Report #1
Submitted on 22 Jun 2020
Anonymous Referee #2

In this version, the authors have compared their results with those in the previous studies, which resulted in a more thorough discussion about the nifH phylotypes in OMZs of the world oceans. The manuscript has been improved significantly with these revisions.

Nevertheless, I still have a major concern about the insufficient dataset in this study. According to the information in Table 1, there were few sequences per sample, some samples even have only 3-4 clones being sequenced. This problem is exactly what I was worried about and that's why I asked the authors to provide more details (Table 1) in the previous review. With such limited dataset, it is hard to say anything about the relative abundance of the OTUs or compare the nifH communities at different depth and OMZs. The authors need to either increase the sequencing depth or convince the audiences that their results can really represent "nifH phylotypes/community in the world OMZs". In the current version, it is obvious that the authors could not draw any convincing conclusions. Besides, I have some specific comments as follows.

Response:
The reviewer is correct that clone libraries cannot provide the depth of sequencing we now expect, and *IF* we were performing the sequencing experiments today, we would use NGS, not clone libraries. The fact is that the samples were collected many years ago and it is not possible to analyze them now using NGS.

We disagree, however, with the reviewer's conclusion that the clone libraries are not sufficient to yield interesting and convincing conclusions. As far as we are aware, there is only one NGS study of nifH in an OMZ region (Cheung et al. 2016, and that one sampled a few depths in the Costa Rica Dome, not one of the major OMZs). Thus our study is unique in being able to compare nifH community composition from all three major oceanic OMZs. Most of the other studies, in whose context we frame our results, were clone library studies, with the same limitations as our own, except that they all considered only one of the OMZs. So they are all suboptimal, but useful and interesting in aggregate.

While the individual sequences are interesting in terms of their phylogenetic affiliation, all analysis of patterns and distributions was done on the basis of OTUs, which were determined from the aggregated sequences. Aggregation was done in order to minimize the kind of bias that derives from small samples size, which is the concern of the reviewer.

Given the reviewer's preference for NGS data, the comparison of the present study with Cheung et al. 2016 is particularly interesting. We analyzed 16 samples for DNA and RNA, and found a total of 59 OTUs (41 for Cluster I, 18 for Clusters II, III, IV). Cheung et al. (2016) analyzed 15 samples for DNA only and found 37 OTUs at the 95% level. All of their samples All of their samples showed strong dominance in the community composition. It's really striking that even with many more sequences, that study found only five abundant, i.e., dominant, types and one of them was also in the top four in our study. The 37 OTUs from Cheung et al. do imply that

deeper sequencing in our samples would result in higher numbers of OTUs. But it does not imply that we would see a different community composition in terms of structure/ dominance.

This reviewer's critique has helped us to improve and expand the comparative analysis in our manuscript. We hope the conclusions are clearer now, as well as the basis upon which they were made (see below).

L16-17: "The OTUs were biogeographically distinct for the most part – there was little overlap among regions," Would it be a result of limited sequencing depth in your study? Based on Table 1, only the most dominant phylotypes could be detected with 3-50 sequences per sample.

Response:
True, only the most dominant phylotypes would be detected at low sequencing depth – which actually adds ecological significance to the statement that there was little overlap among regions – dominance was observed in all three regions (i.e., the assemblage might be more diverse than we can evaluate but the most abundant types differed between locations). That is probably a robust finding, and we have made that explicit in the revised manuscript. That observation is supported by the only relevant NGS study (Cheung et al. 2016 as described in the preceding paragraph) as well in the other limited clone library studies which are available for comparison.

L54: I did not see 32 samples in Table 1.

Response:
We considered the analysis to include 32 samples because 16 samples were analyzed separately for RNA and DNA. We've revised it to make clear that there were 16 water samples but 30 clone libraries.

L88: You should construct maximum-likelihood tree and choose the best-fit model using model test.

Response:
We have taken that suggestion and redone the phylogenetic tree using Maximum Likelihood methods. We used the Poisson model as the one used by previous authors working with nifH (including Cheung et al. 2016). All of the major branch points are identical to the previously presented Neighbor Joining tree for Cluster I, but there were several changes for Clusters II, III, IV, which make the positions more consistent with previous work.

L149: How can you compare community composition and biogeography with 3-50 sequences per sample?

Response:
These comparisons were all made by grouping depths and stations within regions. The OTUs were defined from groups of sequences (Table 2). All the groups contained 86 – 275

sequences, with one exception: 10 sequences for Clusters II, III, IV for the Arabian Sea. None of the comparisons were made on the basis of 3- 50 sequences.

L405-408: It seems over-speculation to me. The "similar metabolic types" was not supported by the result. You only sequenced some clones of nifH gene fragments. Also, without the actual abundance (i.e. gene copy number), you cannot make any statement about "bloom". "Most of the sites/depths, both in this study and in others from OMZ regions, are dominated by one or a few OTUs" could be due to limited sequencing depth or low diversity of nifH phylotypes in the samples.

Response:
As described above, the conclusion that there are probably only a few important OTUs when the sequences are grouped by depth or region is a robust finding. Deeper sequencing would not discover new dominants, it would only make the tail of the distribution longer. Good point about the metabolic types. We have changed the wording to make it clear that the metabolic type is an assumption based on the phylogenetic affiliation of the nifH sequence (Line 410) by inserting the phrase "nifH genes associated with" similar metabolic types...

Report #2
Submitted on 30 Jun 2020
Anonymous Referee #3

Review
Jayakumar and Ward present an intercomparison of the diazotroph community composition of three prominent OMZs (ETNP, ETSP and Arabian Sea). The approach is used to analyse diversity is clone sequencing, which is arguably outdated nowadays and does not provide enough coverage, and the authors should acknowledge this drawback more clearly and be more cautious when interpreting their results accordingly. Nevertheless, the strength of the manuscript is the intercomparison of the three regions, which has not been addressed in previous studies. Hence, I consider that this work is of interest for the community and deserves publication, but first some minor changes are needed. The ecological role of diazotrophs in OMZs is not properly introduced, nor discussed. The results and discussion section look more like a report than a proper discussion. Finally, interpretations and conclusions need to be reformulated and acknowledge the drawbacks of the methods used. Below I provide specific comments.

Response:
We recognize the limitations of the clone library approach (see response to Referee #2) and have taken the advice of both referees to acknowledge that in the text and to point out where different conclusions might derive from deeper sequencing. We discuss our results in the context of all the nifH/OMZ publications we could find and have been careful not to overstate the power of the analysis.

It's not clear what more we can say about the "ecological role of diazotrophs in the OMZs". That question is one of the motivations for the work, but the analysis so far does not point to

an obvious answer to why nifH genes are so prevalent or what role N fixation might play in the OMZ.

Introduction
L26: "sp." should not be italicized.

Response:
Done.

L29: I suggest using sunlit waters or euphotic zone instead to account for the depth range where cyanobacterial diazotrophs are found in (sub)tropical waters.

Response:
We added "sunlit" to the description but kept the word "surface" as well, because we use "surface" as category in the later analysis (L35).

L33: The hypothesis tested in Deutsch et al. 2007 has been extensively turned down by several publications in the past years (e.g. Bonnet et al., 2017; Knapp et al., 2016). Please account for this in your text. As it reads now, the reader perceives that this is a confirmed hypothesis still.

Response:
We have always been skeptical of this hypothesis. However, it was a motivating factor in the research. The sentence simply states that the Deutsch et al. proposed the idea, not that it is true.

L43: But check (Bentzon-Tilia et al., 2015).

Response:
Bentzon-Tilia et al 2015 detected N2 fixation in samples in which heterotrophic diazotrophs were abundant and were able to estimate indirectly that the heterotrophs might have contributed 20 – 50 % to one of the measured rates. That study was in two very shallow estuaries, so it's hard to know how to relate them to this work or the N fixation phenomenon in OMZs. Interestingly, however, some of the non-cyanobacterial nifH OTUs were phylogenetically very similar to the heterotrophic diazotrophs reported here and elsewhere. We can cite the Bentzon-Tilia et al. paper and replace "not" with "rarely".

Materials and Methods
The methods are described in great detail. However, can the authors confirm that a DNase treatment was used and no-RT controls run?

Response:
RT controls were mentioned in the text (implied no-RT) but we have made both of these steps explicit in the revised text (L 82).

L49: Please indicate which are the three major OMZ regions you include in your analysis. Perhaps it would fit better in the introduction, a brief description of the three regions, what they have in common, what they differ on…

Response:
We have added very brief descriptions of the three OMZ regions in the introduction (L55).

L51: For no expert readers, an introduction to what ODZ means (which oxygen thresholds are considered), would be helpful.

Response:
We removed this unnecessary acronym.

L93-94: This is unclear. Each sequence that was 3% different from each other was considered an OTU? Can you provide a reference?

Response:
An OTU is simply defined by some threshold cutoff in identity. There is no standard cutoff, so we looked at several cutoffs (between 3% and 10%) and settled on 3% as the one that makes phylogenetic and biogeographically meaningful distinctions. So yes, sequences that are 3% different would be in different OTUs. Useful OTU cutoffs differ among different genes (e.g., ribosomal genes are less variable than functional genes as a rule) and some functional genes are more highly conserved than others (e.g., RuBisCO is much more conserved that nitrate reductase in diatoms). We cited Schloss and Handlesman (2009) for the OTU threshold method and Gaby et al (2018) for the threshold that is meaningful for nifH.

Results and Discussion
L109-111: The authors should make it clear that this is a compilation of previously published data.

Response:
The papers in which they were previously published are explicitly cited in L135, previously L110.

L174: Groups that are both denitrifiers and N_2 fixers are rather common in low oxygen environments, including coastal sediments. A discussion on the double ecological role of these groups in OMZs would be appreciated.

Response:
True, the cooccurrence of N2 fixers and denitrifiers, even of both capabilities within the same microbe, is often observed. To comment on the ecological role or the biogeochemical significance of that finding would be pure speculation here, as we have no information to go on other than the presence of the sequences.

L205-206: Not sure what the authors mean here, why the separation between alfa and beta was unlear?

Response:
Alpha and Beta -proteobacterial are distinguished on the basis of their 16S rRNA genes, but their nifH sequences don't always follow the same distinctions. So you can't always tell from the nifH alone whether it comes from an Alpha or a Beta -proteobacterium. That must mean a combination of lateral gene transfer and evolution has occurred somewhat independently for the functional gene, which is subject to different selection than is the rRNA gene. In the phylogenetic tree, it is obvious that the nifH sequences do not fall out exactly along the branch pattern predicted from the 16S rRNA tree for the Betaproteobacteria.

L224: Please add references.

Response:
The references can be found by following the accession numbers, which are clearly presented in the tree itself. It is not usually necessary to list the citations for every sequence, when that information is associated with the unique accession numbers. Each of the sequences listed in this paragraph is identified exactly in the tree and the reader can follow the references from the accession numbers. This is the usual convention for citation because not every sequence is associated with a literature publication and not every sequence included in the tree is mentioned in the text.

L251-253: This is a nice hypothesis but it is not sufficiently explained or referenced. How could this be tested? (idem for L355-357). What references do the authors have of noncyanobacterial diazotrophs blooming upon important inputs of organic matter?

Response:
We mention it as a hypothesis, but testing it is far beyond the scope of this work. The analogy of denitrifiers was provided, and we have now added a sentence about a published study in which N2 fixation was stimulated by organic matter addition (L301).

L357-359: Please add a reference at the end of this sentence.

Response:
That sentence is just another statement of the hypothesis mentioned above. It is a suggestion derived from our own work, not a citation.

L366: Replace N for N_2.

Response:
Done.

L367: Or rather for the nitrogen budget of OMZ zones…

Response:
If it is a minor contribution to the N budget of the OMZ, it is certainly a minor contribution to the N budget of the ocean. It is the latter that people have been trying to address by looking for N fixation in OMZs, i.e., the ocean is the box of interest here, not the OMZ as a separate box in the oceanic inventory.

L370: This is the first time that OMZ is included as part of the "dark ocean". No explanation or comparison with non-cyanobacterial diazotrophy in the dark ocean realm are provided in the manuscript.

Response:
We removed reference to the dark ocean.

Tables
Table 1: A reference of the base of the photic zone at each of these stations would be useful.

Response:

Thanks for suggesting that clarification. We have added a column to Table 1 to indicate which samples were considered surface (within the euphotic zone) and which were OMZ (all of which are below the euphotic zone).

References

Bentzon-Tilia, M., Traving, S. J., Mantikci, M., Knudsen-Leerbeck, H., Hansen, J. L. S., Markager, S., & Riemann, L. (2015). Significant N2 fixation by heterotrophs, photoheterotrophs and heterocystous cyanobacteria in two temperate estuaries. ISME Journal, 9(2), 273–285. https://doi.org/10.1038/ismej.2014.119

Bonnet, S., Caffin, M., Berthelot, H., & Moutin, T. (2017). Hot spot of N2 fixation in the western tropical South Pacific pleads for a spatial decoupling between N2 fixation and denitrification. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1619514114

Knapp, A. N., Casciotti, K. L., Berelson, W. M., Prokopenko, M. G., & Capone, D. G. (2016). Low rates of nitrogen fixation in eastern tropical South Pacific surface waters. Proceedings of the National Academy of Sciences, 113(16), 4398–4403. https://doi.org/10.1073/pnas.1515641113

List of all relevant changes made in the manuscript:

1. As suggested by both referees, we have explicitly addressed the limitations of clone libraries for diversity and biogeographic analysis. Referee #2 may still be critical, but we point out where and how the analysis would be affected by increased sequencing depth. Importantly, we point out that one of the major patterns observed is strong dominance of a few OTUs, which is observed in all the clone library studies and in the one NGS study available for comparison. We point out the strength of the current data set, as appreciated by Referee #3, which is the ability to compare all three OMZs using similar kinds of data sets. Because our analysis is based on grouped samples, the small sample size problem emphasized by Referee #2 is minimized within the constraints of the clone library approach.

2. A few sentences were added to the abstract to emphasize the main findings of the paper.

3. A few additional citations have been added, in order to support some of the inferences drawn from our data (e.g., Bonnet et al. 2013, Hamersley et al. 2011).

4. We have redone the phylogenetic trees in figures 3 and 6, using methods suggested by Referee #2 and have implemented minimal changes to the manuscript to accompany the minor changes in the interpretation of the results. Three new citations for the methods used to produce the new trees are included.

5. We have added a brief description of OMZs in the introduction as suggested by Referee #3.

6. A new column has been inserted in Table 1, as suggested by Referee #3, to indicate which samples were considered surface (within the euphotic zone) and which were within OMZs.

Diversity and distribution of Nitrogen Fixation Genes in the Oxygen Minimum Zones of the World Oceans

[1]Amal Jayakumar and [1]Bess B. Ward

[1]Department of Geosciences
Princeton University
Princeton, NJ 08544

Abstract

Diversity and community composition of nitrogen fixing microbes in the three main oxygen minimum zones (OMZs) of the world ocean were investigated using operational taxonomic unit (OTU) analysis of *nifH* clone libraries. Representatives of the all four main clusters of *nifH* genes were detected. Cluster I sequences were most diverse in the surface waters and the most abundant OTUs were affiliated with Alpha- and Gammaproteobacteria. Cluster II, III, IV assemblages were most diverse at oxygen depleted depths and none of the sequences were closely related to sequences from cultivated organisms. The OTUs were biogeographically distinct for the most part – there was little overlap among regions, between depths or between cDNA and DNA. In this study of all three OMZ regions, and from the few other published reports from individual OMZ sites, dominance by a few OTUs was commonly observed. This pattern suggests dynamic response of the components of the overall diverse assemblage to variable environmental conditions. Community composition in most samples was not clearly explained by environmental factors, but the most abundant OTUs were differentially correlated with the obvious variables, temperature, salinity, oxygen and nitrite concentrations. Only a few cyanobacterial sequences were detected. The prevalence and diversity of microbes that harbour *nifH* genes in the OMZ regions, where low rates of N fixation are reported, remains an enigma.

Deleted:

Introduction

Nitrogen fixation is the biological process that introduces new biologically available nitrogen into the ocean, and thus constrains the overall productivity of large regions of the ocean where N is limiting to primary production. The most abundant and most important diazotrophs in the ocean are cyanobacteria, members of the filamentous genus *Trichodesmium* and several unicellular genera, including *Chrocosphaera* sp. and the symbiotic genus *Candidatus* Atelocyanobacterium thalassa (UCYN-A). Although these cyanobacterial species are wide spread and have different biogeographical distributions (Moisander et al. 2010), they are restricted to sunlit surface waters, mainly in tropical or subtropical regions.

Because diazotrophs have an ecological advantage in N depleted waters, and because those conditions occur in the vicinity of oxygen minimum zones, due to the loss of fixed N by denitrification, it has been proposed that N fixation should be favoured in regions of the ocean influenced by OMZs (Deutsch et al. 2007). It has also been suggested that the energetic constraints on N fixation might be partially alleviated under reducing, i.e., anoxic, conditions (Großkopf and LaRoche 2012). In response to these ideas, the search for organisms with the capacity to fix nitrogen has been focused recently in regions of the ocean that contain OMZs. That search usually takes the form of characterizing and quantifying one of the genes involved in the fixation reaction, *nifH*, which encodes the dinitrogenase reductase enzyme. Diverse *nifH* assemblages have been reported from the oxygen minimum zone of the Eastern Tropical South Pacific (Turk-Kubo et al. 2014, Loescher et al. 2016, Fernandez et al. 2011) and the Costa Rica Dome, at the edge of the OMZ in the Eastern Tropical North Pacific (Cheung et al 2016). The search for non cyanobacterial diazotrophs has resulted in discovery of diverse *nifH* genes, but they have rarely been associated

Formatted: Font: Not Italic

Deleted: not

with significant rates of N fixation (Moisander et al. 2017, Bentzon-Tilia et al. 2015). Thus the occurrence and diversity of putative diazotrophs in nitrogen rich aphotic waters remains unexplained.

Here we report on the distribution and diversity of *nifH* genes in all three of the world ocean's major OMZs. The two Pacific OMZs, the Eastern Tropical North (ETNP) and South (ETSP) Pacific, are both highly productive eastern boundary regions. The ETSP is the one of the most productive regions in the world ocean and has an oxygen depleted layer of about 400 m at its greatest depth. The ETNP is less well ventilated and less productive, with an anoxic layer of more than 700 m. The third major OMZ is the Arabian Sea, which is geographically constrained to the northern Indian Ocean. It experiences an annual monsoon cycle but is permanently and stably stratified with a maximum anoxic layer of about 800 m. Both surface and anoxic depths, and both DNA and cDNA (i.e., both presence and expression of the *nifH* genes) were investigated. The approach used here to investigate diazotroph assemblages is based on clone library analysis of *nifH* sequences. Next generation amplicon sequencing would yield greater numbers of sequences, although it might not overcome the primer bias associated with PCR and cloning. The strength of the current study is the inclusion of similar data from all three OMZs. By comparing these results to previous studies using the same and other methods, we find robust biogeographical patterns and community structure among the non-cyanobacterial diazotroph assemblages.

Materials and Methods:

Samples analysed for this study were collected from the three major OMZ regions of the world oceans (16 total samples, Table1) from surface and oxygen minimum zone (OMZ, including oxycline and anoxic) depths. Particulate material from water samples (5 – 10 L), collected using

Deleted: ,

Deleted: including samples from b

Formatted: Font: Italic

Deleted: 32

Deleted: and 2

Deleted:

Deleted: ,

Deleted: oxygen depleted

Deleted: zone (ODZ)

Niskin samplers mounted on a CTD (Conductivity-Temperature-Depth) rosette system (Sea-Bird Electronics), was filtered onto Sterivex capsules (0.2 µm filter, Millipore, Inc., Bedford, MA) immediately after collection using peristaltic pumps. The filters were flash frozen in liquid nitrogen and stored at -80°C until DNA and RNA could be extracted. For samples from the Arabian Sea, DNA extraction was carried out using the PUREGENE™ Genomic DNA Isolation Kit (Qiagen, Germantown, MD) and the RNA was extracted using the ALLPrep DNA/RNA Mini Kit (Qiagen, Germantown, MD). For samples collected from ETNP and ETSP DNA and RNA were simultaneously extracted using the ALLPrep DNA/RNA Mini Kit (Qiagen, Germantown, MD). SuperScript III First Strand Synthesis System (Invitrogen, Carlsbad, CA, USA) was used to synthesise cDNA immediately after extraction following purification of RNA using the procedure described by the manufacturer, including no-RT controls. The extracted DNA was treated with DNase before transcription and no-RT controls verified the absence of *nifH* DNA in the RNA preps. DNA was quantified using PicoGreen fluorescence (Molecular Probes, Eugene, OR) calibrated with several dilutions of phage lambda standards.

PCR amplification of *nifH* genes from environmental sample DNA and cDNA was done on an MJ100 Thermal Cycler (MJ Research) using Promega PCR kit following the nested reaction (Zehr et al. 1998), with slight modification as in Jayakumar et al. (2017). Briefly, 25µl PCR reactions containing 50 pmoles each of outer primer and 20-25ng of template DNA, were amplified for 30 cycles (1 min at 98°C, 1 min at 57°C, 1 min at 72°C), followed by amplification with the inner PCR primers 50 pmoles each (Zehr and McReynolds 1989). Water for negative controls and PCR was freshly autoclaved and UV-irradiated every day. Negative controls were run with every PCR experiment, to minimize the possibility of amplifying contaminants (Zehr et al. 2003). The PCR preparation station was also UV irradiated for 1 hour before use each day and the number of

Deleted: ,

Formatted: Font: Italic

amplification cycles was limited to 30 for each reaction. Each reagent was tested separately for amplification in negative controls. *nifH* bands were excised from PCR products after electrophoresis on 1.2% agarose gel, and were cleaned using a QIAquick Nucleotide Removal Kit (Qiagen). Clean *nifH* products were inserted into a pCR®2.1-TOPO® vector using One Shot® TOP10 Chemically Competent *E. coli*, TOPO TA Cloning® Kit (Invitrogen) according to manufacturer's specifications. This process resulted in 30 clone libraries, 16 of DNA and 14 of RNA from the 16 samples (Table 1).

Inserted fragments were amplified with M13 Forward (-20) and M13 Reverse primers from randomly picked clones. PCR products were sequenced at Macrogen DNA Analysis Facility using Big Dye™ terminator chemistry (Applied Biosystems, Carlsbad, CA, USA). Sequences were edited using FinchTV ver. 1.4.0 (Geospiza Inc.), and checked for identity using BLAST. Consensus *nifH* sequences (359 bp) were translated to amino acid (aa) sequences (108 aa after trimming the primer region) and aligned using ClustalW in MEGA X (Kumar et al. 2018, Stecher et al. 2020) along with published *nifH* sequences from the NCBI database. The alignment was used to construct a maximum likelihood (ML) phylogenetic tree in MEGA X, based on the Poisson model and the phylogenetic tree was edited using iTOL (Letunic and Bork 2016). Bootstrap analysis was used to estimate the reliability of phylogenetic reconstruction (1000 iterations). The *nifH* sequence from *Methanosarcina lacustris* (AAL02156) was used as an outgroup. The accession numbers from GenBank for the *nifH* sequences in this study are Arabian Sea DNA sequences JF429940- JF429973 and cDNA sequences accession numbers JQ358610-JQ358707, ETNP DNA sequences KY967751-KY967929 and cDNA sequence KY967930-KY968089, and ETSP DNA sequences MK408165-MK408307 and cDNA sequences MK408308-MK408422.

Deleted: X (Thompson et al. 1997)

Deleted: Kumar et al 2018

Deleted: produced,

Deleted: Neighbor-joining trees were produced from the alignment using distance matrix methods (PAUP 4.0, Sinauer Associates).

Formatted: Font: Times New Roman

The *nifH* nucleotide alignment (of 787 sequences) was used to define operational taxonomic units (OTUs) on the basis of DNA sequence identity. Distance matrices based on this nucleotide alignment were generated in MOTHUR (Schloss and Handlesman 2009). The relative *nifH* richness within each clone library was evaluated using rarefaction analysis. OTUs were defined as sequences which differed by ≤3% using the furthest neighbor method in the MOTHUR program (Schloss and Handlesman 2009). The 3% OTU definition is similar to the level at which species are conventionally defined using 16S rDNA sequences, so it may overestimate the meaningful diversity of the functional gene. Redundancy analysis was performed in R using the vegan package. Environmental variables were transformed using decostand.

Results and Discussion:

DNA and cDNA sequences (787 in total) derived from the OMZ regions of the Arabian Sea (AS), Eastern Tropical North Pacific (ETNP) and Eastern Tropical South Pacific (ETSP) were subjected to OTU and phylogenetic analyses to compare the diversity and community composition, biogeography and gene expression, of *nifH* possessing microbes among the three OMZ regions. Phylogenetic analysis of the sequences from the AS, ETNP and ETSP were reported separately in previous publications (Jayakumar et al. 2012, Jayakumar et al. 2017, Chang et al. 2019), but the sequences have been combined for additional global analyses here. We compared the threshold OTU definitions at 3 and 10% and found that the number of OTUs decreased, as expected, as the resolution decreased. Even at the 3% threshold, however, OTUs tended to separate by depth and location, indicating a functionally useful distinction at this level. Thresholds of 3 – 5% as the OTU definition correspond to within and between species level

Deleted: previously

distinctions for *nifH* (Gaby et al. 2018). The sequences from the OMZ regions represented three of the four sequence clusters (I, II, III, IV) described by Zehr et al. (1998).

Cluster I *nifH* OTU distributions: Diversity analysis of the *nifH* cluster 1 sequences for the three OMZs based on OTUs using MOTHUR identified 41 OTUs at a distance threshold of 3% (Table 2). The number of sequences and the number of OTUs varied widely among depths and stations, so the results are grouped by region (AS, ETNP, ETSP) or depth horizon (surface or OMZ, including upper oxycline depths) or cDNA vs DNA (Table 2). Grouping the sequences by depth horizon (surface or OMZ), region (AS, ETSP, ETNP) or DNA/RNA, allows the detection of patterns that are not driven by the relatively low number of sequences obtained from some of the individual clone libraries. The OTUs are numbered in order of decreasing abundance in the clone library, i.e., OTU-1 was the most common OTU.

For all regions and depths combined, the number of OTUs detected (41) was less than the sum of OTUs detected when each region was analyzed separately (45), indicating that there was some overlap of OTUs among regions. The overlap was not large, however. Only three of the 12 most abundant OTUs contained sequences from more than one region and none contained sequences from all three regions (Figure 1A). When sequences for all three regions were combined, only four of the 12 most abundant OTUs contained sequences from both depth horizons (Figure 1B). Most OTUs represented a single depth, and many a single sample. This suggests a pattern of dominance, rather than evenness, in the *nifH* assemblage. Deeper sequencing is therefore expected to discover a larger number of rare OTUs, but might not change the picture that emerges here of a small number of abundant clades. Interestingly, Cheung et al. (2016) reported a similar pattern of dominance based on a larger DNA sequence dataset from

Deleted: all

Deleted: Supplemental

Deleted: 1A

Formatted: Font: Italic

only one location. Using 454-pyrosequencing to obtain a similar number of OTUs (37 total) from the Costa Rica Dome, all of the 15 samples investigated by Cheung et al. (2016) were dominated (>50%) by one of five major OTUs.

The Arabian Sea was strikingly less diverse than other regions and sample subsets (Figure 2). For example, when all DNA and cDNA sequences for all depths are grouped together, the Arabian Sea (OTUs = 14, Chao = 21) contains less species richness than the combined surface samples from all three regions (OTUs = 25, Chao = 52), despite having a similar number of total sequences (178 for the Arabian Sea, 198 for all surface samples combined). This lack of diversity in the AS data may be partly due to the preponderance of cDNA sequences, which generally contained less diversity than a similar number of DNA sequences (see below).

Although similar numbers of sequences were obtained for cDNA (255) vs DNA (257), the OTU "density", i.e., number of OTUs per number of sequences analyzed, was higher for DNA (0.136 for DNA, 0.094 for cDNA). The Chao statistic verified this observation for the combined data from each region in predicting higher total numbers of OTUs for DNA (Chao = 42) than for cDNA (Chao = 24). This difference could indicate that some of the *nifH* genes present were not expressed at the time of sampling, but the cDNA sequences were not simply a subset of the DNA community. Half of the 12 most abundant OTUs contained either cDNA or DNA (Figure 1C), meaning that some genes were never expressed and some expressed genes could not be detected in the DNA. Based on a similar number of sequences from each sample (1 – 52 per sample) from the ETSP, Turk-Kubo et al. (2014) also found that DNA and cDNA clones were differently distributed among stations; one phylotype was recovered exclusively from cDNA and only one phylotype occurred in both DNA and cDNA. The relatively low

Deleted: used

Deleted: and

sequencing depth associated with clone library studies limits the sensitivity of this comparison, but it clearly shows that dominant components of the DNA and cDNA libraries frequently represent different subsets of the total assemblage.

For all regions combined, similar numbers of OTUs were detected in surface waters (OTUs = 25) and in OMZ samples (OTUs = 23), although a larger number of sequences was analyzed for the OMZ environment (198 vs. 314 sequences for surface and OMZ depths, respectively). It might be expected that the presence of phototrophic diazotrophs in the surface water would lead to greater diversity there, but only one OTU representing a known cyanobacterial phototroph (OTU-12 = *Katagmynene spiralis* or *Trichodesmium*) was identified, so most of the additional diversity must be present in heterotrophic or unknown sequences.

Rarefaction curves (Figure 2) indicate that sampling did not approach saturation either for region or depth. The Chao statistic also indicated that much diversity remains to be explored, despite the great uncertainty in these estimates. The total number of OTUs detected, the shape of the rarefaction curve and the diversity indicators (Figure 2, Table 2) all indicate that the greatest *nifH* diversity occurred in surface waters, and much of that diversity was in singletons, i.e., not represented in the 12 most abundant OTUs, which represented 441 (86 %) of the total 512 *nifH* Cluster 1 sequences analyzed. Most of that diversity was contained in the ETNP, not solely a function of number of sequences analyzed (Figure 2).

Cluster I *nifH* Phylogeny: Phylogenetic affiliations at both DNA and protein level are shown for the 12 most abundant OTUs in Table 3. The most abundant OTU (129 sequences), OTU-1, contained Gammaproteobacterial DNA and cDNA sequences from both surface and OMZ depths of the ETNP and cDNA sequences from oxycline and OMZ depths in the Arabian Sea (Figure 3). Although very similar to each other, none of these sequences had higher than

91% identity at the DNA level (96% at AA level) with cultivated strains and were most closely related to *Pseudomonas stutzeri*. *P. stutzeri* is a commonly isolated marine denitrifier, but it is also known to possess the capacity for N fixation (Krotzky and Werner 1987). OTU-4, OTU-6 and OTU-8 also contained Gammaproteobacterial sequences. All had high identity with cultivated strains at the protein level but none were >91% identical to cultivated strains at the DNA level.

Gammaproteobacterial sequences with very close identities to *Azotobacter vinelandii* have been reported from the Arabian Sea ODZ and also from the ETSP (Turk-Kubo et al. 2014). This group of *nifH* sequences with close identities to *A. vinelandii* was also retrieved from the English Channel, Himalayan soil, South Pacific gyre, Gulf of Mexico, mangrove soil and many other environments (Figure 3). *Azotobacter*- like sequences were included in OTU-6 but were not closest identity at the DNA level. Although a large number of clones were analyzed here, no sequence that was closely associated with *A. vinelandii* was retrieved from the three regions. None of the g-244774A11 sequences, Gammaproteobacterial relatives that were abundant in the South Pacific (Moisander et al. 2014), were detected in this study.

OTUs-2, 3, 5, 10, and 11 all represented Alphaproteobacterial sequences, with closest identities to various *Bradyrhizobium, Sphingomonas* and *Methylosinus* species. Thus, Alphaproteobacterial sequences (206 sequences) were the most abundant in the clone library. OTU-2 contained almost exclusively ETSP ODZ DNA and cDNA sequences (plus one AS ODZ DNA sequence). OTU-3 contained DNA sequences from ETNP surface waters. OTU-5 contained exclusively Arabian Sea DNA sequences from Station 3, while OTU-10 contained only surface samples from the ETNP. An OTU threshold of 11% grouped all (179 sequences in five OTUs) of

255 these Alphaproteobacterial sequences together, but the 3% threshold is consistent with the
256 phylogenetic tree, which shows small scale biogeographical separation of sequence groups.
257 OTUs-7 and -9 were identified as Betaproteobacteria with closest identities to *Rubrivivax*
258 *gelatinosum* and *Burkholderia*, 91 and 90% respectively at the DNA level. However, at the AA
259 level, these sequences were 99 and 100% identical to *Novosphingobium malaysiense* and *S.*
260 *azotifigens,* both Alphaproteobacteria, and again were biogeographically distinct. OTU-7 contained
261 25 DNA sequences from the ODZ depths in the Arabian Sea, and OTU-9 contained 17
262 *Burkholderia*-like sequences from the oxycline at Station 1 in the Arabian Sea. No
263 Betaproteobacterial *nifH* sequences were detected in the ETNP or ETSP, but sequences similar to
264 *Burkholderia phymatum*, *Cupriavidus sp.* and *Sinorhizobium meliloti* were reported from ETSP
265 previously (Fernandez et al. 2015). Consistent with our previous report, however, there is no clear
266 separation between the alfa and the beta groups in *nifH* phylogeny (Jayakumar et al 2017).
267 Most of the Cluster I ETSP sequences from this study were contained in two OTUs (2 and 4).
268 OTU-2 contained 89 Alphaproteobacterial sequences with >98% identity to *nifH* sequences from
269 *Bradyrhizobium* sp. Uncultured bacterial sequences retrieved from the South China Sea, English
270 Channel, mangrove sediment, wastewater treatment and grassland soil were related to these ETSP
271 sequences. OTU-4 contained 29 Gammaproteobacterial sequences retrieved from both surface and
272 ODZ depths. Four of the remaining ETSP Cluster I sequences were grouped together as OTU-17
273 (Alphaproteobacteria, 89 and 96% identities with *Methyloceanibacter sp.* and *Bradyrhizobium sp.* at
274 the DNA and AA level respectively), three were in OTU-23 (*Bradyrhizobium* 100% identity) and
275 two were singletons. One of the singletons was most closely related to uncultured soil and sediment
276 sequences and to *Azorhizobium sp*. (86%) and one had 97% identity with *Bradyrhizobium*
277 *denitrificans* and many sequences from marine sediments.

OTU-22 represents the Deltaproteobacterial group. This novel group was reported previously from the ETNP (Jayakumar et al. 2017) and has three sequences from Arabian sea (OTU-22) and two singletons from ETNP surface waters. *nifH* possessing Deltaproteobacteria have been reported not only from all the three ODZs but also in several other marine environments including Chesapeake Bay water column, microbial mats from intertidal sandy beach in a Dutch barrier island, Jiaozhou Bay sediment, Rongcheng Bay sediment, Bohai Sea, Mediterranean Sea, Narragansett Bay, and the south Pacific gyre.

Proteobacteria-like sequences, especially Alpha- and Gammaproteobacteria, are the most frequently reported *nifH* sequences from the OMZs studied here and similar environments. Thirty one of 37 OTUs detected by Cheung et al. (2016) in the Costa Rica Dome OMZ were Proteobacteria, the two most common OTUs being closely related to Alphaproteobacterium *Methylocella palustris* and the Gammaproteobacterium *Vibrio diazotrophicus*. Loescher et al. (2014, 2016) also found *V. diazotrophicus*-like sequences, as well as several other Gammaproteobacteria in the ETSP. *V. diazotrophicus* was reported previously in the Arabian Sea (Jayakumar et al. 2012) but was not prominent in the present study. Sequences most similar to various *V. diazotrophicus*, other *Vibrio* species, and other Gammaproteobateria, including *P. stutzeri*, were the most common non-cyanobacterial Cluster I sequences reported for the low oxygen waters of the Southern California Bight (Hamersley et al. 2011). *Bradyrhizobium spp.*, one of the most common genera reported here and in surface waters of the Arabian Sea (Bird and Wyman 2013) and by Fernandez et al. (2011) in the ETSP, were also detected in the Costa Rica Dome OMZ and were the dominant OTU at 1000 m at one station (Cheung et al. 2016). *Bradyrhizobium*-like sequences were the most abundant among those amplified from ODZ incubations in which the N_2 fixation rate was enhanced by the addition of glucose (Bonnet et al 2013). In addition to

Deleted:

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Subscript

Bradyrhizobium-like and *Teredinibacter*-like *nifH* sequences, Turk-Kubo et al. (2014) found four other abundant Gammaproteobactera-like *nifH* sequences, which were entirely novel. The "Gamma A", which are commonly reported non-cyanobacteria diazotroph *nifH* sequences from non-OMZ environments (Langlois et al. 2015, Moisander et al. 2017), were represented by a singleton from the ETNP in the present study.

nifH sequences related to various Alphaproteobacterial methylotrophs are commonly found in OMZs: *Methylosinus trichosporium*-like sequences, which are reported here in OTU-5 from the Arabian Sea at both surface and ODZ depths, were also reported by Fernandez et al. (2011) in the ETSP. *Methylocella palustris*-like *nifH* genes comprised the most common OTU in the ODZ core depths in the Costa Rica Dome (Cheung et al. 2016). *M trichosporium* and *M. palustris* represent obligate and facultative methanotrophs, respectively, both also obligately aerobic. Detection of *nifH* genes closely related to those of methanotrophs does not prove that methanotrophy is present or important in the anoxic environment of the ODZ but the consistency of this finding across sites motivates further investigation of the potential for methane production and consumption in ODZs.

Deleted: on

The pattern of high diversity of *nifH*-bearing mostly heterotrophic microbes, but dominance in each sample by one or a small number of *nifH* OTUs, suggests a bloom and bust pattern of organic matter-supported growth. That is, we suggest that organic matter, which is supplied episodically in the upwelling regimes, stimulates the growth of copiotrophic microbes that respond rapidly in bloom like fashion. This bloom scenario has been described for denitrifying bacteria based on the OTU patterns observed in the *nirS* and *nirK* genes as a function of the stage of denitrification in both natural assemblages and incubated samples from OMZs (Jayakumar et al. 2009). Amino acids and glucose both stimulated N_2 fixation in OMZ samples from the ETSP, and *nifH* sequences associated with Alpha- and Gammaproteobacteria, as well as Cluster III phylotypes,

Formatted: Subscript

Formatted: Font: Italic

were found in a glucose enrichment experiment (Bonnet et al. 2013) The role of *nifH* in these heterotrophic microbes is unclear, especially because rates of nitrogen fixation in these locations in the absence of cyanobacteria or nutrient enrichment is often very low (Turk-Kubo et al. 2014, Loescher et al. 2016, Chang et al. 2019).

Although *Trichodesmium*-like clones have been retrieved from the surface waters of the Arabian Sea and the ETNP OMZs, only ten clones (OTU-12) in the combined clone library analyzed here were related to *Trichodesmium* (98% identity), including both cDNA and DNA from the Arabian Sea and cDNA from the ETNP. These sequences were actually 100% identical to *Katagnymene spiralis*, a close relative of *Trichodesmium* isolated from the South Pacific Ocean. Turk-Kubo et al. (2014) also retrieved only a few cyanobacterial sequences from the ETSP. No other cyanobacterial *nifH* sequences were identified.

Clusters II, III, IV *nifH* OTU distributions: The other three *nifH* clusters were combined for OTU analysis due to the limited number of sequences and OTUs obtained. A total of 18 OTUs were identified in the combined set of 275 sequences with a 3% distance threshold (Table 2). Most of the Cluster II, III, IV sequences were from the ETNP and ETSP. As with the Cluster I sequences, there was very little geographic and depth overlap among these OTUs (Figure 4A, 4B). Only OTU-1 contained sequences from more than one site, the ETNP and the ETSP. OTU-2 contained only cDNA sequences representing ODZ depths at both ETNP stations. OTU-3 contained exclusively ETSP DNA sequences from surface and cDNA sequences from ODZ depths. Only 10 of the Cluster II, III, IV sequences were from the Arabian Sea, and they formed three separate OTUs, a greater "OTU density" than was present at either of the Pacific sites. As observed for Cluster I, most of the OTUs that were detected in the DNA were not being expressed, and those that were expressed were not detected in the DNA (Figure 4C).

Rarefaction curves (Figure 5) indicate that sampling for Cluster II, III, IV did not approach saturation. The Chao statistic also indicated that much diversity remains to be explored, despite the great uncertainty in these estimates. Unlike the Cluster I analysis, there were relatively few singletons in the Cluster II, III, IV data and the assemblages were dominated by a few types.

Clusters II, III, IV *nifH* phylogeny: Four large OTUs (OTU-1, -2, -4 and -6) in Clusters II, III, IV belonged to *nifH* Cluster IV and Alphaproteobacteria/Spirochaeta and Deltaproteobacteria were the dominant phylogenies (Table 3, Figure 6). The largest OTU, OTU-1, contained 88 DNA sequences from the ETNP ODZ depths from both stations and from both depths in the ETSP. This OTU had no similarity to any cultured microbe. OTU-4 contained 30 sequences from the ETSP, all cDNA from one surface station, in *nifH* Cluster IV.

OTU-2 (75 sequences) in Cluster IV contained only cDNA sequences, all from ODZ samples in the ETNP (both stations), and had no close relatives among cultivated species. Although Turk-Kubo et al. (2014) retrieved a few clones identified as belonging to Cluster II from the euphotic zone of the ETSP, we did not find any sequence falling into this Cluster. OTU-3 contained 35 sequences in Cluster III and was dominated by DNA sequences from surface depths of the ETSP. OTU-5 represented Deltaproteobacteria in *nifH* Cluster III and contained 18 identical DNA sequences from 90 m at Station BB1 in the ETNP. Thus, of the five most common OTUs (89% of the total Cluster II, III, IV sequences analyzed), only one could be identified to a closely related genus (i.e., OTU-4 with 90% identity with *R. palustris*) and there was no overlap between DNA and cDNA OTUs from the same depths.

The other 13 OTUs in the Cluster II, III, IV sequences represented either Cluster III or IV. None of these were very closely related to any cultivated sequences. OTU-6 contained both DNA

Deleted: Three

Deleted: I

Deleted: also

and cDNA from the OMZ at one ETSP station. OTU-7 contained four sequences from ETNP surface waters with close identities with a sequence retrieved from Bohai sea. OTU-11, had one DNA and one cDNA sequences from the ETSP. All of the other sequences were less than 84% identical to any sequence in the database and could only be loosely identified as Firmicutes or Proteobacteria.

Although there were few high identities with known species, many of the Cluster II, III, IV sequences (OTUs -2, -5, -7, -9, -10) were most closely affiliated with sulfate reducing clades at either the DNA or protein level. Four OTUs with highest identity to known sulfate reducers were reported by Cheung et al. (2016) and one of them comprised nearly 40% of the sequences in one anoxic sample. *nifH* sequences that cluster with *Desulfovibrio spp.* are often reported from ODZ samples (Turk-Kubo et al. 2014, Loescher et al. 2014, Fernandez et al. 2011). Consistent reports of *nifH* genes associated with obligate anaerobes involved in sulfate reduction suggests a role for this metabolism in the ODZ, again motivating further research on the significance of both sulfate reduction and associated N_2 fixation in ODZ waters.

Biogeography and Environmental Correlations: The dominant factor determining OTU composition and distribution is clearly biogeography (Figure 4). That geographical factor is also evident in the redundancy analysis (Figure 7). (Only sites that contained sequences from one of the top OTUs are represented in the plots, so the number of site symbols is less than 30 for both plots.) For example, Cluster I OTU-5 containing only Arabian Sea surface sequences was positively correlated with both T and S and all of the Arabian Sea samples clustered in the quadrant associated with high T and S (Figure 7A). Surface samples from the ETSP were also in that quadrant, but surface ETNP samples were negatively correlated with S. The surface ETNP samples correlated with OTUs-3. -6, -10 and -11, all of which contained exclusively surface samples. The two largest

Formatted: Subscript

Deleted: 2

Cluster I OTUs were associated with the deep samples from the ETNP and ETSP and correlated positively with nitrite concentration and negatively with oxygen – a signature of the OMZ. Nitrate concentration and depth did not increase the power of the analysis and were omitted from the Cluster I RDA. Most of the sites and five of the most common Cluster I OTUs were not well differentiated by any of the usual environmental parameters.

The Arabian Sea contained very few sequences in Clusters II, III, IV and none of them were in the top six OTUs, so only ETNP and ETSP samples are represented in the RDA for these clusters (Figure 7B). The two largest OTUs in Clusters II, III, IV were negatively correlated with T and S but separated along the second RDA axis, demonstrating opposite relationships with oxygen, nitrite, and nitrate concentrations. OTU-1 included ETSP surface sequences, as well as ODZ sequences from both ETNP and ETSP, while OTU-2 contained only ODZ sequences but both OTUs were phylogenetically related to anaerobic clades (Table 2). Inclusion of all six environmental variables was necessary to obtain maximum separation of the sites and OTUs for Clusters II, III, IV.

Conclusions

The OMZ regions of the world ocean contain substantial *nifH* diversity, both in surface waters and oxygen depleted intermediate depths. Surface waters contained greater diversity for Cluster I, but the ODZ held the highest diversity for Clusters II, III, IV. Cyanobacterial sequences were rare in the combined dataset and were not detected in the ETSP. The ETSP contained the least diversity of Cluster I sequences, while Cluster II, III, IV were least abundant and least diverse in clone libraries from the Arabian Sea. Most of the sequences in all three Clusters of the conventional *nifH* phylogeny were not closely related to any sequences from cultivated Bacteria or Archaea. The most abundant OTUs in Cluster I and in Clusters II, III, and IV could be assigned to the

Deleted: ODZ

Deleted: four

Alphaproteobacteria, followed by the Gammaproteobacteria for Cluster I and Deltaproteobacteria for Clusters II, III, IV sequences. Most of the OTUs were not shared among regions, depths or DNA vs cDNA and sometimes were restricted to individual samples. Some Cluster I sequences had high identity to known species (e.g., *Bradyrhizobium*, *Trichodesmium*) but most of the Cluster II, III, IV sequences were only distantly related to any cultured species.

The assemblage composition of *nifH*-bearing microbes is mainly explained by region, but OTU composition was also consistent with the influence of key environmental parameters such as oxygen and temperature, and reflects association with the secondary nitrite maximum for deep samples. There are few studies that report *nifH* sequences from oceanic OMZs (Jayakumar et al. 2012, Arabian Sea; Fernandez et al. 2011, Loescher et al. 2014, Turk-Kubo et al. 2014, all from the ETSP) or similar environments (Cheung et al. 2016, Costa Rica Dome and Hamersley et al. 2011, hypoxic basins in the Southern California Bight). Combining those reports from individual regions, plus the new sequences from the ETNP reported here, shows that most of the sites/depths, both in this study and in others from OMZ regions, are dominated by one or a few OTUs, which suggests bloom-type dynamics within a diverse background assemblage. Microbes occupying very similar niches and present at low population levels might respond differentially to episodic inputs of organic matter, resulting in spatially and temporally varying dominance by a few clades. Thus we find *nifH* sequences associated with similar metabolic types represented across all the OMZs, although the specific species and genus level affiliations differ. The consistent detection of *nifH* sequences related to those found in known sulfate reducers and methanotrophs suggests the need for further investigation of these pathways in ODZs.

While measurements of N_2 fixation rates are not reported here, the abundance of cDNA sequences suggests that the cells harboring these genes are active. Low, but analytically significant,

Deleted: accounted

Deleted: M

Formatted: Font: Italic

rates have been detected in ODZ depths in the ETNP (Jayakumar et al. 2017) and ETSP (Chang et al. 2019), which suggests that non-cyanobacterial N_2 fixation could make a minor contribution to the nitrogen budget of the ocean. It is therefore important in future work to determine how the diversity described here actually contributes to biogeochemically significant reactions and what environmental and biotic factors might influence or control the activity of diazotrophs in the OMZ.

Formatted: Subscript

Deleted: dark

Deleted: ocean

Figure Legends

Figure 1. Histogram of the 12 most common OTUs from Cluster I *nifH* clone libraries from the three OMZ regions. OTUs were considered common if the total number of sequences in an OTU was ≥2% of the total number of *nifH* clones analyzed (The common OTUs contained 441 of the 512 Cluster I sequences). OTUs were defined according to 3% nucleotide sequence difference using the furthest neighbor method. OTU designation is from most common (OTU-1) to least. A) OTU distribution among regions. B) OTU distribution between OMZ (including core of the ODZ and the upper oxycline depths) and surface depths (oxygenated water). C) OTU distribution of cDNA vs DNA clones.

Figure 2. Rarefaction curve displaying observed OTU richness versus the number of clones sequenced for Cluster I *nifH* sequences (cDNA and DNA). OTUs were defined and designated as in Figure 1. Chao estimators (individual symbols) are shown for each of the same subsets represented in the rarefaction curves.

Figure 3. Maximum likelihood (ML) phylogenetic tree, based on Poisson model of Cluster I, partial *nifH* translated amino acid sequences from DNA and cDNA. Bootstrap values >50% of 1,000 replications are labeled with black circles on the branches. Accession number of reference sequences from NCBI are provided at the end of each reference names. Positions of the OTUs are shown relative to their nearest neighbors from the database. Individual sequence identities comprising each OTU are listed in Table 3.

Deleted: Phylogenetic tree of Cluster 1 *based on* amino acid sequences

Formatted: Font: Italic

Deleted: Supplemental

Deleted: 2

485 Figure 4. Histogram of the 6 most common OTUs from Cluster II, III, IV *nifH* clone libraries
486 from the three OMZ regions. OTUs were considered common if the total number of sequences
487 in an OTU was ≥2% of the total number of *nifH* clones analyzed (the common OTUs contained
488 252 of the 275 Cluster II, III, IV sequences). OTUs were defined according to 3% nucleotide
489 sequence difference using the furthest neighbor method. OTU designation is from most common
490 (OTU-1) to least. A) OTU distribution among regions. B) OTU distribution between OMZ
491 (including core of the ODZ and the upper oxycline depths) and surface depths (oxygenated
492 water). C) OTU distribution of cDNA vs DNA clones.
493
494 Figure 5. Rarefaction curve displaying observed OTU richness versus the number of clones
495 sequenced for Cluster II, III, IV *nifH* sequences (cDNA and DNA). OTUs were defined and
496 designated as in Figure 4. Chao estimators (individual symbols) are shown for each of the same
497 subsets represented in the rarefaction curves.
498
499 Figure 6. Maximum likelihood (ML) phylogenetic tree, based on Poisson model, of Cluster II,
500 III, IV partial *nifH* translated amino acid sequences from DNA and cDNA. Bootstrap values
501 >50% of 1,000 replications are labeled with black circles on the branches. Accession number of
502 reference sequences from NCBI are provided at the end of each reference names. Positions of the
503 OTUs are shown relative to their nearest neighbors from the database. Individual sequence
504 identities comprising each OTU are listed in Table 3.
505
506 Figure 7. RDA plots for (A) Cluster I and (B) Clusters II, III, IV illustrating the relationships
507 among OTUs (green circles containing the OTU number) and sites. DNA = squares; cDNA =

Deleted: Phylogenetic tree of Clusters II, III, IV based on amino acid sequences

Deleted: Supplemental

Deleted: 2

circles. Arabian Sea = cyan (surface) and blue (OMZ); ETNP = pink (surface) and red (deep); ETSP = yellow (surface) and orange (deep). (A) Twelve most abundant OTUs for Cluster I and the four most independent environmental variables. T = temperature, S = salinity, NO2 = nitrite concentration, O2 = oxygen concentration. (B) Six most abundant OTUs for Clusters II, III, IV and all six environmental variables. NO3 = nitrate concentration, Z = depth.

Tables

Table 1. Sampling regions and depths and sequences derived from each depth. S = Surface (within the euphotic zone). OMZ = oxycline or core of the OMZ, all below the euphotic zone.

Table 2. OTU summary for both clusters

Richness and diversity statistics for *nifH* clone libraries from three OMZ regions. ACE and Chao are non-parametric estimators that predict the total number of OTUs in the original sample.

Table 3. OTU identities for both clusters

Cultivated species with closest nucleotide identity to the OTUs identified in the *nifH* clone libraries from three OMZ regions. Only the 12 most common OTUs (out of 41 total) are listed for Cluster 1 sequences, and the six most common (out of 18 total) for the Clusters II, III, IV libraries.

Supplemental

S Table 1A and B. List of sequences in each OTU for both clusters

S Table 2. Distribution of OTUs among sites

References

1. Bentzon-Tilia, M., Traving, S. J., Mantikci, M., Knudsen-Leerbeck, H., Hansen, J. L. S., Markager, S. and Riemann, L., 2015. Significant N-2 fixation by heterotrophs, photoheterotrophs and heterocystous cyanobacteria in two temperate estuaries. Isme Journal. 9, 273-285.
2. Bird, C. and Wyman, M., 2013. Transcriptionally active heterotrophic diazotrophs are widespread in the upper water column of the Arabian Sea. Fems Microbiology Ecology. 84, 189-200.
3. Bonnet, S., Dekaezemacker, J., Turk-Kubo, K. A., Moutin, T., Hamersley, R. M., Grosso, O., Zehr, J. P. and Capone, D. G., 2013. Aphotic N-2 Fixation in the Eastern Tropical South Pacific Ocean. Plos One. 8.
4. Chang, B. X., Jayakumar, A., Widner, B., Bernhardt, P., Mordy, C. M., Mulholland, M. R. and Ward, B. B., 2019. Low rates of dinitrogen fixation in the eastern tropical South Pacific. Limnology And Oceanography. 64, 1913-1923.
5. Cheung, S. Y., Xia, X. M., Guoand, C. and Liu, H. B., 2016. Diazotroph community structure in the deep oxygen minimum zone of the Costa Rica Dome. Journal of Plankton Research. 38, 380-391.
6. Deutsch, C., Sarmiento, J. L., Sigman, D. M., Gruber, N. and Dunne, J. P., 2007. Spatial coupling of nitrogen inputs and losses in the ocean. Nature. 445, 163-167.
7. Fernandez, C., Farias, L. and Ulloa, O., 2011. Nitrogen Fixation in Denitrified Marine Waters. Plos One. 6.
8. Fernandez, C., Lorena Gonzalez, M., Munoz, C., Molina, V. and Farias, L., 2015. Temporal and spatial variability of biological nitrogen fixation off the upwelling system of central Chile (35-38.5 degrees S). Journal of Geophysical Research-Oceans. 120, 3330-3349.
9. Gaby, J. C., Rishishwar, L., Valderrama-Aguirre, L. C., Green, S. J., Valderrama-Aguirre, A., Jordan, I. L. and Kostka, J. E., 2018. Diazotroph community characterization via a high-throughput nifH amplicon sequencing and analysis pipeline. Applied And Environmental Microbiology. 84, eO1512-01517.
10. Großkopf, T. and LaRoche, J., 2012. Direct and indirect costs of dinitrogen fixation in Crocosphaera watsonii WH8501 and possible implications for the nitrogen cycle. Frontiers in Microbiology. 3.
11. Hamersley, M. R., Turk, K. A., Leinweber, A., Gruber, N., Zehr, J. P., Gunderson, T. and Capone, D. G., 2011. Nitrogen fixation within the water column associated with two hypoxic basins in the Southern California Bight. Aquatic Microbial Ecology. 63, 193-205.
12. Jayakumar, A., Al-Rshaidat, M. M. D., Ward, B. B. and Mulholland, M. R., 2012. Diversity, distribution, and expression of diazotroph nifH genes in oxygen-deficient waters of the Arabian Sea. Fems Microbiology Ecology. 82, 597-606.
13. Jayakumar, A., Chang, B. N. X., Widner, B., Bernhardt, P., Mulholland, M. R. and Ward, B. B., 2017. Biological nitrogen fixation in the oxygen-minimum region of the eastern tropical North Pacific ocean. Isme Journal. 11, 2356-2367.

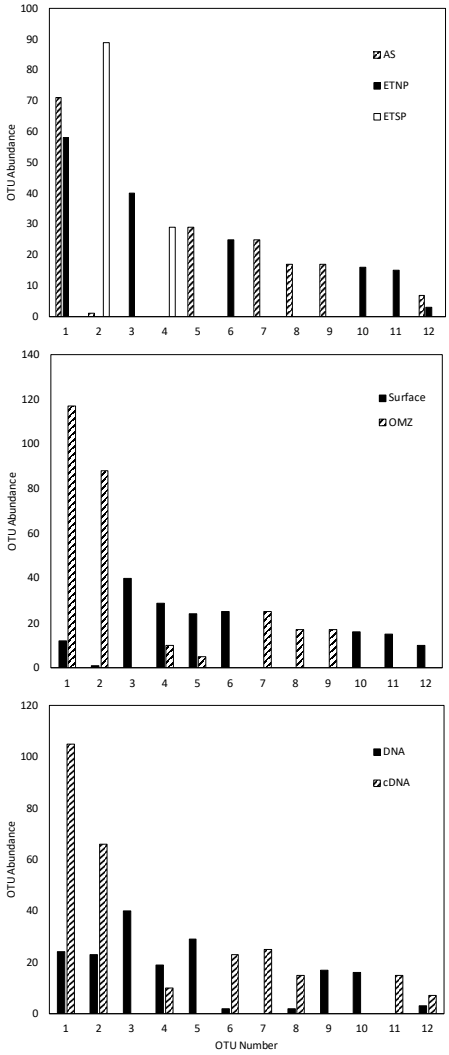
14. Jayakumar, A., O'Mullan, G. D., Naqvi, S. W. A. and Ward, B. B., 2009. Denitrifying bacterial community composition changes associated with stages of denitrification in oxygen minimum zones. Microbial Ecology. 58, 350-362.
15. Krotzky, A. and Werner, D., 1987. NITROGEN-FIXATION IN PSEUDOMONAS-STUTZERI. Archives of Microbiology. 147, 48-57.
16. Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution. 35, 1547-1549.
17. Langlois, R., Grokopf, T., Mills, M., Takeda, S. and LaRoche, J., 2015. Widespread Distribution and Expression of Gamma A (UMB), an Uncultured, Diazotrophic, gamma-Proteobacterial nifH Phylotype. Plos One. 10.
18. Letunic, I. and Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research. 44, W242-W245.
19. Loescher, C. R., Bange, H. W., Schmitz, R. A., Callbeck, C. M., Engel, A., Hauss, H., Kanzow, T., Kiko, R., Lavik, G., Loginova, A., Melzner, F., Meyer, J., Neulinger, S. C., Pahlow, M., Riebesell, U., Schunck, H., Thomsen, S. and Wagner, H., 2016. Water column biogeochemistry of oxygen minimum zones in the eastern tropical North Atlantic and eastern tropical South Pacific oceans. Biogeosciences. 13, 3585-3606.
20. Loescher, C. R., Grosskopf, T., Desai, F. D., Gill, D., Schunck, H., Croot, P. L., Schlosser, C., Neulinger, S. C., Pinnow, N., Lavik, G., Kuypers, M. M. M., LaRoche, J. and Schmitz, R. A., 2014. Facets of diazotrophy in the oxygen minimum zone waters off Peru. Isme Journal. 8, 2180-2192.
21. Moisander, P. H., Beinart, R. A., Hewson, I., White, A. E., Johnson, K. S., Carlson, C. A., Montoya, J. P. and Zehr, J. P., 2010. Unicellular Cyanobacterial Distributions Broaden the Oceanic N-2 Fixation Domain. Science. 327, 1512-1514.
22. Moisander, P. H., Benavides, M., Bonnet, S., Berman-Frank, I., White, A. E. and Riemann, L., 2017. Chasing after Non-cyanobacterial Nitrogen Fixation in Marine Pelagic Environments. Frontiers in Microbiology. 8.
23. Moisander, P. H., Serros, T., Paerl, R. W., Beinart, R. A. and Zehr, J. P., 2014. Gammaproteobacterial diazotrophs and nifH gene expression in surface waters of the South Pacific Ocean. The ISME Journal 8, 1962–1973.
24. Schloss, P. D. and Handlesman, J., 2009. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and Environmental Microbiology. 71, 1501-1506.
25. Stecher, G., Tamura, K. and Kumar, S., 2020. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. Molecular Biology and Evolution. 37, 1237-1239.
26. Turk-Kubo, K. A., Karamchandani, M., Capone, D. G. and Zehr, J. P., 2014. The paradox of marine heterotrophic nitrogen fixation: abundances of heterotrophic diazotrophs do not account for nitrogen fixation rates in the Eastern Tropical South Pacific. Environmental Microbiology. 16, 3095-3114.
27. Zehr, J. P., Crumbliss, L. L., Church, M. J., Omoregie, E. O. and Jenkins, B. D., 2003. Nitrogenase genes in PCR and RT-PCR reagents: implications for studies of diversity of functional genes. Biotechniques. 35, 996-1005.

626 28. Zehr, J. P. and McReynolds, L. A., 1989. Use of degenerate oligonucleotides for amplification
627 of the *nifH* gene from the marine cyanobacterium *Trichodesmium theiebautii*. Applied
628 and Environmental Microbiology. 55, 2522-2526.
629 29. Zehr, J. P., Mellon, M. T. and Zani, S., 1998. New nitrogen-fixing microorganisms detected in
630 oligotrophic oceans by amplification of nitrogenase (nifH) genes. Applied and
631 Environmental Microbiology. 6, 3444-3450.
632

633

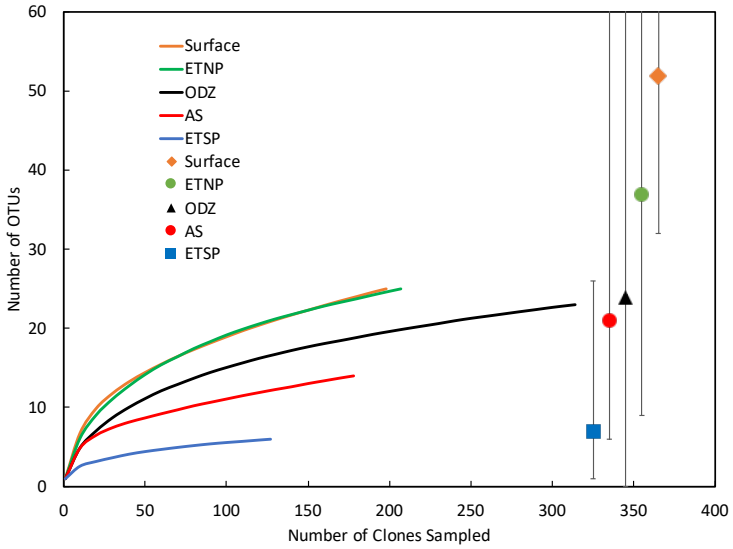
634

635 Figure.1



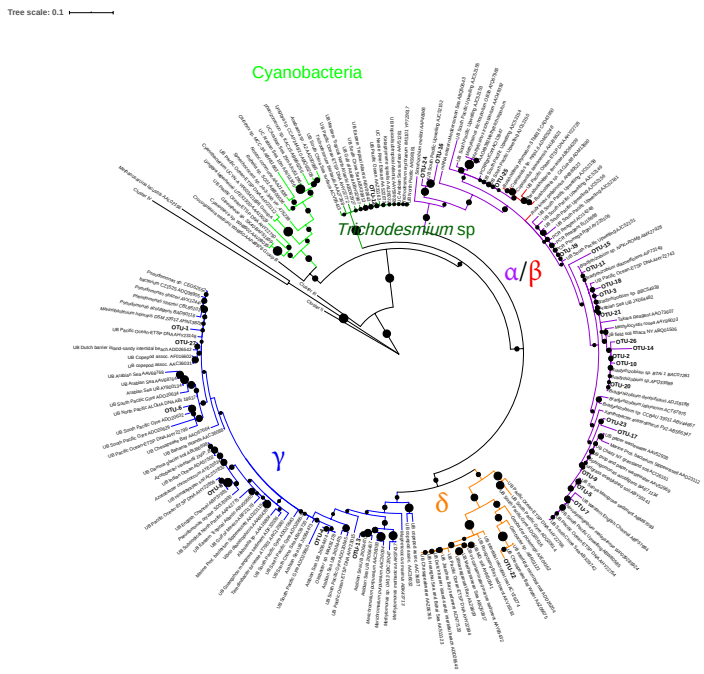
636

637 Figure. 2

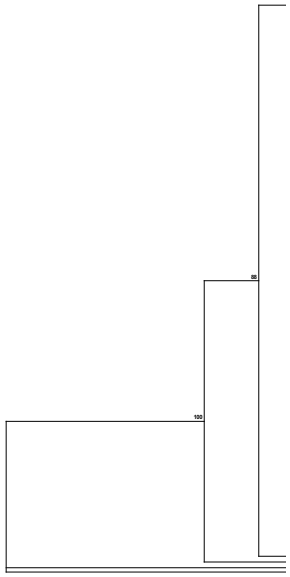


638
639

640 Figure. 3

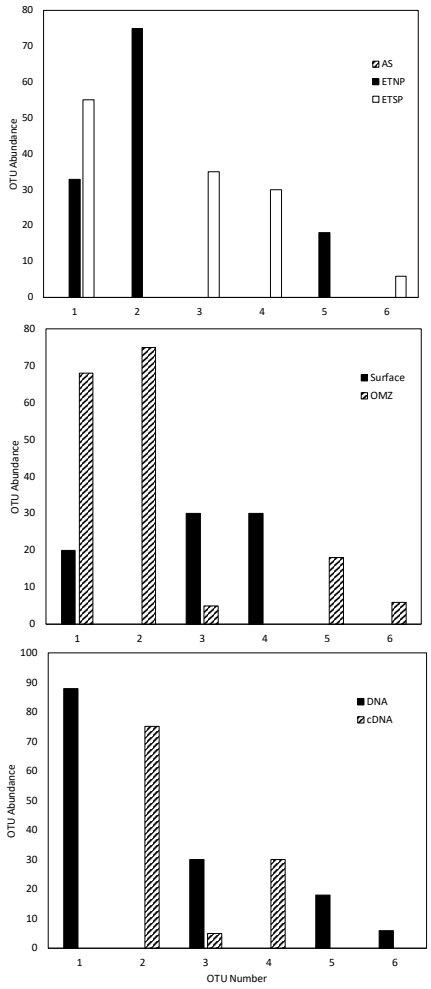


641
642



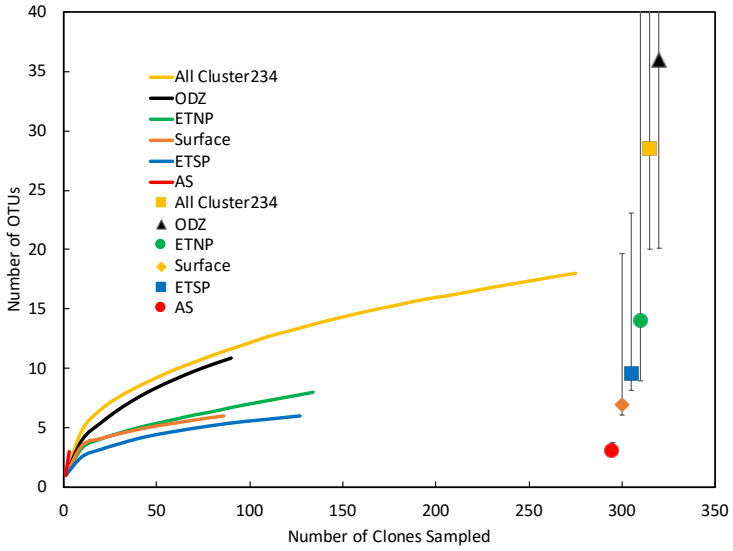
Deleted:

644 Figure. 4



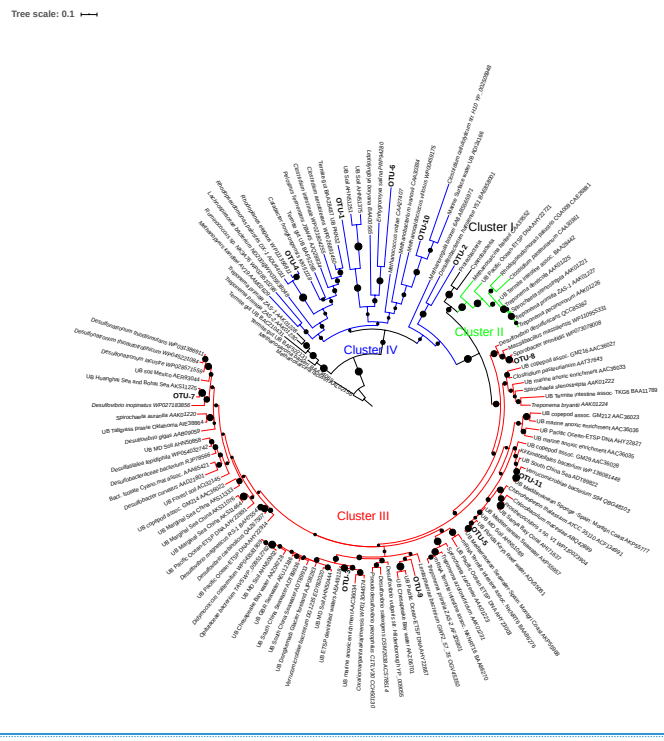
645
646

647 Figure. 5.

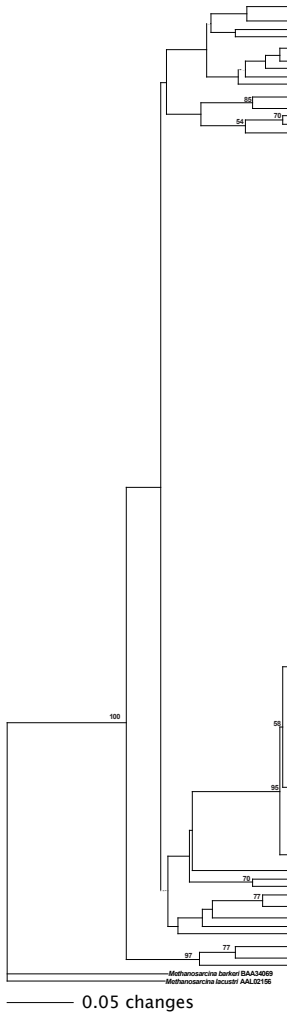


648
649

650 Figure 6.



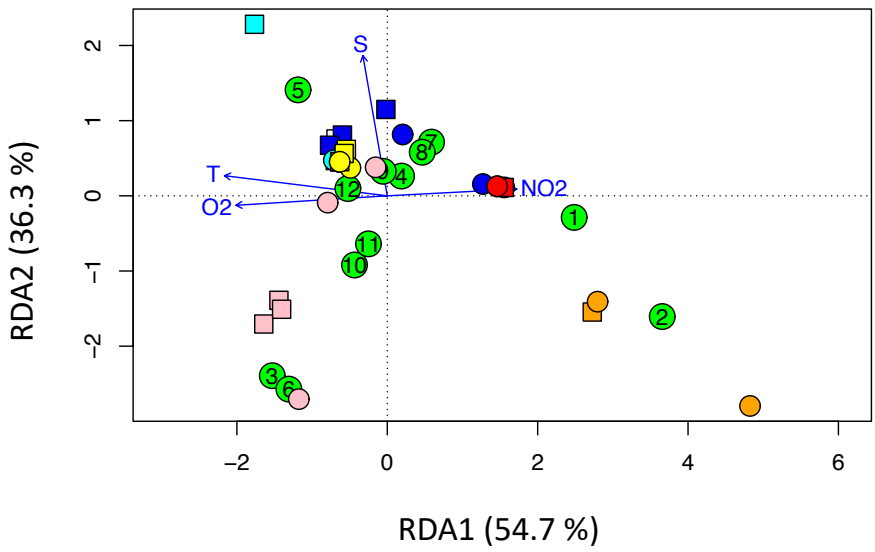
651
652



Deleted:

654

Figure 7A



655

Figure 7B

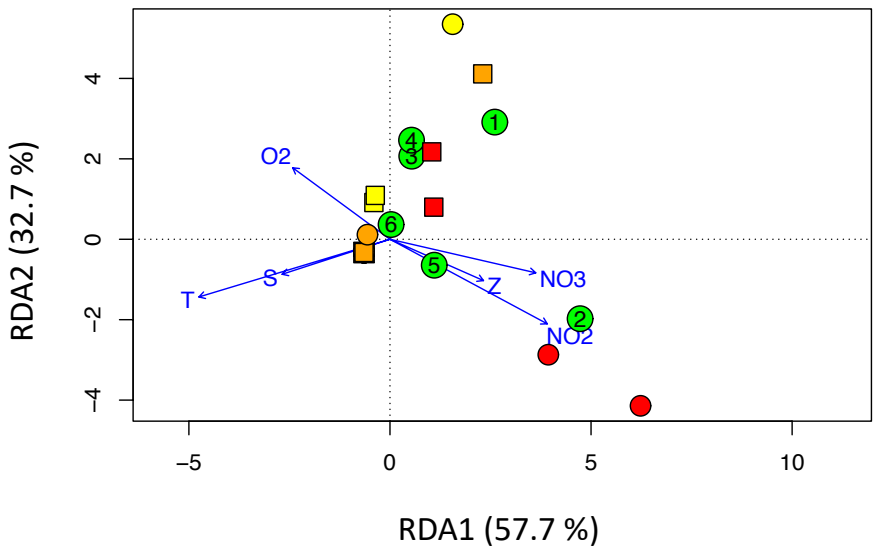


Table 1 Sample and clone library descriptions

OMZ Region	Station	Latitude	Longitude	Depth (m)	Depth Feature	DNA Seqs	cDNA Seqs
Arabian Sea	S1	19°N	67°E	10	S	3	0
Arabian Sea	S1	19°N	67°E	60	S	20	0
Arabian Sea	S1	19°N	67°E	150	OMZ	23	25
Arabian Sea	S1	19°N	67°E	175	OMZ	10	22
Arabian Sea	S2	15°N	64°E	150	OMZ	4	25
Arabian Sea	S3	12°N	64°E	10	S	25	4
Arabian Sea	S3	12°N	64°E	110	OMZ	4	23
ETNP	BB1	20 9.6°N	106°W	0	S	26	5
ETNP	BB1	20 9.6°N	106°W	18	S	24	17
ETNP	BB1	20 9.6°N	106°W	90	OMZ	42	38
ETNP	BB2	16 31°N	107 6.8°W	0	S	40	35
ETNP	BB2	16 31°N	107 6.8°W	150	OMZ	47	67
ETSP	BB1	13 59.9°S	81 12.0°W	2	S	29	1
ETSP	BB1	13 59.9°S	81 12.0°W	130	OMZ	46	44
ETSP	BB2	20. 46.1°S	70 39. 5°W	20	S	45	30
ETSP	BB2	20. 46.1°S	70 39. 5°W	115	OMZ	23	40

Deleted: Sampling position details

Formatted Table

659 Table 2 OTU Summary

Sample subset	Depths, regions included	No. of Sequences	No. of Unique Sequences	No. of OTUs (cutoff ~3)	OTU /seq	Shannon	Simpson	Chao	Ace
Cluster I									
AS	Arabian Sea, all depths	178	36	14	0.079	1.8	0.22	21	45
ETNP	ETNP, all depths	207	80	25	0.121	2.37	0.14	37	34
ETSP	ETSP, OMZ depths	127	51	6	0.047	0.87	0.53	7	8
All ClusterI	Three regions, all depths	512	165	41	0.080	2.7	0.11	59	67
All ClusterI DNA	Three regions, all depths	257	97	35	0.136	2.8	0.08	42	45
All ClusterI cDNA	Three regions, all depths	255	75	24	0.094	1.7	0.25	24	27
All ClusterI Surface	Three regions, surface depths	198	73	25	0.126	2.5	0.10	52	75
All ClusterI OMZ	Three regions, all depths	314	98	23	0.073	0.9	0.23	30	37
Clusters II, III, IV									
AS	Arabian Sea, all depths	10	6	3	0.300	1.09	0.27	3	3
ETNP	ETNP, all depths	134	49	8	0.060	1.19	0.39	14	38
ETSP	ETSP, all depths	131	64	8	0.061	1.37	0.30	9	19
All Clusters II,III,IV	Three regions, all depths	275	117	18	0.065	1.88	0.21	28	26
All Clusters II,III,IV DNA	Three regions, all depths	155	65	12	0.077	1.20	0.37	22	17
All Clusters II,III,IV cDNA	Three regions, all depths	120	56	9	0.075	1.11	0.45	12	15
All Clusters II,III,IV Surface	Three regions, surface depths	86	46	6	0.070	1.32	0.29	7	13
All Clusters II,III IV OMZ	Three regions, OMZ depths	189	76	15	0.079	1.57	0.29	46	24

660
661
662
663

664 Table 3 OTU identities

665

Formatted: Font: (Default) +Body (Calibri), 10 pt

	No. of Sequences	Phylogenetic Affiliation	Closest cultured relative (DNA)	Identity DNA %	Coverage %	Closest cultured relative (Protein)	Identity AA %	Coverage %
Cluster I								
OTU-1	129	Gamma	*Psuedomonas stutzeri*	91	98	*Pseudomonas stutzeri strain SGAir0442*	95.8	99
OTU-2	89	Alpha	*Bradyrhizobium sp*	99	100	*Bradyrhizobium denitrificans strain LMG 8443*	99	99
OTU-3	40	Alpha	*Bradyrhizobium sp. TM124*	94	98	*Bradyrhizobium sp. MAFF 210318*	99	98
OTU-4	29	Gamma	*Marinobacterium lutimaris*	87	100	*Oleibacter sp*	100	99
OTU-5	29	Alpha	*Methylosinus trichosporium*	92	99	*Sphingomonas azotifigens*	99	100
OTU-6	25	Gamma	*Azotobacter chroococcum* strain B3	81	99	*Psuedomonas stutzeri*	94	99
OTU-7	25	Beta/Alpha	*Rubrivivax gelatinosus*	91	99	*Novosphingobium malasiense*	99	100
OTU-8	17	Gamma	*Psuedomonas stutzeri*	91	98	*Azotobacter chroococcum* strain B3	97	100
OTU-9	17	Beta/Alfa	*Burkholderia*	90	100	*Sphingomonas azotifigens*	100	100
OTU-10	16	Alpha	*Bradyrhizobium*	97	98	*Bradyrhizobium sp. ORS 285*	99	99
OTU-11	15	Alpha	*Bradyrhizobium*	97	98	*Bradyrhizobium diazoefficiens*	98	99
OTU-12	10	Cyanobacterium	*Katagnymene spiralis*	100	99	*Trichodesmium erythraeum*	100	99
Clusters II, III IV								
OTU-1	88	Alpha/Spirochaetaceae	*Rhizobium sp*	74	59	*Treponema primitia ZAS-1]*	55	98
OTU-2	75	Delta/Firmicutes	*Geobacter*	73	43	*Desulfitobacterium hafniense*	98	61
OTU-3	35	Verrumicrobia	*Opitutaceae bacterium*	82	99	*Coraliomargarita akajimensis*	95	99
OTU-4	30	Alpha	*Rhodopseudomonas palustris*	90	98	*Rhodoplanes elegans*	96	99
OTU-5	18	Delta/Chlorobi	*Desulfovibrio piezophilus*	79	99	*Prosthecochloris sp. V1, Chloroherpeton thalassium, Chloroherpeton thalassium*	92	99
OTU-6	6	Beta/Delta	*Azoarcus communis*	70	88	*Enhygromyxa salina*	70	74
OTU-7	4	Delta	*Desulfovibrio carbinolicus strain DSM 3852*	81	99	*Desulfovibrio inopinatus*	90	99
OTU-8	4	Delta/Firmicutes	*Desulfovibrio desulfuricans strain IC1*	77	100	*Sporobacter termitidis*	99	99

OTU-9	3	Delta/Lentisphaerae	*Desulfovibrio magneticus RS-1 DNA*	84	100	*Lentisphaerae bacterium GWF2_57_35, Desulfatitalea tepidiphila, Desulfobacteraceae bacterium*	84	100
OTU-10	3	Delta/Methanococci	*Desulfovibrio desulfuricans strain IC1*	77	100	*Methanocaldococcus villosus*	65	99
OTU-11	2	Verrucomicrobia	Verrucomicrobia bacterium S94	87	100	Verrucomicrobia bacterium S94	97	99

666
667
668