MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 242–249 doi: 10.1111/1755-0998.12300

Considering external information to improve the phylogenetic comparison of microbial communities: a new approach based on constrained Double Principal Coordinates Analysis (cDPCoA)

S. DRAY,*† S. PAVOINE‡§ and D. AGUIRRE DE CÁRCER¶
Université de Lyon, F-69000, Lyon, France, †CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, F-69622, Villeurbanne, France, ‡Muséum national d'Histoire naturelle, Département Ecologie et Gestion de la Biodiversité, UMR 7204-CNRS-UPMC, CP 51, 55-61 rue Buffon, 75005 Paris, France, §Mathematical Ecology Research Group, Department of Zoology, University of Oxford, South Parks Road, OX1 3PS, Oxford, UK, ¶Department of Virology and Microbiology, Centro de Biología Molecular Severo Ochoa CSIC/UAM, 28049 Madrid, Spain

Abstract

The use of next-generation sequencing technologies is revolutionizing microbial ecology by allowing a deep phylogenetic coverage of tens to thousands of samples simultaneously. Double Principal Coordinates Analysis (DPCoA) is a multivariate method, developed in community ecology, able to integrate a distance matrix describing differences among species (e.g. phylogenetic distances) in the analysis of a species abundance matrix. This ordination technique has been used recently to describe microbial communities taking into account phylogenetic relatedness. In this work, we extend DPCoA to integrate the information of external variables measured on communities. The constrained Double Principal Coordinates Analysis (cDPCoA) is able to enforce *a priori* classifications to retrieve subtle differences and (or) remove the effect of confounding factors. We describe the main principles of this new approach and demonstrate its usefulness by providing application examples based on published 16S rRNA gene data sets.

Keywords: 16S, 18S, DPCoA, microbial community analysis, QIIME, Unifrac

Received 3 February 2014; accepted 25 June 2014

Introduction

The use of next-generation sequencing technologies is revolutionizing the way researchers study microbial ecosystems. The most common approach to microbial community analysis includes the retrieval from environmental samples of phylogenetic markers by PCR, such as the 16S/18S rRNA gene, using primers carrying a nucleotide tag or barcode specific for each sample (Parameswaran *et al.* 2007). The resulting products are then pooled and sequenced *en masse*, with the barcode sequence later serving as the basis for sample origin determination of each sequence obtained. The use of this approach enables deep phylogenetic coverage of tens to thousands of samples at the same time.

The Unifrac metric (Lozupone & Knight 2005) is commonly used together with ordination or clustering techniques to study the phylogenetic relationships between microbial communities. While this approach is a standard method in microbial ecology, Eckburg *et al.* (2005) used the Double Principal Coordinate Analysis (DPCoA) as an alternative to study the diversity of the human intestinal microbiota using 16S rRNA gene data sets. DPCoA (Pavoine *et al.* 2004) has been developed in community ecology as an ordination technique able to integrate a matrix describing differences among species in the analysis of a species abundance matrix. DPCoA output relies on a simultaneous representation of species and community points in a reduced space (factorial map) where distances among species preserve the original differences and are used to compute the positions of the community points. The main differences between the two approaches (Unifrac and DPCoA) are the definition of phylogenetic distances among communities (see below) and the possibility with DPCoA to visualize the organisms that drive the phylogenetic dissimilarities among communities, as both communities and species can be plotted in the same space. In DPCoA, the phylogenetic distance among two sequences can be defined as

Correspondence: Daniel Aguirre de Cárcer, Fax: +34-911964420; E-mail: daniel.aguirre@cbm.csic.es

the square root of sum of branch lengths in the smallest path that connects them on the phylogenetic tree (de Vienne *et al.* 2011). Then the phylogenetic distance between two communities is defined as the average phylogenetic distance among two individuals from the pool of the two communities minus the average phylogenetic distance among two individuals drawn from the same community. On the other hand, the Unifrac metric measures the difference between two communities as the fraction of the branch length of the phylogenetic tree of sequences that is unique to one community or the other.

In many cases, external variables are recorded on communities (e.g. habitat type, host health status), and it should be useful to consider this information to improve the ecological interpretation of microbial community comparisons. Standard (partial) canonical ordination methods (e.g. redundancy analysis or canonical correspondence analysis) could be useful in this context, but they focus only on differences in species compositions and thus implicitly assume that species characteristics are similar as they are not able to consider phylogenetic information in the analysis. In this paper, we show the usefulness of the constrained DPCoA (cDPCoA) that allows constraining the phylogenetic-based study of microbial community data by highlighting or removing the effect of external variables. As DPCoA is a particular case of a general statistical framework (duality diagram theory, Dray 2007), it can be easily extended to introduce external information on communities while retaining the ability to visualize species contributions (see below for mathematical details). Here, we consider only the case where external variables are qualitative, that is communities are grouped into levels of a factor (e.g. countries or habitats). DPCoA can be constrained in two different ways. The first approach, referred to as a 'between-class analysis' (BCA), compares community compositions by highlighting the differences among the levels of the factor (analysis of the average phylogenetic differences among, for example habitats). In contrast, the second approach referred to as a 'within-class analysis' (WCA) compares community compositions by controlling the effect of the external variable; the average (phylogenetic) differences among the levels of the factor are removed and the analysis focuses on the residual (phylogenetic) differences among the communities. In this second approach, also known as partial analysis (see for instance Wesuls *et al.* 2012), the external factor is considered as a covariable (cofactor), while it is considered as an explanatory variable in the first approach. Note that both approaches could also be mixed if two factors are considered. For instance, the between-class analysis can be applied to highlight the effect of a factor after the effect of a cofactor has been partialled out (ter Braak 1988) (see also Pavoine *et al.* 2013). Such approaches could also probably be developed with the Unifrac metric, but to our knowledge, have so far not been reported.

Methods

Mathematical notations

Consider $m_\bullet$ communities that contain individuals belonging to S species. Consider that the communities are clustered into r groups (e.g. r levels of a factor); the i-th group contains m_i communities ($m_\bullet = \Sigma_{i=1}^{r} m_i$). Let n_{ijk} be the abundance of species k in a community j belonging to group i; $n_{ij\bullet} = \Sigma_{k=1}^{S} n_{ijk}$ the number of individuals in the community; $n_{i\bullet\bullet} = \Sigma_{j=1}^{m_i} n_{ij\bullet}$ the number of individuals in group i as a whole; and $n_{\bullet\bullet\bullet} = \Sigma_{i=1}^{r} n_{i\bullet\bullet}$ the total number of individuals in the data set. The proportion of species k in community j associated with group i may be defined as $p_{ijk} = n_{ijk}/n_{ij\bullet}$. The weight attributed to community j of group i is $\mu_{ij} = n_{ij\bullet}/n_{i\bullet\bullet}$ and the weight attributed to group i, $\lambda_i = n_{i\bullet\bullet}/n_{\bullet\bullet\bullet}$. We have $\Sigma_{j=1}^{m_i} \mu_{ij} = 1$ for all i, *and* $\Sigma_{i=1}^{r} \lambda_i = 1$. *Let* $\mathbf{p}_{ij} = (p_{ij1}, \ldots, p_{ijk}, \ldots, p_{ijS})^t$ (t stands for transpose) *be the vector of proportions of the species in community* j of group i, with $\mathbf{p}_{ij}{}^t\mathbf{1} = 1$; $\mathbf{p}_{i\bullet} = \Sigma_{j=1}^{m_i} \mu_{ij} \mathbf{p}_{ij}$ the vector of proportions of the species in group i as a whole ($\mathbf{p}_{i\bullet} = (p_{i\bullet 1}, \ldots, p_{i\bullet k}, \ldots, p_{i\bullet S})^t$); and $p_{\bullet\bullet} = \Sigma_{i=1}^{r} \lambda_i \mathbf{p}_{i\bullet}$ the vector of proportions of the species in the whole data set ($\mathbf{p}_{\bullet\bullet} = (p_{\bullet\bullet 1}, \ldots, p_{\bullet\bullet k}, \ldots, p_{\bullet\bullet S})^t$). Let $\mathbf{\Delta}$ be a matrix of (e.g. phylogenetic) distances among all species.

Common space for species, communities and groups

Our approach is restricted to matrices $\mathbf{\Delta} = (\delta_{kl})$, for all k, $l = 1 \ldots S$, with Euclidean properties so that it is possible to define a Euclidean space in which each species k will be positioned at a point M_k with $\|M_k M_l\| = \delta_{kl}$ for all k and l. Many distances used in biology satisfy these properties (Gower & Legendre 1986). For example, the phylogenetic distance between two species, defined as the square root of the sum of branch lengths in the smallest path that connects them on the phylogenetic tree, has Euclidean properties (de Vienne *et al.* 2011). Alternatively, for any distance matrix that does not have Euclidean properties, simple transformations exist to render the matrix Euclidean (e.g. Caillez 1983, Lingoes 1971). Let $\mathbf{M}$ be the $S \times n$ matrix of coordinates, with points (species) as rows and principal axes as columns (n is the dimension of the Euclidean space). The space is obtained by a weighted principal coordinate analysis, with species' weights given by vector $\mathbf{p}_{\bullet\bullet}$ (Gower 1984). Each community is then positioned at the centroid of the species points that occur in it: a community j from group i with the vector of proportions $\mathbf{p}_{ij}$ is positioned at point C_{ij} with the vector of coordinates $\mathbf{c}_{ij} = \mathbf{p}_{ij}^t \mathbf{M}$. Each group

is positioned at the centroid of the species points that belong to it: a group i with the vector of proportions $\mathbf{p}_{i\bullet}$ is positioned at point G_i with the vector of coordinates $\mathbf{g}_i = \mathbf{p}_{i\bullet}^t\mathbf{M}$. It is thus possible to obtain a Euclidean space in which all species, communities and groups are positioned simultaneously. In this space, all clouds of points are centred: $\mathbf{M}\mathbf{p}_{\bullet\bullet} = 0$, $\Sigma_{i=1}^{r}\lambda_i\Sigma_{j=1}^{m_i}\mu_{ij}\mathbf{c}_{ij} = 0$, $\Sigma_{i=1}^{r}\lambda_i\mathbf{g}_i = 0$, where $\mathbf{0}$ is the $n \times 1$ vector of zeros. Let $\mathbf{C}$ be the $m_\bullet \times n$ matrix providing the coordinates of all communities and $\mathbf{G}$ be the $r \times n$ matrix providing the coordinates of all groups. These coordinates thus depend on which species compose the communities and groups and how (phylogenetically) distant they are.

DPCoA

To analyse the differences among all communities without consideration of the groups, one can obtain the principal axes of the community points weighted by $\lambda_i\mu_{ij}$ (by a weighted principal component analysis). This approach corresponds to DPCoA (Pavoine *et al.* 2004). Let $\mathbf{W}_C = \mathrm{diag}\{\lambda_i\mu_{ij}\}$ be the diagonal matrix with community weights and $\mathbf{C}^t\mathbf{W}_C\mathbf{C} = \mathbf{U}\Upsilon\mathbf{U}^t$. Υ is the diagonal matrix with eigenvalues, and $\mathbf{U}$ contains eigenvectors which define the principal axes of the community points. The coordinates of the communities are given by $\mathbf{CU}$, and the coordinates of the species are given by $\mathbf{MU}$.

Between-DPCoA

To analyse the differences between the groups, one can obtain the principal axes of the group points (averages per group) weighted by λ_i. We designate this approach 'Between-DPCoA'. Let $\mathbf{W}_G = \mathrm{diag}\{\lambda_i\}$ be the diagonal matrix with group weights and $\mathbf{G}^t\mathbf{W}_G\mathbf{G} = \mathbf{V}\Psi\mathbf{V}^t$. Ψ is the diagonal matrix with eigenvalues, and $\mathbf{V}$ contains eigenvectors, for example the new principal axes. All coordinates are projected on these principal axes: the coordinates of the groups, communities and species are thus given by $\mathbf{GV}$, $\mathbf{CV}$ and $\mathbf{MV}$, respectively.

Within-DPCoA

Another interesting approach consists in analysing the distances among communities after having partialled out the distances among groups. The first step of this analysis is to modify the positions of the communities as follows: the new position of community j in group i has coordinates $\mathbf{c}_{ij} - \mathbf{g}_i = (\mathbf{p}_{ij} - \mathbf{p}_{i\bullet})^t\mathbf{M}$. In this new configuration, the groups are now all positioned at the origin of the space, whereas the positions of the species are unchanged. To analyse the differences between the communities within the groups, one can obtain the principal axes of these new community points weighted by $\lambda_i\mu_{ij}$. We designate this approach 'Within-DPCoA'. Let $\mathbf{N}$ be the $m_\bullet \times n$ matrix providing the new coordinates of all communities. Let $\mathbf{N}^t\mathbf{W}_C\mathbf{N} = \mathbf{Z}\Xi\mathbf{Z}^t$. Ξ is the diagonal matrix with eigenvalues, and $\mathbf{Z}$ contains eigenvectors (the new principal axes). The communities and the species are projected on these principal axes: their coordinates are $\mathbf{NZ}$ and $\mathbf{MZ}$, respectively.

Particular case of two factors

Now consider that the communities are clustered according to the levels of two factors A and B. Here μ_{ij} is the weight attributed to the community ij associated with level i of factor A and level j of factor B. These weights have to be positive and to sum to 1. In our approach, some combinations of the two factors can be missing. For example, in one of our case studies (see below), there is no shrubland community with pH lower than 4. The approach starts with the matrices $\mathbf{M}$ and $\mathbf{C}$ defined above that contain preliminary coordinates for the species and communities, respectively. It also considers a new writing of matrix $\mathbf{W}_C$ that contains the global weights attributed to the communities: $\mathbf{W}_C = \mathrm{diag}\{\mu_{ij}\}$. With the notations defined above, μ_{ij} is equal to $n_{ij\bullet}/n_{\bullet\bullet\bullet}$ (the relative number of individuals observed in community ij).

Consider that we intend to remove the effects of a factor B before analysing the effects of a factor A. A solution was introduced in the context of partial redundancy analysis (Sabatier *et al.* 1989), partial canonical correspondence analysis (ter Braak 1988). This solution is also discussed in Pavoine *et al.* (2013). We extend it here to the case where some combinations of the two factors can be missing.

Let $\mathbf{U}_A$ be the matrix with communities as rows and levels of factor A as columns. The entry at the i^{th} row and j^{th} column contains 1 if the community i is associated with the j^{th} level of factor A and 0 otherwise. Let $\mathbf{U}_B$ be the matrix with communities as rows and levels of factor B as columns. The entry at the i^{th} row and j^{th} column contains 1 if the community i is associated with the j^{th} level of factor B and 0 otherwise.

Let $\mathbf{P}_{\mathbf{U}_B}^{\perp} = \mathbf{I}_{rm} - \mathbf{U}_B\left(\mathbf{U}_B^t\mathbf{W}_C\mathbf{U}_B\right)^{-1}\mathbf{U}_B^t\mathbf{W}_C$ the projector on the orthogonal complement of the subspace generated by $\mathbf{U}_B$. The approach consists in projecting all points (species, communities and levels of factor A) on the subspace (A/B) generated by the matrix $\mathbf{P}_{\mathbf{U}_B}^{\perp}\mathbf{U}_A$. The projector on this subspace is $\mathbf{P}_{(A/B)} = \mathbf{P}_{\mathbf{U}_B}^{\perp}\mathbf{U}_A\left(\mathbf{U}_A^t\mathbf{W}_C\mathbf{P}_{\mathbf{U}_B}^{\perp}\mathbf{U}_A\right)^{-1}\mathbf{U}_A^t\mathbf{W}_C\mathbf{P}_{\mathbf{U}_B}^{\perp}$. The remaining steps are similar to those of the Within-DPCoA. Let $\mathbf{N}_{(A/B)} = \mathbf{P}_{(A/B)}\mathbf{C}$ be the matrix providing the new coordinates of all communities (grouped by levels of factor A). Let $\mathbf{N}_{(A/B)}^t\mathbf{W}_C\mathbf{N}_{(A/B)} = \mathbf{Z}_{(A/B)}\Xi_{(A/B)}\mathbf{Z}_{(A/B)}^t$. $\Xi_{(A/B)}$ is the diagonal matrix with eigenvalues, and $\mathbf{Z}_{(A/B)}$ contains

eigenvectors (the new principal axes). The communities and the species are projected on these principal axes: their coordinates are $\mathbf{N}_{(A/B)}\mathbf{Z}_{(A/B)}$ and $\mathbf{MZ}_{(A/B)}$, respectively.

Testing procedure

As in several other ordination methods, a permutation testing procedure (Manly 1997) could be used to evaluate the statistical significance of the effect of the constraining factor. The ratio of the total inertia of the Between-DPCoA divided by the total inertia of the DPCoA (i.e. the ratio of the sum of its eigenvalues: $trace(\Psi)/trace(\Upsilon)$) is used as the statistic of the test. It measures the part of the total phylogenetic variability of the communities that is explained by the external variable. To test the null hypothesis that the constraining factor has no effect on the phylogenetic variability, the observed value of the statistic is compared to the distribution of values obtained after a random permutation of the communities in the groups. Note that, as in standard ANOVA, the total inertia is fully decomposed in an additive ways in two components corresponding to between- and within-groups inertia (i.e. we have $trace(\Upsilon) = trace(\Psi) + trace(\Xi)$).

Data analysis

We provide examples of the utility of cDPCoA with 16S rRNA gene-based microbial community data sets employing Lauber and co-workers' data from 88 soil communities (2009), and Dethlefsen & Relman's data on antibiotic-driven perturbations of the distal gut microbiota (2011). In the first case, raw sequences were obtained from Genbank's SRA, then processed using QIIME (Caporaso *et al.* 2010). Briefly, sequences were filtered for appropriate length and quality values, clustered into OTUs at 0.95 distance threshold and assigned to each original sample according to its barcode information. The resultant table (OTUs abundance by samples) was subsampled to a common depth to avoid bias, and singletons in each sample were removed to decrease the complexity of the overall data set (Aguirre de Cárcer *et al.* 2011). Finally, the most abundant sequence from each OTU was chosen as representative of that cluster. Its taxonomic affiliation was obtained using QIIME, and a distance matrix between all representative sequences was obtained using MOTHUR (Schloss *et al.* 2009) with default parameters. In the case of the gut microbiota data set, we directly used the OTU table made available by the authors as SI, and employed the accession numbers defined for each OTU to obtain their reference sequences from the database. Then, we subsampled to a common depth, removed singletons and obtained a distance matrix between representative sequences as above. Both abundance and distance matrices were subjected to DPCoA and cDPCoA (using, as external variables, pH range and habitat type defined according to EnvO ontology 'feature' http://environmentontology.org/ for the soils data set, while subject and [treatment] stage were used in the case of the gut microbiota data set). We used the R package *ade4* (Dray 2007), including the newly developed bca.dpcoa (Between-DPCoA), wca.dpcoa (Within-DPCoA) and bwca.dpcoa (for two factors analysis) functions. Data and R scripts to reproduce all analyses are available at ftp://pbil.univ-lyon1.fr/pub/datasets/dray/MER2014/.

Results

The initial DPCoA revealed (Fig. 1a) a community distribution very similar to that previously reported by Lauber *et al.* (2009) in which soil pH is the main driver of community composition. As shown in Fig. 1b, a particular group within the *Acidobacteriaceae* (A) decreases with increasing pH, with several other groups showing the opposite behaviour. Lauber *et al.* did not find overall community differences driven by vegetation type (EnvO ontology 'feature') using a nonmetric multidimensional scaling based on Unifrac distances. However, such differences could lay hidden beneath the stronger pH effect. To test this hypothesis, partial canonical correspondence analysis can be applied to remove the pH effect and evidence vegetation type driven effects (Fig. 2). Differences between soil types were detected, but it was difficult to identify key groups characterizing each soil type (Fig. 2b).

We applied Between-DPCoA using vegetation type as a constraining factor to study the effect of other environmental parameters on community composition. This analysis explained 21.5% ($p = 0.001$) of the variation of the phylogenetic diversity among communities (sum of eigenvalues for Between-DPCoA divided by the sum of eigenvalues for DPCoA). However, the results were obscured by the observed strong pH effect (see Fig. 1c,d). Thus, we applied Within-DPCoA to remove the pH effect. The part of the total variation in community composition not explained by the pH was equal to 54.5% (sum of eigenvalues for Within-DPCoA divided by the sum of eigenvalues for DPCoA). Obviously, no pH nor vegetation type effect were identified (Fig. 1, panels e and f). Both Between-DPCoA and Within-DPCoA can be combined to focus on differences between vegetation types after removing pH effect (Fig. 1, panels g and h). This new analysis focused on 5.45% ($p = 0.001$) of the total variation in phylogenetic composition and corresponded to differences between vegetation types that were not explained by the pH effect. It

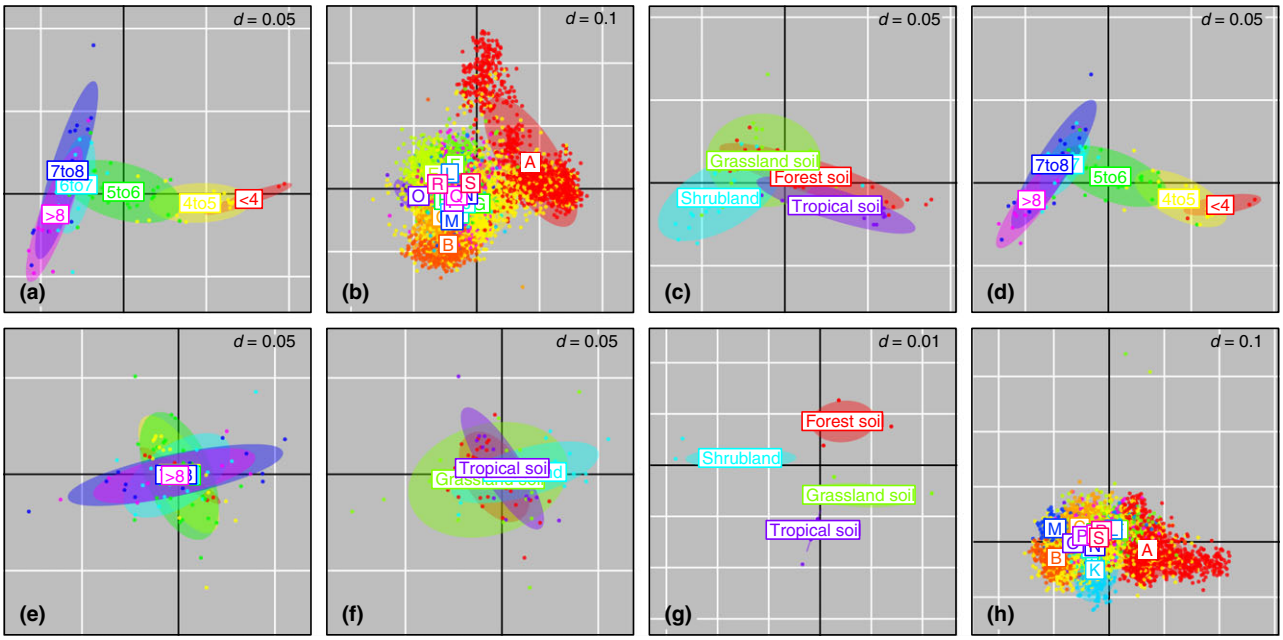

Fig 1 Panel (a); DPCoA of Lauber *et al*. data set. Soil communities have been grouped by pH range, and the ellipsoids represent their variability. Panel (b); Same DPCoA as in panel (a), but the effect of each OTU in community variation is shown. OTUs are coloured according to RDP II classifier results. Panels (c) and (d); Results of a cDPCoA using vegetation type as constraining factor (Between-DPCoA) and with communities grouped by vegetation type (c) or pH (d). Panel D demonstrates that pH range is still driving the vegetation type-related variation. Panels (e) and (f); cDPCoA using pH range as constraint, and retaining the residuals (Within-DPCoA, hence effectively removing pH-related variation). Communities are grouped by pH (e) or vegetation type (f). The group ellipsoids in panel E demonstrate that no pH-related variation remains in the data, while panel (f) shows no apparent vegetation type-related effect. Panels (g) and (h); Results of DPCoA applied to two factors using vegetation type as constraint applied to the data set while removing pH-related variation. OTUs are coloured according to RDP II classifier results. Code for OTUs classes: A-*Acidobacteriaceae*, B-*Actinobacteria*, C-*Alphaproteobacteria*, D-*Bacteria*, E-*Bacteroidetes*, F-*Betaproteobacteria*, G-*Chloroflexi*, H-*Cyanobacteria*, I-*Deinococcales*, J-*Deltaproteobacteria*, K-*Firmicutes*, L-*Gammaproteobacteria*, M-*Gemmatimonadaceae*, N-*Nitrospiraceae*, O-OD1, P-*Planctomycetaceae*, Q-*Proteobacteria*, R-TM7, S-*Verrucomicrobiales*

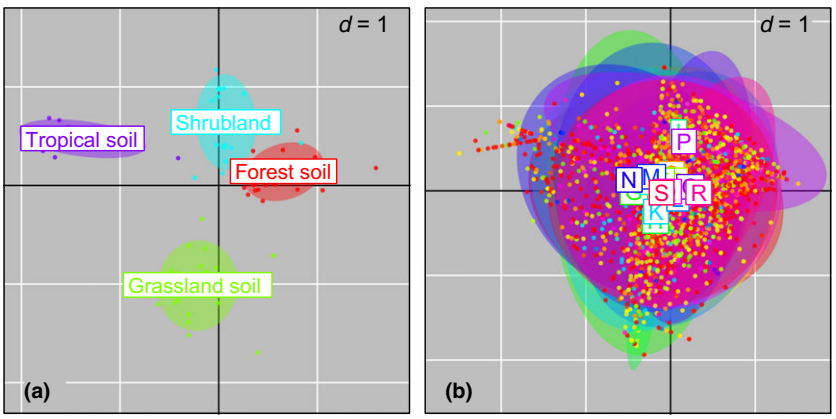

Fig 2 Decomposition of the soils dataset variability according to vegetation type. The diagram shows the result of a between-class analysis applied to a correspondence analysis using vegetation type as the constraint on the data, in which pH effect has been previously removed. (a) The ellipsoids represent the collective variance of each type. (b) OTUs are coloured according to RDP II classifier results (see Fig. 1).

segregated the different types of soil, and, contrary to partial canonical correspondence analysis, it allowed the identification of the bacterial taxa responsible for such separation. For instance, there was an association between overall *Firmicutes* (K) abundance and tropical soils, whereas *Actinobacteria* (B) mainly occurred in shrubland soils.

In their study of the distal gut microbiota, Dethlefsen & Relman (2011) used principal coordinates analysis of unweighted unifrac metrics to observe antibiotic-driven perturbation, intermingled with strong inter-subject differences. Here, cDPCoA could permit to focus solely on differences due to antibiotic treatment while controlling for inter-subject variability, and additionally provide

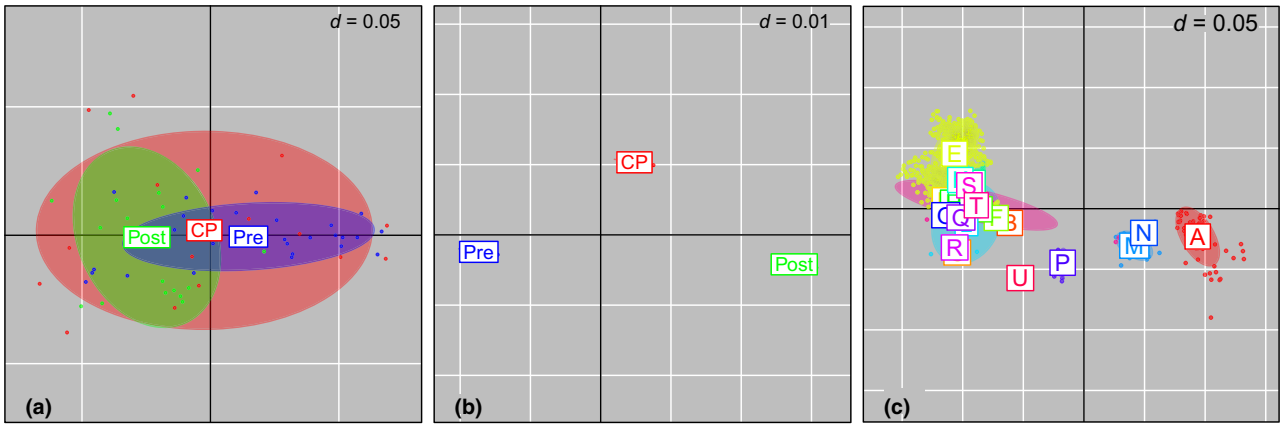

Fig 3 Panel (a); DPCoA of Dethlefsen & Relman data set. Distal gut communities have been grouped by sample categories (Pre; before antibiotic treatment. CP; during treatment. Post; after antibiotic treatment), and the ellipsoids represent their variability. Panel (b); cDPCoA of the same data set using sample category (treatment) as constraint after the removal of inter-subject variability. Panel (c); Same cDPCoA as in panel (b), but the effect of each OTU in community variation is shown. OTUs are coloured according to RDP II classifier results. Code for OTUs classes: A-*Bacteroidaceae*, B- *Bifidobacteriaceae*, C-*Burkholderiales*, D-*Caulobacterales*, E-*Clostridiales*, F-*Denitrobacterium*, G-*Desulfovibrionales*, H-*Enterobacteriales*, I-*Holdemania*, J-Incertae Sedis, K-*Lactobacillales*, L-other, M-*Porphyromonadaceae*, N-*Prevotellaceae*, O-*Rhodospirillales*, P-*Rikenellaceae*, Q-*Sphingomonadales*, R-*Streptophyta*, S-*Turicibacter*, T-uncultured, U-*Verrucomicrobiaceae*.

visualization of the organisms that drive the phylogenetic dissimilarities among sample classes. A DPCoA of the data (Fig. 3a) showed evidence of differences between microbial communities *pre-* and *post*-antibiotic treatment, but with considerable overlap. We followed a similar strategy as for the soils data set by removing inter-subject variability and focusing on treatment classes, which revealed much finer segregation between sample classes (Fig. 3b), and allowed a projection of the groups driving the phylogenetic segregation (Fig. 3c). In this case, the analysis showed an increase in the relative abundance of *Bacteroidaceae* (A), *Porphyromonadaceae* (M) and *Prevotellaceae* (N) as the main phylogenetic signature of the antibiotic treatment (Fig. 3c).

Discussion

cDPCoA is an extension of DPCoA to analyse the relative effects of two factors on the phylogenetic diversity of communities or to remove variability due to a factor before focusing on the other. As an illustration, we analysed 16S rRNA bacterial sequences representing soil and distal gut communities. As expected from a previous analysis (Lauber *et al*. 2009), differences in the phylogenetic composition of soil communities were mostly driven by soil pH. cDPCoA allowed the swift identification of the groups of species driving these differences, with a subgroup of *Acidobacteriaceae* decreasing with increasing pH, a pattern previously reported (Sait *et al*. 2006; Lauber *et al*. 2009). Differences among habitats were thus hidden by differences in soil pH among communities. While it is beyond the scope of the present report to re-analyse the test data set, we have made use of it to show how cDPCoA offers the means to overcome this obstacle but focusing the analyses on phylogenetic differences among habitats. Using cDPCoA, the pH effect can be removed to identify residual structures or to focus on the effect of another variable (here the habitat, or more precisely EnvO ontology 'feature'). Whereas partial canonical ordinations can be used to focus and/or remove the effect of some variables, these approaches do not take into account the phylogenetic relationships among species (here OTUs) as Unifrac and DPCoA do. Taking into account phylogenetic information permits to observe correlations between external variables and phylogenetic community composition, but also to make better sense of data sets with thousands of OTUs by considering that two distant species induce more differences in community composition than two close species. To date, no suitable method allowed the use of explanatory variables and cofactors in combination with ordination methods including the phylogenetic relatedness among species, which prompted our implementation of the cDPCOA. Similarly, cDPCoA was shown to be useful in removing existing strong inter-subject variability between the gut microbiota of different subjects before analysing treatment effects. Here, the use of cDPCoA clearly improves the biological interpretation and allows to identify the main phylogenetic signature of the antibiotic treatment after partialling out the subject variation.

We believe this brief exercise helps demonstrate how cDPCoA approaches can be very useful in the analysis of diversity patterns in microbial communities. Compared to DPCoA, cDPCoA enforces *a priori* classifications to

focus on subtle differences and (or) removing confounding effects, such as the well-reported inter-subject variability in the study of the human microbiome (Eckburg *et al.* 2005; Aguirre de Carcer *et al.* 2011). A priori classifications are frequent in the analysis of microbial communities and should thus be considered during the statistical analysis.

Here, we have followed the most common approach in dealing with high-throughput 16S rRNA gene sequences. This includes clustering sequences at a predefined similarity threshold to produce OTUs, which then serve as proxies for taxonomic ranks. This procedure presents two important pitfalls. First, sequence errors in the reads (e.g. PCR errors, sequencing errors, chimeric sequences) can lead to an overestimation of the number of OTUs. While this bias cannot be easily overcome, several studies have produced benchmarked tools and approaches able to reduce its impact (Schloss *et al.* 2011; Bragg *et al.* 2012; Bokulich *et al.* 2013). The second issue relates to the comparison of the physiological and ecological diversity captured within OTUs defined at arbitrarily predefined similarity threshold. In this respect, Philippot *et al.* (2010) have postulated that ecological coherence within taxonomic groups arises from the presence of a characteristic and unique set of gene families associated with environmental interactions, with the number of signature genes being negatively correlated with taxonomic rank. In that context, Zaneveld *et al.* (2010) recently showed that for the genomes of the main bacterial groups residing in the gut, gene conservation was correlated with 16S rRNA distance. The bacterial species concept remains a hot topic of discussion and represents an active area of research (Mende *et al.* 2013).

This study shows how cDPCoA, designed to reveal relationships between the structure of complex communities, differences among their species, and environmental factors, can be coupled with powerful molecular techniques to help clarify the scope and relative impact of deterministic and stochastic factors in microbial communities (Dethlefsen *et al.* 2006). This study also opens the way to new research on describing patterns in phylogenetic diversity. Instead of simply providing quantification of *how much* different communities are as with Unifrac (Lozupone & Knight 2005), DPCoA answers the question of *how* different communities are. Indeed, DPCoA provides a direct link between the compositions of the communities in terms of which species they contain and the quantification of phylogenetic differences among these communities (Pavoine *et al.* 2004). As a result the, species driving the phylogenetic differences among communities can be precisely identified. In comparison with PCoA applied to Unifrac distances, DPCoA not only allows plotting communities and species onto the same space, but was also found to be robust to the small fluctuations around zero, which are the main sources of noise to be expected from Amplicon sequencing data (Fukuyama *et al.* 2012).

Further research could be carried out now to include a wider variety of variables that can affect communities. For example, the raw value of soil pH, here used to classify communities, could be considered in a more continuous way as a quantitative explanatory variable. Sets of more than two explanatory factors could also be envisaged. Overall, we hope that the methods suggested will aid ecologists in attaining an understanding of the factors driving microbial community structure. To favour the use of DPCoA and related methods, we provide R functions and code to reproduce the analyses presented in the paper and to apply these new statistical techniques to other real data sets.

Acknowledgements

We thank Christian Lauber for data set and metadata information. Daniel Aguirre de Cárcer was supported by the Marie Curie IIF Grant PIIF-GA-2012-328287. The authors declare no conflict of interest.

References

Aguirre de Carcer D, Cuiv PO, Wang T *et al.* (2011) Numerical ecology validates a biogeographical distribution and gender-based effect on mucosa-associated bacteria along the human colon. *ISME Journal*, **5**, 801–809.

Aguirre de Cárcer D, Denman SE, McSweeney C, Morrison M (2011) Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. *Applied and Environmental Microbiology*, **77**, 8795–8798.

Bokulich NA, Subramanian S, Faith JJ *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, **10**, 57–59.

ter Braak CJF (1988) Partial canonical correspondence analysis. In: *Classification and Related Methods of Data Analysis* (ed. Bock HH), pp. 551–558. North Holland, Amsterdam.

Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW (2012) Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nature Methods*, **9**, 425–426.

Cailliez F (1983) The analytical solution of the additive constant problem. *Psychometrika*, **48**, 305–308.

Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

Dethlefsen L, Relman DA (2011) Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences USA*, **21**, 4554–4561.

Dethlefsen L, Eckburg PB, Bik EM, Relman DA (2006) Assembly of the human intestinal microbiota. *Trends in Ecology & Evolution*, **21**, 517–523.

Dray S (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22**, 1.

Eckburg PB, Bik EM, Bernstein CN *et al.* (2005) Diversity of the human intestinal microbial flora. *Science*, **308**, 1635–1638.

Fukuyama J, McMurdie PJ, Dethlefsen L, Relman DA, Holmes S (2012) Comparisons of distance methods for combining covariates and

abundances in microbiome studies. *Pacific Symposium Biocomputing* 213–224.
Gower JC (1984) Distance matrices and their Euclidean approximation. In: *Data Analysis and Informatics III* (eds Diday E, Jambu M, Lebart L, Pagès J, Tomassone R), pp. 3–21. Elsevier, Amsterdam.
Gower JC, Legendre P (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5–48.
Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*, **75**, 5111–5120.
Lingoes JC (1971) Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, **36**, 195–203.
Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71**, 8228–8235.
Manly BFJ (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London, UK.
Mende DR, Sunagawa S, Zeller G, Bork P (2013) Accurate and universal delineation of prokaryotic species. *Nature Methods*, **10**, 881–884.
Parameswaran P, Jalili R, Tao L *et al.* (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research*, **35**, e130.
Pavoine S, Dufour A-B, Chessel D (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology*, **228**, 523–537.
Pavoine S, Blondel J, Dufour AB, Gasc A, Bonsall MB (2013) A new technique for analysing interacting factors affecting biodiversity patterns: crossed-DPCoA. *PLoS One*, **8**, e54530.
Philippot L, Andersson SG, Battin TJ *et al.* (2010) The ecological coherence of high bacterial taxonomic ranks. *Nature Review Microbiology*, **8**, 523–529.
Sabatier R, Lebreton JD, Chessel D (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In: *Multiway Data Analysis* (eds Coppi RB, Bolasco S), pp. 341–352. Elsevier, Amsterdam.
Sait M, Davis KER, Janssen PH (2006) Effect of pH on isolation and distribution of members of subdivision 1 of the phylum Acidobacteria occurring in soil. *Applied and Environmental Microbiology*, **72**, 1852–1857.
Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.
Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE*, **6**, e27310.
de Vienne DM, Aguileta G, Ollier S (2011) Euclidean nature of phylogenetic distance matrices. *Systematic Biology*, **60**, 826–832.
Wesuls D, Oldeland J, Dray S (2012) Disentangling plant trait responses to live-stock grazing from spatio-temporal variation: the partial RLQ approach. *Journal of Vegetation Science*, **23**, 98–113.
Zaneveld JR, Lozupone C, Gordon JI, Knight R (2010) Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Research*, **38**, 3869–3879.

All three authors conceived the work, S.D. and S.P. developed the mathematical concepts and the R scripts, D.A. carried out all sequence preprocessing and analysis. All authors co-wrote the paper.

Data Accessibility

Data and R scripts to reproduce all analyses are available at ftp://pbil.univ-lyon1.fr/pub/datasets/dray/MER2014/.