of the variation in the data matrix keeping in mind that this number will vary across data sets.

The first option is to use available plots from *adegraphics*, which, as we discuss below, are somewhat restrictive. A second option is to use the graphics available from the EMPeror project [23] which provides tools for 3 D interactive visualization as an extension of the QIIME [24] pipeline designed for 16S rRNA data. This interactive tool allows for plots of up to ten PCs and the user can change settings to display information by covariate classes of interest. However, the ordination methods that feature prominantly in QIIME are PCA, PCoA and Procrustes analysis. We developed code for linking output from the R-based methods available in the *ade4* package with the interactive graphical techniques available in EMPeror via the QIIME interface. This allows the practitioner to combine the extended ordination methods available in R with the interactivity of EMPeror graphics. In addition, for those who use different pipelines in the first steps of microbiome data analysis including using packages written in R, this provides an interface for the results to be viewed in EMPeror.

Finally, as we explored some vaginal microbiome data sets, we developed intensity plots that reveal aspects of the data that neither the *adegraphics* nor EMPeror plots were capable of elucidating. For instance, based on PCA, one can use intensity plots to subset the samples into groups as a function of their distance from the average, and these intensity plots can also be examined as increasing numbers of PCs are included accounting for more variation. Similarly, one can examine taxa intensity plots which can point us towards taxa that behave unusually or whose intensity changes dramatically with the number of PCs. We provide explanations for these visualizations with R code in supplementary materials.

We illustrate the above visualizations via (i) a published dataset on the vaginal microbiome of reproductive age women [25] with PCA as the ordination method, and (ii) a dataset that was generated as part of the Vaginal Human Microbiome Project [26] at Virginia Commonwealth University, where the ordination method is PCoA, and the covariate subsetting the data is diagnosis of bacterial vaginosis (BV). The first data set has $n > p$, namely, more samples than taxa, while the second one has $n < p$. We provide code for ordination analysis in R with *ade4*, along with the interface to the QIIME pipeline for graphics in EMPeror and for the intensity plots in R. The code and visualization methods presented here can be extended to other ordination methods since they use the output of the second step, the singular value decomposition, which is common to all ordination methods. Moreover, ideas for visualization can also be extended to understand other aspects of microbiome studies.

Methods

As our goal is to explore different ways of visualizing results from ordination methods, we first present brief explanations for PCA and PCoA as methods for one OTU table, and for an extension of PCoA to co-inertia analysis (CoIA) for two OTU tables.

Among the various packages in R that implement ordination methods, *ade4* [17, 18, 19] has the most complete set of methods which leads us to use it for the first step in our analyses. Central to the implementation is the eigen-decomposition of a matrix $\mathbf{H}$, where formulation of $\mathbf{H}$ from the data matrix $\mathbf{X} = \mathbf{X}_{n \times p}$, for n samples and p taxa, is specific to the ordination method being used. Here we discuss how $\mathbf{H}$ is constructed for PCA, PCoA and CoIA with an emphasis on linking matrix operations to the quantities for plotting and exploring the data. Finally, we present options for visualizations beyond those available in *ade4* and the associated *adegraphics*.

One table methods

Principal components analysis (PCA)

In this well known technique, the p taxa are expressed as linear combinations of new variables called *principal components* by decomposing the matrix of either taxa covariances or correlations. For covariance PCA, the data matrix $\mathbf{X}$ is centered by taxa (column) means, while the centered matrix is further transformed by scaling via taxa standard deviations for correlation PCA. The transformed matrix, denoted by $\tilde{\mathbf{X}}$, is combined with $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$ containing row weights, and a column weights matrix $\mathbf{Q} = \mathbf{I}_p$, with $\mathbf{I}_p$ a $p \times p$ identity matrix; note that $\mathbf{Q}$ is not always the identity, for example in correspondence analysis. The matrix to be eigen-decomposed, $\mathbf{H}$, and its eigen-decomposition are

$$\mathbf{H} = \begin{cases} \tilde{\mathbf{X}}^T\mathbf{D}\tilde{\mathbf{X}}\mathbf{Q} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T & \text{if } n > p \\ \tilde{\mathbf{X}}\mathbf{Q}\tilde{\mathbf{X}}^T\mathbf{D} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T & \text{if } n \leq p. \end{cases} \tag{1}$$

In many microbiome applications, the number of samples n can be much smaller than the number of taxa p so that the separate eigen-decompositions in eq. (1) provide computational ease (see [17, 27, 28] for details and theoretical properties). The eigen-decomposition yields two sets of axes used in visualization of ordination results in the reduced dimensional space. Specifically, in *ade4*,

- for $n > p$, the set of *right eigenvectors* $\mathbf{U}$ from eq. (1) are scaled to obtain a set of axes

$$\begin{cases} \mathbf{A} = \mathbf{Q}^{-1/2}\mathbf{U} & \text{taxa loadings} \\ \mathbf{K} = \tilde{\mathbf{X}}\mathbf{Q}^{1/2}\mathbf{U}\boldsymbol{\Lambda}^{-1/2} & \text{sample scores,} \end{cases}$$

- for $n \leq p$, eigen-decomposition of $\mathbf{H}$ produces a set of *left eigenvectors* $\mathbf{V}$ in eq. (1) that in *ade4* are scaled to obtain a set of axes

$$\begin{cases} \mathbf{A} = \tilde{\mathbf{X}}^T\mathbf{D}^{1/2}\mathbf{V}\boldsymbol{\Lambda}^{-1/2} & \text{taxa loadings} \\ \mathbf{K} = \mathbf{D}^{-1/2}\mathbf{V} & \text{sample scores.} \end{cases}$$

Principal coordinates analysis (PCoA)

PCoA, usually called MDS in the statistics literature, is an eigen-decomposition of a distance matrix. Pairwise distances between samples (rows) of $\mathbf{X}$ are computed with respect to some distance metric; popular metrics are Bray-Curtis, Jaccard, Jensen-Shannon, UniFrac, and weighted UniFrac distances. The distance matrix, denoted here by the $n \times n$ matrix $\mathbf{B}$, is then squared element-wise, $\mathbf{B}^{(2)}$, and centered according to a procedure proposed by Gower [29], to give $\mathbf{H}$ as

$$\mathbf{H} = -\frac{1}{2}\mathbf{D}^{1/2}\mathbf{F}\mathbf{B}^{(2)}\mathbf{F}\mathbf{D}^{1/2} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T, \text{ where } \mathbf{F} = (\mathbf{I}_n - \mathbf{D}\mathbf{J}), \tag{2}$$

where $\mathbf{D}$ is the row weights matrix from above and $\mathbf{J}$ is an $n \times n$ matrix of ones. The $\mathbf{H}$ in eqn. (2) can be rewritten as $\mathbf{H} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ with $\tilde{\mathbf{X}} = \mathbf{V}\boldsymbol{\Lambda}^{1/2}$ giving *approximate Euclidian distances for each sample from the origin.* Since PCoA eigen-decomposes a standardized matrix of distances between samples, only sample scores as a function of the eigenvectors are available for further examination and information about individual taxa is unavailable,

$$\begin{cases} \mathbf{A} = \mathbf{I} \\ \mathbf{K} = \mathbf{D}^{-1/2}\mathbf{V} \quad \text{sample scores,} \end{cases}$$

with the taxa loadings matrix $\mathbf{A}$ as the identity matrix.

General Framework

As shown in Figure 2, the ordination method-specific matrices of sample scores $\mathbf{K}$, taxa loadings $\mathbf{A}$ and eigenvalues $\boldsymbol{\Lambda}$ derived from the eigen-decomposition of $\mathbf{H}$ correspond to the singular value decomposition of $\tilde{\mathbf{X}}$. The results of this one–table decomposition in *ade4* contain a table of sample scores, denoted $l1, and a table of taxa loadings, denoted $c1. The scores and loadings are used to construct lower dimensional visualizations of samples and taxa, especially via biplots. The entries in $\boldsymbol{\Lambda}^{\mathbf{1/2}}$ (these are standard deviations in PCA) reflect the importance of the dimensions in the singular value decomposition; mathematically they are the lengths of the axes. For visualization purposes they are used to scale $l1 and $c1 to obtain the scaled sample scores ($li) and scaled taxa loadings ($co) which represent the configuration of samples or taxa in various plots.

Two-table methods

Co-inertia analysis (CoIA)

Co-inertia analysis allows for simultaneous exploratory analysis of two multidimensional tables by linking the common rows (samples) so that data across the columns (eg. taxa) of the two tables can be visualized. CoIA can thus be used for visualization of microbiome data collected longitudinally on the same samples at two different times. Implementations of CoIA in the literature [30] rely on PCA, CA, or NSCA. Covariance PCA and NSCA are highly biased towards abundant taxa, while correlation PCA and CA give a lot more weight to rare taxa. Distance based methods, especially those relying on UniFrac or Jensen-Shannon divergence, are less prone to either of these extremes, often have biological interpretation, and are widely used in microbiome data analyses. Here we extend CoIA using distance matrices derived from OTU tables on the same subjects at two different time periods.

First, from the OTU tables at two different times (eg. baseline and disease state, or different stages of pregnancy), we construct distances matrices from separate PCoAs to give $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$. Next, as suggested in [30], the two matrices are coupled to give $\mathbf{H}$ which is then eigen-decomposed,

$$\mathbf{H} = (\tilde{\mathbf{X}}_2^T\mathbf{D}\tilde{\mathbf{X}}_1)^T(\tilde{\mathbf{X}}_2^T\mathbf{D}\tilde{\mathbf{X}}_1) = \mathbf{U}\Lambda\mathbf{U}^T = (\mathbf{V}\Lambda^{1/2}\mathbf{U}^T)^T(\mathbf{V}\Lambda^{1/2}\mathbf{U}^T), \tag{3}$$

with $\mathbf{D}$ as the row weights matrix. The eigen-decomposition in eq. (3) is used to obtain co-inertia axes that are *common* to both OTU tables,

$$\begin{cases} \mathbf{V} = (\tilde{\mathbf{X}}_2^T\mathbf{D}\tilde{\mathbf{X}}_1)\mathbf{U}\mathbf{\Lambda}^{-1/2} & \text{OTU 1 co-inertia axes} \\ \mathbf{U} & \text{OTU 2 co-inertia axes.} \end{cases}$$

Finally, the common representation of samples from time 1 to 2 are given by the co-inertia sample scores

$$\begin{cases} \mathbf{G}_1 = \tilde{\mathbf{X}}_1\mathbf{V} & \text{OTU 1 co-inertia sample scores} \\ \mathbf{G}_2 = \tilde{\mathbf{X}}_2\mathbf{U} & \text{OTU 2 co-inertia sample scores.} \end{cases} \tag{4}$$

The distance between the i^{th} sample's co-inertia sample scores is obtained by comparison, on the r axes, of the co-inertia sample scores for time periods 1 and 2,

$$d_i = \sqrt{\Big(\sum_{j=1}^{k}(g_{ij}^1 - g_{ij}^2)^2\Big)}, \tag{5}$$

where $\boldsymbol{g}_i^1 = (g_{i1}^1, \ldots, g_{ir}^1)$ and $\boldsymbol{g}_i^2 = (g_{i1}^2, \ldots, g_{ir}^2)$ are the i^{th} rows of sample score matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ respectively. Smaller values for the distance d_i indicates more similarity in i^{th} sample's distance information across the two time periods.

Exploring ordination results visually

Visualization using ade4 and adegraphics

The list of plots possible for the output from one table using *ade4* and *adegraphics* include scree plots, two dimensional (based on 2 PCs) sample and taxa plots as well as biplots, and confidence ellipses that represent inertia of the groups of samples under the bivariate normality assumption. The enhanced *adegraphics* allows one to color and rescale each point on a samples plot according to information from a continuous covariate. One can also zoom in to concentrate on a subset of variables. For CoIA, the plots provided with *adegraphics* help visualize the difference between each row's observations in two tables by plotting the distances $d_i, i = 1, \ldots, n$ based on the first two co-inertia axes. One main drawback of *ade4* and *adegraphics* is that they are not interactive and visualization is restricted to two dimensions. The plots work extremely well for data with a smaller number of taxa than in human microbiome data. With over 50 points for samples or taxa, the plots can be cluttered and difficult to interpret; see Additional File 1: Figure S1.

Inputting ade4 output into QIIME and EMPeror for other plots

For a dataset decomposed using *ade4*, we provide code (Additional File 2: Code S1) that formats the *ade4* output and covariate information and makes it readable by QIIME for interactive visualization in EMPeror; note that *ade4* is written in R while QIIME is written in Python. There are various interactive plots available in EMPeror. One can obtain 3-dimensional plots of either sample scores or taxa loadings on any 3 of the first 10 PCs. In these plots, information from up to 3 categorical covariates can also be included in the form of scale, transparency, and color of the points. Also available are 3-dimensional biplots where taxa and sample information can be plotted on the first 3 PCs along with information from up to 3 categorical covariates. While these plots are very useful, they are, obviously, restricted by the inability to plot in more than 3 dimensions. Parallel coordinates plots are available where the coordinates are the first 10 PCs with sample scores (or taxa loadings) on the 10 PCs yielding a line for each subject that can be color-coded by covariate classes of interest.

R code for intensity and other plots

The first decision in any ordination analysis is to choose how many dimensions to explore. In the statistics literature, a scree plot is routinely examined to either find a break after which the eigenvalues are very small so that the contribution of additional dimensions is negligible, or in the absence of a break, retain enough dimensions to obtain a good approximation of the matrix involved. In the PCA literature, this is often described as keeping enough PCs to explain a high percentage of the variation, usually above 80%. For human microbiome data, the percentage of variation explained by 2–3 PCs is usually quite low. Our approach is to examine different aspects of sample scores and taxa loadings as we increase the number of PCs or more generally, the dimension of the approximation.

As discussed in the next section, we provide code for an informative, color-coded scree plot (Additional File 3: Code S2). We use intensity plots where the intensity is a function of the sample scores or taxa loadings plotted as the number of PCs is increased thereby increasing the dimension of the approximation. Different functions give different information, and we generally examine the changes in intensity for groups of samples or taxa as the number of PCs changes. The plots can be broadly classified into ‘intensity plots’ and ‘inclusion in an intensity decile plots’.

Intensity plots for samples have samples (or subgroups of samples) on the vertical axis while the horizontal axis has an increasing number of PCs. The intensity of the points in a row indicates the distance of the sample from the origin or ‘center’ or ‘average’ as the number of PCs increases. Smaller distances are coded with lighter colors indicating closeness to the center, and increasing distance from origin implies that the sample has a relatively different taxa profile compared to other samples. Similar plots for taxa, as opposed to samples, with an increasing number of PCs on the horizontal axis display lower intensity with lighter colors implying that most samples have the taxon at about the same level, while darker color implies variation between the samples with respect to this taxon.

For the ‘inclusion in an intensity decile’ plots using samples, the intensity or distance from the origin is fixed using a specific decile (eg. 1st or 10th), and the samples on the vertical axis are either included or excluded in that decile as the number of PCs increases; analogous plots can be made for taxa. Again, color coding the samples by covariate classes or taxa by biological groups can add another dimension of information allowing for visual comparison across groups. We provide code (Additional File 4: Code S3) that takes the *ade4* output and generates intensity plots and ‘inclusion in an intensity decile’ plots.

Results

We now present examples of visualization of ordination results using two different vaginal microbiome data sets. The examples were chosen to illustrate methods for one OTU table and for two OTU tables that are obtained from the same subjects as is the case in longitudinal analyses. While we use the *ade4* package in R to conduct the ordination, we illustrate the results using *adegraphics*, EMPeror and our contribution of intensity plots. We provide code that can be adapted to output from other ordination methods, and also suggest extensions.

Vaginal microbiome data: one table with $n > p$

Using PCA as the ordination method, we explore the 16S rRNA microbiome OTU table with 247 taxa from $n = 394$ reproductive age women analyzed in Ravel et al. [25]. We used a traditional filtering method implemented in the '*kOverA()*' function in the R package *genefilter* [31], where taxa with an expression measure above k (we used 0) in less than m (we used 5) samples are filtered out; leaving us with $p = 135$ taxa. Covariates available include race/ethnic information (Asian, Black, Hispanic, White), and those important for diagnosis of and predisposition towards bacterial vaginosis (BV), namely, Nugent score and vaginal pH levels. One of the goals in [25] was to explore the composition of the vaginal microbiota in relation to BV. Microbial diversity is often summarized by clustering microbiome profiles into community state types (CSTs). In [25], PCA based on the covariance matrix of relative abundances of the taxa was used to visualize membership in different CSTs.

Vaginal microbiome data sets are often dominated by highly abundant *Lactobacillus* species along with a lot of rare taxa. In such cases where the scales of measurement vary widely, the general recommendation in the statistics literature is to base PCA on the correlation matrix (eg. [2] among others). In Figure 3, we present scree plots using both correlation and covariance based PCA. With correlation PCA, 61 PCs were required to explain 90% of the variation, much larger than the 4 using covariance PCA. This result is typical for sparse data which contain a small number of dominant values (abundant taxa) that explain the bulk of the variation in covariance PCA but not nearly so with correlation PCA. In supplementary materials (Additional File 3: Code S2), we provide R code for making similar scree plots where the color-coding is for user-specified ranges in cumulative amounts of variation explained; eg. $\leq 40\%$, $\leq 70\%$, $\leq 80\%$, and $\leq 90\%$ in Figure 3. Coloring the screeplot in this manner is an easy step in exploring the number of PCs to use for further data visualization.

We conducted PCA using the correlation matrix. As discussed in the previous section, 3-dimensional interactive plots are available in EMPeror [23], we present such a plot (Additional File 5: Figure S2) where the points, shown as spheres, are colored by ethnicity and scaled by Nugent score category. However, the 3 correlation-based PCs for this data only explain about 19% of the total variation. Similarly, the interactive biplot (Additional File 6: Figure S3) shows the subjects represented by solid spheres and taxa within the same plot, with the twenty most prevalent taxa as transparent gray spheres whose diameter reflects taxa abundance; spheres representing the various *Lactobacilli* species are close to each other. Individuals with low scores also have highly abundant *Lactobacilli* and cluster together. While the interactivity and ability to color these plots is extremely useful, they are limited by visualization in up to 3 dimensions.

A parallel coordinates plot from EMPeror (Additional File 7: Figure S4). shows, on the left, 108 samples classified as CST IV, while on the right are the 286 from CSTs I–III and V, dominated by various *Lactobacilli* species. Sample scores on 10 individual PCs provide the coordinates, each sample is represented by a line which can be subset further by another covariate, or examined individually. The plots in EMPeror, though very useful in terms of interactivity and ability to visualize groupings according to covariate classes, are limited by the inability to examine results in more than 3 dimensions or for parallel coordinates plots for more than 10 dimensions.

Given these limitations, and the inability to assess the cumulative effect of PCs, we explored other types of plots. For the correlation PCA of these data, in Figure 4 we present plots of the inclusion of samples in the 1^{st} and 10^{th} decile as the number of PCs included is increased from 1 to 61. Only the women who were present in at least one column are included in the plots with the samples grouped by race/ethnic covariate information. Of the 394 women, 176 are in the 1^{st} decile with 76 in the 10^{th} and 5 appearing in both. The 1^{st} decile is useful in defining a 'center' or 'average' taxa profile since proximity to the origin is a measure of how close a sample is to the mean measurements of the variables, here the taxa proportions. However, this can change as the number of PCs is changed. For instance, subject 250 in the last row would be considered 'average' if only one PC is used, while Subject 359 in the top row is consistently identified as 'average' when at least 10 PCs are used.

The 10^{th} decile helps to identify samples with more dissimilar taxa profiles, as they are farthest from the origin. In the plots, inclusion stabilizes as at least 31 PCs explaining $\approx 70\%$ of the variation are included.

Table 1 displays a summary of how many samples are included if up to 5 PCs are used versus more than 45 with 'Intersection' as the number of samples at the intersection of $\leq$ 5 and $>$ 45 PCs used within that decile. The intersection is negligible for the first decile indicating that the samples with 'average' profile change as the number of PCs changes. However, from the 10^{th} decile, the samples with dissimilar taxa profiles are more stable.

Similarly, 5 women (S129, S224, S230, S288, S307) appear in both deciles and examining subject 230 who is representative of the 5, we see that she seems close to the origin when 1 or 2 PCs are used, but is found to be far from the origin when 16 PCs are used. Inspection of the OTU table reveals that subject 230 has the largest and second largest proportions, respectively, of the relatively rare taxa *Lactobacillales.1* and *Lactobacillus.3*. Because her taxa values are close to the mean for more abundant taxa, she does not appear to be unusual when a small number of PCs are used, as the more abundant taxa load heavier on the first couple of PCs. Her covariate information identifies her as Black, in the low Nugent score category, and with a microbiome dominated by *L. gasseri*. The inclusion plots make identification of such subjects much easier compared to the other plots presented earlier.

In Figure 5 we present a taxa intensity plot using correlation PCA for the subset of the 135 taxa that are identified in [25] as dominating the five CSTs and 3 other taxa, specifically, *Ureaplasma*, *Lactobacillus vaginalis*, and *Rothia*; in Figure 6, the corresponding plot using covariance PCA is presented. In the latter, 4 PCs explain up to 90% of the variation, and the taxa that contribute the most to variability (darker colors) correspond to those used to define the CSTs. However, the abundance data (proportions multiplied by the total number of reads) indicate that the 4 most abundant taxa are the four species of *Lactobacilli* that largely characterize CSTs I, II, III and V; the taxa that contribute the most to variability are confounded with the highly abundant taxa, as is often the case when covariance PCA is used. In fact, all the taxa from the list are among the most abundant. From Figure 5 we see that the intensity patterns for these taxa change as the number of PCs are increased with correlation PCA. Of the additional taxa in Figure 5, *Ureaplasma*, associated with the diagnosis of BV, was detected in small proportions in 147 subjects with only 3 women having proportions greater than 10% and one of those had a high Nugent score. *L. vaginalis*, although not one of the dominant *Lactobacilli* species, is considered protective against BV and was detected in small proportions in about a third of the women (124), with only one having proportions greater than 10%. *Rothia* is a common oral bacterium rarely found in vaginal microbiome samples and among the six women with *Rothia*, none had a high Nugent score.

Vaginal Human Microbiome Project (VaHMP) data example

The second vaginal microbiome dataset is from the Human Microbiome Project II at Virginia Commonwealth University (VCU) [26]. The data, from female volunteers recruited during regular visits to a VCU clinic, are obtained from vaginal swab samples and health questionnaires. The full data set contains 1–8 visits on 4118 women, here we use a subset of 36 non-pregnant women who transitioned from a state of 'Pre BV' to 'BV' during consecutive visits that were 19 to 765 days apart with a median of 192 days; BV was diagnosed using the Amsel criterion. A total of 378 taxa were identified; we used the R package *DESeq* ([32]) to adjust for library size and to filter the data while keeping the two OTU tables together. For the *DESeq* library size factor estimation, a pseudo-count of 1 was added to each element of the OTU tables, rows of the original OTU tables were then adjusted using the estimates. As in the previous example, taxa were filtered out if they were present in less than 5 samples using the *genefilter* package in R, resulting in $p = 94$ taxa for our analysis. The final data are still sparse with a large number of zeros.

Given data from two consecutive visits from the same women, we conduct CoIA after PCoA, with the distance as the Jensen-Shannon divergence metric [33] computed with the R package *entropy* [32]. Computation of this metric uses logarithms of frequencies, which is problematic with zeros. We used the common solution that estimates OTU frequencies from normalized counts using a James-Stein type shrinkage estimator implemented in the function *KL.shrink()* in the *entropy* package.

In Figure 7, we present a plot of the distances from CoIA, using r=2 from eq. (5). The distances are calculated and plotted using the first two axes from the eigendecomposition, with the coordinates from the appropriate entries of $\mathbf{G}_1$ and $\mathbf{G}_2$ in eq.(4); the two axes together explain about 74% of the variation. We highlight 5 women, subjects labeled 3, 8, 12, 21, and 26 to illustrate different patterns. Subjects 3, 12 and 21 cluster close together at the 'Pre BV' state, and while 12 and 21 remain close in the 'BV' state, subject 3 changes dramatically. Of the remaining two, 8 transitions in the opposite direction as 12 and 21, while 26 appears to be outlying in both states. The value of such a plot, code for which is in (Additional File 8: Code S4), is in revealing potential clustering patterns and in highlighting subjects whose data could be explored further. Though this plot indicates changes in clustering, it should be noted that it does so based on 2 components explaining 74% of the variation. One idea is to incorporate (cumulative) information to obtain a 'distance-intensity' plot as k or the number of axes increases and the cumulative percentage of the eigenvalues is closer to 90%. Such a plot with the samples colored according to some important covariate class would give information about the nature of the change from one state to the next.

As mentioned earlier, PCoA and CoIA after PCoA are based on distances between samples. A possibility is to look for clusterings of women, eg. 12 and 21 vs 3, and examine their OTU profiles by looking at the OTU frequencies. Admittedly, examining proportions for 94 taxa is cumbersome, which once again leads back to examining intensity plots for the subjects based on PCA, and examining the taxa with unusually high loadings and their role in the subgroup of women.

Since there are two OTU tables, we now examine the taxa intensity profiles as the number of PCs increases for 8 taxa of interest for BV. In Figure 8, the top line for each taxon is the intensity in the 'Pre BV' state, while the second intensity profile is for the 'BV' state; the coloring for the second state indicates whether the range of OTU counts increased or decreased according to paired t-test results at significance level 0.05 when comparing proportions in the pre-BV and BV states. The top two taxa, the *Lactobacilli* species are both indicative of a healthier vaginal microbiome, and the taxa intensity plots as well as the changes in proportions demonstrate this. For both states, these have intensities that place the taxa closer to the origin or average, and the proportion decreases significantly from Pre BV to BV for *L.iners*. The next 5 have been suggested for use in diagnosis of BV. For most, the intensity profiles indicate a shift in the taxa from away from the origin (darker colors) in the Pre BV state to closer to the origin (lighter color) in the BV state, indicating that in the BV state, these taxa are part of the 'average' profile; the one exception is

Atopobium vaginae. In addition, the proportions increase significantly for 3 of the 5. The 8^{th} taxon, *Tenericutes OTU M1* is not diagnostic for BV but is considered to be strongly associated [34]; the profile for the 'No BV' state indicates it is in the lower deciles or closer to the 'average' taxa profiles while it moves closer to the upper deciles in the 'BV' state, though there is no significant change in proportions.

Discussion

Visualization of data is the first step in any data analysis, and visualization of sparse large dimensional data as contained in an OTU table obtained for microbiome studies presents special challenges. Specifically, visualization is restricted to lower-dimensions but examining only the lowest dimensional results for a large dimensional data set can result in loss of information. Given this possibility, we advocate using various plots that can help glean information at different levels. Species and taxa plots as well as biplots as available in *adegraphics* and EMPeror, though very informative, can only be viewed in up to 3 dimensions. Parallel coordinate plots can be very useful especially if they are interactive as in EMPeror, but they can get cluttered. None of these can simultaneously also give a gauge of contribution loss or change as the dimension is changed, eg. as more PCs are added or removed. Intensity plots, though two-dimensional in their current formulation, are a means of assessing this change in contribution. As all these plots elucidate different aspects of the ordination analysis, we advocate using all of them to reveal patterns in the data.

Consider our first example where the findings of Ravel et al. [25] indicate that the vaginal microbiome of Hispanic and Black women is not dominated by the *Lactobacilli* species as compared to that of white and Asian women. Traditionally, a high proportion of *Lactobacilli* was associated with 'healthy' vaginal microflora, and one of the goals in [25] was characterizing composition of 'normal' microbiome communities for each race/ethnic group. Subjects close to the origin in PCA plots can be viewed as women with 'average' taxa profiles. Therefore examining taxa present in the majority of these samples is a step towards characterizing a race/ethnic group's taxa profile. These can serve as a benchmark of what is considered 'normal' for each group.

Depending on the number of PCs used, the samples considered as 'average' for each group appear to be different. In Figure 3, the 1^{st} decile plot illustrates that samples viewed as close to the origin based on $1-3$ PCs are different from ones that are close to the origin when a sufficient amount of variability is taken into account. Among Hispanic women, the women considered 'average' using $1-3$ PCs are not 'average' when using 61 components, with the possible exception of subject $S142$. Table 2 compares the number of women and taxa identified as having 'average' taxa composition based on the $1-3$ and 61 PCs.

The list of 'average' taxa present in each race/ethnic group along with the number of women whose OTUs contain these taxa is tabulated in (Additional File 9: Table S1). For Asian, Black and White women, the taxa identified as 'average' using 61 PCs corresponding to 90% variability are a subset of the much longer list of taxa when $1-3$ PCs are used. For Hispanic women, 4 taxa identified as 'average' using 61 PCs that do not appear when $1-3$ PCs are used. The 9 taxa common to all groups using 61 PCs are *Lactobacillales 5, Lactobacillales 2, Lactobacillus iners, Ureaplasma, Lactobacillus crispatus, Lactobacillus vaginalis, Lactobacillus jensenii, Streptococcus* and *Lactobacillus gasseri*. These are the core vaginal microbiome taxa, and their presence among the 'average' taxa common to all ethnic groups is highlighted by our use of inclusion in 1^{st} decile plots. All these taxa, with the exception of *Ureaplasma*, produce lactic acid [25]; bolstering the notion of conservation of lactic acid production regardless of species composition. Finally, despite the detrimental role *Ureaplasma* may play in neonatal outcomes [34], it has been estimated that the bacteria can be found in the vaginal microbiomes of 40–70% of asymptomatic, sexually active women [35].

Taxa intensity plots as presented in Figure 5 for the first example and Figure 8 for the second example illustrate that the position of a taxon relative to the others, ie. its ordination, can change as the number of PCs or dimension of the reduction changes. Dimension should not be reduced so drastically that substantial information loss occurs. While deciding on how much reduction to settle on is a difficult question, it should not be driven by our inability to visualize in more than 3 dimensions. Plots for fixed dimensions give a snapshot of the relationships between taxa, while intensity plots give an idea of the changing contributions of each taxon. Examining intensity plots to see how many PCs 'stabilize' the plots might be a good way of helping decide on the reduction of the dimension. Once that is decided on, plots that capture relationships in up to 3 dimensions can be examined.

Conclusions

For large-dimensional data, dimension reduction is necessary for any visualization. We develop intensity plots for both taxa and subjects that capture their changing positions as the dimension of the reduction is changed. We illustrate these using PCA, arguing that with very different magnitudes of abundance data, correlation PCA should be used. With covariance-based PCA for such data, the largest contributors to overall variation are the taxa with the largest abundances, thus confounding dissimilarity pattern with the abundance pattern. For longitudinal data from two or more time points, questions of how to reduce dimension remain. In addition to using techniques such as CoIA, examining the intensity plots for taxa or subjects in each time period also provides insight into their changing roles over time. While existing R packages and add-ons to the QIIME pipeline provide useful visualizations, augmenting those with intensity plots provides a glimpse into patterns that might potentially be missed or over-interpreted if additional dimensions are not included.

List of abbreviations

BV: bacterial vaginosis
CA: correspondence analysis
CoIA: co-inertia analysis
CST: community state type
MDS: multidimensional scaling
NSCA: non-symmetric correspondence analysis
OTU: Operational taxonomic unit
PC: principal component
PCA: principal components analysis
PCoA: principal coordinate analysis
VaHMP: Vaginal Human Microbiome Project
VCU: Virginia Commonwealth University

Ethics approval and consent to participate

Neither data set came from trials of health care interventions. Data set 1 is publicly available. Data set 2 was generated as part of the VCU NIH Microbiome II project. Specifically, the Institutional Review Board at Virginia Commonwealth University approved this study under protocol HM12169.

Consent for publication

Not applicable.

Availability of data and materials

There are 2 datasets supporting the conclusions of this article.
The data from Ravel, et al. [25] is available at
http://www.pnas.org/content/suppl/2010/06/03/1002611107.DCSupplemental/st04.xlsx.

The sequencing reads for the 16S rRNA data from the VaHMP project have been deposited into NIH's Short Read Archive under BioProjectID PRJNA46877. Clinical diagnostic data is available through the Research Alliance for Microbiome Science (RAMS) Registry at Virginia Commonwealth University (http://vmc.vcu.edu/ramsregistry).
The link for the BioProject is https://www.ncbi.nlm.nih.gov/bioproject/PRJNA46877
The materials supporting the conclusions of this article are included within the article and its additional files.
Project name: MB-Intensity-plots
Project home page: https://github.com/GenieJx/MB-Intensity-plots
Archived version: https://doi.org/10.5281/zenodo.167450
Operating system: Platform independent
Programming language: R

Competing interests

The authors declare that they have no competing interests.

Funding

EJ, ES and SH were funded by NSF–DMS–1100615. EJ and SH were funded by a pilot grant from the Mountain West Consortium for Clinical and Translational Infrastructure grant from National Institute of General Medical Sciences (P20GM103440) - NIH. AM was partially funded by the Office of Research and Economic Development at the University of Wyoming. All co-authors were at the U of WY when the work was conducted.

Authors' contributions

EJ conceptualized and programmed the initial visualizations; both the plots and interpretations were further developed with input from the other authors, in particular, AM. ES made the ordination methods accessible and clarified notation and output from *ade4*, and extended CoIA using PCOA. EJ and ES conducted the data analyses, and with AM contributed to interpretation. SH developed the background ideas and context, constantly questioned every detail, and with input from others, wrote the manuscript. All authors read and approved the final manuscript.

Authors' information

E. Jackson, S. Huzurbazar, Department of Statistics, University of Wyoming. E. Smirnova, Department of Mathematical Sciences, University of Montana A. Menon, Shionogi, Inc. New Jersey, USA.

Acknowledgements

The authors with to thank Jennifer Fettweis and Paul Brooks at VCU for useful insights and use of data, and Ed Bedrick at U of AZ for useful input and encouragement on the ideas in the manuscript.

References

1. Legenrde, P., Legendre, L.: Numerical Ecology. Elsevier, Amsterdam (1998)
2. Izenman, A.: Modern Multivariate Statistical Techniques. Springer, New York (2008)
3. Sellitto, M., Bai, G., Serena, G., Fricke, W.F., Sturgeon, C., Gajer, P., White, J., Koenig, S.S.K., Sakamoto, J., Boothe, D., Gicquelais, R., Kryszak, D., Puppa, E., Catassi, C., Ravel, J., Fasano, A.: Proof of concept of microbiome-metabolome analysis and delayed gluten exposure on celiac disease autoimmunity in genetically at-risk infants. PLoS ONE **7**(3), 33387 (2012)
4. Gong, H.-L., Shi, Y., Zhou, L., Wu, C.-P., Cao, P.-Y., Tao, L., Xu, C., Hou, D.-S., Wang, Y.-Z.: The composition of microbiome in larynx and the throat biodiversity between laryngeal squamous cell carcinoma patients and control population. PLoS ONE **8**(6), 66476 (2013)
5. Kraneveld, E.A., Buijs, M.J., Bonder, M.J., Visser, M., Keijser, B.J.F., Crielaard, W., Zaura, E.: The relation between oral candida load and bacterial microbiome profiles in dutch older adults. PLoS ONE **7**(8), 42770 (2012)
6. Boutin, S., Graeber, S.Y., Weitnauer, M., Panitz, J., Stahl, M., Clausznitzer, D., Kaderali, G. L andEinarsson, Tunney, M.M., Elborn, J.S., Mall, M.A., Dalpke, A.H.: Comparison of microbiomes from different niches of upper and lower airways in children and adolescents with cystic fibrosis. PLoS ONE **10**(1), 0116029 (2015)
7. Claud, E.C., Keegan, K.P., Brulc, J.M., Lu, L., Bartels, D., Glass, E., Chang, E.B., Meyer, F., Antonopoulos, D.A.: Bacterial community structure and functional contributions to emergence of health or necrotizing enterocolitis in preterm infants. Microbiome **1**(20) (2013)

8. Shankar, V., Agans, R., Holmes, B., Raymer, M., Paliy, O.: Do gut microbial communities differ in pediatric ibs and health? Gut Microbes **4**(4), 347–352 (2013)
9. Dickson, R.P., Erb-Downward, J.R., Prescott, H.C., Martinez, F.J., Curtis, J.L., Lama, V.N., Huffnagle, G.B.: Cell-associated bacteria in the human lung microbiome. Microbiome **2**(28) (2014)
10. Pourabedin, M., Guan, L., Zhao: Xylo-oligosaccharides and virginiamycin differentially modulate gut microbial composition in chickens. Microbiome **3**(15) (2015)
11. Lu, K., Abo, R.P., Schlieper, K.A., Graffam, M.E., Levine, S., Wishnok, J.S., Swenberg, J.A., Tannenbaum, S.R., Fox, J.G.: Arsenic exposure perturbs the gut microbiome and its metabolic profile in mice: An integrated metagenomics and metabolomics analysis. Environmental Health Perspectives **122**(3), 284–291 (2014)
12. Langille, M.G., Meehan, C.J., Koenig, J.E., Dhanani, A.S., Rose, R.A., Howlett, S.E., Beiko, R.G.: Microbial shifts in the aging mouse gut. Microbiome **2**(50) (2014)
13. Hansen, C.H.F., Andersen, L.S.F., Krych, L., Metzdorff, S.B., Hasselby, J.P., Skov, S., Nielsen, D.S., Buschard, K., Hansen, L.H., K, H.A.: Mode of delivery shapes gut colonization pattern and modulates regulatory immunity in mice. The Journal of Immunology **193**(3), 1213–1222 (2014)
14. Fouts, D.E., Pieper, R., Szpakowski, S., Poh, H., Knoblach, S., Suh, M.-J., Huang, S.-T., Ljungberg, I., Sprague, B.M., Lucas, S.K., Torralba, M., Nelson, K.E., Groah, S.L.: Integrated next-generation sequencing of 16s rdna and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. Journal of Translational Medicine **10**(174) (2012)
15. Cribbs, S.K., Uppal, K., Li, S., Jones, D.P., Huang, L., Tipton, L., Fitch, A., Greenblatt, R.M., Kingsley, L., Guidot, D.M., Ghedin, E., Morris, A.: Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in hiv infection. Microbiome **4**(3) (2016)
16. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015). R Foundation for Statistical Computing. https://www.R-project.org/
17. Dray, S., Dufour, A.B.: The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software **22**(4), 1–20 (2007)
18. Dray, S., Dufour, A.B., Chessel, D.: The ade4 package-ii: Two-table and k-table methods. R News **7**(2), 47–52 (2007)
19. Chessel, D., Dufour, A.B., Thioulouse, J.: The ade4 package-i-one-table methods. R News **4**, 5–10 (2004)
20. Dray, S., Siberchicot, A., Thioulouse, J., Julien-Laferrière, A.: Adegraphics: An S4 Lattice-Based Package for the Representation of Multivariate Data. (2016). R package version 1.0-5. http://CRAN.R-project.org/package=adegraphics
21. Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H.: Vegan: Community Ecology Package. (2016). R package version 2.3-4. http://CRAN.R-project.org/package=vegan
22. McMurdie, P.J., Holmes, S.: Phyloseq: a bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pp. 235–246 (2012). NIH Public Access
23. Vazquez-Baeza, Y., Pirrung, M., Gonzalez, A., Knight, R.: Emperor: a tool for visualizing high-throughput microbial community data. Gigascience **2**(1), 16 (2013)
24. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Gonzalez Pena, A., Goodrich, J.K., Gordon, J.I., G.A., H., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R.: Qiime allows analysis of high-throughput community sequencing data. Nature Methods **7**(5), 335–336 (2010)
25. Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S.K., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., Brotman, R.M., Davis, C.C., Ault, K., Peralta, L., Forney, L.J.: Vaginal microbiome of reproductive age women. Proc Natl Acad Sci, USA **108**(1), 4680–4687 (2011)
26. Fettweis, J.M., Alves, J.P., Borzelleca, J.F., Brooks, J.P., Friedline, C.J., Gao, Y., Gao, X., Girerd, P., Harwich, M.D., Hendricks, S.L., et al.: The vaginal microbiome: Disease, genetics and the environment (2011)
27. De la Cruz, O., Holmes, S.: The duality diagram in data analysis: examples of modern applications. The annals of applied statistics **5**(4), 2266–2277 (2011)

28. Yata, K., Aoshima, M.: Effective pca for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. Journal of Multivariate Analysis **101**, 2060–2077 (2010)

29. Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**(3/4), 325–338 (1966)

30. Dray, S., Chessel, D., Thioulouse, J.: Co-inertia analysis and the linking of ecological data tables. Ecology **84**(11), 3078–3089 (2003)

31. Gentleman, R., Carey, V., Huber, W., Hahne, F.: Genefilter: Genefilter: Methods for Filtering Genes from Microarray Experiments. (2016). R package version 1.54.2

32. Hausser, J., Strimmer, K.: Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research **10**, 1469–1484 (2009)

33. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (1991)

34. Fettweis, J.M., Brooks, J.P., Serrano, M.G., Sheth, N.U., Girerd, P.H., Edwards, D.J., Strauss III, J.F., Jefferson, K.K., Buck, G.A., Consortium, V.M., *et al.*: Differences in vaginal microbiome in african american women versus women of european ancestry. Microbiology **160**(10), 2272–2282 (2014)

35. Remington, J.S.: Infectious Diseases of the Fetus and Newborn Infant. Saunders/Elsevier, Philadelphia, PA (2011)

Figures

Figure 1

Flow chart showing options for ordination methods for one OTU table.

Figure 2

Relationship between transformed matrix $\tilde{\mathbf{X}}$ and sample scores, taxa loadings and singular values, using *ade4* notation.

Figure 3

Scree Plots for the data from Ravel et al. [25] using covariance and correlation for PCA.

Figure 4

Inclusion in intensity deciles plots. 1^{st} decile indicates closest to the origin or average taxa profile while 10^{th} indicates farthest or most variable. The vertical lines correspond to the % variation explained using the scree plot for correlation PCA in Figure 3. Note that inclusion appears to stabilize past 31 PCs corresponding to at least 70% of variation.

Figure 5

Taxa Intensity plots using correlation PCA for taxa that dominate the CSTs (I, II, III, IV) as found in [25] and 3 other taxa; note that colors are alternated for labels for this order.

Figure 6

Taxa Intensity plots using covariance PCA for taxa that dominate the CSTs (I, II, III, IV) as found in [25] and 3 other taxa; note that colors are alternated for labels for this order.

Figure 7

Co-inertia plot: the two axes are the first two eigenvectors, each line is for a woman with the line directed from the ‘Pre BV’ to the ‘BV’ state.

Figure 8

Intensity plots for some taxa of interest for BV. For each taxon, top rows have intensity for ‘Pre BV’ and bottom row has ‘BV’ intensities, where red indicates an increase in the proportion of taxa while blue indicates a decrease. The decrease or increase was inferred using an 80% confidence interval for the differences in proportions.

Table 1 Number of samples included in deciles as up to 5 versus more than 45 PCs are used

	1^{st} decile (near origin)				10^{th} decile (far from origin)			
	Total	1–5 PCs	> 45 PCs	Intersection	Total	1–5 PCs	> 45 PCs	Intersection
Asian	44	17	15	0	21	15	16	11
Black	43	29	4	0	21	13	6	4
Hispanic	41	20	19	4	22	17	15	10
White	48	28	9	0	12	9	8	5

Table 2 Comparison of the number of women and taxa identified as 'average' for each ethnic group based on $1 - 3$ and 61 PCs.

	$1 - 3$ PC		61 PC	
	number of women	number of taxa	number of women	number of taxa
White	23	95	8	26
Hispanic	15	78	16	38
Black	28	99	4	12
Asian	14	70	12	22
number of taxa present in all ethnic groups		50		9

Tables

Additional Files

Additional File 1: Figure S1

Correlation PCA was performed on the data from Ravel [25]. We used the *adegraphics* package to produce a scatter plot of the taxa loadings for the first two PCs and used the *addhist* method to add univariate marginal distributions.

Additional File 2: Code S1

Code to format single table $ade4$ output and covariate information for interactive visualization in EMPeror through QIIME.

Additional File 3: Code S2

Code to produce color-coded scree plots as in Figure 2.

Additional File 4: Code S3

Code to take $ade4$ output and generate intensity plots and 'inclusion in an intensity decile plots'.

Additional File 5: Figure S2

EMPeror: 3-d plot of 3 covariance-based PCs: colors represent ethnicity and scaling represents Nugent score category with green (Black), turquoise (Hispanic), coral (Asian) and pink (White) for sphere colorings.

Additional File 6: Figure S3

EMPeror: Biplot with 3 PCs as axes, 20 most abundant taxa are shown and labeled along with all 394 samples. Samples are colored by ethnicity and scaled by Nugent category, with low Nugent scores presented as larger spheres.

Additional File 7: Figure S4

EMPeror: Parallel coordinates plot on 10 PCs for individuals from CST IV on the left. Contrast this plot with the parallel coordinates plot on the right for individuals from all other CSTs, those dominated by *Lactobacillus* species.

Additional File 8: Code S4

Code to take $ade4$ output and generate co-inertia plot.

Additional File 9: Table S1

Table of 'average' taxa present in each race/ethnic group along with the number of women whose OTUs contain these taxa.