1 General

Comment Overall, the study is of good quality and a better understanding of the nature of bacterial communities and of their drivers of diversity in the environment was achieved. In this sense, the study represents an interesting contribution to the scope of Biogeosciences by revealing the interplay of multiple factors on the diversity of specific pelagic bacterial communities in the Atlantic.

The work is clearly presented. Yet, the abstract is rather descriptive of the experimental strategy and results, and the main ecological hypotheses or research questions need to be stated more clearly. There were only few typos, which are indicated below. Tables and figures are appropriate.

The two important points of the study were: 1) Detailed description of vertical and spatial variation in community structure of bacterial assemblages in the water column of the eastern Atlantic ocean. 2) Beyond traditional community ecology statistical methods, using Bayesian inference to investigate OTU patterns, as a way to complement the classical community ecology approach.

The molecular methods that were used have become the gold standard in the field, and the main questions and comments I have mostly concern the statistical treatment of the data, as this is one of the main points of the study. In particular the use of Bayesian inference, which offers an interesting option to look at the data, would need to be better justified and explained.

The results could also incorporate some recent work that has been published and which compare pelagic and benthic bacterial communities based on the same molecular approach (Zinger et al. 2011 Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems PLoS ONE), and also work that investigate the impact of rarity or dominance on ecological interpretation of community patterns (e.g. Gobet et al. 2010 Nucleic Acids Research).

Response We are very appreciative of the time and efforts of the referee. The thoughtful comments will help us present a more clear manuscript that will be appreciated by a broader audience. Particularly, we acknowledge Alban Ramette's willingness not to remain anonymous. We do acknowledge that in the spirit of brevity and simplicity, many technical details and required assumptions made for the use of Bayesian inference were not included in the manuscript. We have thoroughly tested and taken into consideration each of the referee's comments and suggestions and feel our methodology and results are justified.

We are confident the nature of Bayesian hypothesis testing is well-suited to model the stochastic nature of the divergence of these communities in space and time, allowing for biogeographical hypothesis testing.

2 Ecological statistics

Comment 1 L8-9P117: Using ACE or Chao1 estimators make use of singletons in their formulae. Therefore if they are removed (L18 P116) from the data, what kind of effects this may have on your reported diversity estimates?

Response 1 A consequence of removing singleton reads (those reads which are found only once across all samples) is that there are no OTUs which occur only once across all samples. In other words, the minimum abundance of an OTU in our data is 2. This fact could result in an even greater underestimation of diversity, and is one of the main reasons we calculate the CatchAll diversity estimate which is reported to be robust against such effects (Bunge, 2011).

Comment 2 If the authors goal was to explain community variation as a function of environmental vs. spatial variables, the use of indirect gradient analysis (i.e. PCoA axes calculated on the community data, which are in a second step further explained by environmental and spatial variables) or (partial or simple) Mantel tests might not be appropriate. Those methods are not recommended when one wants to determine the respective effects of space vs. environment in explaining changes in beta diversity (see Legendre et al. 2005 Ecol Monographs; Peres-Neto et al. 2006 Ecology; Ramette and Tiedje 2007 PNAS for an example in microbial ecology). If biotic and environmental variations share a common spatial structure, spatial processes

must generally be considered when examining the effects of environment on biotic variation (Borcard et al. 1992 Ecology). Instead, variation partitioning using constrained ordination (e.g. redundancy analysis) should be preferred because the amount of covariation between groups of factors is also quantified, and this is not the case when using (simple or partial) Mantel tests. The second reason is that the variance of a dissimilarity matrix among sites is not the variance of the community composition table nor a measure of beta diversity (Legendre et al. 2005 Ecol Monographs), so what is represented by PCoA is the variation of the variation in community structure, while the variance of the community composition table is a measure of beta diversity. The Mantel approach is, however, still appropriate for testing other hypotheses, such as the variation in beta diversity among groups of sites. It is not clear if the authors were aware of the implications the choice of the numerical strategies has on their ecological interpretation.

Response 2 Prior to submitting the original manuscript, and again since receiving the reviews, we have carefully considered the potential of applying CCA to analyze our data. Our goal in reporting the correlations between PCoA axis scores and the environment (Table 3) was simply to highlight the most obvious ecological drivers (e.g., depth), but not to quantify the relative contribution of each environmental factor in terms of individual partitioned variance. While we appreciate the referees suggestion that this additional information could be obtained via CCA, there are several issues that lead us to question the applicability of CCA to our particular dataset. Our specific concerns are as follows:

(1) CCA is a form of direct gradient analysis, which means that it only allows for ordination of the samples based on the subset of the variation in the species data that can also be explained by the environmental variables. This approach is best suited to community data sets where one can be relatively certain that the most important underlying environmental variables have been measured. Gradients are assumed to be known and represented by the measured variables or their combinations, while species abundance or occurrence is considered to be a response to those gradients (Ramette, 2007). Due to the wide spatial extent of our samples, and the fact that our sampling represents several habitat types (considering depth & province), we are not comfortable making the assumption that our small set of environmental variables is adequate to describe the drivers of community composition across this diversity of habitats. Further, use of CCA would mean that not all the variation in species data would be displayed in the ordination diagram. Preliminary analysis using CCA indicates that only $\sim$10% of the variance would be presented, compared to $\sim$50% in our PCoA (using the first two axes of each). Part of our goal with this study is to show the overall patterns across samples such that future studies may target oceanic regions of particular interest or develop sampling schemes to best survey for large-scale ecological drivers. Thus the exclusion of unexplained variability in the species data (via CCA) would restrict a reader to interpreting community data that only corresponds to the environmental parameters we assayed. By correlating the PCoA results to the environmental data, we identified target environmental parameters that may be of future interest, but that is not the main focus of the paper.

(2) Our data set consists of 16 samples (only 14 of which have full environmental data) and 11 environmental parameters. The use of CCA is discouraged when the number of environmental variables approaches the number of samples (McCune et al., 2002). This is because the more environmental variables that are used, the weaker the "constraints" on the axes become, and thus CCA approaches CA. This results in an inflated estimate of the percent of variance in the species environmental relationship that is explained, and misinterpretation of any correlation between a given species and set of environmental conditions. Thus, one of the referees main suggested advantages to CCA (to gain insight to the respective contribution of spatial vs environmental factors) would not be achieved in our particular case due to an artificial inflation of the community-environmental correlation matrix.

(3) Our current approach displays the simple correlation coefficient between each environmental variable and each axis of the PCoA. We did this in an effort to highlight potentially important ecological drivers, but did feel it was inappropriate to quantitatively compare these correlations given the small number of samples and the multicollinearity among our environmental variables. If we applied CCA to our data, we expect a low predictability of the model due to intercorrelations among the explanatory variables (Ramette, 2007; Legendre and Legendre, 1998). When we endeavor to account for this using only the environmental

variables that contribute most significantly to the variation in the OTU table (e.g., using forward or stepwise selection), we further reduce our pool of candidate environmental predictors, making it even less appropriate to assume that we have pinpointed key environmental drivers (an assumption of CCA, see #1).

Regarding the referees suggestion that variation partitioning using constrained ordination (e.g., CCA) would be preferable, compared to our use of Mantel tests. We explored variation partition and implemented the approach outlined in Borcard et al. (1992). Upon doing so, we discovered that most of the variability in our community data set was unexplained (50-60%, depending on normalization procedures; performed using CANOCO version 4.5). This suggests that we have, in fact, not adequately sampled the environmental factors that correlate with microbial community structure, and supports our choice to avoid constrained (direct) forms of ordination.

Instead, we elected to compare the PCoA output to additional environmental variables using Spearman's rank correlation coefficients, an approach suggested in Ramette (2007), who further cite Legendre and Legendre (1998). In the discussion section of our paper, we are cautious in our interpretation of these results, as we are aware that this is an analysis that considers variation of a dissimilarity matrix and not the variance of the community composition per se. Another advantage to our approach is that it allowed us to identify OTUs associated with PCoA separation (see Table 4), which revealed that the driving force separating our samples in ordination space is associated with less abundant OTUs (an important result given the recent debate about the significance of the rare biosphere in microbial ecology).

With respect to the Mantel tests, we used them only to determine the presence of spatial structure in our dataset after accounting for any spatial patterns shared with the environmental data matrix (pg 129, line 5). The use of Mantel and partial Mantel tests has become very popular in this context (e.g., Parker and Spoerke (1998); Cho and Tiedje (2000); Horner-Devine et al. (2004); Scortichini et al. (2006), all cited in Ramette (2007)). The analysis was done as a prelude to the biogeographical signal uncovered using the Bayesian approach. We agree with the referees statement that the Mantel approach does not allow us to determine the respective effects of space vs. environment in explaining changes in beta diversity but, as stated above, a direct qualitative comparison of these two sets of variables was not among the goals of this study, and we do not report any results that extend the Mantel approach to explore community-environment relationships.

Comment 3 Before computing the environmental dissimilarity matrix needed for the Mantel test, it seemed that standardization of the variables to the same variance was not performed. Standardization to a variance of unity gives the same weight to each variable in the matrix. If not, variables measured on larger scales will have more impact on the resulting in ecological interpretation.

Response 3 By using the Gower coefficient to calculate the environmental dissimilarity matrix, this standardization in done automatically as each term is normalized according to its individual range (Gower, 1971).

Comment 4 The study uses the 3D location of the samples (latitude, longitude, depth). Why log-transforming variable depth and not the other ones (L4 P118)? Keeping those spatial data on the same scale would enable a more homogeneous description of spatial pattern

Response 4 We chose to log transform depth because of the skewness of the depth distribution. Thirteen of the 16 samples were at depths of 200 m or less and the others were at 1100, 1300, and 4600 m.

3 Bayesian inference

Comment 5 Hypothesis ii (L24, P118) the distribution and abundance pattern for any OTU is independent of the pattern of any other OTU is difficult to assume,given the way data were collected and the fact that

bacteria in communities do interact. If this hypothesis is true, it would correspond to bacterial cells inactively floating in the water column, without any function or interactions with their neighbors.

Response 5 The need for the assumption of independence can be explained based on both biological and statistical principles. Ecology defines communities as an interacting group of populations of different species in a same location and the effect of those interactions on the structure and organization of the community. However, how one defines and takes into account those interactions can be problematic, if not mathematically intractable. The bacterial cells sampled in this study can potentially interact not only with other bacteria, but also with members of the other two domains of life, Archaea and Eukarya (that were not sampled in this study). Even if all the possible interacting species are sampled, there is no information that will allow us to determine those interactions. Based on those uncertainties the most parsimonious solution is to assume independence and treat the OTUs as random variables. Statistically, the independently and identically distributed assumption is used to describe the joint distribution of two or more random variables when stochastic processes are modeled. If independence is not assumed, taking into account the dependency among OTUs makes the computational methods intractable. Specifically, in a Bayesian statistical framework where the likelihood function is used, the independence assumption allows determining probabilities (likelihood) at each position of the abundance matrix and then multiplying across all positions. This property is not unique to Bayesian inference, and would be an assumption in nearly all methods which rely on multiplication of probabilities. As biologists, were are aware that mathematical models are useful but not always perfect. If the members of the community are assumed to diverge independently from each other, then the resulting mathematical analysis is clear enough to allow useful biological conclusions. Statistically, it is assumed the method is robust to assumption violations and the data will reject the model if inappropriate.

Comment 6 L8-9P119: traditional approaches loose some information, but by transforming the data into ranks, is it not the same that was achieved?

Response 6 There was some confusion here that will be changed in the revised manuscript. The data was not ranked but rather range standardized. By implementing the gap weighting algorithm described in Schols et al. (2004), and Thiele (1993) we are accounting for both the order and distribution of OTU abundances across samples. The steps between the resulting weighted values are therefore proportional to the distance between the original OTU abundances.

Comment 7 L6-7P120: How does the recoding affect the result? In fact, by rank-transforming the data and recoding them, the initial raw data are therefore very different from this new set. It would be good to apply the same approaches to both sets to see how different the results might be.

Response 7 Below is a toy example (Figure 1) of how the coding takes place for a set of five OTUs in six samples, and illustrates the differences between the data as they are recoded rather than ranked. The minimum value of an OTU across samples is given the value of 0, and the highest 9. Other values are coded proportionally in the range. An important distinction is that with the weighting, we can approximate of the original abundance distribution that would be lost in a simple rank transform. As stated in the paper, these recoded values are then transformed to their four-bit binary equivalents in order to represent their magnitude in the input matrix to the Bayesian inference machinery (David Swofford, personal communication, Woods Hole, MA, 2011).

Comment 8 Why was Bayesian inference required here? For instance, could the authors make use of prior knowledge? L10-11P120: what does across-site rate variation following a gamma distribution have to do with the current problem? There is no rate of evolution here, so are those parameters and model adequate for the data at hand? Generally MrBayes can be used to infer phylogenies based on sequence data so phylogenetic or evolutionary models need to be supplied. If the authors used the default settings, one should ask if this is really appropriate given the type of data (relative sequence abundance) that is used in the study

Response 8 The ultimate goal of comparing and relating community composition and abundance is to draw possible conclusions about biogeographical patterns. As defined, biogeography is the study of the distribution of biodiversity over space and time; based on this definition we have to assume the difference in the composition and abundance of the members of the communities (OTUs) sampled is due to processes determined by space and time (e.g., ecological drift and evolution). This assumption of community divergence in space and time implies those communities are changing at certain rate. The overall rate of change of those communities is determined by the effect of those spatial and temporal processes on each one of the members (OTUs) of those communities. A priori, we can assume those processes do not affect all members of the community equally, resulting in variation in the rate of change across the members of the community. This is the justification for assuming across-site rate variation following a gamma distribution (and is corroborated by the Bayes factor testing approach described below). We have considerable expertise using tree topology reconstruction algorithms and have been extremely careful in selecting the most appropriate model and parameters for the data we are trying to model. Our intention was not to present a technical methods manuscript although considerable testing, simulations, and randomizations of the data were performed allowing us feel very confident that our approach is well-suited for this data type. Bayesian tree inference, as implemented in MrBayes, can reconstruct tree topologies using four different types of data: nucleotide, protein, restriction enzymes, and standard morphology character data. Any type of analysis using OTU abundance data to determine relationships between communities implicitly or explicitly assumes that the OTU distribution data provides information that uniquely characterize the community, and we assume that this can be equated with phenotypic or morphological character data.

We have modeled the OTU abundance data using the model for standard morphology data type in MrBayes, this model allows to describe the "characters" utilizing up to ten discrete character states (0-9). Character data is defined as information about the attributes of the objects under study and those characters can be visualized as a set of independent variables existing in a set of mutually exclusive character states. In our case, each OTU is considered a character and its abundance represents one of the many different character states the data can assume. Because MrBayes input can accommodate up to 10 character states, the OTU abundance is converted to a score between 0 and 9 by range-standardizing the data according formula 1 described in materials and methods as in Thiele (1993) and Schols et al. (2004) (see also response to comments 6 and 7).

Additionally, we used a Bayes factor to quantify the degree to which a model incorporating gamma rate variation would better fit the data. To do this, we ran two new iterations of the Bayesian machinery (1 000 000 generations), one with an equal rate model (0) and one modeled with gamma (1). After confirming that the new tree run with gamma rate variation was identical to the original tree in the paper, a Bayes factor was calculated by comparing the logarithms of the harmonic means of the likelihoods from the MCMC for two models. The harmonic mean from the MCMC in this case approximates the marginal likelihood of the model. We find that there is strong evidence to suggest that the model incorporating gamma should be strongly preferred ($B_{10} = -51\,811.84 + 52\,162.33 = 350.5$) (Kass and Raftery, 1995).

Comment 9 L9P130: PCoA is designed to summarize the variation in a data matrix onto few axes (generally up to three components) to facilitate the interpretation of the main patterns of variation in a dataset. The Mantel test is designed to correlate two matrices and test for the signicance of the correlation coefficient by considering the fact that the data originates from matrices (i.e. data not freely exchangeable in the permutation scheme). Therefore, with both approaches the idea is clearly not to visualize individual OTU variation. Yet, the information from individual OTU variations is considered in the final matrix. Those approaches are not dissimilar to what the authors acknowledged (L19- 20P130) as being their goal: to identify the optimal tree topology that best explains the relationships based on overall patterns of OTU abundance. One could also obtain the same idea by applying a clustering algorithm on the Bray-Curtis dissimilarity matrix, and use a bootstrapping approach to test for the reliability and support of the branches in the dendrogram. So again, it is not clear why and how the Bayesian approach is better than the classical approach to deal with such dataset.

Response 9 The classical approach to the analysis of multivariate data for community comparisons is to compute a matrix of pairwise dissimilarities or distances (e.g., Bray-Curtis dissimilarity) between samples, followed by the generation of a dendrogram produced using a cluster analysis (e.g, UPGMA) on that matrix. This methodology one of the many numerical classification techniques and concepts developed by Sokal and Sneath (1963; 1973) to classify organisms based on phenetics principles. More recently in their widely used book, Numerical Ecology, Legendre and Legendre (1998) apply the techniques of numerical taxonomy to ecological groupings. The grouping or numerical classification of the objects or organisms under study, is based on the similarities of the phenotypic or morphological characters identified. The distance-based dendrogram reconstruction methodology is widely used in community ecology due to the ease of interpretation and computational efficiency (speed). However, its application determines only one tree without requiring the evaluation of the massive amount of competing trees or alternative hypothesis required and tested by character-based methods, like Bayesian inference. By reducing the original data matrix to a pairwise distance matrix, traditional clustering methods do not make direct use of all the character information available and may not have the resolution power to tease out all the possible patterns encoded in the data (please see also comment 10 illustrating the potential bias of Bray-Curtis dissimilarity). Please also see the response to comment 10 of Referee 2.

It has also been demonstrated that the cluster analysis of any random set of similarity data will produce a dendrogram, and the application of nonparametric bootstrapping to such data will always generate the same tree, leading to the erroneous conclusion that the tree is extraordinarily reliable (Swofford et al., 1996). BI is considered a character-based method of tree reconstruction, and it employs an optimality criteria to explore the tree space (in our case, approximately 10^{23} possible trees) and decide whether the data support a given tree or trees. The process computes the likelihood function for every column of the data matrix which is used to calculate the posterior probability of the proposed tree. Although much more computationally expensive, it uses the information contained at each column of the data matrix to test not only if a tree structure is the best representation of the relationships of the samples under study but also to determine which tree topology is the one supported by the data.

Comment 10 L3-5P127: The hypothesis that the less (sequence) abundant members drive community change can be tested with the data. Considering only parts of the data to investigate the role of dominance or rarity has already been explored elsewhere (e.g. Gobet et al. 2010 Multivariate Cutoff Level Analysis (MultiCoLA) of Large Community Datasets. Nucl Acids Research).

Response 10 We performed MultiCoLA (Gobet et al., 2010) analysis of our data and found that even after removing 80% of the least abundant OTUs,that we maintained approximately 80% Procrustes correlation between the truncated and original data NMDS ordinations (Figure 2). However, we are concerned that this number is inflated due to the influence of most abundant species in the Bray-Curtis distance calculation. As a test, we generated many combinations of two samples, each containing 500 OTUs with random abundances drawn from a Gamma distribution ($\alpha = 1.0, \theta = 1000$). The OTUs were ordered according to increasing total abundance and sequentially removed from the the data. At each step, a Bray-Curtis distance was calculated, and the results for one of these simulations were plotted in Figure 3. We see that the distance between the samples remains relatively constant until a certain point, confirming that the Bray-Curtis distance is robust to the removal of rare OTUs. As the NMDS ordination uses the Bray-Curtis distance matrix as input, we expect that correlations would always remain high as rare OTUs are removed when compared in this way.

These results highlight the importance of the less abundant tail of the OTU distribution when considered in light of full evidence methods, such as the Bayesian one that we present here. We are able to recover the major depth gradient using distance-based ordination, but the tail is what allows us to hone in the biogeography signal that gets lost when the OTU abundances are reduced to a simple distance.

4 Minor Comments

Comment 11 The title should probably indicate Bacterial instead of Microbial community diversity to be more in line with what has been done in the study.

Response 11 We agree and will change in the revised manuscript.

Comment 12 Results: Many values are reported with some indication of variation. It would be useful to know if these are standard deviation, and also how many observations were used in each case to calculate the sd.

Response 12 These values are standard deviation. In the text, it should be clear that when we refer to distribution per sample (e.g, L21P121) that N = 16. When we talk about distribution in the water layers, it would be useful to note the sample numbers here, and this will be done in the revised manuscript.

Comment 13 L10P110: 16s rDNA should be changed to 16S rRNA gene.

Response 13 Both are correct, and we feel this is a matter of style. However, we will make this change in the revised manuscript.

Comment 14 L25P110: suggest

Response 14 Will be changed in revised manuscript.

Comment 15 L9P115: PCR reactions change to PCR

Response 15 "PCR reactions" is correct as written.

Comment 16 L3P116: Why using 95% bootstrap support? Provide a reference or experimental justication.

Response 16 95% is a common cutoff, but chosen arbitrarily here to allow some slack in the classifier. In the data, however, there are no sequences that are classified as "Bacteria" that fall into the 95-100% range. The revised manuscript will be adjusted to say that sequences not classified as bacteria with 100% bootstrap were excluded from further analysis.

Comment 17 L10P116: reformulate "sequence containing 'N'"

Response 17 Will be changed in the final manuscript (e.g., following the removal of any sequencing containing ambiguous bases (e.g., "N"))

Comment 18 L3P117: why not using OTU richness instead of species richness as the reason for using OTU is to avoid dealing with the microbial species concept in microbes?

Response 18 We will use OTU richness in the revised manuscript.

Comment 19 L28P118: dene what M and N are here.

Response 19 We will add "sample (row) by OTU (column)" to the revised manuscript.

Comment 20 L3-5P127: Make sure here that your readership understands that you are talking about relative sequence abundance and not actual counts.

Response 20 We will address this in the revised manuscript.

Comment 21 L14-15P127: Chlorophyll-a is written in two different ways.

Response 21 We will address this in the revised manuscript.

Comment 22 L6P130: "experimental observations sounds like different experiments were carried. Maybe use "values instead?

Response 22 We will address this in the revised manuscript.

References

Borcard, D., Legendre, P., and Drapeau, P.: Partialling Out the Spatial Component of Ecological Variation, Ecology, 73, 1045–1055, 1992.

Bunge, J. A.: Estimating the number of species with CatchAll, Pac Symp Biocomput, 2011.

Cho, J. and Tiedje, J. M.: Biogeography and degree of endemicity of fluorescent Pseudomonas strains in soil, Applied And Environmental Microbiology, 66, 5448–5456, 2000.

Gobet, A., Quince, C., and Ramette, A.: Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets., Nucleic Acids Research, 38, e155–, 2010.

Gower, J.: A general coefficient of similarity and some of its properties, Biometrics, 1971.

Horner-Devine, M., Lage, M., Hughes, J., and Bohannan, B. J. M.: A taxa-area relationship for bacteria, Nature, 432, 750–753, 2004.

Kass, R. and Raftery, A.: Bayes Factors, Journal of the American statistical association, 90, 773–795, 1995.

Legendre, P. and Legendre, L.: Numerical Ecology, Volume 24, Second Edition (Developments in Environmental Modelling), Elsevier Science, 1998.

McCune, B., Grace, J. B., and Urban, D. L.: Analysis of ecological communities, Mjm Software Design, 2002.

Parker, M. and Spoerke, J.: Geographic structure of lineage associations in a plant-bacterial mutualism, Journal of Evolutionary Biology, 11, 549–562, 1998.

Ramette, A.: Multivariate analyses in microbial ecology., Fems Microbiology Ecology, 62, 142–160, 2007.

Schols, P., D'hondt, C., Geuten, K., Merckx, V., Janssens, S., and Smets, E.: MorphoCode: coding quantitative data for phylogenetic analysis, Phyloinformatics, 4, 1–4, 2004.

Scortichini, M., Natalini, E., and Marchesi, U.: Evidence for separate origins of the two Pseudomonas avellanae lineages, Plant Pathology, 55, 451–457, 2006.

Sokal, R. R. and Sneath, P. H. A.: Principles of Numerical Taxonomy (A Series of Books in Biology), W. H. Freeman and Company, 1st edn., 1963.

Sokal, R. R. and Sneath, P. H. A.: Numerical Taxonomy: The Principles and Practice of Numerical Classification., W. H. Freeman and Company, 1973.

Swofford, D., Olsen, G., and Waddell, P.: Molecular Systematics, Phylogenetic Inference, Sinauer Associates Inc, 2 edn., 1996.

Thiele, K.: Thiele K. 1993. The holy grail of the perfect character: the cladistic treatment of morphometric data. Cladistics 9: 275-304., Cladistics, 9, 275–304, 1993.

$$\underset{Original}{\begin{pmatrix} 62 & 31 & 25 & 75 & 86 \\ 13 & 9 & 85 & 38 & 8 \\ 73 & 43 & 80 & 92 & 35 \\ 5 & 32 & 21 & 38 & 20 \\ 67 & 29 & 40 & 30 & 29 \\ 19 & 2 & 45 & 27 & 11 \end{pmatrix}} \xrightarrow{gap\ weighting} \underset{Compressed}{\begin{pmatrix} 8 & 6 & 1 & 7 & 9 \\ 1 & 2 & 9 & 2 & 0 \\ 9 & 9 & 8 & 9 & 3 \\ 0 & 7 & 0 & 2 & 1 \\ 8 & 6 & 3 & 0 & 2 \\ 2 & 0 & 3 & 0 & 0 \end{pmatrix}}$$

Figure 1: Gap weighting example for six samples and five OTUs. The matrix on the left represents the original OTU abundances; the matrix on the right represents the compressed abundances.

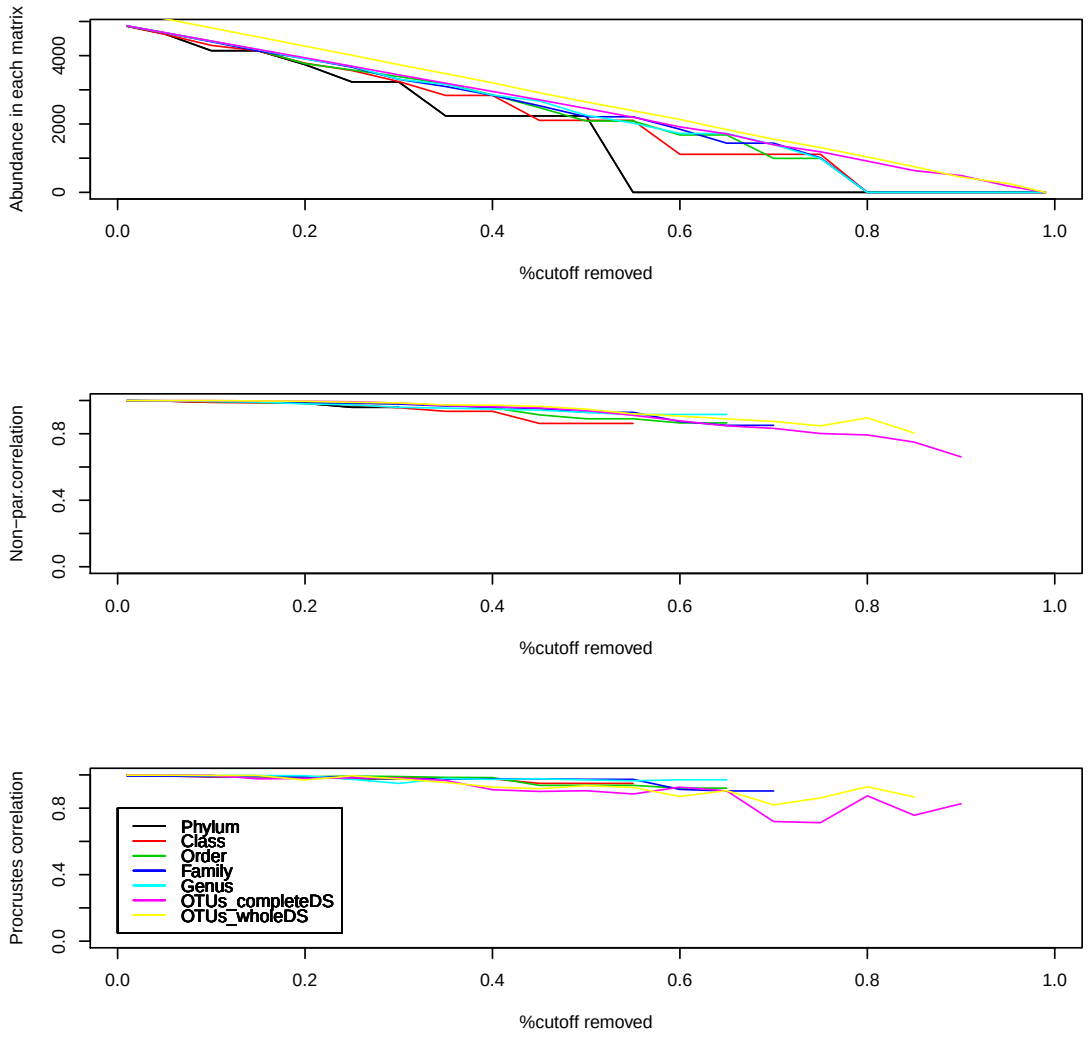

Figure 2: MultiCoLA results

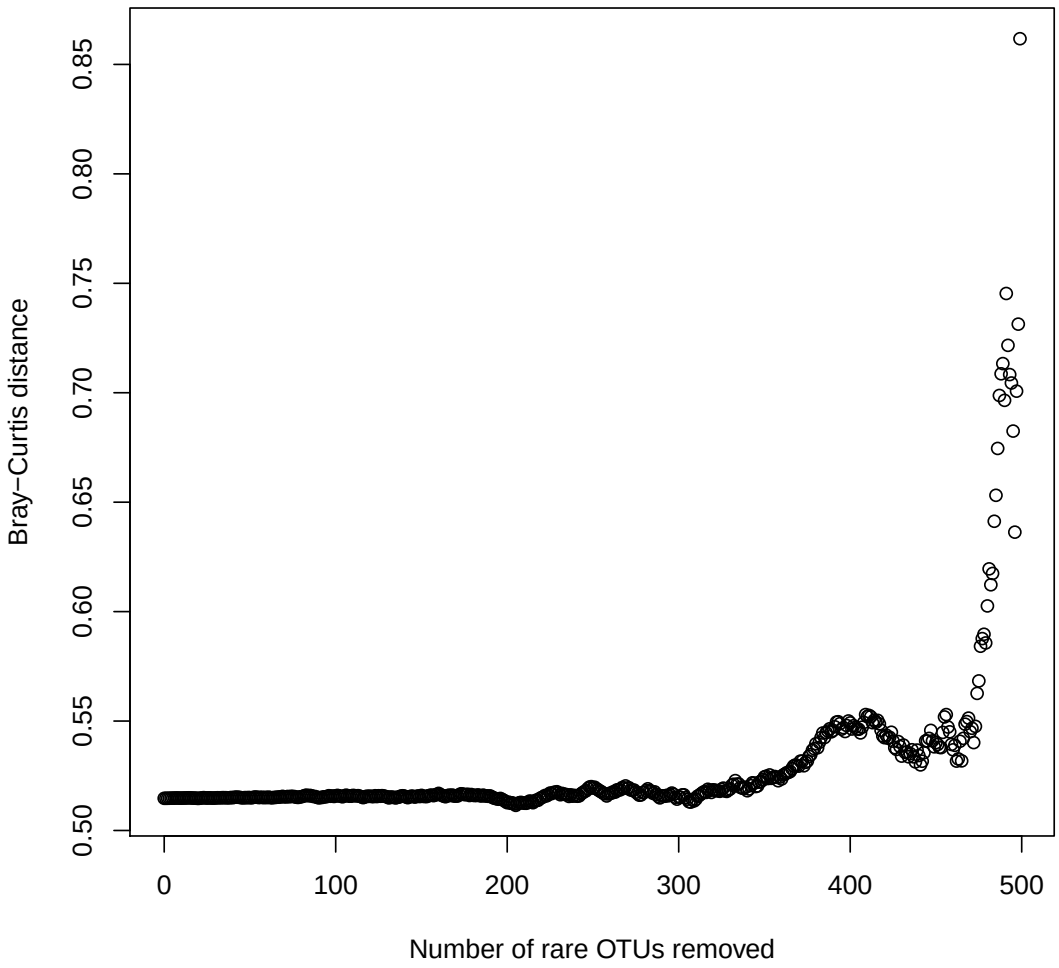

Figure 3: The effect on Bray-Curtis distance as rare OTUs are removed from a random, two-sample, 500 OTU dataset. The points on the plot represent the Bray-Curtis distance between the two samples after OTU removal.