COMPARISONS OF DISTANCE METHODS FOR COMBINING COVARIATES AND ABUNDANCES IN MICROBIOME STUDIES

JULIA FUKUYAMA, PAUL J. MCMURDIE
Statistics Department, Stanford University,
Stanford, CA 94305, USA
**E-mail:{jfukuyama,mcmurdie}@stanford.edu*

LES DETHLEFSEN, DAVID A. RELMAN
Departments of Microbiology & Immunology, and Medicine
Stanford University, VA Palo Alto Health Care System 154T
3801 Miranda Avenue, Palo Alto, CA 94304

SUSAN HOLMES*
Statistics Department, Stanford University,
Stanford, CA 94305, USA
**E-mail:susan@stat.stanford.edu*
www-stat.stanford.edu/˜susan/

This article compares different methods for combining abundance data, phylogenetic trees and clinical covariates in a nonparametric setting. In particular we study the output from the principal coordinates analysis on UNIFRAC and WEIGHTED UNIFRAC distances and the output from a double principal coordinate analyses DPCOA using distances computed on the phylogenetic tree.

We also present power comparisons for some of the standard tests of phylogenetic signal between different types of samples. These methods are compared both on simulated and real data sets. Our study shows that DPCoA is less robust to outliers, and more robust to small noisy fluctuations around zero.

Keywords: Distances. DPCOA. UNIFRAC. Phylogenetic Signal. Nonparametric testing.

1. Introduction

The activities of complex microbial communities are increasingly recognized as vital aspects of the biosphere, often with direct relevance for humans. Since the transition from cultivation- and microscopy-based techniques of traditional microbiology to the nucleotide sequence-based, cultivation-independent techniques predominant today, and especially with the advent of high throughput next generation sequencing technologies, microbial ecologists have had to contend with high dimensional datasets representing the abundance of hundreds or thousands of marker gene sequence variants across tens or hundreds of samples.

This challenge has resulted in the development of preprocessing pipelines such as `qiime`[1] and `mothur`[2] that deliver output in the form of abundance tables of various *phylotypes* or OTUs (An Operational Taxonomic Unit is defined only by sequence data, lacking the physiological characterization required to establish and name a traditional taxon, but serving as a proxy for a microbial species, genus or other taxonomic entity). We will not go into the difficulties of correctly assigning the sequence reads to OTUs. This is most often done by using a x% similarity rule of thumb, e.g., x in the range 97-99 for the 16S rRNA gene sequences. The assignments often include a complex denoising procedure dependent on the actual sequencing technology (454 Life Sciences, Illumina). The goals of these studies include the comparison of

bacterial communities from samples subjected to an experimental intervention, or chosen to represent a natural contrast of interest. Thus, values for a number of clinical or environmental covariates are generally associated with each sample. The analysis of these studies needs to be multivariate to capture complex high dimensional interactions. All current methods rely on computations of relevant distances between communities and their representation using standard multidimensional scaling (MDS/PCoA).

1.1. Challenges in including side information for contingency tables

The output from the standard pipelines mentioned above are contingency tables with abundances in the cells, *species*, *phylotypes* or OTUs are the rows of the tables and columns of the table represent the sampling locations, often different patients at different times. However, complementary side information is available, both about the relationships between the OTUs and about clinical/environmental covariates measured on the sampling locations. From a practical perspective this is handled by the `phyloseq`[3] package. The advantage of a specific structured data approach is that we can use many of the packages already developed for ecological and multivariate data analysis; the `R`[4] packages `ape`,[5] `picante`,[6] `ade4`,[7] `vegan`,[8] and `phyloseq`[3] were used in the current paper. Here, we will show a comparative study of some of the multivariate visualization and testing procedures available in these packages, concentrating on ways to incorporate side information effectively.

The most common statistical approach to date is to use either unweighted or weighted UniFrac[9] distances between communities. Here we compare these to DPCoA[10] (defined in section 2.2) a two step process that combines phylogenetic and abundance data for PCR sequenced phylotypes in a geometric framework.

We also provide complementary visualizations of multivariate biases and a review of available nonparametric testing procedures in the presence of important covariates.

2. Description of Techniques

2.1. UniFrac

UniFrac[11] is a distance between microbial communities for which phylogenetic information about the OTUs is available. The UniFrac distance between community A and community B is defined as the fraction of branches of the phylogenetic tree that lead to members of community A or community B but not both. This definition only considers whether an OTU is present or absent in a community and not how abundant it is.

Weighted UniFrac[12] incorporates abundances and is defined as $wUF(A,B) = \sum_i b_i |A_i/A_T - B_i/B_T|$ where the sum is over the branches of the phylogenetic tree, b_i is the length of the ith branch, A_T is the overall abundance of OTUs in community A, and A_i is the number of OTUs in community A that correspond to descendants of branch i. It is also possible to normalize weighted UniFrac by the average distance of members of the two communities to the root. This normalization can help correct for unequal sampling effort or different evolutionary rates between taxa, but for the purposes of this paper, we will take weighted UniFrac to be the raw (unnormalized) weighted UniFrac distance given above.

2.2. Double Principal Coordinates Analysis

DPCoA[13] is based on work by Rao[14] that aimed to integrate diversity and dissimilarity measures. Rao's description of diversity and distance starts with a distance between individuals and builds up to a measure of the diversity of a distribution and a dissimilarity between distributions. Briefly, he defines the diversity within a population to be the average distance between members of that population. Similarly, the diversity between two populations is the average distance between members of the two populations. He then notes that we expect the average distance between members of two different communities to be larger than the average distance between members within the individual communities. Therefore a natural distance between community i and community j is

$$RD_{ij} = H_{ij} - (H_i + H_j)/2 \tag{1}$$

where H_{ij} is the average distance between members of community i and community j, and H_i is the average distance between members of community i. These definitions of diversity within a community and distance between communities are fairly natural, and they are useful because they allow for a decomposition of diversity of a group of communities similar to the decomposition of variance in ANOVA. If we have k communities with frequencies $\lambda_1, \ldots, \lambda_k$, the diversity of all k communities taken together can be written as

$$H_0 = \sum_{i=1}^{k} \lambda_i H_i + \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i \lambda_j D_{ij} = H(w) + D(b) \tag{2}$$

where the first term is the weighted sum of the internal diversities of the communities and the second term is the weighted sum of the pairwise distances between the communities.

DPCoA is based on the distance in formula (1). The idea is to define a typology for which the inertia of the cloud of points representing the OTUs and communities decomposes the same way the quadratic entropy does. To do this, the OTU points are first positioned in a high-dimensional space in such a way that the distances between them are the same as the patristic distances defined by the tree. Then, if the community points are put at the barycenter of their OTU profiles, the distance between the community points will be the square root of Rao's distance, formula (1).

Distances between individuals in Rao's definition of diversity and distance do not have to come from a tree, so in some sense, DPCoA is more general than UniFrac and weighted UniFrac. Here we will compare the two step DPCoA approach using the patristic distance between OTUs to the standard approach of computing the weighted/unweighted UniFrac and then using this in a MDS/PCoA plot.

2.3. A single framework

Although UniFrac, weighted UniFrac, and the community distances in DPCoA come from very different theoretical perspectives, we find it useful to present them in a single framework as weighted sums over the branches of a tree. Recall that the weighted UniFrac distance is defined as

$$wUF(A, B) = \sum_i b_i |A_i/A_T - B_i/B_T| \tag{3}$$

Method	Original description	New formula	Properties
DPCoA	square root of Rao's distance based on the square root of the patristic distances	$[\sum_i b_i(A_i/A_T - B_i/B_T)^2]^{1/2}$	Most sensitive to outliers, least sensitive to noise, upweights deep differences, gives OTU locations
wUniFrac	$\sum_i b_i \lvert A_i/A_T - B_i/B_T \rvert$	$\sum_i b_i \lvert A_i/A_T - B_i/B_T \rvert$	Less sensitive to outliers/more sensitive to noise than DPCoA
UniFrac	fraction of branches leading to exactly one group	$\sum_i b_i \mathbf{1}\{\frac{A_i/A_T - B_i/B_T}{A_i/A_T + B_i/B_T} \geq 1\}$	Sensitive to noise, upweights shallow differences on the tree

Table 1: Summary of the methods under consideration. "Outliers" refers to highly abundant OTUs, and noise refers to noise in detecting low-abundance OTUs (see the text for more detail).

Evans and Matsen[15] showed that weighted UniFrac was the first Wasserstein distance on a tree and that the second Wasserstein distance on a tree was the quantity in equation (1), although they did not link that observation to Rao or to DPCoA. However, it follows from their work that we can write the distances between communities in DPCoA as

$$[DPCoA(A,B)]^2 = \sum_i b_i(A_i/A_T - B_i/B_T)^2 \tag{4}$$

with one caveat: the branch lengths of the tree that we sum over are slightly different in DPCoA compared to weighted UniFrac. This comes from the fact that in DPCoA, the inertia is supposed to decompose the same way that quadratic entropy does.

Finally, it is informative to rewrite the formula for unweighted UniFrac so that it is as similar as possible to our formulae for weighted UniFrac and DPCoA. Weighted UniFrac and DPCoA have been written in terms of b_i, the branch lengths, and A_i/A_T and B_i/B_T. Unweighted UniFrac can be written, in terms of those same variables, as

$$UF(A,B) = \begin{cases} \sum_i b_i \mathbf{1}\{\left|\frac{A_i/A_T - B_i/B_T}{A_i/A_T + B_i/B_T}\right| \geq 1\} & A_i/A_T + B_i/B_T > 0 \\ 0 & A_i/A_T + B_i/B_T = 0 \end{cases} \tag{5}$$

where $\mathbf{1}$ is the indicator function (i.e. evaluates to 1 if its argument is true and 0 otherwise). This seems like an unnecessarily complicated way of writing unweighted UniFrac, but it is useful because it puts unweighted UniFrac in a form that is more comparable to weighted UniFrac and DPCoA.

Comparing these three formulae (see table 1) makes clear the differences between the three ordination methods. Weighted UniFrac and DPCoA are quite similar: DPCoA is slightly less robust to outliers (in our case, an OTU is an outlier if it is much more abundant than the other OTUs) than weighted UniFrac, but both suppress small "noisy" fluctuations around zero. This

noise can be thought of as the noise inherent in measuring OTUs that are present in abundances near the detection limit. For example, if an OTU is present in all samples but has a very low abundance, it might be only be detected in half the samples. Unweighted UniFrac, on the other hand, is quite different. It puts much more weight on shallow branches than either DPCoA or weighted UniFrac. This makes it more sensitive to the kind of noise discussed before, but also allows it to pick up shallower differences that the other two methods suppress. We will see these properties illustrated in simulated and real data sets.

2.4. Runtimes

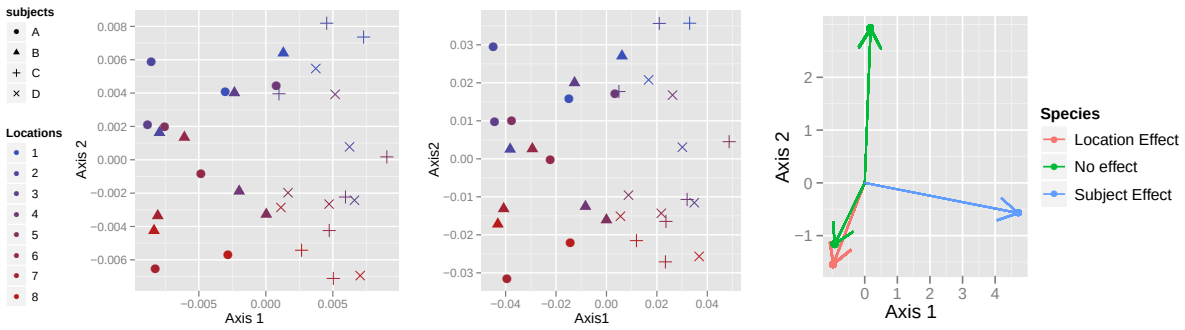
DPCoA's runtime is quadratic in the number of OTUs but is not that dependent on the number of samples. UniFrac's runtime (as implemented in `picante`), in contrast, is linear in the number of OTUs but super-linear in the number of samples. See figure S4 in the supplementary section. The fact that DPCoA is $O(n^2)$ in the number of OTUs could potentially be a problem, but we have performed DPCoA on abundance matrices with as many as 2,500 OTUs in forty minutes in a 32 CPU linux cluster.

3. Simulations

As a first pass at comparing DPCoA with PCoA using weighted UniFrac, we looked at the results of the two methods on simulated data. For the simulation, we imagine that we have four subjects (A through D), each of whom is sampled in eight locations (1 to 8). For each combination of subject and location, we have abundances for 300 species. The relationship between the species is described by a random coalescent tree. The model is that one clade varies along the location gradient (it is overrepresented in location 1 and underrepresented in location 8) and another clade varies between two groups of patients (it is overrepresented in patients A and B relative to patients C and D). This is a very simple data set, but it allows us to look at both continuous and categorical covariates.

Figure 1 shows the ordination of the simulated data by both PCoA with weighted UniFrac (a) and by DPCoA ((b) shows the community points and (c) shows the species points). MDS/PCoA with weighted UniFrac and DPCoA both show the subject effect (in both cases subjects A and B are to the left of subjects C and D) and the location effect (we see that the locations are arranged roughly in order along the second axis in both methods). DPCoA gives us some additional information, however: the locations of the species points. Since the species points and the community points are built in the same space, the plots of the communities and the species could be positioned on the same figure (we have separated them for readability purposes). In figure 1(c), blue points indicate species that we modeled as being overrepresented in subjects A and B relative to C and D, red points indicate species that we modeled as being over- or under-represented along a gradient according to location, and green indicates all other species. It looks like there are only four points in figure 1(c), but there are actually 300, many of which are located in almost exactly the same place because the structure of the simulated data is particularly simple.

Notice that the vector pointing to the center of the red (location effect) points is nearly orthogonal to the vector pointing to the center of the blue (species effect) points. This is in line with the fact that the species effect and the location effect are independent of each other.



(a) UniFrac + PCoA (b) DPCoA communities (c) DPCoA species

Fig. 1: (a) shows ordination of community points by PCoA with the UniFrac metric; (b) by DPCoA. (c) shows ordination of species points by DPCoA: blue indicates species whose levels change in subjects A/B versus C/D, red indicates species whose levels change along a gradient related to location, and green indicates all other species.

Since we can see from the community plot that the first axis is primarily a subject effect and the second axis is primarily a location effect, we could have guessed that species points that lie close to the first axis would represent species related to the subject effect and species points that lie close to the second axis would represent species related to the location effect. Since we have simulated data, we know this is true.

3.1. Robustness to noise

One form of noise present in OTU abundance tables can be thought of as noise around zero: OTUs that are actually present might not be detected if they are present in very low abundances. To look at the robustness of the two methods to this sort of noise, we looked at another simulated data set. We simulated two groups of locations and 100 species related to each other by a tree, each of which is either present or absent in each location. Splitting the species approximately in half at the root, one half of the species are primarily present in group "a" and the other half are primarily present in group "b". We have three sets of simulations, each one with a different amount of noise (see figure S1).

We then used DPCoA and PCoA/MDS with UniFrac to analyze these data sets. The ordination method should correctly form two clusters and recognize that this is essentially a one-dimensional problem (each community is in one of two groups). The results can be seen in figure 2. Both of the techniques separate the two groups along the first axis for all noise levels, but DPCoA keeps most of the variance along the first axis, while in PCoA/MDS with UniFrac, increasing the noise causes the data to spread out along the higher axes. As predicted by our results from section 2.3, DPCoA is more robust to noise than UniFrac. This simulation demonstrates how strongly UniFrac upweights shallow differences compared to DPCoA. We ran a second simulation to look at the kind of noise that is more likely to come up in 454 sequencing (that is, noise resulting in fake OTUs), and we saw a similar effect (see figure S2). Considering the comparative formulae in table 1, we should not be surprised that many different kinds of noise are suppressed by DPCoA.

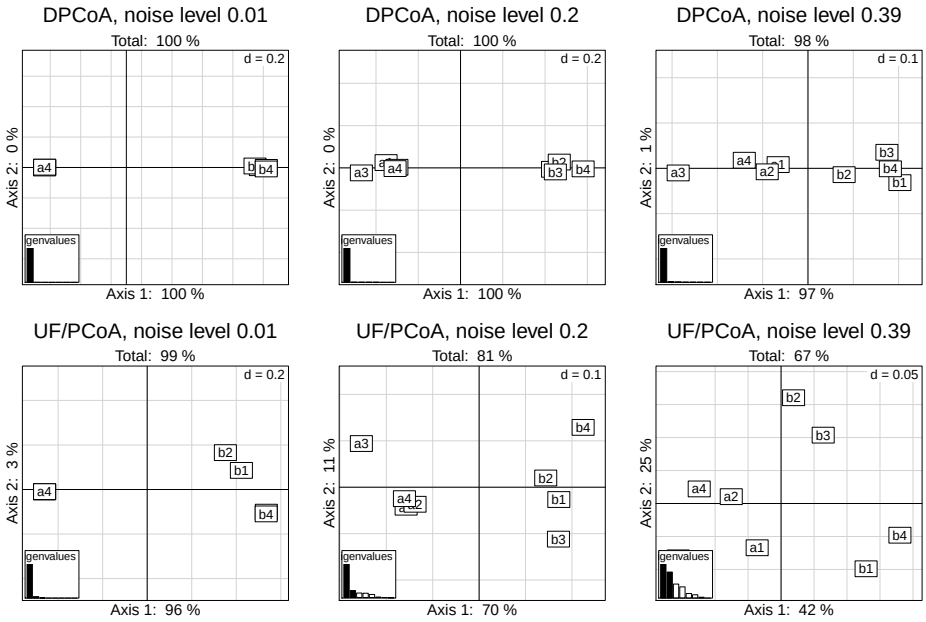


Fig. 2: Comparisons of DPCoA (row 1) and PCoA/MDS with unweighted UniFrac (row 2) for different noise levels. Columns correspond to data sets with noise levels .01, .2, and .39.

4. A Real Dataset

We then looked at the two methods on a real dataset. The data come from stool samples of three subjects, each of whom took two courses of *ciprofloxacin* over the course of ten months. Each patient was sampled about fifty times over those ten months, and we have abundance data for about 2500 OTUs for each time point and patient. The OTUs are related to each other by a phylogenetic tree, and the time points are categorized as `pre-cipro`, `1st cipro`, 1st week post *cipro*, interim, `2nd cipro`, 2nd week post *cipro*, and `post-cipro`. We looked at the data using unweighted UniFrac with PCoA/MDS, weighted UniFrac with PCoA/MDS, and with DPCoA.

4.1. PCoA with UniFrac

We can see from figure 3 that the unweighted UniFrac distance emphasizes different aspects of the data from weighted UniFrac. Unweighted UniFrac separates the subjects into fairly distinct clusters, while weighted UniFrac shows much less of a subject effect. We know that the primary difference between unweighted and weighted UniFrac is that unweighted UniFrac upweights shallow differences, and we can therefore infer from the plots that the differences between the subjects that we see in the unweighted UniFrac plot are probably due to shallow differences in species composition. To show that the difference between unweighted and weighted UniFrac is not merely due to the fact that weighted UniFrac takes into account abundances, we also show the result of using weighted UniFrac on presence/absence data. Comparing figures 3(c) and (b) shows us that the fact that weighted UniFrac can "see" abundances whereas unweighted UniFrac cannot is not the source of the difference between the two methods.

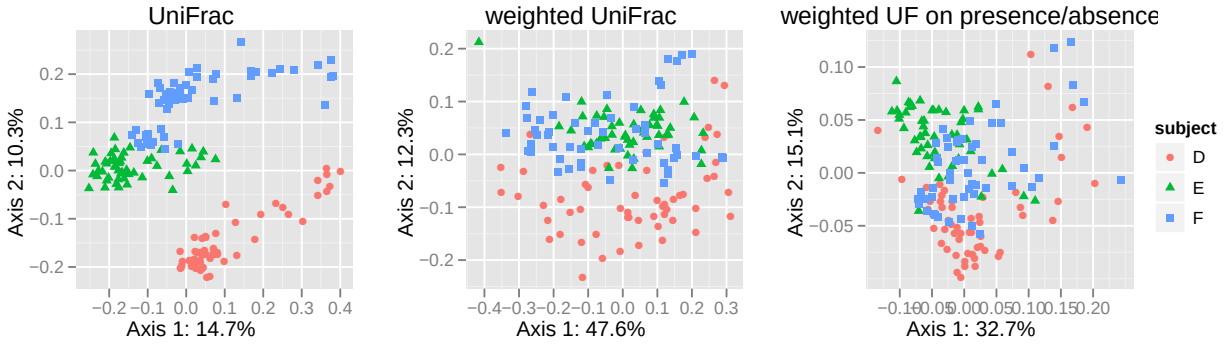


Fig. 3: Comparing the UniFrac variants. From left to right: PCoA/MDS with unweighted UniFrac, with weighted UniFrac, and with weighted UniFrac performed on presence/absence data extracted from the abundance data used in the other two plots.

4.2. DPCoA

The result of DPCoA can be seen in figure 4 (b) and (c). Since, as we saw from equation 4, DPCoA is more sensitive to outliers than either weighted or unweighted UniFrac, we had to remove any outliers, which would otherwise dominate the ordination. The result of DPCoA on the full abundance matrix is given in figure S3 for comparison. The second axis seems to separate the three subjects to a greater extent than weighted UniFrac, but not as much as unweighted UniFrac. The OTU plot can give us some more insight into the ordination given

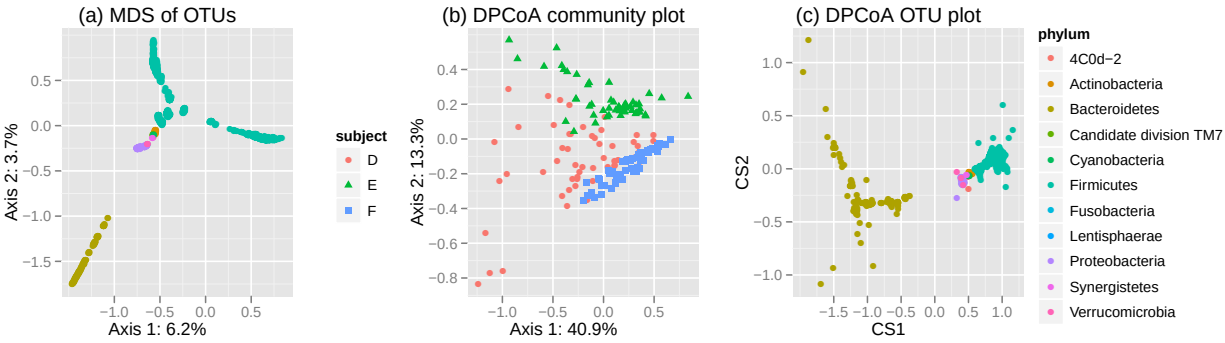


Fig. 4: (a) PCoA/MDS of the OTUs based on the patristic distance, (b) community and (c) species points for DPCoA after removing two outlying species.

by DPCoA, particularly when we compare it to PCoA/MDS of the OTUs. Since the OTU plot in DPCoA shows the OTUs from the perspective that maximizes the inertia of the community points, while a simple PCoA/MDS of the OTUs would show the OTUs from the perspective that maximizes the inertia of the OTUs, groups of OTUs that are more spread out in DPCoA compared to PCoA/MDS are likely to be important to the ordination. When we compare the OTU plot from DPCoA versus the OTU plot from PCoA/MDS, we see that in the DPCoA OTU plot, *Bacteroidetes* are much more spread out in the DPCoA plot, and *Firmicutes* are much

closer together. This indicates that the first axis primarily represents the difference betwen *Bacteroidetes* and the rest of the tree, and the second axis primarily represents differences in the specific *Bacteroidetes* OTUs present in the different communities.

4.3. Antibiotic Stress

We next wanted to visualize the effect of the antibiotic. Figure 5 shows the ordinations of the communities due to DPCoA and UniFrac with information about the whether the community was stressed or not stressed (pre cipro, interim, and post cipro were considered "not stressed", while first cipro, first week post cipro, second cipro, and second week post cipro were considered "stressed"). We see that for UniFrac, the first axis seems to separate the stressed communities from the not stressed communities. DPCoA also seems to separate the out the stressed communities along the first axis (in the direction associated with *Bacteroidetes*), although only for subjects D and E. Since UniFrac emphasizes shallow differences on the tree and since

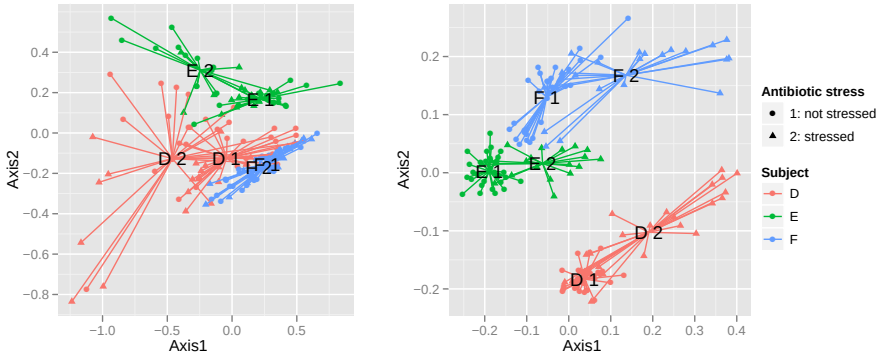


Fig. 5: Community points as represented by DPCoA (left) and PCoA/MDS with unweighted UniFrac (right). The labels represent subject plus antibiotic condition. The ordinations are the same as in figures 4 and 3.

PCoA/MDS with UniFrac seems to separate the subjects from each other better than the other two methods, we can conclude that the differences between subjects are mainly shallow ones. However, DPCoA also separates the subjects and the stressed versus non-stressed communities, and examining the community and OTU ordinations can tell us about the differences in the compositions of these communities.

5. One step further: confirmatory analyses

The previous methods were purely exploratory and allow us to visualize the effect of the covariates in principal coordinate plots. In reality we would like to be able to test the effect of the covariate through a non parametric test. There are a number of methods available for this sort of testing. One example is the UniFrac test, which tests whether two groups have significantly different species compositions by calculating the amount of branch length unique to each environment, permuting the labels, and re-calculating the amount of unique branch length. The p-value is then the fraction of trials for which the permuted unique branch length was longer than the observed unique branch length. The downside of this test is that it only

tests differences between individual locations or communities, not groups of communities, and so it is more useful when you are working with small numbers of communities.

`adonis` (in the package `vegan`) provides a nonparametric multivariate analysis of variance using distances. Given a set of distances between observations, it decomposes the weighted sums of distances according to a linear model and calculates p-values associated with those decompositions by permutation of the labels.

Mantel's test[16] and the RV test (both implemented in `ade4`) test correlation between matrices. Mantel's test takes two distance matrices X and Y and uses as a test statistic $\sum_{i<j} X_{ij}Y_{ij}$. The null distribution is generally estimated by a Monte Carlo method: keeping the entries of one matrix fixed, permuting the entries in the rows and columns of the other matrix, and recomputing the test statistic. The RV test is similar, it is based on a multi-table generalization of the correlation coefficient. ($RV(A,B) \propto Trace(t(A)B)$).[17] This nonparametric test estimates the null distribution of the RV coefficient between two matrices by permuting the rows of one matrix and recomputing the RV coefficient for each permutation.

Finally, the Abouheif test tests traits for a phylogenetic signal. It was originally described as a version of a test for serial independence on the leaves of a phylogenetic tree where the null distribution was determined by permutation of the leaves,[18] but was later shown to be a a version of Moran's I with a certain distance on the tree.[19]

5.1. Testing our data

Our ordination of the simulated data suggested two hypotheses to test: first, that subjects A and B are different from subjects C and D, and second, that the species composition of each subject changes along a gradient associated with the location. Since the simulated data looked fairly similar in DPCoA and PCoA with the UniFrac distance, we will do all the testing on the DPCoA results. To test whether the species composition changes along the location gradient, we can use Mantel's test, the RV test, or `adonis`. For Mantel's test and the RV test, we compare the distances between communities given by DPCoA with distances between the covariates (for example, if we were testing a location effect, we could test the DPCoA distances with the distances between the locations). For `adonis`, we will be testing whether location explains a significant amount of the distances between the communities (distances again as given by DPCoA). Figure 6 shows the p-values we get from each test for varying magnitudes of the effect (0 is no location effect, and 5 is a pretty strong location effect), and we can see that the Mantel is the most sensitive here, followed by `adonis`, followed by the RV test.

To test whether species composition is different in the group containing subjects A and B compared to the group containing subjects C and D, we can use adonis, the Abouheif test, or the UniFrac test. For the Abouheif test, we can look for a phylogenetic signal in the difference in the abundance of each species between subjects A/B and C/D. Adonis, as before, will look at the amount of variance explained by the groups. For the UniFrac test, if we just want to test the difference between the two groups of subjects, we need to take the average species composition over all the samples from subjects A and B and compare it to the average species composition over all samples from subjects C and D. If the distance between the two averages is significant, we can say that the difference between the groups is significant. Figure 6(b) shows the p-values for the Abouheif, adonis, and UniFrac tests at different magnitudes of the

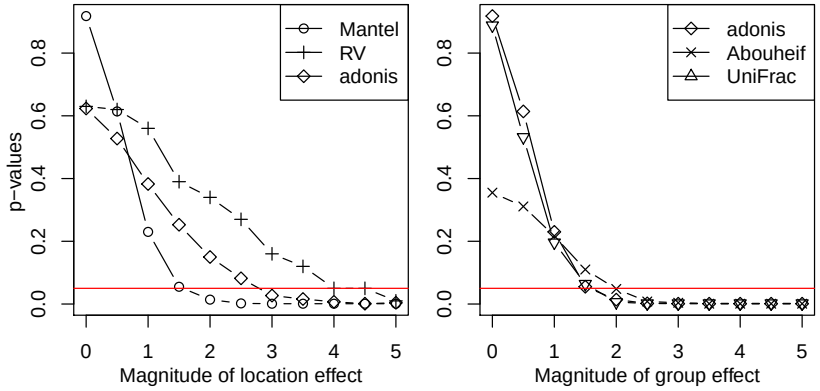


Fig. 6: The left shows the p-values of the Mantel (circles), RV (pluses), and `adonis` (diamonds) tests for a location effect at different magnitudes of the effect. The right shows the p-values of the Abouheif (x's), `adonis` (diamonds), and UniFrac (triangles) tests at different magnitudes of the group effect.

group effect. We see that all three tests have similar performance on these data, and the adonis and UniFrac tests give remarkably similar results.

Our ordination of the real data suggested that we should test whether there is a difference in intestinal bacteria composition between times of antibiotic stress and no stress. Adonis seems to be the most appropriate method in presence of a patient effect, an antibiotic stress effect, and maybe an interaction between the two. Carrying out this analysis gives the result that, for both the distances given by DPCoA and the distances given by UniFrac, the patient effect and the antibiotic stress effect are significant at the .001 level. The interaction term between patient and antibiotic stress is significant at the .05 level for the UniFrac distances and at the .001 level for the DPCoA distances. We can also try the UniFrac test on this data, aggregating by subject or by stress condition as before. We find that the unweighted UniFrac test gives a significant result for both the subject effect and the antibiotic stress effect, but the weighted UniFrac test gives significant results for neither.

6. Summary

This article presents comparisons of some of the current techniques available for analyzing abundance tables in the presence of side information on the OTUs and covariates on the samples. The classical framework for analyzing abundance tables using parametric models or multivariate analysis of variance does not apply in the microbiome studies and recent papers have used UniFrac[20] and DPCoA[21] with success. Both approaches are similar in their use of phylogenetic distances between OTUs and projections using PCoA. However whereas the UniFrac approach allows the comparison of two samples at a time, DPCoA provides a more complex framework enabling different categorical and continuous covariates to be taken into account. In our comparisons of weighted and unweighted UniFrac approaches the unweighted UniFrac seems to distinguish a much clearer separation between subjects for instance, both on simulated and real data. We compared DPCoA and unweighted UniFrac on simulated data to test their performances in the presence of noise. The results show that when there is a

simple one dimensional contrast between the data, DPCoA is impervious to a substantial noise component, whereas the first axis in an unweighted UniFrac PCoA plot has a much smaller eigenvalue. The consequence in more complex studies is that the subsequent components of variation can be interleaved with the main phylogenetic effect leading to difficulties in interpreting the results.

Acknowledgments

We thank Elisabeth Purdom, Sam Pimentel, Kris Sankaran, Yana Hoy, Katie Shelef as well as two anonymous referees for useful comments and motivating questions. J.F. was partially supported by a NSF-DMS-VIGRE grant. P. J. M. and S. H. were supported by the NIH Grant NIH-5-R01GM086884 D. A. R. is funded by the Doris Duke Charitable Trust and a National Institutes of Health Pioneer Award (DP1OD000964), and by the Thomas C. and Joan M. Merigan Endowment at Stanford University.
Supplementary information: `http://stat.stanford.edu/~susan/projects/psb2012.pdf`.

References

1. J. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. Bushman, E. Costello, N. Fierer, A. Peña, J. Goodrich, J. Gordon and R. Knight, *Nature methods* **7**, 335 (2010).
2. P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn and C. F. Weber, *Appl. and environmental microbiology* **75**, 7537 (November 2009).
3. P. J. McMurdie and S. Holmes, Phyloseq: A bioconductor package for handling and analysis of high-throughput phylogenetic sequence data.
4. R. Ihaka and R. Gentleman, *Journal of Computational and Graphical Statistics* **5**, 299 (1996).
5. E. Paradis, J. Claude and K. Strimmer, *Bioinformatics* **20**, 289 (2004).
6. S. Kembel, P. Cowan, M. Helmus, W. Cornwell, H. Morlon, D. Ackerly, S. Blomberg and C. Webb, *Bioinformatics* **26**, 1463 (2010).
7. D. Chessel, A. Dufour and J. Thioulouse, *R News* **4**, 5 (2004).
8. J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, R. G. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens and H. Wagner, *vegan:Community Ecology Package*, (2010).
9. M. Hamady, C. Lozupone and R. Knight, *The ISME Journal* (Jan 2009).
10. E. Purdom, *Annals of Applied Statistics* (Jul 2010).
11. C. Lozupone and R. Knight, *Applied and environmental microbiology* **71**, p. 8228 (2005).
12. C. Lozupone, M. Hamady, S. Kelley and R. Knight, *Applied and environmental microbiology* **73**, p. 1576 (2007).
13. S. Pavoine, A. Dufour and D. Chessel, *Journal of theoretical biology* **228**, 523 (2004).
14. C. R. Rao, *Theoretical Population Biology* **21**, 24 (1982).
15. S. N. Evans and F. A. Matsen, *arXiv* **q-bio.PE** (Jan 2010).
16. N. Mantel, *Cancer research* **27**, p. 209 (1967).
17. P. Robert and Y. Escoufier, *Jour of the Roy. Stat. Soc.. Series C* **25**, 257 (1976).
18. E. Abouheif, *Evolutionary Ecology Research* **1**, 895 (1999).
19. S. Pavoine, S. Ollier, D. Pontier and D. Chessel, *Theoretical population biology* **73**, 79 (2008).
20. R. Ley, M. Hamady, C. Lozupone, P. Turnbaugh, R. Ramey, J. Bircher, M. Schlegel, T. Tucker, M. Schrenzel, R. Knight *et al.*, *Science* **320**, p. 1647 (2008).
21. P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson and D. A. Relman, *Science* **308**, 1635 (Jun 2005).