# 1 General

**Comment** This paper is another in a series of studies from the ICoMM project which supported pyrosequencing of the V6 region of the 16S rRNA gene extracted from bacterial communities in various marine habitats around the world. The authors here present data from a transect in the Atlantic Ocean, from about 50 N to 31 S, covering about 7700 km. The main message of the study is that bacterial communities in the surface layer, deep chlorophyll maximum (DCM), and deep waters are all different. The study provides important new information about bacterial biogeography in the oceans.

The paper could be improved in several respects. First, the authors overstate the novelty of their study in the Introduction, although the Results/Discussion has a bit more complete review. The authors should work harder at identifying the truly novel aspects of their study. Second, the presentation of the study's data in the figures and tables could be improved. Specic problems are discussed below. Finally, I question some of the approaches used by the authors to analyze their results, and they need to explore other techniques commonly used to analyze relationships between community structure and environmental properties. Again, specifics are given below.

Another general issue is that the authors dont compare their results with other papers from the ICoMM project. A complete review is beyond the scope of this paper, but it seems the authors could touch on a few issues. Do they see the same level of diversity as in other waters? What about the relative abundance of rare vs. abundant taxa? Although the rare biosphere may be overhyped, there are some interesting aspects to the question that the authors could address. The rare biosphere question is hardly touched on in this paper. Are the same taxa found in these waters as elsewhere? The paper mentions SAR11 but thats about it.

**Response** We appreciate all the suggestions of the referee to improve the quality of our manuscript and will incorporate the overall findings of other ICoMM projects and findings from other publications in the revised version. Commonly used techniques were carefully tested, when our choices differ from the suggestions of the referee we have explained our rationales in the specific comments section below. Particularly the use of Bayesian inference is justified with more details in these comments and in the comments to Referee 1. We are confident the approach is appropriate to the data analyzed and takes full advantage of the extensive amount of data in this study.

# 2 Specific comments

**Comment 1** Abstract: The Abstract is a bit on the long side, with too much on data which don't need to be highlighted in an Abstract (e.g. the total number of sequences) and on results that are not particularly novel. The text starting at line 11 through 19 can be reduced substantially, with some things deleted completely. The potentially novel results start on Line 28.

**Response 1** The abstract will be amended in the revised manuscript

**Comment 2** P110, line 4 and elsewhere: Eubacteria is archaic and should not be used. Bacteria is sufficient.

**Response 2** We are confused by this comment from the referee, the International Committee on the Systematics of Prokaryotes (ICSP) (`http://www.bacterio.cict.fr/classifphyla.html`) recognizes two prokaryotic domains (or empires), Bacteria (or Eubacteria) and Archaea (or Archaebacteria). If referring to the proper names of the domains of the tree of life as describe by Woese et al. (1990), then those names are Bacteria and Archaea and should be spelled by capitalizing the first letter, but if referring to the two groups of prokaryotes, the common names eubacteria and archaebacteria as we understand are not considered archaic. A quick search on the Web of Science for the use of the term eubacteria on publications of the year 2011 generates a list of over 6000 articles.

**Comment 3** Introduction: There is a lot of general information in this Introduction that any reader of this journal and article will know already. This includes much of the first paragraph. Many readers will quibble with the percentages (they seem too high) and the references, detracting from the novel aspects of the paper. The authors overstate the lack of information about the biogeography of bacteria in the oceans; what about the GOS studies and others? The Results/Discussion section has a bit more comprehensive review of what is known about these issues. The authors should minimize the general stuff (readers will know bacteria are important) and get to the specic questions of this study.

**Response 3** We are pleased with and will follow the advice of the referee to give more prominence to the"novel aspects of the paper. Yes, the GOS studies have performed the most comprehensive ocean biodiversity survey, but also the 2010 "An Ocean and Biogeochemistry Scoping Workshop (Anderson et al., 2010) concludes the lack of knowledge of marine microbial biogeography prevents from answering many important questions in biogeochemistry.

**Comment 4** Page 112, line 3: The sentence here hints that the authors dont use the term metagenomic in the way most others do. Contrary to what is implied by the authors, the approach does not rely on PCR. The authors tag sequencing approach is not a metagenomic one. Methods and Materials: Overall, this section is too long as the authors have included lots of extraneous details. Some specic candidates for deletion are given below.

**Response 4** The metagenome refers to the whole genomic component isolated from environmental samples and its characterization (i.e., metagenomics studies) has followed different approaches. Two of the most common approaches are 16s-based community profiling (tag sequencing) and whole-community shotgun sequencing. The field in general is divided on the use of this term to describe all aspects of what might be termed metagenomics, either as a discipline or a set of techniques. We refer to it, as in Wooley et al. (2010) in the sense of using genomic techniques to study microbial communities in natural environments. Originally, the process of doing metagenomics included bacterial cloning (Handelsman et al., 1998; Handelsman and Smalla, 2003), though that is not common practice any more given next-generation sequencing technology. However, we will adjust this language in the revised manuscript to avoid any confusion.

**Comment 5** Page 116, line 26: It seems that the authors did not normalize their sequences per sample, i.e. by calculating the relative frequency of the OTUs (actual read numbers/total sequences in the sample). Rather, it seems they randomly resampled their data so that each sample had same number of sequences as the sample with the lowest number of sequences. How was this done? How many times did they resample?

**Response 5** In fact, this is exactly what we do. We say on page 116, line 26 "The abundances of the resulting OTUs were normalized using the smallest sample number (n = 6687)." We take the relative frequency of each OTU and multiply that number by the lowest sample number. No random sampling was performed.

**Comment 6** Page 117, line 1-5: This can be deleted. Its not appropriate for this section, and readers will already know all this.

**Response 6** We will address this in the revised manuscript.

**Comment 7** Page 117, line 16: The authors dont need to give reason for doing a PCA (PCA is probably a more common abbreviation than PCoA). Again, not appropriate and not necessary.

**Response 7** We feel that switching to PCA would be confusing to the readers, as that is the abbreviation most commonly used for Principal *Components* Analysis. The two methods (PCoA and PCA) are distinct, and because the application of PCoA in microbial ecology is relatively new, we wish to avoid any confusion by maintaining the distinct nomenclature. Our use of the abbreviation PCoA is consistent with the major texts for multivariate data analysis in ecology as well as recent publications in the field. We also believe that providing some justification for statistical methods is appropriate and recommended, though not often implemented in practice.

**Comment 8** Rather than a PCA, which depends on a parametric measure of distance, the authors should do (at least try) a nonmetric multidimensional scaling analysis which depends on fewer assumptions.

**Response 8** Non-metric multidimensional scaling (NMDS) places samples in ordination space based on the ranked distances between sample pairs; these distances can be determined using essentially any dissimilarity coefficient or distance measure appropriate for the original data type. NMDS is an excellent technique when the goal of data analysis is to provide a graphical representation of community relationships. The disadvantage to NMDS is that because it uses ranks to map the objects in ordination space, the original distances between objects are not reflected by differences in the axis scores in a scalar way. Thus no information can be obtained from NMDS about the species that drive the separation in ordination space, and the correlation of NMDS axis scores to environmental variables would not be valid.

PCoA can start with the same distance matrix as NMDS, so assumptions about the underlying distribution behind each technique are similar and accommodated in the selection of the distance metric, not the actual ordination. Both PCoA and NMDS differ from PCA, which requires multivariate normality.

PCoA is different from NMDS in that PCoA uses a linear (Euclidean) mapping of the distance (dissimilarity) between objects onto the ordination space, instead of ranks. Therefore the axis scores that result from PCoA ordination maintain a signal that corresponds to the original distance (dissimilarity) between communities, which means it is possible to correlate these scores with the original variables to assess their contribution to the ordination (Ramette, 2007). Because we wished to consider the OTUs that cause separation among samples and potential environmental drivers of sample separation, we chose to use PCoA over NMDS.

**Comment 9** Page 117, line 23: The authors should also do a Canonical Correspondence Analysis (CCA) before getting into Mantel and partial Mantel tests.

**Response 9** Please see response to comment 2 from Referee 1

**Comment 10** Page 118, Bayesian Inference. The authors drew up a similarity tree using an approach based on numerical taxonomy. First, other cluster type analyses are commonly done with these data. This web site has a succinct description of them: http://cran.r-project.org/web/views/Environmetrics.html. The authors need a few words to argue why this new approach is necessary, but I dont see why it is necessary. It seems inappropriate to use an approach designed for taxonomy (and evolution) to examine similarities among samples, as the authors do here. Other approaches have been devised specifically for what the authors want to do. Second, its not clear if anything really new is learned from any cluster or tree type of analysis. The PCA says it all.

**Response 10** Our apologies for the confusing wording, we will better justify the use of our methodology in the final manuscript. Our reference to numerical taxonomy is to acknowledge that in the historical, methodological, and philosophical principles our approach compares with the classical methodology. The methods employed in ecology, as synthesized in Numerical Ecology (Legendre and Legendre, 1998), are derived from different fields. One of those is numerical taxonomy or numerical classification which includes ordination, clustering, and phenetic resemblance methods. In the preface to Numerical Ecology this contribution was summarized. They note that the first synthesis was published by Sokal and Sneath (1963), who established numerical taxonomy as a new discipline (Legendre and Legendre, Numerical Ecology, 2nd edition, page xiii). Classical numerical classification or grouping methodology requires the computation of a measure of resemblance and the generation of a distance matrix (e.g. Bray-Curtis dissimilarity between all pairs of samples) from the data (OTU abundance data matrix), followed by the application of numerical methods (cluster analysis by UPGMA) to establish and define the clusters of similar entities from the resemblance or distance matrix. The results of the clustering can be represented by a dendrogram. The general principles and assumptions of numerical taxonomy can be summarized as follows: (1) The greater the information content and the more characters on which a classification is based the better the classification will be. (2) Every character is of equal weight. (3) Overall similarity of two entities is a function of their individual similarities in each of the many characters used for comparison. (4) Distinct groups can be recognized by correlations of character differences. (5) Classifications are based on overall (phenetic) similarity.

Our approach conceptually follows the general principles of numerical taxonomy, but applies a Bayesian hypothesis testing framework instead of clustering analysis to generate the dendrogram. Please see also comments to referee 1, section 3. Although next-generation sequencing technologies allow the sampling and identification of thousands of OTUs (characters), classical metrics like Bray-Curtis dissimilarity effectively use only a small fraction of those OTUs to discriminate between samples. Although this semimetric distance is the most commonly used measurement to express relationships in ecology, it is heavily influenced by the most abundant OTUs, implying these OTUs are the most likely to discriminate between the samples. We feel the classical approach is not well suited for the depth of sampling obtained with this new technology and statistical approaches like Bayesian inference are better suited to take full advantage of the large amount of character data while still preserving the general assumptions and philosophical principles of numerical classification.

The PCoA (Fig. 3) only recovers the major discontinuity in the ocean, depth, probably because only few, highly abundant OTUs are influencing the difference between the samples. The tree topology (Fig. 4), however, reveals this depth signal and a biogeography signal, which is both ecologically meaningful and imperceptible from the PCoA alone.

**Comment 11** Page 120, line 14: The authors begin their Results/Discussion with a lengthy paragraph that basically outlines the study again. This should have been done already in the Introduction. This paragraph should be cut by 50

**Response 11** We will address this in the revised manuscript.

**Comment 12** Page 122, line 2: to down bias is not a verb and its not clear. Please rewrite.

**Response 12** We will address this in the revised manuscript.

**Comment 13** Page 122, line 5: The authors mention that Analysis of the rarefaction curves suggests that some of the deep-water communities are more diverse (e.g. 2, 8, and 12) are more diverse than the rest of the communities. This is an important observation, but its unclear whether overall deep-water communities are more diverse or whether samples 2, 8, and 12 are exceptions. Its impossible to see the sample depth in Figure 2. The comparison between shallow and deep-water community diversity should be tested statistically.

**Response 13** In the revised manuscript, we will add depth information to the figure legend to make this more clear.

**Comment 14** Page 123-124: The authors here talk about the overall composition of the three parts of the water column, but no figure or table summarizes these data. To do this, the authors should either modify Table 2 or devise a new table. That is, it would look like Table 2, but with data for each part of the water column. See further comments below about Table 2.

**Response 14** We will address this in the revised manuscript. See response to comment 17.

**Comment 15** Page 124 line 10-17: This definition and overview of the deep chlorophyll maximum can be reduced by 50

**Response 15** We will address this in the revised manuscript

**Comment 16** Table 2: What is n versus Mean? These should be explained.

**Response 16** N is the number of reads in the OTU, percent is percent of the total analyzed reads. Mean/SD/Range describes the distribution of reads in the OTU across samples. We will make this more clear in the revised manuscript.

**Comment 17** A few entries in this table occur more than once for unclear reasons. For example, Pelagibacter occurs five times. What are these different "Pelagibacters?" The authors should give more information. Even less clear are the two entries called Bacteria. If they cant be classied to a finer taxonomic level, why arent they lumped together? Similarly for Proteobacteria. I suppose they are different OTUs which cannot be classied to a ner level. If so, this should be said. I wonder if they should be lumped together (e.g. the two Bacteria entries) with an explanation in the table caption. Although I can see an argument for this table, an argument against it is that it lumps together very different samples. To make it more useful, this table could be modied to summarize the composition of the most abundant taxa found in the three parts of the water column. That is, the table could give the top five taxa in each part of the water column and give the percentages for the relative abundance of the taxa in the other parts. In the extreme case of each water column section having a different top five, there would be 15 taxa listed in the table. To t these data, some of the columns now given in Table 2 can be deleted without any loss. These include the range and the mean (or N or both, depending on what these are). The percentages (+/- SD) are enough.

**Response 17** In the revised manuscript, we will provide a new table which enumerates the ranks, in each part of the water column, of the overall top 25 OTUs.

The numerical identifier of each OTU is shown in parentheses to differentiate OTUs with identical taxonomic labels, as is stated in the figure legend. That different OTUs have identical taxonomic classification points to the importance of this taxonomy-independent approach. Lumping together of OTUs with the same name undermines the importance performing a taxonomy-independent clustering. In effect, we are letting the sequences speak for themselves. The sequences in the Bacteria OTUs are different enough (3% on average) to be separated indicating that they are likely from phyla that do not appear in the sequencing classification database.

**Comment 18** Figure 2: Why use the normalized data and cut off samples with more sequences than the minimum? An advantage of this analysis is that all of the data can be plotted for looking at the shape of individual curves. To compare samples, the reader then can go to the sample with the fewest sequences and read up and down the figure.

**Response 18** What we are illustrating is that even with normalization, there is still evidence for incomplete sampling (higher degree of undiscovered diversity). Because the number of sequences varied so widely between samples, this was also an aesthetic choice. However, in looking at both sets of curves, the interpretation is the same.

**Comment 19** Figure 3: This figure should be deleted or replace with something that can be seen without magnification. This is pretty much impossible to digest as a figure, even if readers expand it several fold on their computer screens. A table would have a slightly better chance of showing the data more effectively. But the authors could concentrate on the most abundant taxa or lump taxa together and present totals of higher phylogenetic levels.

**Response 19** The 72 families are also identified by their corresponding 18 classes, which are summaries of the underlying OTUs. Even without magnification, the figure is compelling, showing highly differential distribution of taxa across our samples.

**Comment 20** Figure 4: More taxonomic information should be given for Family 2

**Response 20** The full taxonomy of the 15 sequences in the training set used by the mothur classifier for the Family II taxonomic "family" is: Bacteria/Cyanobacteria/Cyanobacteria/Cyanobacteria_order_incertae_sedis/ Family_II/GpIIa; We will provide more information on this in the revised manuscript.

**Comment 21** Figure 5: The solid and dashed lines should be explained.

**Response 21** We will make this clear in the figure as stated in the text on 126, lines 8-12

## References

B. Anderson, K. Arrigo, B. Ward, K.M. UCI, and B. Van Mooy. The Molecular Biology of Biogeochemistry: Using Molecular Methods to Link Ocean Chemistry with Biological Activity. November 2010.

Jo Handelsman and K Smalla. Conversations with the silent majority. *Current Opinion In Microbiology*, 6 (3):271–273, January 2003.

Jo Handelsman, MR Rondon, SF Brady, J Clardy, and RM Goodman. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, 5(10): R245–R249, 1998.

Pierre Legendre and Louis Legendre. *Numerical Ecology, Volume 24, Second Edition (Developments in Environmental Modelling)*. Elsevier Science, December 1998.

Alban Ramette. Multivariate analyses in microbial ecology. *Fems Microbiology Ecology*, 62(2):142–160, November 2007.

Robert R Sokal and Peter H A Sneath. *Principles of Numerical Taxonomy (A Series of Books in Biology)*. W. H. Freeman and Company, 1st edition, 1963.

C R Woese, O Kandler, and M L Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, June 1990.

John C. Wooley, Adam Godzik, and Iddo Friedberg. A Primer on Metagenomics. *Plos Computational Biology*, 6(2):–, 2010.