Response to reviewers

*****Revised draft of manuscript and figures follows. Substantial changes are highlighted in yellow.*****

Reviewer #1

General comments

The article Trichodesmium physiological ecology and phosphate reduction in the western Tropical South Pacific by Frischkorn et al. is one of a series of articles on the scientific results of the OUTPACE project, a field study focused on the undersampled Western Tropical South Pacific. The authors used a combination of environmental metagenomics, metatranscriptomics, nutrient analyses, and analysis of Trichodesmium holobiont phosphate physiology to shed some light on the physiological interactions between and combined metabolic potential of Trichodesmium and its associated microbiome. The study focuses on phosphorus physiology, and touches on iron limitation as well.

I find the article generally very well written and easy to follow. The results presented are relevant to the understanding of how Trichodesmium and its microbiome shape nutrient availability in the regions where they thrive. Studies on Trichodesmium as a holobiont are being published increasingly, but there is still a lack of data, especially from the field. The authors present not only omics data, but also physiological data on an understudied region of the ocean, and nicely combine their data to reach meaningful conclusions. Thus, I generally find this study of interest to the Biogeosciences scientific community.

The article falls a bit short on some of the methods details in the meta-omics sections, which I specify further down. Figure 3 could be presented better and discussed more, as some of the results are not discussed at all.

I highly recommend this article for publication after minor revisions.

Specific comments

Trichodesmium clade distribution:

I understand that the concentrations of Trichodesmium cells is derived from the qPCR of rnpB. This should be noted in the methods part 2.2 (state that absolute numbers were derived, and briefly how it was done, rather than just referring to Rouco et al) and at the beginning of the respective results section, as well as in figure 2. In the reference given for the qPCR, Rouco et al used known cell concentrations as templates for calibration. I just want to remind the authors that Trichodesmium seems to be highly polyploid with hundreds of genomes per cell, and that this differed by up to factor 6 between lab cultures (which I think were used for the calibration here) and the field (Sargent et al, 2016). If the authors share my view here, I suggest discussing the resulting error briefly.

At the reviewer's suggestion, we have updated the manuscript methods to include a description of the qPCR methodology. In the discussion section we also now include a caveat pointing out the potential for polyploidy in *Trichodesmium* and how this might influence interpretation of our results. In short, the counts presented in this paper are tabulated using a standard curve generated from cell counts performed on cultures of Clade I and Clade III (Rouco et al., 2014, 2016). This analysis approach yields C_T values that take into account polyploidy, unlike gene standard approaches, but the method would be prone to variation in the cell calculation if there were discrepancies between ploidy levels in the field and culture for example. This is acknowledged in the revised text. To be conservative we do not draw conclusions based on the absolute cell number, only the patterns. We

note we identified trends in relative abundance that were consistent with the patterns observed with other methods employed on the cruise transect (e.g. Stenegren et al., 2017).

In the text, I would appreciate depth-integrated cell counts for Trichodesmium per station, instead of averages across multiple depths and stations. I think this would be more interesting given that Trichodesmium vertical distribution might change within one day through vertical migration.

The text in this section has been edited to clarify our main points about the *Trichodesmium* distribution. We agree that vertical migration of *Trichodesmium* over day-night cycles is an interesting facet of this organism's ecology, and in the revised version of the manuscript we now include a supplemental table that includes the qPCR counts from all stations and all depths, thus giving interested parties the option to view or use these results in future studies. We have edited the text in this section of the results to better reflect the depths and locations where samples were obtained when specific cell concentrations are highlighted. Considering the scope of our study, the fact that *Trichodesmium* sampling occurred at roughly the same time each day, that biomass was concentrated in the upper 20 m across the entire transect, and that net tow sampling used for all other analyses captures colonies within this 20 m range, in subsequent analyses where cell count values are used, we have continued to use average concentration from the top 20 m, which includes 3 discrete depths captured by the Rosette sampling device.

Trichodesmium colony sampling

Here, I am missing data on the speed of the net tow and the total volume filtered per tow. It is reasonable to assume that with this mesh size and a certain speed, the outer filaments of colonies are lost.

We now include the time it took to perform the net tows (15 minutes total, per sample). We have also edited the text to clarify that these net tows were done by hand (rather than being towed behind the boat) and were performed gently while the ship was stopped to avoid shearing colonies. On hand net tows such as this, we do not include a flow meter and we are unable to determine the exact volume filtered per tow, but we now note in the manuscript that each net tow likely filters thousands of liters of water. We perform these net tows similarly across all field studies, and so the sampling here is consistent with previous work that has consistently recovered *Trichodesmium* and its microbiome (e.g. Rouco et al., 2016; Frischkorn et al., 2017, 2018).

Trichodesmium OGs from the metagenome

The large number of OGs in the Trichodesmium metagenomics bin leaves me a little skeptical. While this is discussed in the paper in section 4.2, I do not think that this high number of unique OGs (9,790) "could potentially be attributed to a greater diversity of Trichodesmium ecotypes in the WTSP" (p17,20-21). The published T. erythraeum genome results in 2,982 OGs according to the authors, so this would mean a tripling of unique protein-coding gene groups within the genus Trichodesmium compared to T. erythraeum, which already is a very large genome with a wealth of unique functions, and is not known to have genomic islands that would facilitate frequent gene shuffling, like for example in Prochlorococcus. While this of course does not mean that this finding is necessarily incorrect, I think it might be the case that multiple OGs are indeed the same protein group, but represent only partial proteins and are thus not grouped together in one OG when they should. To test this, I suggest mapping OG consensus sequences to the T. erythraeum genome with a high mismatch/gap tolerance and checking if multiple OGs often map to the same gene (especially for long genes).

We agree with the reviewer that the high number of unique *Trichodesmium* OGs is surprising and given that we can only really speculate on what may be driving this we have substantially edited this section of the manuscript. The *T. erythraeum* IMS101 genome (Clade III) is not representative of field populations, which are dominated in the WTSP by Clade I, hence the necessity of using metagenome assemblies and *Trichodesmium*-identified genome bins for metatranscriptome read mapping. Clustering proteins using the MCL algorithm only considers proteins longer than 70 amino acids so as to avoid partial sequences, and uses a sequence comparison approach that groups partial proteins into the same OG if there is any sequence overlap. This approach should mitigate generating multiple OGs for the same protein. We have mapped OGs back to the *T. erythraeum* IMS101 genome as suggested. Even with a high mismatch allowance, roughly 60% of the genes from *Trichodesmium*-identified genome bins do not align to the IMS101 genome. Our group has seen similar low mapping of field *Trichodesmium* samples to the IMS101 genome (Rouco et al. 2018, *ISME J*), leading us to believe that this is not an assembly error but instead accurately reflects divergences between the dominant clades in the environment and the Clade III *T. erythraeum* represented by IMS101. Furthermore, examination of some of the key genes discussed in the paper does not show multiple OGs annotated as the same gene. This suggests that the number of OGs is not primarily driven by an artifact of the clustering. As highlighted above, the discussion of the apparent OG expansion has been heavily edited in the revised text.

Metatranscriptomics

Why is the expression for the microbiome in almost all KEGG categories presented in Figure 3 the lowest in sample LDB 3. This is not counterbalanced by higher transcription in the Trichodesmium bin either so it seems that reads are missing here. Could this be a normalization artefact? I understand that RSEM was used for mapping and normalization yielding TPM as normalized counts. These TPM values were additionally normalized within each OG to give all OGs equal weight. The mapping was done only to protein coding regions of the metagenomics scaffolds. Was the fraction of reads mapping to those regions similar for all samples? Maybe reads were lost to other phototrophs (not included in the metagenomics reference) or due to a very high expression of some non-coding RNAs (also not included). I would appreciate some mapping statistics to be able to evaluate this part (total and fraction of reads mapped per sample).

We thank the reviewer for bringing up this aspect of the data to our attention. In brief, we reexamined this analysis and found that overall, similar numbers of metatranscriptome reads mapped to the metagenome across all stations. To address any potential variability however, we normalized the KEGG signals within each sample individually. Even after this normalization, the distribution of reads across KEGG submodules over the samples from LDB was noticeably different than the SD stations, thus when each submodule is visualized in a heatmap as rowwide averages, these samples look different. We maintain that this normalization and display is the best way to visualize changes in *Trichodesmium* and microbiome metabolism across this WTSP transect, and have added a caveat to the revised manuscript that better explains the trends in the data. Additionally, we now note in the revised text that the broad physiological shifts visualized through KEGG module expression could be symptomatic of the crashing phytoplankton bloom that took place during the time when we sampled LDB (Stenegren et al.; de Verneil et al., Valdés et al., *this issue*). An in depth analysis of such dynamics are beyond the scope of this paper (but provide an exciting avenue for inquiry in on-going work on this dataset), however in the revised text we briefly explain this situation as further rationalization for the gene expression trends.

KEGG submodules, Orthologous Groups (OGs), and their annotation:

Annotating metatranscriptomes with de-novo assembled metagenome bins is always challenging. I think the idea of binning proteins into OGs, mapping the transcripts to those OG bins and then trying to classify those OGs functionally is generally a nice way to do it. I am still left a little confused with the overall process. A Diamond search against UniRef90, KEGG annotation, and clustering into OGs are mentioned as three separate annotation methods (."… annotation was also carried out using …", "Proteins … were also clustered …"). I would thus greatly appreciate a workflow diagram as a Suppl. Figure that answers the following questions: Where/How exactly did you derive the functional classification of a protein and an OG? How does this relate to the KEGG module expression profiles used in Figure 3?

We agree with the reviewer that in our initial submission our annotation pipeline was not clear, likely because the multi-pronged approach was indeed difficult to explain with text alone. Thanks to the reviewer's helpful suggestion we now include a workflow diagram in the Supplemental Figures that outlines the pipeline we followed for assembly, annotation and analysis. We have also heavily edited the methods section detailing these protocols so that with the aid of the new diagram our approach will be more clear.

The KEGG categories presented in Figure 3 are mostly subcategories of parent categories named PATHWAY MODULE, STRUCTURAL COMPLEX, FUNCTIONAL SET, and SIGNATURE MODULE, but the links to these high-level parent categories are missing, which makes some of the presented subcategories hard to interpret (the hierarchy can be found here: http://www.genome.jp/kegg-bin/get_htext#C56). Additionally, the submodule "metabolic capacity" has no real meaning unless the reader knows what metabolic capacities are lumped together within this term. Other categories also presented in Figure 3 do not appear within this hierarchy at all (e.g., sugar metabolism, phenylpropanoid and flavonoid biosynthesis). Where were they derived from? If they are children categories, why were only a few of this hierarchy presented? Please clarify and note all parent categories in figure 3. It also seems dubious that *Trichodesmium* has no proteins in the category "carbohydrate metabolism". I would double check that. I thus suggest a revision of the mapping to KEGG categories, deleting those that are neither meaningful nor mentioned from the figure (like "drug resistance" or "metabolic capacity"), stating the parent hierarchy and a general clarification to the reader.

We agree with the reviewer that the way we presented the KEGG analysis in our initial submission was not as clear as it should have been. KEGG definitions were obtained from the modules within the "Pathway module" and "Structural complex" categories and the submodules within these categories (available at https://www.genome.jp/kegg-bin/get_htext?ko00002.keg). We have also edited the figure to remove submodules that are not meaningful, at the reviewer's request. This method and organization scheme was adapted from a method outlined in Alexander et al. 2015 (PNAS) and in our revised manuscript reflects this more cohesive and easy to follow organization. We note finally that it is not uncommon for certain KEGG categories to not be detected, for example the absence of the "carbohydrate metabolism" category from *Trichodesmium*. We also noted the absence of this category in our previous metagenome work (Frischkorn et al., 2017, ISME J) and attribute it to the fact that KEGG is unfortunately not optimized for non-model systems or marine microbes. We highlight however the presence and expression of several other submodules attributed to *Trichodesmium* in the carbohydrate and lipid metabolism module that are depicted in revised Fig. 3.

Phosphonate synthesis gene cluster

This is a nice finding. To make it bulletproof that this assembled scaffold indeed originated from a Trichodesmium genome, I suggest to add a suppl. figure that shows an alignment of the scaffold with all raw

reads that cover the transition from the non-coding region of the T. erythraeum IMS101 genome to the ppm gene. Please also provide the sequence of the scaffold in a suppl. file.

At the reviewer's request, we now include a supplemental figure that shows reads from all three metagenome samples aligned to the phosphonate biosynthesis scaffold. This new figure shows that in all three metagenomes, reads span the non-coding region of the *T. erythraeum* IMS101 genome to the ppm gene, and mate pairs also connect the IMS101 portion to the ppm gene. We also now include an additional supplemental file that includes the full sequence of the *ppm*-containing scaffold.

Technical corrections
P1, 25ff: Split into two sentences to facilitate reading.
P1, 29-30: I find this sentence too generic.
P3, 7: add: taxonomically conserved across ocean basins?
P4, 14: "Nutrient analyses were collected" – change to Samples for nutrient analyses were collected
P5, 9: (qPRC) should be (qPCR)
P5, 17: "gentle vacuuming" – How gentle? Please state the mbar pressure if possible.
P6, 1: please clarify what this means: "after some sets of Trichodesmium incubations"
P6, 13: Please state how many colonies per sample were extracted.
P7, 3: What about associated phototrophs? Were those sequences ignored for all further analysis?
P7, 17ff: Does this mean: For the majority of all successfully annotated proteins within each orthologous group, the UniRef and KEGG annotations matched? Please re-write the annotation method to make a little clearer (see also Specific Comment section above).
P8, 6: Add "Maximum likelihood phylogenetic analysis"
P8, 15: I assume RNA was also extracted from isolated Trichodesmium colonies, but please state this here,a nd also mention how many colonies per sample were extracted.
P8, 25: Please state fragmentation method (ultrasound or chemical, if any), resulting fragment size, HiSeq instrument version, and specify read length.
P9, 12: "tags per million" should be transcripts per million (TPM). Please correct throughout the manuscript.
P 10, 1-2: I cannot find stations LDB1-3 mentioned anywhere. Figure 1 only shows one station: LDB. Maybe the authors mean 3 individual samples from station LDB? Please see also my comment on figure 1.
P10, 10ff: With this sentence, the authors start using cells/mL for the rest of the paragraph, instead of the cells/L that is used before and in figure 2a. This should be corrected.
In line 11, 9,000 – 58,000 cells mL-1 are given as a concentration range for "stations west of 170°W and LDB". However, looking at figure 2a, there are clearly concentrations above 100,000 cell/L (orange; assuming units should be the same). I am not sure whether I may be reading the figure wrongly or this is due to the ODV extrapolation, or simply a mistake.
Also, to clarify east/west for the reader across the 180° longitude and not confuse it with the x-axis labels, I suggest to add in line 11: ("stations west/left of 170°W and LDB")
P11: Both sections are titled "3.3 …"
P12, 8-10: I do not see this statement reflected in figure 3.
P18, 8: Which one is the "broadest KEGG annotation level"? Please name it.
P22, 15ff: Were those phosphate reduction rates in the western North Atlantic not explainable so far? If yes, I would state this to help the reader understand this sentence better. I would also move this sentence to the end of the paragraph and re-write it. It feels a little fuzzy ("may in part underpin").
P22, 28: add "in cultured Trichodesmium isolates from Clade III, specifically T. erythraeum, (Dyhrman et al, 2009)"

Figure 1: Not clear to me which ones are the SD and which ones the LD stations. “SD” does not appear in the figure, but caption says “denoted SD”. Please make that clear in the figure or text. Please increase font size of the station tags.

Figure 2: Also, station LDB is indicated, but has no data points. Is this correct? It would be nicer if the panels a, b, and c all had the same x-axis, so that all the stations are indeed directly beneath each other. The difference seems to be that the x-axis in panel a is not linear.

Figure 3: Is the color code indicating relative expression (low to high) linear or log? I find it a bit difficult to find the functions mentioned in the Results part (p12, 5ff) in the Figure. I suggest to add a number to all categories in the figure and refer to those numbers in the text instead of just writing for example “lipid, vitamin, glycan, .. metabolism”. That way the reader can easily find what the text refers to. Also, I would change the color code for the KEGG modules, and, as mentioned in the comment above, indicate the full KEGG hierarchy, and get rid of non-telling categories.

Figure4: The asterisks in LDB** are not mentioned in the caption. What do they mean?

We have revised the manuscript, making all of the above edits, correcting mistakes and omissions and adding the requested data. We thank the reviewer for editing the manuscript so carefully.

References

Sargent, E.C., Hitchcock, A., Johansson, S.A., Langlois, R., Moore, C.M., LaRoche, J., Poulton, A.J., and Bibby, T.S. (2016). Evidence for polyploidy in the globally important diazotroph Trichodesmium. FEMS Microbiology Letters fnw244.

Reviewer #2

This is an impressive dataset that has a high potential to offer tantalizing insight into the gene expression of the *Trichodesmium* holobiont in a relatively understudied environment. In spite of my low rating, I actually think it is not far from living up to this potential. That said, I think there is some remaining work that needs to be done with respect to analysis and presentation of these data.

Minor Technical Issues: What methods were used for the biogeochemical measurements? Which of the Outpace articles are these data originally presented? Note: the link to the data from the cruise requires a login and there is no indication in this manuscript as to where these data are published (if they are). If they are not published in the other Outpace manuscripts, the methods should be clearly presented here. If they are in other manuscripts, make it clear which manuscript contains the relevant information. What were the methods for the phosphate turnover measurements? I assume these are only presented in this manuscript. While the phosphate uptake measurements are given, the phosphate turnover details are left out.

We agree with the reviewer that these methods should have been presented in the first draft and regret that the link we provided was not available to the public. In the edited manuscript we have revised our geochemical analyses using new data from a broader set of iron measurements (Guieu et al., 2018 *Scientific Reports,* in press) as well as using new nM phosphate measurements detected with a low level phosphate assay provided by our co-authors Guieu and Louis. In the revised version of our manuscript we now delineate these methods as requested, and incorporate these data into our analyses. These values are also reflected in the revised supplemental table.

Results Interpretation: What is different about the 3 samples from LDB? Were they different times? Different colonies? Were they supposed to be replicates? It seems there is as much if not more variability in these three samples as is seen in samples from the other stations, especially with respect to the microbiome, but also with respect to the Trichodesmium expression results. This needs to be addressed somewhere in the manuscript.

We appreciate the reviewer pointing out that we omitted discussing this trend in our initial submission. The samples from LDB were collected on three different days in the same water mass, during the 6 days that the expedition spent at this "long duration" station. All samples were collected at the same time each morning. We have edited the text surrounding this LDB sampling in the revised manuscript to make this more clear. Furthermore, in the revised text we also discuss how LDB samples were taken during a declining phytoplankton bloom (de Verneil et al. *Biogeosci,* this issue), which might explain the high degree of variability. We discuss the implications of this in the revised version of the paper.

What is known about the overall expression levels of the genes that are discussed as marker genes? In other words, is the fact that these genes are among the highest expressed a sign that they are upregulated or because they are constitutively expressed at a high level? The statement on p17 that marker genes are not detectable in cultures grown under replete conditions is false. At least in the citations listed it appears the marker genes are detectable (though significantly downregulated) in replete conditions.

Can you use the expression of a housekeeping gene to normalize results in some way? The mix of iron response genes listed are concerning as high expression of all of them is not actually suggestive of iron limitation. Yes, the are all linked to iron metabolism, but previous work has shown that some are upregulated and others are down regulated in situations of iron limitation, and some have shown inconsistent results. For example, Polyvou et al 2018 found that bfr, ftn, fur3, and nifH gene expression was not regulated in response to iron. While examining protein expression, not transcripts, Snow et al 2015 found that ftn was only abundant under iron replete conditions (and absent under iron limitation) and nifH was similarly higher under iron replete conditions as would be expected. Actually, the high nifH expression and high rates of nitrogen fixation measured on the cruise as referenced in the other articles in this special issue would suggest that maybe the Trichodesmium are doing ok. It's very hard to assess the level of iron-limitation, phosphorus-limitation, or co-limitation of the Trichodesmium based on the transcriptome data without some relative measure or other metric to assess what these expression levels mean. It seems as though the authors have potentially relevant data that they could mine to address these issues.

We regret that our discussion of the *Trichodesmium* expression patterns was not clear. In the revised version of the paper we have added some additional analysis (Figure 4), as well as clarified, and streamlined the discussion of this section as requested. With the feedback from the reviewer, we have removed what was initially Figure 4a from the analysis. Although the high expression of many of our gene targets was tantalizing, the reviewer is correct that we can't confirm that this is driven by limitation scenarios without a comparison to a more replete condition. Rather, we focus our discussion on relative patterns of expression across the SD stations and the LD time-series. The focus on relative OG expression patterns across the sample set did not necessitate normalization to a reference gene, and our method for this analysis approach (Frischkorn et al., 2018, *ISME J*) is clarified in the revised text.

In the revised manuscript, we have limited our nutrient responsive gene analysis to only those that have been experimentally shown to respond with significantly increased transcript/protein expression in response to limitation conditions resulting in the low P responsive gene set (*sphX, ptxD, phoX*) (Dhyrman et al., 2006; Orchard et al., 2009; Polyviou et al., 2016), the low Fe responsive set (*fld1/2, isiA, fbaA, idiA*) (Chappell and Webb 2010; Chappel et al., 2012; Snow et al. 2015), and the co-limitation responsive set (*aroB, metE, FMO, OXSM1*) (Walworth et al., 2017). In our revised Figure 4, we show how that the expression of the majority of the P and Fe responsive OGs track together, showing similar expression dynamics across the transect. Notable exceptions are the expression of *phoA,* which clustered with the Fe responsive set and away from its counterpart *phoX* (potentially because P responsive *phoX* has an Fe-Ca cofactor that has repressed expression relative to *phoA* in low Fe environments (Rouco et al., 2018)), and *phnJ*. In the revised analysis we include only Fe and P OGs that track together across stations in the K-S testing to evaluate significant enrichment patterns (Figure 4b). The results of this new clustering show an enrichment in Fe responsive gene expression at SD2, an enrichment in P responsive gene expression at the first day from LDB, and mirroring in overall patterns of enrichment for low P responsive and co-limitations responsive gene sets. Because expression of these genes/proteins are experimentally verified to increase during limitation, and because of the similar expression patterns previously discussed, we feel that the patterns and dynamics of these sets—rather than the overall high expression, which we no longer discuss—are potent markers of the physiological state of *Trichodesmium* in a typically under-sampled region in the oligotrophic ocean, thus adding valuable insight to this organism's ecology and physiology.

We have also edited the text surrounding the discussion of *nifH*—in the revised discussion and analysis we now show that the expression of this iron requiring enzyme subunit clusters with the Fe-marker genes, a potential indicator of the tight modulation of cellular Fe demand in *Trichodesmium*. We are unable to directly compare specific *Trichodesmium nifH* signals with measurements of bulk water column N_2 fixation, however both the data presented in the revised text, and studies from the special issue, consistently suggest that Fe and P are important drivers of phytoplankton physiology in this region of the WTSP. The manuscript now better integrates OUTPACE papers and their findings into the discussion.

One of the most exciting bit about the paper is the presence of the Trichodesmium gene cassette that appears to allow the organism to perform phosphonate biosynthesis. Right now that exciting result is lost in the weeds of a convoluted discussion of micro-nutrient limitation.

At the reviewer's request we have streamlined the results and discussion of earlier sections of the paper (see comments above).

Additionally, in the conclusions the authors suggest that the variable limitation could be influencing growth and nitrogen fixation. It seems that there's a lot of data from this cruise they could examine to see if this is the case. It would be nice to see them include some concrete statements referencing the other manuscripts from the cruise.

In the revised text we now include a more integrated discussion referencing other studies to put our observations from the WTSP into greater context. We note however that the other samples discussed in this special issue were taken from bulk water rather than the method we employed of isolating low density *Trichodesmium* colonies. Therefore, measurements of N_2 fixation and other parameters reflect free living planktonic microbes and are not directly comparable to genomic, gene expression, or activity data that we present in this manuscript.

Trichodesmium physiological ecology and phosphate reduction in the western Tropical South Pacific

Kyle R. Frischkorn[1,2], Andreas Krupke[3*], Cécile Guieu[4,5], Justine Louis[4**], Mónica Rouco[2***], Andrés E. Salazar Estrada[1,2], Benjamin A. S. Van Mooy[3], and Sonya T. Dyhrman[1,2****]

[1]Department of Earth and Environmental Sciences, Columbia University, New York, NY, U.S.A.
[2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, U.S.A.
[3]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, U.S.A.
[4]Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, F-06230 Villefranche-sur-mer, France, Villefranche-sur-mer, France
[5]The Center for Prototype Climate Modeling, New York University in Abu Dhabi, Abu Dhabi, UAE

*Present address: Thermo Fisher Scientific, Life Science Solutions, 180 Oyster Point Blvd., South San Francisco, CA, U.S.A.
**Present address: University Rennes 1, Saint-Brieuc, France
***Present address: K=1 Project, Center for Nuclear Studies, Columbia University, New York, NY, U.S.A.
****Correspondence to: Sonya T. Dyhrman (sdyhrman@ldeo.columbia.edu)

Abstract

N_2 fixation by the genus *Trichodesmium* is predicted to support a large proportion of the primary productivity across the oligotrophic oceans, regions that are considered among the largest biomes on Earth. Many of these environments remain poorly sampled, limiting our understanding of *Trichodesmium* physiological ecology in these critical oligotrophic regions. *Trichodesmium* colonies, communities that consist of the *Trichodesmium* host and their associated microbiome, were collected across the oligotrophic western tropical South Pacific (WTSP). These samples were used to assess host clade distribution, host and microbiome metabolic potential, and functional gene expression, with a focus on identifying *Trichodesmium* physiological ecology in this region. Genes sets related to phosphorus, iron, and phosphorus-iron co-limitation were dynamically expressed across the WTSP transect, suggestive of the importance of these resources in driving *Trichodsmium* physiological ecology in this region. A gene cassette for phosphonate biosynthesis was detected in *Trichodesmium,* the expression of which co-varied with the abundance of *Trichodesmium* Clade III*,* which was unusually abundant relative to Clade I in this environment. Coincident with the expression of the gene cassette, phosphate reduction to phosphite and low molecular weight phosphonate compounds was measured in *Trichodesmium* colonies. The expression of genes that enable use of such reduced phosphorus

compounds were also measured in both *Trichodesmium* and the microbiome. Overall, these results highlight physiological strategies employed by consortia in an undersampled region of the oligotrophic WTSP, and reveal the molecular mechanisms underlying previously observed high rates of phosphorus-reduction in *Trichodesmium* colonies.

1 Introduction

The oligotrophic oceans extend over approximately 70% of the Earth and are characterized by chronically low nutrient concentrations that limit primary productivity (Moore et al., 2013). Within oligotrophic marine environments, N_2 fixing microorganisms can serve as a source of "new" nitrogen that is bioavailable to other organisms. Among these marine diazotrophs, the colonial, filamentous cyanobacterium *Trichodesmium* plays a disproportionately large role in the cycling of carbon, phosphorus and nitrogen: it supplies fixed carbon through photosynthesis, was recently found to be a hotspot of phosphate reduction (Van Mooy et al., 2015), and has been estimated to be responsible for approximately half of the biologically fixed N_2 in the ocean (Bergman et al., 2013; Capone et al., 1997). As such, the efficiency of the biological pump in sequestering carbon in the deep ocean is dependent in part on the distribution and activities of diazotrophic organisms like *Trichodesmium,* and an understanding of how this organism's physiology and ecology varies across diverse environments is a critical aspect of understanding present, and future, global biogeochemical cycles.

Diazotrophy frees *Trichodesmium* from nutrient limitation by nitrogen. As such, the distribution and activities of this cyanobacterium are predominantly influenced by the availability of phosphorus and iron in the surface ocean, which vary depending on the ocean basin and its proximity to supply of these resources (Moore et al., 2013; Sohm et al., 2011). Evidence of the intense competition for phosphorus and iron is evident in the suite of physiological strategies that this organism is known to employ. These strategies include the production of transporters and enzymes that take up and hydrolyze diverse organic and reduced phosphorus compounds (Dyhrman et al., 2006; Orchard et al., 2009; Polyviou et al., 2015), or enable the uptake and storage of organic and inorganic iron (Polyviou et al., 2018; Snow et al., 2015). The genes encoding these functions are expressed *in situ* across diverse environments, indicating that

competition for these resources is a critical aspect of *Trichodesmium* physiology (Chappell et al., 2012; Dyhrman et al., 2006; Rouco et al., 2018a). Recent evidence from culture studies also suggests that *Trichodesmium* employs a unique set of physiological strategies to cope with co-limitation of phosphorus and iron that differs from that of either resource alone (Walworth et al., 2017).

Survival in oligotrophic environments might also be enabled by biological interactions within *Trichodesmium* colonies. *Trichodesmium* has long been known to occur with tightly associated bacteria that are unique from those free-living in the water column (Hmelo et al., 2012; Paerly et al., 1989). Recent evidence suggests that these interactions are ubiquitous and taxonomically conserved across ocean basins (Lee et al., 2017; Rouco et al., 2016a) and that this epibiotic bacterial community, referred to as the *Trichodesmium* microbiome, contains a large amount of metabolic potential that exceeds and complements that of the *Trichodesmium* host in populations from the western North Atlantic (Frischkorn et al., 2017). Coordinated gene expression patterns within the holobiont (*Trichodesmium* and its microbiome) suggest an interdependence of the microbiome on host-derived fixed carbon, N_2 and vitamins, and suggests microbiome respiration could create conditions that favour continued diazotrophy and photosynthesis (Frischkorn et al., 2018; Paerl and Bebout, 1988). The stability of these relationships in the future ocean is unknown, but they are likely to change. For example, incubation of cultured *Trichodesmium* colonies with an elevated carbon dioxide concentration resulted in significant changes in microbiome nutritional physiology (Lee et al., 2018). These microbiome changes have the potential to alter the amount of fixed N_2 and carbon that transfer from the colony to the environment at large. Overall, the continued appreciation of the importance of the microbiome in *Trichodesmium* ecology underscores that investigations must consider these microbial communities as a holobiont in order to fully understand and predict their role in the future environment.

Geochemical drivers of *Trichodesmium* distribution and N_2 fixation are increasingly well characterized in regions of the ocean where either phosphorus or iron are limiting such as the North Atlantic and North Pacific Subtropical Gyre (Rouco et al., 2018a; Sañudo-Wilhelmy et al., 2001; Sohm et al., 2011). The western tropical South Pacific (WTSP) represents and understudied region of the world's oceans

(Bonnet et al., 2018) with conditions characterized by chronically low concentrations of both iron and phosphate (Moore et al., 2013; Sohm et al., 2011). Despite low resources, this region can support high levels of N_2 fixation, with rates exceeding 700 umol m^{-2} d^{-1} where this resource transfers across diverse ecological groups and ultimately supplies up 90% of to the photic zone with new nitrogen (Bonnet et al., 2017; Caffin et al., 2017, 2018). In this study, metagenomic and metatranscriptomic sequencing was leveraged along with taxonomic distribution, physiological activities and geochemical measurements to better understand *Trichodesmium* physiological ecology in an under-sampled but important region of the oligotrophic ocean.

2 Materials and Methods

2.1 Biogeochemical analyses

Samples were collected across a transect of the western Tropical South Pacific (WTSP) during the OUTPACE cruise (Oligotrophy to UlTra-oligotrophy PACific Experiment, DOI: http://dx.doi.org/10.17600/15000900) aboard the *R/V L'Atalante* during austral summer (February – April) of 2015 (Fig. 1, Moutin et al., 2017). Samples for nutrient analyses were collected using a Titanium Rosette mounted with 24 Teflon-coated 12L GoFlos and operated along a Kevlar cable. Samples were filtered directly from the GoFlos through 0.2-µm cartridges (Sartorius Sartrobran-P-capsule with a 0.45-µm prefilter and a 0.2-µm final filter) inside a clean van and analysed for dissolved inorganic phosphorus (DIP) and total dissolved iron concentrations (DFe). DIP was analysed directly on board using a 2-m length Liquid Waveguide Capillary Cells (LWCC) made of quartz capillary following the method described in Pulido-Villena et al. (2010). Briefly LWCC was connected to a spectrophotometer and the measurements were performed in the visible spectrum at 710nm. The 2-m length LWCC allowed for a detection limit of 1 nM and a relative standard deviation of less than 10%. DFe concentrations were measured by flow injection with online preconcentration and chemiluminescence detection using the exact protocol, instrument, and analytical parameters as described by Blain et al., 2008.

Water column phosphate uptake rate was determined as described by Van Mooy et al., (2015), and is briefly outlined here. First, 50 mL aliquots of whole seawater collected in Niskin bottles were decanted into acid-washed polycarbonate vials. Next, 1 µCi of carrier-free ^{33}P-phosphoric acid was added to the bottles, which represented an amendment of approximately 10 pmol L^{-1} of phosphoric acid. Then, bottles were incubated in an on-deck incubator for 2-4 hr. Finally, the seawater in the bottles was filtered through a 25 mm diameter, 0.2 µm poresize polycarbonate membrane, and the radioactivity of the membrane was determined by liquid scintillation counting. The phosphate turnover rate in each incubation was calculated as the quotient of the ^{33}P-radioactivity of the membrane and the total ^{33}P-radioactivity of the amendment, divided by the time duration of the incubation. The phosphate uptake rate was calculated as the product of the phosphate turnover and the DIP concentration. Data were visualized and contoured using Ocean Data View 4.6.2 with the DIVA grid method (R. Schlitzer; http://odv.awi.de).

2.2 *Trichodesmium* clade sampling and analysis

Samples for *Trichodesmium* clade distribution analysis were obtained at selected short duration stations across the transect from water depths ranging from 5 m to ~150 m using 12 L of water for each depth obtained from a Rosette sampling device filtered through 47 mm 10 µm pore size polycarbonate filters. Filters were flash frozen and stored in liquid nitrogen until processing. Quantification of absolute cell numbers in these samples was performed following a previously described protocol (Rouco et al., 2014). Briefly, DNA was extracted from filters and the abundance of *Trichodesmium* clade I (which encompasses *T. thiebautii, T. tenue, T. hildebrandtii,* and *T. spiralis*) and III (which encompasses *T. erythraeum* and *T. contortum*), was determined with quantitative polymerase chain reactions (qPCR) targeting the *rnpB* gene using clade specific primer sets (Chappell and Webb, 2010). Amplification of standards, no template controls (RNase-free water), and field samples were run in triplicate on a Bio-Rad CFX96 Real-Time System C1000 Touch Thermal Cycler using Bio-Rad SYBR Green SuperMix (Bio-Rad Laboratories, Inc., Hercules, CA, USA). Standard curves were generated from DNA extracts

performed on filters with known concentrations of *T. erythraeum* IMS101 and *T. thiebautii* VI-I. Concentrations were previously determined by 10 replicates of cell counting using a Sedgwick Rafter slide after Rouco et al. (2014). Reactions were run in final volumes of 25 µL, encompassing 12.5 µL SuperMix, 2 µL template, 9.5 µL sterile water, and 200 nmol L^{-1} forward and reverse primers. Reaction conditions were as follows: 2 min. at 50°C, 10 min. at 95°C; 40 cycles of 15 sec. at 95°C, and 1 min at 55°C with a fluorescence measurement. Resulting C_Ts were averaged across the triplicates and compared against the standard curve to calculate the abundance of the *rnpB* gene, which we interpret as absolute cell number. Clade distribution data was displayed and contoured using Ocean Data View 4.6.2 with the DIVA grid method (R. Schlitzer; http://odv.awi.de).

2.3 *Trichodesmium* colony sampling

Trichodesmium colonies were sampled across the transect at approximately the same time (between 8AM and 10:50AM local time). *Trichodesmium* samples were obtained with six manual hauls of a 130 µm mesh size net hand towed through surface sea water. The total time for six hauls of the net tow was approximately 15 minutes and likely filtered thousands of liters of seawater. Colonies were skimmed from the concentrated net tow sample with a Pasteur pipette from the surface layer of net towed samples and then washed two times with 0.2 µm sterile-filtered surface seawater on 5 µm, 47 mm polycarbonate filters with gentle vacuuming (<170 mbar) to remove non-tightly associated microorganisms. All colony samples were cleaned and processed within 15 minutes of collection. Filters with colonies were flash frozen and stored in liquid nitrogen for DNA or RNA extraction.

2.4 Phosphate reduction in *Trichodesmium* colonies

Phosphate uptake and synthesis rates of low-molecular-weight (LMW) reduced phosphate (P(+3)) compounds in *Trichodesmium* colonies were determined as described by Van Mooy et al., (2015). Briefly, phosphate uptake by *Trichodesmium* colonies was determined by filling acid-washed polycarbonate 50 mL bottles with filter-sterilized surface seawater and approximately 20 *Trichodesmium* colonies. Incubation in on-deck incubators and measurement by liquid scintillation counting proceeded as previously described above for the whole water analyses. In parallel, to measure

the synthesis rates of LMW P(+3) compounds, at stations SD2, LDA, SD9, SD11, and LDB *Trichodesmium* incubations were not immediately measured by liquid scintillation counting. Instead, colonies were placed in a cryovial containing 1 mL of pure water and flash frozen in liquid nitrogen. These samples were then transported to the lab ashore and subjected to numerous freeze-thaw cycles to extract intracellular LMW P(+3) compounds. The LMW P(+3) compounds in the extracts were then isolated by preparative anion chromatography. Two fractions were collected in retention time windows consistent with retention times of pure standards of 1) methylphosphonic acid, 2-hydroxy ethylphosphonic acid, and 2-amino ethylphosphonic acid; and 2) phosphorous acid. The ^{33}P radioactivity in these two operationally-defined fractions is ascribed to LMW phosphonates and phosphite, respectively.

2.5 DNA extraction and metagenome sequencing

Genomic DNA was extracted from *Trichodesmium* colony samples (~40 colonies per sample) obtained from one day each during LDA and LDB, as well as at station SD5 (Fig. 1) using the MoBio Power Plant Pro DNA Isolation Kit (MoBio Laboratories, Inc., Carlsbad, CA, USA) following the manufacturer instructions. Genomic DNA extracts were sequenced at the Argonne National Lab (Lemont, IL, USA) following a *Trichodesmium* consortium protocol previously described (Frischkorn et al., 2017). Briefly, DNA was sheared with a Covaris Sonicator (Woburn, MA, USA), transformed into libraries with WaferGen Apollo324 automated library system (Clonetech Laboratories, Mountain View, CA, USA) and Illumina compatible PrepX ILMN DNA kits (San Diego, CA, USA) following manufacturer instructions. An average insert size of ~750 base pairs was targeted. Sage BluePippin (Beverly, MA, USA) was used to size select libraries prior to sequencing all three samples split across one 2 x 100 bp lane of the Illumina HiSeq2000. Metagenomic reads from these three samples are available on the NCBI Sequence Read Archive under BioProject number PRJNA435427.

2.6 Metagenomic sequence assembly and analysis

Raw reads were trimmed assembled into scaffolds and subsequently analysed following the protocol previously reported in Frischkorn et al., (2017) and summarized here in an assembly and analysis

methods pipeline (Supplemental Fig. 1). Briefly, reads were trimmed with Sickle (https://github.com/najoshi/ sickle), converted into fasta format, merged together and co-assembled with IDBA-UD (Peng et al., 2012) so as to create a South Pacific *Trichodesmium* holobiont genomic template to which future metatranscriptomic reads could be mapped. Assembled scaffolds were partitioned between *Trichodesmium* and heterotrophic bacteria (hereafter referred to as the microbiome) and after clustering into genome bins using MaxBin 2.0 with default parameters (Wu et al., 2015). The taxonomic partitioning of binned scaffolds was carried out after translation of each scaffold into protein coding genes with the metagenomic setting of Prodigal (Hyatt et al., 2010), annotation of resultant proteins with the blastp program of DIAMOND (Buchfink et al., 2015) against the NCBI nr database, and classification with MEGAN6 (Huson et al., 2013) based on the phylogenetic classification of the majority of proteins within a genome bin. Scaffolds from bins identified as phototrophic bacteria or eukaryotes were excluded from further analyses. Functional annotations for translated proteins in the *Trichodesmium* and microbiome identified genome bins were obtained by DIAMOND against the UniRef90 database (Suzek et al., 2007) with an e-value cut-off of $1x10^{-3}$. Functional annotation was also carried out using the Kyoto Encyclopedia of Genes and Genomes (KEGG) with the online Automatic Annotation Server using the bi-directional best-hit method, the GHOSTX search program, and the prokaryote representative gene set options. KEGG definitions were obtained from the modules within the "Pathway module" and "Structural complex" categories and the submodules therein (available at https://www.genome.jp/kegg-bin/get_htext?ko00002.keg).

Proteins from the merged assembly were also clustered into gene families of similar function or orthologous groups (OGs) following a previously reported pipeline (Frischkorn *et al.,* 2017; Supplemental Fig. 1). Briefly, reciprocal blast of translated proteins greater than 70 amino acids were performed with the DIAMOND blastp program. Blast results were processed using the program MCL (Markov cluster algorithm) set to an inflation parameter or 1.4. UniRef was used for the consensus functional annotation of each OG. The final UniRef annotation and KEGG annotations for each OG represent the individual annotation that the majority of proteins within that OG were assigned to. Homologs to the phosphonate biosynthesis gene phosphoenolpyruvate phosphomutase (*ppm*) were

found by screening against manually annotated and reviewed Ppm proteins from the Swiss-Prot database. These verified proteins were aligned with Muscle version 3.8.425 with default parameters (Edgar, 2004), converted into HMM profiles with hmmbuild and hmmpress and used as a the database for hmmsearch, all using HMMER version 3.1 (Eddy, 1995). This HMMER approach was used to screen proteins generated from this study, as well as a previously published *Trichodesmium* consortia metagenome assembly from the western tropical North Atlantic Ocean (Frischkorn et al., 2017) and protein sequences derived from the genome sequence of *Trichodesmium erythraeum* IMS101 (https://genome.jgi.doe.gov/portal/TrieryIMS101_FD/TrieryIMS101_FD.info.html). Sequence alignments were visualized using Geneious version 11.0.3 (http://www.geneious.com/, Kearse et al., 2012) and important residues were obtained from previous crystal structure analyses (Chen et al., 2006). A reassembly of *Trichodesmium*-identified bins was performed to lengthen scaffolds in an attempt to provide genomic context to the *ppm*-containing scaffold. The subset of all metagenomic reads mapping to *Trichodesmium* bins was selected using Samtools (Li et al., 2009) and the subseq program in Seqtk (https://github.com/lh3/seqtk) and then reassembled using IDBA-UD as described previously. Maximum likelihood phylogenetic analysis of Ppm and related proteins was performed using the FastTree plugin in Geneious with default settings (Price et al., 2010), following a protocol previously employed for the annotation of environmental Ppm proteins (Yu et al., 2013). The sequences used to generate the tree were composed of proteins with homology to the identified *Trichodesmium* Ppm proteins, as well as similar sequences pulled from the NCBI nr database after online blastp analysis. The Interactive Tree of Life program was used to edit phylogenetic trees (Letunic and Bork, 2016). Proteins were also screened for MpnS, a protein that produces methylphosphonate, using the same protocol described above.

2.7 RNA extraction and metatranscriptomic sequencing

Prokaryotic RNA was extracted and sequenced from *Trichodesmium* colony samples obtained from SD2, SD6 and SD9 as well as colonies collected on three separate days at LDB (~40 colonies per sample) (Fig. 1) following a protocol described previously (Frischkorn et al., 2018). Briefly, the Qiagen RNeasy Mini Kit (Qiagen, Hildern, Germany) was used to extract total RNA following manufacturer

instructions, with the addition of 5 minutes of bead beating with ~500 µL 0.5 mm zirconia/silica beads after addition of Buffer RLT. On column DNase digestion (RNase-free DNase Kit, Qiagen) was performed. A MICROBEnrich Kit (ThermoFisher Scientific, Waltham, MA, USA) was used to enrich the prokaryotic RNA fraction and ribosomal RNA was removed with the Ribo-Zero Magnetic kit for bacteria (Illumina), both following manufacturer instructions. Concentration and integrity of mRNA was assessed using a BioAnalyzer and the RNA 600 Nano Kit (Agilent Technologies, Santa Clara, CA, USA). Library preparation and sequencing was performed at the JP Sulzberger Genome Center at Columbia University. Libraries were generated with the Illumina TruSeq RNA sample preparation kit. Samples were chemically fragmented using the Fragment, Prime, Finish Mix reagent (Illumina) which generates fragments of 140-220 bp. An Illumina HiSeq 2500 was used to sequence 60 million paired end 100 bp reads for each sample. Metatranscriptomic reads from these six samples are available on the NCBI Sequence Read Archive under BioProject number PRJNA435427.

2.8 Metatranscriptomic sequence analysis

Metatranscriptomic reads were trimmed, normalized and mapped as previously described (Frischkorn et al., 2018) and are summarized in the analysis and assembly pipeline (Supplemental Fig. 1). Briefly, raw reads were pre-processed following the Eel Pond Protocol for mRNAseq (Brown et al., 2013). Cleaned reads were mapped with RSEM (paired-end and bowtie2 options selected) (Li and Dewey, 2011) to the protein coding regions of the metagenomic scaffolds previously partitioned across *Trichodesmium* and the microbiome with an average of 10 million reads mapping per sample. Read counts were summed separately for *Trichodesmium* and microbiome fractions for all genes in an OG. Counts were normalized in each sample by calculating the transcript reads mapped per million (TPMs) separately for the *Trichodesmium* and microbiome fractions. In each sample, percentages of reads recruiting to KEGG submodules were normalized individually, and for *Trichodesmium* and the microbiome separately. KEGG heatmaps are visualized as row-wide averages for each submodule. Comparisons of relative enrichment across sets of nutrient responsive genes were made using Kolmogorov-Smirnov tests to examine the null hypothesis that the expression of gene sets at a given station did not deviate

significantly from the average expression of that set across the transect. Prior to testing, TPM expression values for each OG were normalized to the average abundance of that OG across the six samples. This normalization equalized the relative contribution of individual OGs to the gene set as a whole, thereby avoiding bias caused by highly expressed individual OGs. *P* values less than 0.05 were considered significant. Pairwise correlation coefficients between clade abundance, phosphate and OG expression (TPMs) were calculated using the cor function in R, and *p* values less than 0.05 were considered significant. In the case of the OG expression, LDB samples from separate days were averaged. The data did not span enough coincident samples to do pairwise correlations between OG expression and iron. *Trichodesmium* OG TPMs were hierarchically clustered with the Broad Institute Morpheus program (https://software.broadinstitute.org/morpheus/) using the "one minus Pearson correlation" metric and the "Average" linkage method. Only OGs with an average expression greater than 2 TPM across all samples were included in this analysis.

3 Results

3.1 Biogeochemistry

Across the study transect the DIP concentration in surface water (10 m depth, with the exception of the sample from LDA which was collected at 30 m) ranged between 2.3 nmol L^{-1} and 230 nmol L^{-1} (Fig. 1, Supplemental Table 1) and averaged 36 nmol L^{-1}. The phosphate turnover time in the water column microbial community was variable across the transect ranging between approximately 2 hours and 800 hours (Fig. 1, Supplemental Table 1), averaging ~220 hours across all stations sampled. The water column phosphate uptake rate was similarly variable, ranging from 0.006 nmol L^{-1} hr^{-1} to 0.68 nmol L^{-1} hr^{-1} with lower uptake measured at stations where turnover time was high and vice versa (Fig. 1). Iron concentrations in the surface water (10 m depth, with the exception of the sample from SD12 which was collected at 30 m) ranged between 0.21 nmol L^{-1} and 1.16 nmol L^{-1} (Fig. 1, Supplemental Table 1) and averaged 0.6 nmol L^{-1}.

3.2 *Trichodesmium* clade distribution

Trichodesmium (combined cell counts of Clade I and Clade III) was detected at every station and at all depths sampled across the transect with a maximum estimated concentration of over 172,000 cells L^{-1} at 11 m at SD6 (Fig. 2a). Overall, abundance was markedly greater at the stations in the western half of the transect (stations west/left of 170°W and LDB (Fig. 2a). In this western region of the transect, *Trichodesmium* was concentrated in the surface ocean between approximately 5 - 20 m depth, where the average concentration was ~53,000 cells L^{-1}. To the east/right of station LDB and the transition into the ultra-oligotrophic region, the concentration of *Trichodesmium* dropped and the average concentration at the surface between approximately 5 - 20 m was 75 cells L^{-1} (Fig. 2a). All cell counts are provided in Supplemental Table 2. In addition to overall abundance of *Trichodesmium,* the contribution of Clade I (*T. thiebautii, T. tenue, T. hildebrandtii,* and *T. spiralis*) and Clade III (*T. erythraeum* and *T. contortum*) to the total *Trichodesmium* abundance at each station was determined (Fig. 2b). Across the transect, the *Trichodesmium* communities were dominated by Clade I which made up approximately 80% of the cells measured on average at each station, while Clade III made up approximately 20% on average (Fig. 2b). Although Clade I was dominant overall, the percentage of Clade III rose to nearly 50% of the measured *Trichodesmium* community in some samples, and was highest between 4 and 40 m at SD4, SD6 and SD11 (Fig. 2b). Clade III was not detected at any depth in the ultra-oligotrophic subtropical gyre stations (SD14 and SD15) (Fig. 2b).

3.3 Metagenomic characterization of WTSP *Trichodesmium* consortia

A merged metagenomic assembly of *Trichodesmium* consortia reads from selected stations across the OUTPACE transect (Fig. 1) yielded 801,858 scaffolds in total. Taxonomic binning partitioned scaffolds into 48 genome bins with similar read coverage and tetranucleotide frequency. After phylogenetic analysis with MEGAN6, 18 of the bins were classified as *Trichodesmium,* 23 as heterotrophic bacteria, while the remaining bins had the majority of their proteins phylogenetically classified as Eukaryotes or other photosynthetic cyanobacteria. Subsequently, all *Trichodesmium*-identified scaffolds were merged and considered as the *Trichodesmium* fraction of the WTSP consortia sampled. Similarly, all heterotrophic bacteria-identified bins were merged together and considered as the microbiome fraction. Together, these taxonomically verified scaffolds were translated into 198,156 proteins, which clustered

into 75,530 gene families of putatively similar function, or orthologous groups (OGs). Within the WTSP consortia, *Trichodesmium* and their microbiome possessed 9,790 and 68,538 OGs respectively. Each OG likely represents a unique function or variant of a gene, as the majority of key functions examined subsequently were represented by one OG. The majority of these OGs were unique to the microbiome, with 2,798 (3.7%) of the total OGs composed of proteins found in both the *Trichodesmium* and microbiome genome bins.

Functional annotation of OGs showed that the microbiome contained nearly 10 times more unique KEGG IDs (Supplemental Fig. 2). The greatest differences in functional gene capacity between *Trichodesmium* and the microbiome were found in the Environmental Information Processing and Carbohydrate and Lipid Metabolism modules (Supplemental Fig. 2). Unique microbiome KEGG submodules included glycan, fatty acid, and carbohydrate metabolism functions. In the Environmental Information Processing category the microbiome possessed unique functions pertaining to peptide, nickel, phosphate, amino acid, and ABC transporters as well as an enrichment in proteins related to bacterial secretion systems (Supplemental Fig. 2).

3.4 Expression profiling of WTSP *Trichodesmium* consortia

A total of 7,251 *Trichodesmium* OGs and 21,529 microbiome OGs recruited metatranscriptome reads from at least one sample. Hierarchical clustering of expression patterns in KEGG annotated OGs identified variable patterns in *Trichodesmium* and the microbiome gene expression over space and time (Fig. 3). Submodules in the carbohydrate and lipid metabolism, genetic information processing, and nucleotide and amino acid metabolism modules had elevated expression in the SD stations relative to LDB. Similar patterns were generally observed in the expression of microbiome modules, most strikingly in the nucleotide and amino acid metabolism and environmental information processing modules. In *Trichodesmium,* exceptions to the trend included OGs in the ATP synthesis, nitrogen metabolism, phosphate and amino acid transport, two-component regulatory systems, and certain amino acid metabolism submodules, all of which had elevated relative expression during the three days of

samples from LDB (Fig. 3). Summed TPM for each OG in the *Trichodesmium* and microbiome fractions along with their most abundant UniRef annotations are provided in Supplemental Table 3.

Hierarchical clustering of expression patterns in *Trichodesmium* OGs identified shifts in *Trichodesmium* gene expression from station to station, and over the three days of sampling at LDB (Fig. 4a). The nitrogenase enzyme subunit *nifH* OG peaked at SD2 (Fig. 4a), while a RuBisCO OG peaked at LDB 1 (Fig. 4a). The expression of *Trichodesmium* OGs known to be responsive to low nutrient conditions also showed variation in expression pattern across the transect (Fig. 4a). The low phosphorus (low P) responsive set was composed of OGs previously shown to have increased relative expression under conditions of phosphate stress and included the alkaline phosphatases *phoA* (Tery_3467) and *phoX* (Tery_3845; proteins identified as PhoX clustered into two separate OGs, as has been previously observed in natural populations of *Trichodesmium* (Rouco et al., 2018)), the high affinity phosphate binding protein *sphX* (which is homologous to and clustered into one OG with *pstS*) (Tery_3534), phosphite dehydrogenase *ptxD* (Tery_0368) and the carbon-phosphorus lyase gene marker *phnJ* (Tery_5000) (Dyhrman et al., 2006; Orchard et al., 2009; Polyviou et al., 2015). The low iron (low Fe) responsive set included OGs previously shown to have increased transcript or protein expression in experimental low iron cultures and *in situ* in low iron environments and included the flavodoxins *fld1* and *fld2* which clustered into the same OG (Tery_1666, Tery_2559), fructose bisphosphate aldolase class II *fbaA* (Tery_1687), the iron-stress induced protein *isiA* (Tery_1667), and the iron-stress induced gene *idiA* (Tery_3377) (Chappell et al., 2012; Chappell and Webb, 2010; Snow et al., 2015; Webb et al., 2001). Also assayed were a suite of OGs recently shown to be significantly enriched in cultures of *T. erythraeum* IMS101 following prolonged maintenance in co-limiting concentrations of phosphorus and iron (herein called the co-limitation responsive set) (Walworth et al., 2017). These OGs included the flavin-containing monoxygenase *FMO* (Tery_3826) that hydrolyzes organic nitrogen, 5-methyltetrahydropteroyltriglutamate—homocysteine methyltransferase *metE* (Tery_0847), the 3-dehydroquinate synthase *aroB* (Tery_2977), and beta-ketoacyl synthase (*OXSM*, Tery_3819 and Tery_3821 which clustered into one OG). All OGs included in the three nutrient responsive *Trichodesmium* sets were expressed at each station In general, low P and low Fe responsive gene set

expression patterns tracked together from sample to sample. Two exceptions to this trend were the alkaline phosphatase *phoA* and the phosphonate lyase *phnJ* (Fig. 4a). There were no significant correlations between the low P responsive OGs and the phosphate concentration. The majority of the OGs (*aroB*, *metE*, *FMO*) in the co-limitation responsive set had similar expression patterns to the majority of the low P responsive set, with the exception of *OXSM* (Fig. 4a). Expression of the RuBiSCO OG was also modulated with this pattern (Fig. 4a). The expression of the *nifH* OG tracked with the low Fe responsive set as well as *phoA,* and *OXSM* (Fig. 4a).

The relative enrichment of the low P responsive gene set, low Fe responsive gene set, and co-limitation gene set were examined from station to station, and over the three days of sampling at LDB using Kolmogorov-Smirnov tests (Fig. 4b) This approach estimated enrichment based on normalized expression levels of the resource responsive genes at individual stations relative to the average expression of the set across all stations sampled. In testing enrichment of the low P responsive set, only the four OGs whose expression clustered together across the transect (Fig. 4a) were used. The signal of the low P responsive set was significantly enriched during the first day sampled at LDB, while the signal in the low Fe responsive set was significantly enriched at SD2 (Fig. 4b). The co-limitation set was not found to significantly deviate from the average expression levels at any station sampled, although the overall pattern in expression was similar to that of the low P responsive set (Fig. 4b).

3.5 Evidence of phosphonate biosynthesis in *Trichodesmium*

A ~7,400 bp scaffold containing a cassette of genes encoding a phosphonate biosynthesis pathway was found in the *Trichodesmium* partitioned genome bins (Fig. 5a). This scaffold contained 6 protein coding regions. The first three genes were annotated as phosphoenolpyruvate phosphomutase (*ppm*), phosphonopyruvate decarboxylase (*ppd*), and 2-aminoethylphosphonate-pyruvate transaminase (*2-AEP-TA*). These first two genes are most similar to homologs in the UniRef databaste that belong to the non-heterocystous diazotroph *Planktothrix agardhii, a* member of the order Oscillatoriales along with *Trichodesmium.* The *2-AEP-TA* is homologous to a gene in a Gammaproteobacterial *Beggiatoa sp.* The fourth, fifth and sixth genes in this scaffold were annotated as a methyltransferase, a

cytidylyltransferase (both homologs to genes in *P. agardhii*), and a group 1 glycosyl transferase (homologous to a gene from a *Tolypothrix* sp., a freshwater cyanobacterium in the order Nostocales).

In order to verify genome binning of the *ppm*-containing scaffold into the *Trichodesmium* fraction, reassembly was attempted to lengthen this scaffold. This effort extended the length of the scaffold upstream of the *ppm* cassette and resulted in the addition of 44 bp to the 5' end with 95% homology to a non-coding region in the *T. erythraeum* IMS101 genome, as well as a protein coding gene fragment with no known annotation (Fig. 5a; sequence included in Supplemental Table 4). Metagenomic reads mapped to this scaffold showed extensive coverage of this IMS101 region as well as mate pairs connecting it to genes downstream in the *ppm*-containing cassette (Supplemental Fig. 3).

Comparison of the putative *Trichodesmium ppm* gene's amino acid sequence against experimentally verified phosphonate bond forming enzymes in other organisms showed high sequence identity (Supplemental Fig. 4). Across the length of the peptide sequence, the *Trichodesmium* Ppm was 72% and 66% similar to sequences from the freshwater cyanobacteria *Planktothrix sp.* and *Moorea producens*, 65% similar to that of the freshwater ciliate *Tetrahymena,* 63% similar to the blue mussel *Mytilus edulis*, and 35% similar to that of a bacterium in the *Streptomyces* genus. Furthermore, closer inspection of the sequences of all organisms showed 100% identity across residues involved in cofactor and substrate interactions, as well as strong conservation across other key residues involved in the enzyme's tertiary structure as determined by crystal structure analysis and comparison to similar enzymes (Supplemental Fig. 4) (Chen et al., 2006). Phylogenetic analysis of *Trichodesmium* Ppm placed it in a phylogenetic branch along with sequences from other N_2 fixing freshwater cyanobacteria (Supplemental Fig. 5).

A metagenome derived from *Trichodesmium* consortia in the western tropical North Atlantic (Frischkorn et al., 2017) was also re-screened for proteins with homology to the *Trichodesmium* Ppm sequence from this dataset. Five homologous sequences were recovered from this North Atlantic metagenome, with two proteins falling in the branch of verified phosphonate producing Ppm proteins.

One protein was located on a branch adjacent to the *Trichodesmium* Ppm sequences recovered from this South Pacific dataset (Supplemental Fig. 5). With the exception of the closely related North Atlantic sequence adjacent to the Ppm sequence recovered from this dataset, the other similar proteins from the South Pacific and North Atlantic consortia did not exhibit amino acid identity at the key conserved residues determined from crystal structure analysis (Chen et al., 2006).

The *T. erythraeum* IMS101 genome was also screened for homologs to the Ppm protein recovered here, but no sequences with conserved amino acid identity across important conserved residues were detected. Furthermore, another enzyme responsible for biosynthesis of methylphosphonate, MpnS, was not detected in proteins from the *Trichodesmium* or microbiome fractions.

All 6 genes in this scaffold recruited reads after metatranscriptome mapping of each sample, indicating that all genes in the phosphonate biosynthesis cassette were expressed at each station, with the exception of the methyltransferase, which recruited no reads from LDB1 (Fig. 5b). Relative expression of *ppm* at the three SD stations and the average expression across the three LDB samples was significantly positively correlated with surface water column average abundance (~5-25 m) of Clade III *Trichodesmium* ($R = 0.99$, $p = 0.006$, Pearson correlation). Furthermore, the concentration of DIP (10 m) was positively correlated with both Clade III abundance ($R = 0.99$, $p = 0.009$, Pearson correlation) and expression of *ppm* ($R = 0.97$, $p = 0.03$, Pearson correlation). There was no significant correlation between *ppm* relative expression and Clade I abundance ($p = 0.1$), or between DIP concentration and Clade I abundance ($p = 0.13$). Coincident with expression of phosphonate biosynthesis genes at all stations sampled, phosphate reduction by *Trichodesmium* colonies was also detected in each sample analysed (Fig. 5c). At the five stations tested, approximately 2% of the radiolabeled phosphate taken up by colonies was reduced to either a LMW phosphonate compound (methylphosphonate, phosphonoacetylaldehyde, or 2-aminoethylphosphonate) or phosphite (PO_3^{3-}) (Fig. 5c). Coincident with phosphonate production, both *Trichodesmium* and the microbiome possessed and expressed markers for reduced phosphorus metabolism, including the phosphonate (C-P) lyase *phnJ* (Tery_5000 in the *T. erythraeum* IMS101 genome) and *ptxD* (Tery_0368 in the *T. erythraeum* IMS101 genome), a gene

identified as phosphite dehydrogenase that is implicated in oxidation of phosphite to phosphate (Polyviou et al., 2015).

4 Discussion

4.1 *Trichodesmium* distributions in the oligotrophic WTSP

The WTSP is considered to be among the most oligotrophic environments in the global ocean due to low concentrations of critical resources like nitrogen, phosphorus and iron, coupled with intense stratification that prevents upwelling of remineralized nutrients (Moutin et al., 2008). In spite of chronically depleted resources, a diverse assemblage of free-living and symbiotic diazotrophs thrive in this region (Stenegren et al., 2017). In the WTSP, *Trichodesmium* is typically abundant, and the *Trichodesmium* distribution determined here by clade-specific qPCR agreed with previous analyses along this transect using genus specific qPCR that showed high abundance in the west, decreasing sharply at the transition into the gyre (Stenegren et al., 2017). The absolute value of *Trichodesmium* cells estimated here must be interpreted with some caution, as cultures and field samples of *Trichodesmium* are known to exhibit polyploidy (Sargent et al., 2016). However, the counts presented here are tabulated using a standard curve generated from cell counts performed on cultures of Clade I and Clade III (Rouco et al., 2014, 2016b), which would yield C_T values that take into account polyploidy, unlike gene standard approaches. Regardless, the consistent trends in relative abundance observed here and with other methods (Stenegren et al., 2017) suggest that the observed patterns are robust and that the level of polyploidy does not likely vary drastically in *Trichodesmium* populations across the WTSP.

In general, low *Trichodesmium* relative abundance at the easternmost stations (SD14 and 15) sampled along this transect were found despite high DIP concentrations detected in this region. This could be due to the low and homogeneous iron concentrations throughout the entire 0 to 500 m profile (0.1-0.3 nM DFe, Guieu et al., 2018) at these eastern stations, limiting growth. This is corroborated by the fact

that rates of water column N_2 fixation were lowest in the easternmost stations along this transect (Bock et al., 2018; Bonnet et al., 2018). Further, *Trichodesmium* distribution and overall rates of N_2 fixation in the water column were positively correlated with iron (Bonnet et al., 2018), the source of was from shallow hydrothermal vents west of the Tonga arc (~175°W, near SD11) (Guieu et al., 2018). *Trichodesmium* biomass was too low to evaluate gene expression patterns for the easternmost stations, but regardless, these data are consistent with the importance of iron concentration as a driver of *Trichodesmium* distribution and activities in the WTSP. In addition to iron, phosphorus also likely exerts a strong influence over *Trichodesmium* in the WTSP. The surface water concentration of DIP measured along this transect was low in the context of the global ocean (Sohm et al., 2011 and references therein) and measured phosphate turnover times in the water column at some stations on the order of hours indicated there was intense competition for phosphate (Van Mooy et al., 2009). The WTSP is poorly sampled relative to other oligotrophic ocean basins (Bonnet et al., 2018; Luo et al., 2012), and little is definitively known about the canonical resource controls that characterize this environment over prolonged time periods (Sohm et al., 2011). Herein, the *Trichodesmium* distributions were consistent with the potential roles of both iron and phosphorus in driving the physiological ecology of this genus in the WTSP.

4.2 Expression of metabolic potential in *Trichodesmium* and its microbiome

Trichodesmium does not exist in isolation, as filaments and colonies are associated with an assemblage of epibiotic microorganisms that co-occur ubiquitously in the environment (Lee et al., 2017; Rouco et al., 2016a) and contribute a large amount of metabolic potential that could underpin success in oligotrophic, low nutrient environments (Frischkorn et al., 2017). At the broadest functional level, the microbiome contained approximately 10 times the unique KEGG functions found within the *Trichodesmium* fraction of the WSTP, and these functions were largely consistent with those observed previously in the western North Atlantic (Frischkorn et al., 2017). The presence of unique microbiome transporter functions, especially those related to the transport of phosphate and metals including iron, reflect the importance of these resources within the colony microenvironment that is likely depleted in these key resources. The enrichment of microbiome functions related to the transport and subsequent

metabolism of sugars, carbohydrates and lipids could reflect the transfer of fixed carbon from host to microbiome, as the genes encoding these functions are known to oscillate over day night cycles in lockstep with *Trichodesmium* photosynthesis and carbon fixation genes (Frischkorn et al., 2018). These oscillations may support respiration processes that help maintain an environment favourable for N_2 fixation, and the functional enrichment observed here could underpin interactions within the holobiont that help maintain N_2 fixation in the WTSP, despite the scarcity of resources.

In addition to distinct functions in the metagenome, the expression of broad functional categories varied for *Trichodesmium* and the microbiome. Expression of OGs that belonged to carbohydrate metabolism and nucleotide and amino acid metabolism KEGG modules were elevated at the three SD stations sampled, relative to the three days of samples obtained from LDB. Arrival at this long duration station coincided with the decline of a phytoplankton bloom that had been at this location for approximately two months (De Verneil et al., 2017). Coincident with this decline in sea surface chlorophyll *a,* heterotrophic bacterial populations at LDB differed from other stations both taxonomically and in their response during experimental incubation with increased dissolved organic compounds from copepods (Valdés et al., 2018). In *Trichodesmium,* the decreased relative expression at LDB samples in functions related to carbon fixation, DNA and RNA replication, and a suite of amino acid metabolic functions that require nitrogen relative to the SD stations suggest shifts away from these energy intensive and cell-division processes with bloom demise. Similarly, in the microbiome, decreased relative expression of functional categories over the course of sampling at LDB could reflect shifts in physiology away from reliance on *Trichodesmium* or changes in community structure away from common colonizers of *Trichodesmium* to opportunistic or saprophytic species. For example, in the coral reef system, pulses of organic carbon similar to what could be released during a declining phytoplankton bloom, led to activity shifts in associated bacterioplankton including the increased expression of virulence factors (Cárdenas et al., 2018). Taken together, these data continue to reinforce that the microbiome both possesses and expresses unique metabolic potential relative to *Trichodesmium* alone, and as such, could play an important role in the physiological ecology of this important diazotroph.

4.3 Expression of resource-related signals in *Trichodesmium*

The expression of genes that lead to changes in activities like nitrogen fixation (*nifH*), or resource acquisition (e.g. *phoX*) can be used to assess the physiology of *Trichodesmium in situ*. Genes responsive to low phosphorus and iron conditions are particularly well-studied in *Trichodesmium* (e.g. Chappell et al., 2012; Chappell and Webb, 2010; Dyhrman et al., 2006; Orchard et al., 2009; Snow et al. 2015) and a recent culture study assessed the physiological response of *Trichodesmium* to coupled low phosphorus and low iron conditions, yielding a set of genes with significantly elevated expression under co-limitation conditions (Walworth et al., 2017). Expression of many of these marker genes is heavily repressed in cultures grown under replete conditions (Chappell and Webb, 2010; Orchard et al., 2009) or in field samples with relatively high concentrations of resources like iron (e.g. Chappell et al., 2012). Expression of these resource-responsive OG sets was detected in *Trichodesmium* across all samples, indicating that there was intense scavenging of phosphorus and iron, consistent with the low levels of these resources at stations analysed for gene expression.

The expression of P-responsive OGs related to phosphate uptake (*sphX*), phosphoester hydrolysis (*phoX*) and the metabolism of phosphite (*ptxD*) tracked together, with RuBisCO, and were significantly enriched at LDB 1. This pattern may indicate increased P stress at LDB 1 relative to other stations on the transect, and may be an indicator of shifts in *Trichodesmium* physiology associated with modulating carbon fixation and or the declining bloom in this region (De Verneil et al., 2017). For example, increased expression of *ptxD* may indicate increased metabolism of phosphite. Notably, the expression of this low P responsive set, which included OGs related to phosphate uptake (*sphX*), was significantly enriched at LDB 1 where the phosphate turnover time was among the lowest observed. Although *phnJ* has been shown to be regulated by phosphorus concentration in culture studies (Dyhrman et al., 2006), the *phnJ* OG here deviated from the expression pattern of the other P-responsive genes like *phoX*. As a result, there may be some variability in *Trichodesmium* processing of phosphoesters and phosphonates over these stations. Regardless, the expression of *phoX*, *ptxD*, and *phnJ* OGs underscores the

importance of organic phosphorus compounds, and phosphite in supporting *Trichodesmium* growth across the WTSP.

The OGs in the low Fe responsive set were also detected at all stations consistent with the sub-nanomolar concentrations of iron observed across the transect. This set was significantly enriched at SD2, where iron concentration was roughly half that of LDB. Although a larger dataset would be needed to fully resolve patterns of iron stress, these data are suggestive of an increase in *Trichdoesmium* iron stress at SD2 compared to the other stations. Given that rates of N_2 fixation in the water column were positively correlated with iron (Bonnet et al., 2018), and *Trichodesmium* was commonly a dominant diazotroph with the high cell-specific N_2 fixation activity (Stenegren et al., 2017; Bonnet et al., 2018), these patterns of gene expression are consistent with iron control of *Trichodesmium* N_2 fixation at SD2. Strikingly, expression of the alkaline phosphatase *phoA*, and the nitrogenase subunit *nifH* had similar expression patterns to this low Fe responsive set. Nitrogenase requires iron and its expression in *Trichodesmium* is tightly synchronized with iron processes (Frischkorn et al., 2018). Conversely, the enzyme PhoA does not require iron, instead using a zinc-magnesium cofactor, as opposed to PhoX which has an iron-calcium cofactor (Luo et al., 2009; Yong et al., 2014). In low iron environments, *Trichodesmium phoA* is known to show enriched expression relative to that of *phoX* (Rouco et al., 2018), a strategy that could free up iron for use in photosynthetic or nitrogen fixation enzymes. Collectively, these results suggests there is intense scavenging of iron by *Trichodesmium* in the WSTP, and in this and other environments where multiple resources can be low or co-limiting, co-factors like iron could play a role in the phosphorus acquisition strategies employed by *Trichodesmium*.

The OGs in the co-limitation responsive set were detected in all samples. The expression patterns of these co-limitation responsive OGs clustered among both the low P and low Fe sets, and there were no significant patterns of enrichment between stations. The lack of significant enrichment in the co-limitation set is consistent with the fact that the low P and low Fe responsive sets were not simultaneously enriched at the same station. Broadly however, the expression patterns between low P and co-limitation sets were more similar to each other than that shown by the low Fe responsive set,

suggesting that across these samples phosphorus was a driver of expression of co-limitation OGs. More field observations over a greater range in iron and phosphorus might further resolve these putative co-limitation signals, which have not been previously tracked in field populations. Collectively, the OGs expressed across this WTSP transect are suggestive of the importance of both iron and phosphorus in driving *Trichodsmium* physiological ecology in this region. These results provide physiological and metabolic context to the abundance and activity of *Trichodesmium* observed in surface waters across the OUTPACE transect. Together with studies of *Trichodesmium* in regions with contrasting geochemical characteristics like the North Atlantic and North Pacific (Rouco et al., 2018), these observations of *Trichodesmium* from the under-sampled WTSP enhance understanding of the constraints on this geochemically important diazotroph across diverse oligotrophic environments.

4.4 Phosphonate biosynthesis by *Trichodesmium* in the WTSP

Phosphate exists in vanishingly low concentrations in the oligotrophic surface ocean and the activity of diazotrophs increases the demand for phosphorus by relieving nitrogen stress—a process that is enhanced by periodic increases in iron availability (Moutin et al., 2005). In low phosphate environments, marine microbes, like *Trichodesmium,* can hydrolyse phosphate from organically bound compounds like phosphoesters and phosphonates, the concentration of which far surpasses phosphate in the oligotrophic ocean (Dyhrman et al., 2007). The production and hydrolysis of reduced compounds like phosphonates are of particular interest because the hydrolysis of methylphosphonate has the potential to release methane, a potent greenhouse gas (Karl et al., 2008; Repeta et al., 2016). Previous studies showing biosynthesis of phosphonates by certain *Trichodesmium* isolates (Dyhrman et al., 2009) as well as rapid phosphate reduction to phosphonate and phosphite and release by *Trichodesmium* colonies in the environment (Van Mooy et al., 2015) implicate this diazotroph as an important player in phosphonate biogeochemistry, yet the molecular mechanisms underlying phosphonate biosynthesis are poorly understood for this genus.

A *Trichodesmium* scaffold containing the full set of genes necessary to synthesize phosphonate compounds was recovered from metagenomes assembled from this WTSP transect. The *Trichodesmium* origin of this scaffold is supported by tetranucleotide frequency and metagenomic read mapping coverage, as well as the presence of a stretch of non-coding DNA with homology to the *T. erythraeum* IMS101 genome. Furthermore, the protein in this scaffold identified as phosphoenolpyruvate phosphomutase (Ppm), the enzyme that carries out the formation of a carbon-phosphorus bond using phosphoenolpyruvate as a substrate (McGrath et al., 2013), was phylogenetically most similar to Ppm sequences from cyanobacteria like *Planktothrix* that are closely related to *Trichodesmium*. The phylogenetic distance between the *Trichodesmium* Ppm and those of heterotrophic bacteria further support that this scaffold was recovered from a *Trichodesmium* genome and not from a member of the microbiome.

The molecular machinery necessary to synthesize phosphonates is evolutionarily conserved and the biosynthesis of phosphonoacetaldehyde is the starting point from which a diverse suite of organic phosphonate compounds can be produced (McGrath et al., 2013). Based on the genes in this *Trichodesmium* scaffold, synthesis begins with the formation of the carbon-phosphorus bond after molecular rearrangement of phosphoenolpyruvate to phosphonopyruvate, catalysed by Ppm. Next, phosphonopyruvate decarboxylase (Ppd), the protein encoded by the following gene in the cassette, likely performs the irreversible conversion of phosphonopyruvate to phosphonoacetaldehyde which prevents reversion to the ester bond structure. Finally, the presence of the gene for 2-aminoethylphosphonic acid pyruvate-transaminase (*2-AEP-TA*) suggests that phosphonoacetaldehyde is further converted to 2-aminoethylphosphonate (2-AEP), the organophosphonate that occurs most commonly in the environment (McGrath et al., 2013). The *mpnS* gene mediates the production of methylphosphonate down stream of *ppm* in the marine microbes where it has been detected (Metcalf et al., 2012). There was no evidence of *mpnS* in *Trichodesmium* or the microbiome, but *Trichodesmium*-derived phosphonates could potentially be further modified to methylphosphonate by organisms not associated with colonies.

The *ppm* gene can be found in approximately 7% of microbial genome equivalents recovered from the Global Ocean Survey, and of these *ppm*-containing genomes, 20.6% are estimated to be cyanobacterial in origin (Yu et al., 2013). A protein with homology to Ppm was previously detected and attributed to *Trichodesmium* in metagenomic samples from the western North Atlantic (Frischkorn et al., 2017), though this gene was not found to be part of a 2-AEP synthesis cassette. OGs derived from genes on the *ppm*-containing *Trichodesmium* scaffold identified here recruited metatranscriptomic reads from each sample sequenced, suggesting active use of these enzymes across the transect. Furthermore, reduced phosphonate compounds (which would include 2-AEP) were produced from radiolabeled phosphate taken up by *Trichodesmium* colonies at each station analysed. Together, these results clearly illustrate a pathway by which *Trichodesmium* synthesizes phosphonates, and that this pathway is active in the WTSP and likely other environments like the western North Atlantic, where high rates of phosphate reduction in *Trichodesmium* colonies has also been measured (Van Mooy et al., 2015).

The relative expression of the *ppm* OG across the transect was significantly correlated with the abundance of *Trichodesmium* Clade III. Although this is consistent with the detection of phosphonates in cultured *T. erythraeum* isolates from Clade III (Dyhrman et al., 2009), no clear homologs of *ppm* have been found in the *T. erythraeum* IMS101 genome assembly. At this juncture, this may be attributable to an incomplete genome, or the pathway could be mediated with a different mechanism, and more work would be required to screen *Trichodesmium* isolates for *ppm*. Regardless, these field data are consistent with culture studies that suggest that phosphonate biosynthesis may be variably present in different *Trichodesmium* clades or species. Across this WTSP transect, Clade III accounted for up to 47.5% of the total *Trichodesmium* population in some locations, and in general composed nearly 20% of the population on average from station to station. This percentage of Clade III is considerably higher than that observed in the western North Atlantic (Rouco et al., 2014) and higher than in the North Pacific subtropical gyre (Rouco et al., 2016a). If phosphonate biosynthesis is consistently a unique feature of Clade III, then *Trichodesmium*-mediated phosphonate production could be higher in the South Pacific than other ocean basins. Furthermore, models project that future ocean conditions will select for the dominance of *T. erythraeum* (Clade III) over that of Clade I species like *T.*

thiebautii which are most abundant in the present ocean (Hutchins et al., 2013). Such an expansion of *T. erythraeum* could lead to a subsequent increase in phosphonate biosynthesis and an increase in the availability of these compounds in the water column.

The production of a recalcitrant form of phosphorus and its potential release into the oligotrophic environment could have important consequences for microbial communities. Phosphonates are a critical source of phosphorus in the oligotrophic ocean, and the ability to utilize this resource could influence microbial ecology in low nutrient environments. Across this transect, both *Trichodesmium* and their microbiome contained and expressed OGs related to phosphonate catabolism, including the marker of the C-P lyase enzyme complex, *phnJ*. *Trichodesmium* and microbiome *phnJ* genes have also been detected and expressed in *Trichodesmium* communities from the chronically low phosphate western North Atlantic ocean as well as the North Pacific subtropical gyre (Dyhrman et al., 2006; Frischkorn et al., 2017, 2018). In addition to the production and hydrolysis of phosphonate compounds, we also detected evidence of the use of other forms of reduced phosphorus. The expression of the *ptxD* gene which is responsible for the oxidation of phosphite, another reduced phosphorus compound (Polyviou et al., 2015), was also expressed by *Trichodesmium* and the microbiome at all stations sampled. This finding suggests active transformation and exchange of reduced phosphorus compounds between consortia members. In the low phosphorus western North Atlantic, up to 16% of the phosphate taken up by *Trichodesmium* colonies has been shown to be reduced and subsequently released from cells, an amount of phosphorus cycling that rivals the amount input to marine systems from allochthonous or atmospheric sources (Van Mooy et al., 2015). The evidence of utilization of these traits across additional geochemical environments, like in the WTSP, and the large quantities of phosphorus they recycle suggests that phosphonate cycling composes an integral facet of the *Trichodesmium* holobiont's physiology, yet the reasons behind this cycling remain enigmatic. Not all microbes can metabolize phosphonates (Villarreal-Chiu, 2012), therefore it could also be plausible that formation of such compounds creates a cryptic phosphorus pool that would in part restrict access to this critical nutrient by other microbes. Uptake and reduction to a more recalcitrant form as a mechanism of luxury storage exclusive to members of the *Trichodesmium* holobiont is supported by the significant positive

correlation between expression of *ppm* and DIP concentration. In short, exchange of these compounds within the *Trichodesmium* holobiont, especially if through a cryptic pool, could help support N_2 fixation in *Trichodesmium* by modulating access to bioavailable phosphorus in the oligotrophic WTSP (Van Mooy et al., 2015).

Marine N_2 fixation is expected to increase in future oceans that are predicted to have higher temperatures and CO_2 concentration (Hutchins et al., 2007), and *Trichodesmium* cultures incubated in high CO_2 conditions exhibited irreversibly increased rates of N_2 fixation (Hutchins et al., 2015). In such conditions, cycling of phosphonate compounds that are not accessible to the full microbial community could support enhanced N_2 fixation if enough iron is available, the release of new nitrogen into the water column and subsequently fuel primary production. In future studies it will be important to assay how future ocean conditions will alter the clade distribution of *Trichodesmium* in the environment, as this could play a role in determining the potential flux of phosphonate compounds from colonies to the water column.

5 Conclusion

Marine microbes interact and alter the environment through abiotic transformations as well as through biotic interactions with one another and across trophic levels, and these processes work in tandem to influence global biogeochemical cycles. Understanding these processes *in situ* is of paramount importance to forecasting the ocean's role in the future climate, yet challenges persist with sampling remote locations and filling knowledge gaps surrounding the ecology and physiology of key species. The OUTPACE research expedition afforded a unique opportunity to investigate communities of the keystone N_2 fixer *Trichodesmium* and their microbiome in the under-sampled South Pacific. Metagenomic and metatranscriptomic data showed a majority of unique physiological functions within the microbiome, many of which were expressed *in situ*, and these functions may be important to *Trichodesmium* physiological ecology in this environment. Patterns of OG expression in low Fe responsive, low P responsive, and co-limitation sets suggested that iron and phosphorus are highly scavenged and that *Trichodesmium* variably experienced changes in these resources, which could

modulate growth and N_2 fixation *in situ*. A *Trichodesmium* gene cassette for the biosynthesis of the phosphonates, its expression, and corresponding phosphate reduction rate measurements suggested that *Trichodesmium* is producing reduced phosphate in the WTSP. This finding expands the environments where phosphate reduction has been detected, and confirms the role of *Trichodesmium* in this poorly understood aspect of phosphorus biogeochemistry. Collectively, these data underscore the importance of iron and phosphorus, and the microbiome, in jointly driving the physiological ecology of this key diazotroph in the WTSP.

Competing interests

The authors declare they have no conflicts of interest.

Acknowledgments

This is a contribution to the OUTPACE (Oligotrophy to UlTra-oligotrophy PACific Experiment) project which was funded by the French national research agency (ANR-14-CE01-0007-01), the LEFE-CyBER program (CNRS- 8 INSU), the GOPS program (IRD) and the CNES (BC T23, ZBC 4500048836). The authors thank Thierry Moutin and Sophie Bonnet for leading the OUTPACE expedition and for the invitation to participate. The authors also thank the OUTPACE science party and the captain and crew of the R/V *L'Atalante* for their assistance and camaraderie while underway. Special thanks to Andrea Caputo and Marcus Stenegren of Stockholm University for assistance while at sea. The authors also acknowledge the Carbonate cluster of the National Center for Genome Analysis Support, and Carrie Ganote in partiulcar, for bioinformatics assistance. Grants from the National Science Foundation to STD (OCE-1332912) and BASVM (OCE-1536346 and OCE-1332898) supported this research. KRF is partially supported by a National Science Foundation Graduate Research Fellowship (DGE-16-44869). This research was also funded by the Simons Foundation's Simons Collaboration on Ocean Processes and Ecology (SCOPE) (SCOPE award ID 329108 to STD and BVM), and is a contribution of SCOPE.

References

Bergman, B., Sandh, G., Lin, S., Larsson, J. and Carpenter, E. J.: *Trichodesmium*--a widespread marine cyanobacterium with unusual nitrogen fixation properties., *FEMS Microbiol. Rev.*, 37(3), 286–302, doi:10.1111/j.1574-6976.2012.00352.x, 2013

Blain S., Bonnet S., & Guieu C., Dissolved iron distribution in the tropical and sub tropical South Eastern Pacific, *Biogeosciences*, **5**, 269-280 (2008).

Bock, N., Van Wambeke, F., Dion, M. and Duhamel, S.: Microbial community structure along trophic and N_2 fixation gradients in the Western Tropical South Pacific, *Biogeosciences*, 1–24, 2018.

Bonnet, S., Caffin, M., Berthelot, H. and Moutin, T.: Hot spot of N_2 fixation in the western tropical South Pacific pleads for a spatial decoupling between N_2 fixation and denitrification, *Proc. Natl. Acad. Sci.*, 114(14), E2800–E2801, doi:10.1073/pnas.1619514114, 2017.

Bonnet, S., Caffin, M., Berthelot, H., Grosso, O., Benavides, M., Helias-Nunige, S., Guieu, C., Stenegren, M. and Foster, R. A.: In depth characterization of diazotroph activity across the Western Tropical South Pacific hot spot of N_2 fixation, *Biogeosciences*, 1–30, doi:10.5194/bg-2017-567, 2018.

Brown CT, Sheneman LS, Camille CM, Rosenthal JH, Adina C. khmer-protocols documentation. 2013. https://khmer-protocols. readthedocs.io/en/latest/mrnaseq/index.html 19.

Buchfink, B., Xie, C. and Huson, D. H.: Fast and sensitive protein alignment using DIAMOND, *Nat. Methods*, 12, 59–60, doi:10.1038/nmeth.3176, 2015.

Caffin, M., Moutin, T., Foster, R. A., Bouruet-Aubertot, P., Doglioli, A. M., Berthelot, H., Grosso, O., Helias-Nunige, S., Leblond, N., Gimenez, A., Petrenko, A. A., de Verneil, A. and Bonnet, S.: Nitrogen budgets following a Lagrangian strategy in the Western Tropical South Pacific Ocean: the prominent role of NN_2 fixation (OUTPACE cruise), *Biogeosciences Discuss*., 1–34, doi:10.5194/bg-2017-468, 2017.

Caffin, M., Berthelot, H., Cornet-Barthaux, V. and Bonnet, S.: Transfer of diazotroph-derived nitrogen to the planktonic food web across gradients of N_2 fixation activity and diversity in the Western Tropical South Pacific, *Biogeosciences*, 2018.

Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B. and Carpenter, E. J.: *Trichodesmium*, a globally significant marine cyanobacterium, *Science*, 276, 1221–1229, 1997.

Cárdenas, A., Neave, M. J., Haroon, M. F., Pogoreutz, C., Rädecker, N., Wild, C., et al. (2018). Excess labile carbon promotes the expression of virulence factors in coral reef bacterioplankton. *ISME J.* 12, 59–76. doi:10.1038/ismej.2017.142.

Chappell, P. D. and Webb, E. A.: A molecular assessment of the iron stress response in the two phylogenetic clades of *Trichodesmium*, *Environ. Microbiol.*, 12, 13–27, doi:10.1111/j.1462-2920.2009.02026.x, 2010.

Chappell, P. D., Moffett, J. W., Hynes, A. M. and Webb, E. A.: Molecular evidence of iron limitation and availability in the global diazotroph *Trichodesmium*, *ISME J.*, 6, 1728–39, doi:10.1038/ismej.2012.13, 2012.

Chen, C. C. H., Han, Y., Niu, W., Kulakova, A. N., Howard, A., Quinn, J. P., Dunaway-Mariano, D. and Herzberg, O.: Structure and kinetics of phosphonopyruvate hydrolase from *Voriovorax* sp. Pal2: New insight into the divergence of catalysis within the pep mutase/isocitrate lyase superfamily, *Biochemistry*, 45, 11491–11504, doi:10.1021/bi061208l, 2006.

Clark, L. L., Ingall, E. D. and Benner, R.: Marine phosphorus is selectively remineralized, *Nature*, 393, 426, 1998.

Dyhrman, S. T., Chappell, P. D., Haley, S. T., Moffett, J. W., Orchard, E. D., Waterbury, J. B. and Webb, E. A.: Phosphonate utilization by the globally important marine diazotroph *Trichodesmium*, *Nature*, 439, 68–71, 2006.

Dyhrman, S. T., Ammerman, J. W., Van Mooy, B. A. S.: Microbes and the marine phosphorus cycle, *Oceanography,* 20, 110-116, 2007.

Dyhrman, S. T., Benitez-Nelson, C. R., Orchard, E. D., Haley, S. T. and Pellechia, P. J.: A microbial source of phosphonates in oligotrophic marine systems, *Nat. Geosci*., 2, 696–699, doi:10.1038/ngeo639, 2009.

Eddy, S. R.: Multiple alignment using hidden Markov models., *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 3, 114–120, doi:7584426, 1995.

Edgar, R. C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, 32, 1792–1797, 2004.

Frischkorn, K. R., Rouco, M., Van Mooy, B. and Dyhrman, S. T.: Epibionts dominate metabolic functional potential of *Trichodesmium* colonies from the oligotrophic ocean, *ISME J.,* 11, 2090-2101, doi:10.1038/ismej.2017.74, 2017.

Frischkorn, K. R., Haley, S. T. and Dyhrman, S. T.: Coordinated gene expression between *Trichodesmium* and its microbiome over day–night cycles in the North Pacific Subtropical Gyre, *ISME J.*, 12, 997-1007, doi:10.1038/s41396-017-0041-5, 2018.

Guieu, C., Bonnet, S., Petrenko, A., Menkes, C., Chavagnac, V., Desboeufs, K., Maes C. & Moutin, T. (2018). Iron from a submarine source impacts the productive layer of the Western Tropical South Pacific (WTSP). Scientific reports, 8(1), 9075.

Hmelo, L., Van Mooy, B. and Mincer, T.: Characterization of bacterial epibionts on the cyanobacterium *Trichodesmium*, *Aquat. Microb. Ecol.*, 67, 1–14, doi:10.3354/ame01571, 2012.

Huson, D. H., Hadidi, M. El, Ruscheweyh, H., Huson, D. H. and Meta, S.: Improved metagenome analysis using MEGAN5, 2013. http://ab. inf.uni-tuebingen.de/software/megan5/ 26.

Hutchins, D. A., Fu, F.-X., Zhang, Y., Warner, M. E., Feng, Y., Portune, K., Bernhardt, P. W. and Mulholland, M. R.: CO_2 control of *Trichodesmium* N_2 fixation, photosynthesis, growth rates, and elemental ratios: Implications for past, present, and future ocean biogeochemistry, *Limnol. Oceanogr.*, 52, 1293–1304, doi:10.4319/lo.2007.52.4.1293, 2007.

Hutchins, D. A., Fu, F.-X., Webb, E. A., Walworth, N. and Tagliabue, A.: Taxon-specific response of marine nitrogen fixers to elevated carbon dioxide concentrations, *Nat. Geosci.*, 6, 1–6, 2013.

Hutchins, D. A., Walworth, N. G., Webb, E. A., Saito, M. a., Moran, D., McIlvin, M. R., Gale, J. and Fu, F.-X.: Irreversibly increased nitrogen fixation in *Trichodesmium* experimentally adapted to elevated carbon dioxide, *Nat. Commun.*, 6, 8155, doi:10.1038/ncomms9155, 2015.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W. and Hauser, L. J.: Prodigal: prokaryotic gene recognition and translation initiation site identification., *BMC Bioinformatics*, 11, doi:10.1186/1471-2105-11-119, 2010.

Karl, D. M., Beversdorf, L., Björkman, K. M., Church, M. J., Martinez, A. and Delong, E. F.: Aerobic production of methane in the sea, *Nat. Geosci*., 1, 473–478, doi:10.1038/ngeo234, 2008.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. and Drummond, A.: Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data, *Bioinformatics*, 28, 1647–1649, doi:10.1093/bioinformatics/bts199, 2012.

Lee, M. D., Webb, E. A., Walworth, N. G., Fu, F., Held, N. A., Saito, M. A., and Hutchins, D. A.: Transcriptional activities of the microbial consortium living with the marine nitrogen-fixing cyanobacterium *Trichodesmium* reveal potential roles in community-level nitrogen cycling, *Appl. Environ. Microbiol*., 84, 1–16, 2018.

Lee, M. D., Walworth, N. G., Mcparland, E. L., Fu, F., Mincer, T. J., Levine, N. M., Hutchins, D. A. and Webb, E. A.: The *Trichodesmium* consortium : conserved heterotrophic co-occurrence and genomic signatures of potential interactions, 11, 1813–1824, doi:10.1038/ismej.2017.49, 2017.

Letunic, I. and Bork, P.: Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees, *Nucleic Acids Res*., 44, W242–W245, doi:10.1093/nar/gkw290, 2016.

Li, B. and Dewey, C. N.: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome., *BMC Bioinformatics*, 12, doi:10.1186/1471-2105-12-323, 2011.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.: The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, 25, 2078–2079, doi:10.1093/bioinformatics/btp352, 2009.

Luo, H., Benner, R., Long, R. A., and Hu, J. (2009). Subcellular localization of marine bacterial alkaline phosphatases. *Proc. Natl. Acad. Sci.* 106, 21219–21223. doi:10.1073/pnas.0907586106.

Luo, Y.-W., Doney, S. C., Anderson, L. A., Benavides, M., Berman-Frank, I., et al.: Database of diazotrophs in the global ocean: abundance, biomass and nitrogen fixation rates, *Earth Syst. Sci. Data*, 4, 47–73, doi:10.5194/essd-4-47-

2012, 2012.

McGrath, J. W., Chin, J. P. and Quinn, J. P.: Organophosphonates revealed: new insights into the microbial metabolism of ancient molecules., *Nat. Rev. Microbiol*., 11, 412–9, doi:10.1038/nrmicro3011, 2013.

Metcalf, W. W. and van der Donk, W. A.: Biosynthesis of phosphonic and phosphinic acid natural products, *Annu Rev Biochem*, 78, 65–94, doi:10.1146/annurev.biochem.78.091707.100215.Biosynthesis, 2009.

Metcalf, W. W., Griffin, B. M., Cicchillo, R. M., Gao, J., Janga, S. C., Cooke, H. A., Circello, B. T., Evans, B. S., Martens-Habbena, W., Stahl, D. A. and Donk, W. A. van der: Synthesis of methylphosphonic acid by marine microbes: A source for methane in the aerobic ocean, *Science*, 337, 1104–1107, 2012.

Moore, C. M., Mills, M. M., Arrigo, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., Galbraith, E. D., Geider, R. J., Guieu, C., Jaccard, S. L., Jickells, T. D., La Roche, J., Lenton, T. M., Mahowald, N. M., Marañón, E., Marinov, I., Moore, J. K., Nakatsuka, T., Oschlies, A., Saito, M. A., Thingstad, T. F., Tsuda, A. and Ulloa, O.: Processes and patterns of oceanic nutrient limitation, *Nat. Geosci*., 6, 701–710, doi:10.1038/ngeo1765, 2013.

Moutin, T., Van Den Broeck, N., Beker, B., Dupouy, C., Rimmelin, P. and Le Bouteiller, A.: Phosphate availability controls *Trichodesmium* spp. biomass in the SW Pacific Ocean, *Mar. Ecol. Prog. Ser*., 297, 15–21, doi:10.3354/meps297015, 2005.

Moutin, T., Karl, D. M., Duhamel, S., Rimmelin, P., Raimbault, P., Mooy, B. A. S. Van, Claustre, H., Luminy, C. De and Hole, W.: Phosphate availability and the ultimate control of new nitrogen input by nitrogen fixation in the tropical Pacific Ocean, *Biogeosciences*, 5, 95–109, 2008.

Moutin, T., Michelangelo Doglioli, A., De Verneil, A. and Bonnet, S.: Preface: The Oligotrophy to the UlTra-oligotrophy PACific Experiment (OUTPACE cruise, 18 February to 3 April 2015), *Biogeosciences*, 14, 3207–3220, doi:10.5194/bg-14-3207-2017, 2017.

Orchard, E. D., Webb, E. A and Dyhrman, S. T.: Molecular analysis of the phosphorus starvation response in *Trichodesmium* spp., *Environ. Microbiol*., 11, 2400–11, doi:10.1111/j.1462-2920.2009.01968.x, 2009.

Paerl, H. W. and Bebout, B. M.: Direct measurement of O_2-depleted microzones in marine *Oscillatoria*: Relation to N_2 fixation, *Science*, 241, 442–445, 1988.

Paerl, H. W., Bebout, B. M. and Prufert, L. E.: Bacterial associations with marine *Oscillatoria* sp. (*Trichodesmium* sp) populations: ecophysiological implications, *J. Phycol*, 25, 773–784, 1989.

Peng, Y., Leung, H. C. M., Yiu, S. M. and Chin, F. Y. L.: IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth., *Bioinformatics*, 28, 1420–8, doi:10.1093/bioinformatics/bts174, 2012.

Polyviou, D., Hitchcock, A., Baylay, A. J., Moore, C. M. and Bibby, T. S.: Phosphite utilisation by the globally important marine diazotroph *Trichodesmium*, *Environ. Microbiol. Rep*., 7, 827-830, doi:10.1111/1758-2229.12308, 2015.

Polyviou, D., Baylay, A. J., Hitchcock, A., Robidart, J., Moore, C. M. and Bibby, T. S.: Desert dust as a source of iron to the globally important diazotroph Trichodesmium, *Front. Microbiol*., 8, 1–12, doi:10.3389/fmicb.2017.02683, 2018.

Price, M. N., Dehal, P. S. and Arkin, A. P.: FastTree 2 - Approximately maximum-likelihood trees for large alignments, *PLoS One*, 5, doi:10.1371/journal.pone.0009490, 2010.

Pulido-Villena, E., Rérolle, V., Guieu, C.: Transient fertilizing effect of dust in P-deficient LNLC surface ocean. *Geophys. Res. Letters,* 5, doi:10.1029/2009GL041415, 2010.

Repeta, D. J., Ferrón, S., Sosa, O. A., Johnson, C. G., Repeta, L. D., Acker, M., Delong, E. F. and Karl, D. M.: Marine methane paradox explained by bacterial degradation of dissolved organic matter, *Nat. Geosci*., 9, doi:10.1038/ngeo2837, 2016.

Rouco, M., Joy-Warren, H., McGillicuddy, D. J., Waterbury, J. . J. B. and Dyhrman, S. T.: *Trichodesmium* sp. clade distributions in the western North Atlantic Ocean, *Limnol. Oceanogr.*, 59, 1899–1909, doi:10.4319/lo.2014.59.6.1899, 2014.

Rouco, M., Haley, S. T. and Dyhrman, S. T.: Microbial diversity within the *Trichodesmium* holobiont, *Environ. Microbiol*., 18, 5151–5160, doi:10.1111/1462-2920.13513, 2016.

Rouco, M., Frischkorn, K. R., Haley, S. T., Alexander, H. and Dyhrman, S. T.: Transcriptional patterns identify resource controls on the diazotroph *Trichodesmium* in the Atlantic and Pacific oceans, *ISME J*., in press, doi:10.1038/s41396-018-0087-z, 2018.

Sañudo-Wilhelmy, S. A., Kustka, A. B., Gobler, C. J., Hutchins, D. A., Yang, M., Lwiza, K., Burns, J., Capone, D. G.,

Raven, J. A. and Carpenter, E. J.: Phosphorus limitation of nitrogen fixation by *Trichodesmium* in the central Atlantic Ocean., *Nature*, 411, 66–9, doi:10.1038/35075041, 2001.

Sargent, E. C., Hitchcock, A., Johansson, S. A., Langlois, R., Moore, C. M., LaRoche, J., et al. (2016). Evidence for polyploidy in the globally important diazotroph Trichodesmium. *FEMS Microbiol. Lett.* 363, 1–7. doi:10.1093/femsle/fnw244.

Snow, J. T., Polyviou, D., Skipp, P., Chrismas, N. A. M., Hitchcock, A., Geider, R., Moore, C. M. and Bibby, T. S.: Quantifying integrated proteomic responses to iron stress in the globally important marine diazotroph *Trichodesmium*, *PLoS One*, 10, e0142626, doi:10.1371/journal.pone.0142626, 2015.

Sohm, J. A., Webb, E. A. and Capone, D. G.: Emerging patterns of marine nitrogen fixation., *Nat. Rev. Microbiol.*, 9, 499–508, doi:10.1038/nrmicro2594, 2011.

Stenegren, M., Caputo, A., Berg, C., Bonnet, S. and Foster, R. A.: Distribution and drivers of symbiotic and free-living diazotrophic cyanobacteria in the Western Tropical South Pacific, *Biogeosciences Discuss*., 1–47, doi:10.5194/bg-2017-63, 2017.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. and Wu, C. H.: UniRef: comprehensive and non-redundant UniProt reference clusters, *Bioinformatics*, 23, 1282–1288, doi:10.1093/bioinformatics/btm098, 2007.

De Verneil, A., Rousselet, L., Doglioli, A. M., Petrenko, A. A. and Moutin, T.: The fate of a southwest Pacific bloom: Gauging the impact of submesoscale vs. mesoscale circulation on biological gradients in the subtropics, *Biogeosciences*, 14, 3471–3486, doi:10.5194/bg-14-3471-2017, 2017.

Van Mooy, B. A. S., Fredricks, H. F., Pedler, B. E., Dyhrman, S. T., Karl, D. M., Koblížek, M., Lomas, M. W., Mincer, T. J., Moore, L. R., Moutin, T., Rappé, M. S. and Webb, E. A.: Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity, *Nature*, 457(7234), 69–72, 2009.

Van Mooy, B. A. S., Krupke, A., Dyhrman, S. T., Fredricks, H. F., Frischkorn, K. R., Ossolinski, J. E., Repeta, D. J., Rouco, M., Seewald, J. D. and Sylva, S. P.: Major role of planktonic phosphate reduction in the marine phosphorus redox cycle, *Science*, 348, 783–785, 2015.

Villarreal-Chiu, J. F.: The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine

environment, *Front. Microbiol.*, 3, 1–13, doi:10.3389/fmicb.2012.00019, 2012.

Walworth, N., Pfreundt, U., Nelson, W. C., Mincer, T., Heidelberg, J. F., Fu, F., Waterbury, J. B., Glavina del Rio, T., Goodwin, L., Kyrpides, N. C., Land, M. L., Woyke, T., Hutchins, D. A., Hess, W. R. and Webb, E. A.: *Trichodesmium* genome maintains abundant, widespread noncoding DNA in situ, despite oligotrophic lifestyle, *Proc. Natl. Acad. Sci.*, 201422332, doi:10.1073/pnas.1422332112, 2015.

Walworth, N. G., Fu, F.-X., Lee, M. D., Cai, X., Saito, M. A., Webb, E. A. and Hutchins, D. A.: Nutrient co-limited *Trichodesmium* as nitrogen source or sink in a future ocean., *Appl. Environ. Microbiol.*, 84, AEM.02137-17, doi:10.1128/AEM.02137-17, 2017.

Wu, Y., Simmons, B. A. and Singer, S. W.: MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets, *Bioinformatics*, 1–2, doi:10.1093/bioinformatics/btv638, 2015.

Yong, S. C., Roversi, P., Lillington, J., Rodriguez, F., Krehenbrink, M., Zeldin, O. B., et al. (2014). A complex iron-calcium cofactor catalyzing phosphotransfer chemistry. *Science.* 345, 1170–1173. doi:10.1126/science.1254237.

Yu, X., Doroghazi, J. R., Janga, S. C., Zhang, J. K., Circello, B., Griffin, B. M., Labeda, D. P. and Metcalf, W. W.: Diversity and abundance of phosphonate biosynthetic genes in nature, *Proc. Natl. Acad. Sci.*, 110, 20759–20764, doi:10.1073/pnas.1315107110, 2013.

Figures

Figure 1. Surface water column dissolved inorganic phosphorus (DIP) concentration, community phosphate uptake rate, community phosphate turnover, and total dissolved iron concentration measured at stations across the OUTPACE (Oligotrophy to UlTra-oligotrophy PACific Experiment) transect during austral summer (February – April) of 2015. All samples were obtained from 10 m depth, with the exception of the DIP measurement from LDA and the DFe measurement from SD12 (30 m). Numbers above and below indicate the short duration (denoted SD) or long duration (denoted LDA, LDB or LDC) stations where samples were obtained across the transect. In the top panel, * indicates stations where metagenomic or metatranscriptomic samples were obtained and † indicates stations where *Trichodesmium* clade distribution samples were obtained.

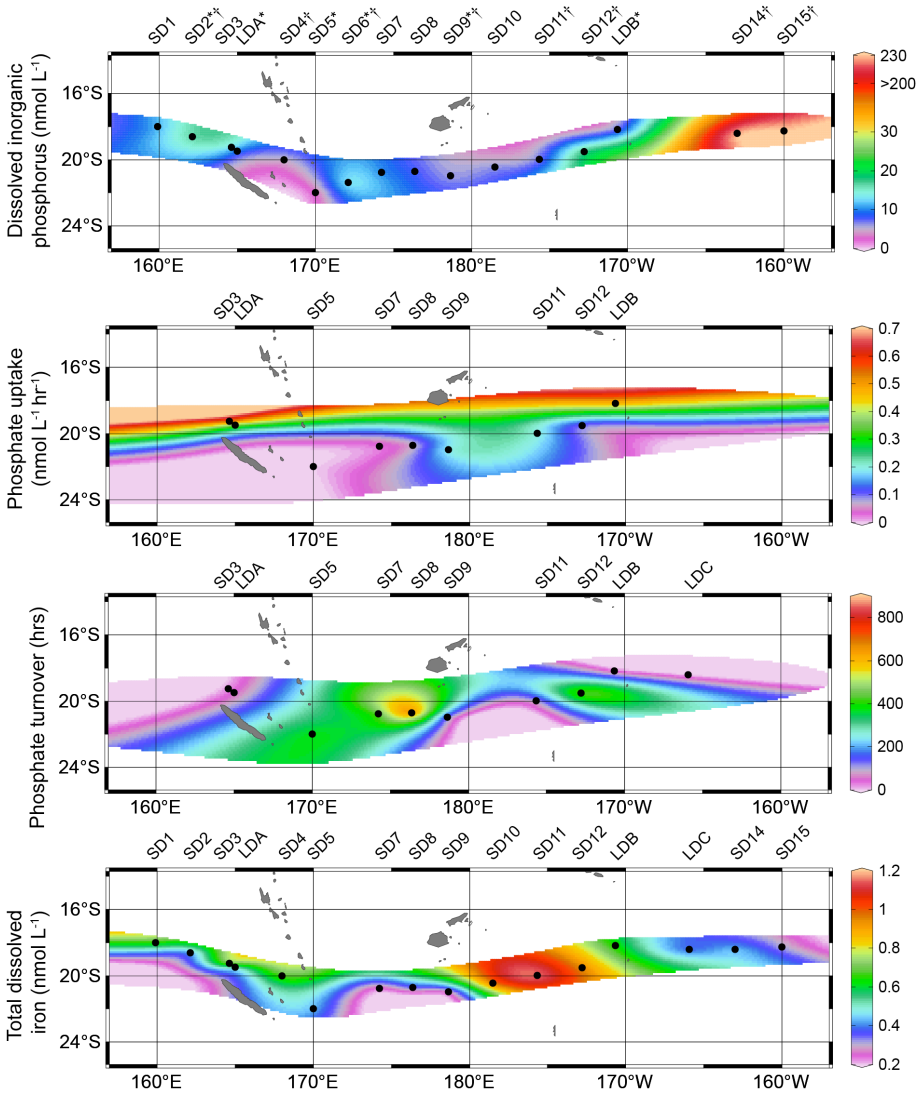


Figure 2. Abundance and clade distribution of *Trichodesmium.* Black dots denote the depths at which samples were taken, while station number is indicated above the panels. (a) Concentration of total *Trichodesmium* cells estimated from qPCR of the rnpB gene (cells L^{-1}). (b) Relative proportion of *Trichodesmium* Clade I (top panel) and Clade III (bottom panel) across the transect. Clade I includes *T. thiebautii, T. tenue, T. hildebrandtii,* and *T. spiralis*. Clade III includes *T. erythraeum* and *T. contortum*. * indicates stations where metatranscriptomic sequences were sampled. Clade samples were not obtained at LD stations.

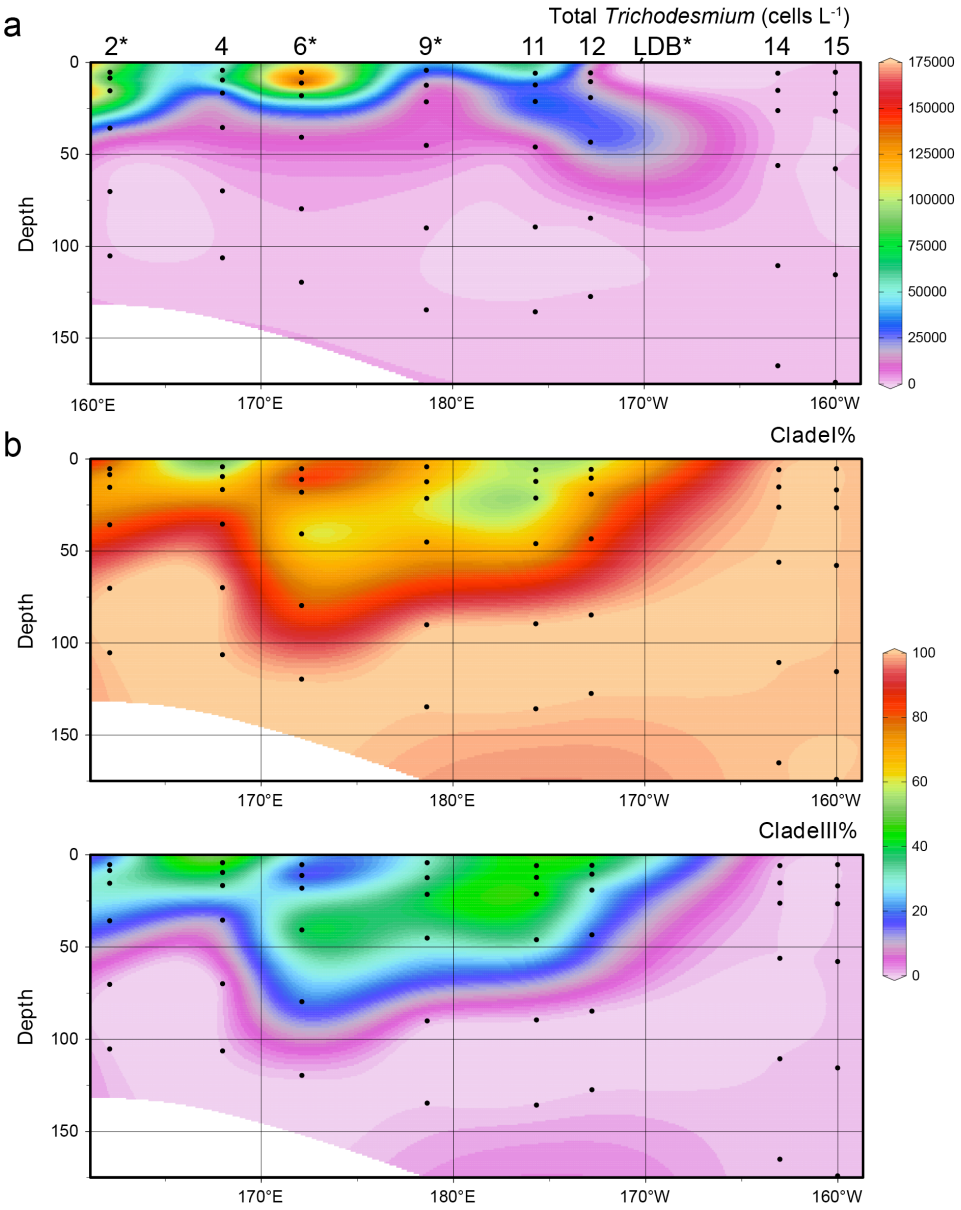


Figure 3. Heatmap of the summed relative expression of *Trichodesmium* and microbiome orthologous groups (OGs) belonging to KEGG modules (large categories) and submodules at the short duration (SD) and over three days at the long duration (LDB) stations. Relative expression values are normalized to row averages for either *Trichodesmium* or the microbiome for that particular OG. The carbohydrate and lipid metabolism module was simplified as follows: carbohydrate metabolism is the sum of carbohydrate metabolism, central carbohydrate metabolism, glycan metabolism and other carbohydrate metabolism submodules; lipid metabolism is the sum of fatty acid and lipid metabolism submodules, The genetic information processing module was simplified as follows: RNA processes is the sum of the RNA processing and RNA polymerase submodules No microbiome OGs were detected in the other amino acid metabolism submodule.

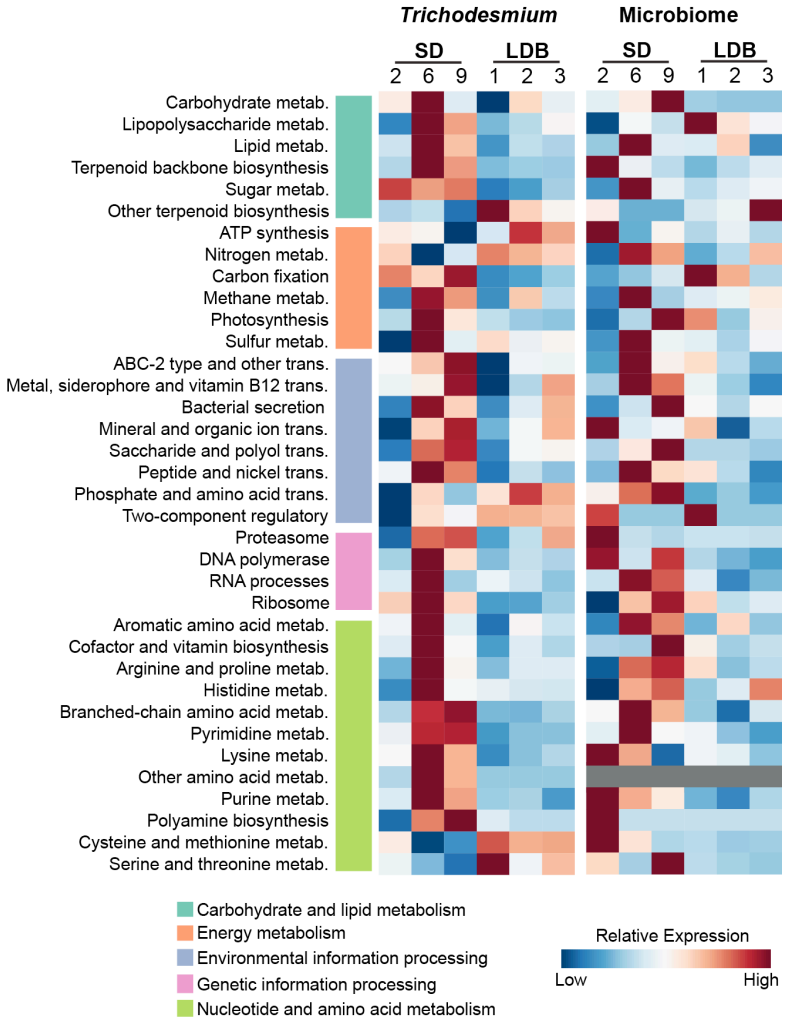


Figure 4. *Trichodesmium* orthologous group (OG) expression patterns. (a) Hierarchically clustered expression patterns of *Trichodesmium* OGs with key genes called out. * Indicates OGs used in subsequent Kolmogorov-Smirnov tests. Colored text indicates phosphorus (green), iron (brown), co-limitation (blue), nitrogen (purple), or carbon fixation (black) related OGs. (b) Distribution of expression patterns in OG sets known to be significantly responsive in *Trichodesmium* to low phosphorus (P), low iron (Fe), and P/Fe co-limiting conditions. Distributions for each set at each station were compared to the average distribution across all six stations using Kolmogorov-Smirnov tests to examine the null hypothesis that the expression of gene sets at a given station did not deviate significantly from the average expression of that set across the transect. Whiskers show the normalized enrichment level for the least and most enriched OG in that set. Boxes denote the upper and lower 25th percentiles, while the line indicates the median enrichment expression value. Asterisks indicate significance: * $p < 0.05$, ** $p < 0.005$. Black triangles denote whether significant stations were increased or decreased relative to the mean.

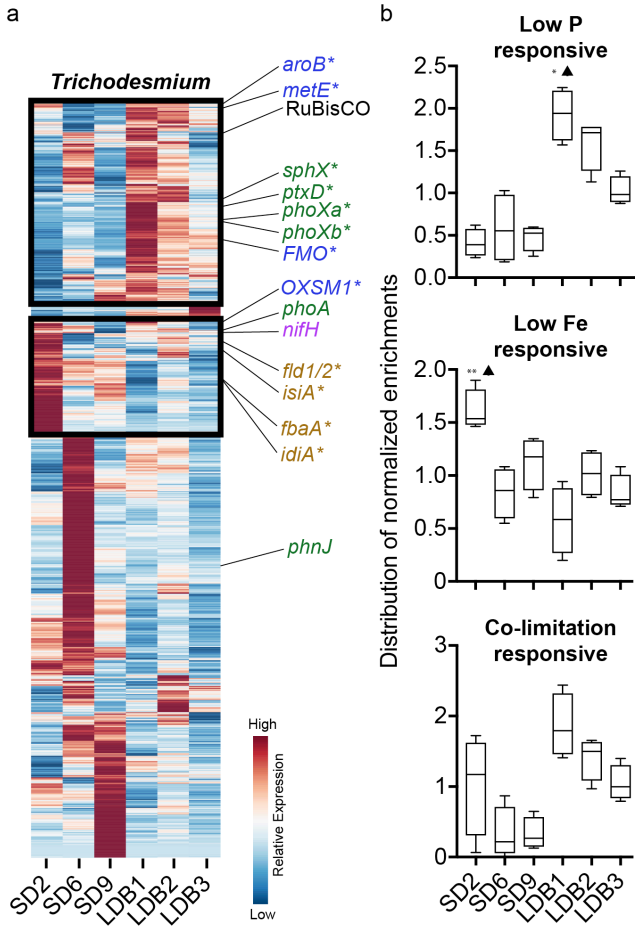
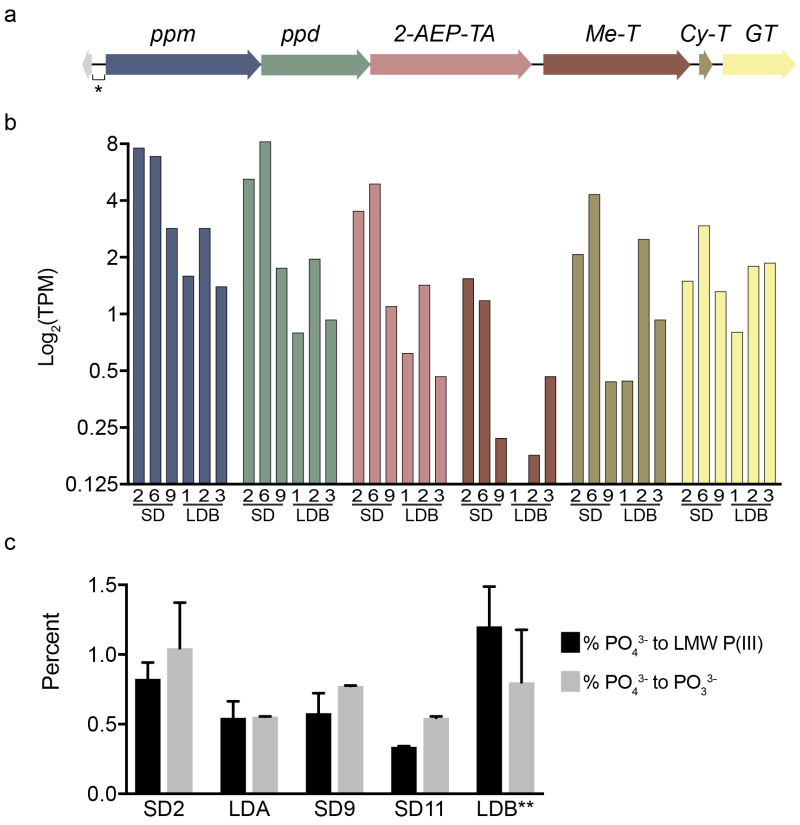


Figure 5. Annotations, topology and expression levels of a scaffold containing a phosphonate biosynthesis cassette recovered from a *Trichodesmium* identified metagenome bin. (a) Gene organization and annotations across the scaffold where arrows represent direction of transcription. * denotes a region with 95% homology to a non-coding region of the *T. erythraeum* IMS101 genome. (b) Normalized expression (transcripts per million, TPM) of each OG on this *ppm*-containing scaffold at each short duration (SD) or long duration (LDB) station sampled. Column colours match those of the top panel (a). (c) Percentage of total radiolabeled phosphate taken up by *Trichodesmium* colonies and reduced into low molecular weight (LMW) phosphonate compounds (including methylphosphonate, phosphonoacetylaldehyde, and 2-aminoethylphosphonate) and phosphite. Error bars denote standard deviation. Abbreviations: *ppm,* phosphoenolpyruvate phosphomutase; *ppd,* phosphonopyruvate decarboxylase; *2-AEP-TA,* 2-aminoethylphosphonate-pyruvate transaminase; *Me-T,* SAM-dependent methyltransferase; *Cy-T,* cytidylyltransferase; *GT,* group 1 glycosyltransferase. **LDB values represent an average of three measurements obtained during sampling at this long duration station.



Supplemental /tables

Supplemental Table 1. Summary of water column (10 m) geochemical data collected across the OUTPACE cruise transect. *Measurement obtained at 30 m.

Station	Lat	Lon	Water colum dissolved inorganic phosphorus (nmol/L) at 10 m		Water column phosphate turnover (1/hr)		Phosphate turnover time (hrs)	Water column phosphate uptake rate (nmol/L/hr)	Water column iron (nmol/L) at 10 m	*Tricho.* colony: % of PO4 uptake as MPn		*Tricho.* colony: % PO4 uptake as PO3		Sum MPn PO3 (%)
			avg.	stdev	avg	stdev				avg	stdev	avg	stdev	
SD1	-18.0	159.9	11.2	0.5376					0.67					
SD2	-18.6	162.1	16.4	0.7872					0.36	0.83	0.12	1.05	0.33	1.87
SD3	-19.5	165.0	8.3	0.3984	0.0251	0.0013	39.7986	0.209	0.38					
LDA	-19.2	164.6	13.8*	0.6624	0.0494	0.0031	20.2525	0.681	0.85	0.55	0.12	0.55	0.00	1.10
SD4	-20.0	168.0	4.8	0.2304					0.69					
SD5	-22.0	170.0	2.3	0.1104	0.0026	0.0004	383.4898	0.006	0.44					
SD6	-21.4	172.1	13.4	0.6432										
SD7	-20.8	174.3	10.21	0.49008	0.0029	0.0000	340.9558	0.030	0.21					
SD8	-20.7	176.4	8.2	0.3936	0.0012	0.0009	827.7212	0.010	0.38					
SD9	-21.0	178.6	6.1	0.2928	0.0389	0.0004	25.7207	0.237	0.22	0.58	0.14	0.77	0.005	1.35
SD10	-20.5	-178.5	6.6	0.3168					0.97					
SD11	-20.0	-175.7	6.6	0.3168	0.0366	0.0002	27.3563	0.241	1.16	0.34	0.00	0.55	0.01	0.88
SD12	-19.5	-172.8	18.7	0.8976	0.0019	0.0008	519.0585	0.036	0.94*					
LDB	-18.2	-170.7	7	0.336	0.0694	0.0044	14.4041	0.486	0.65	1.20	0.28	0.80	0.38	2.00
SD13	-18.2	-169.1												
LDC	-18.4	-165.9			0.5041	0.2785	1.9839		0.37					
SD14	-18.4	-163.0	214	10.272					0.46					
SD15	-18.3	-160.0	230	11.04					0.31					

Supplemental Table 2. qPCR counts of *Trichodesmium* Clade I and Clade III obtained across the transect.

	Clade I (VI-I type)			Clade III (IMS-101 type)		
Station #	Depth (m)	Cells/Liter	SD	Depth (m)	Cells/Liter	SD
2.1	105.33	49.22	3.54	105.33	0.00	0.00
2.2	70.32	256.52	46.19	70.32	0.00	0.00
2.3	35.88	16935.72	3178.65	35.88	4241.04	187.77
2.4	15.45	78277.16	11013.84	15.45	34467.80	2706.92
2.5	8.53	40360.93	6311.34	8.53	17495.42	1212.01
2.6	5.45	78681.67	1277.21	5.45	21452.43	2785.86
4.1	106.42	63.98	8.21	106.42	0.00	0.00
4.2	69.98	1222.95	285.07	69.98	0.00	0.00
4.3	35.46	12498.82	1218.82	35.46	1457.31	183.75
4.4	16.73	11120.73	1051.08	16.73	5596.86	161.16
4.5	9.64	44795.69	8852.79	9.64	28478.09	4533.35
4.6	4.30	24450.06	4167.23	4.30	22152.05	1674.11
6.1	119.68	581.66	154.77	119.68	0.00	0.00
6.2	79.68	1456.09	99.28	79.68	391.01	49.60
6.3	40.70	6462.53	376.48	40.70	4515.22	99.35
6.4	18.17	40658.71	8811.30	18.17	13284.30	1015.35
6.5	11.23	149154.78	10230.38	11.23	22935.19	3787.89
6.6	5.38	82856.35	13163.61	5.38	23110.81	6456.73
9.1	134.66	108.49	15.83	134.66	0.00	0.00
9.2	90.08	44.12	10.07	90.08	0.00	0.00
9.3	45.18	5761.77	1220.23	45.18	2483.23	194.07
9.4	21.53	4628.82	1255.19	21.53	3010.94	171.89
9.5	12.52	6162.76	345.63	12.52	3067.23	136.50
9.6	4.331	20587.22	717.25	4.331	9164.47	2635.38
11.1	135.83	14.17	4.03	135.83	0.00	0.00
11.2	89.59	72.95	4.20	89.59	0.00	0.00
11.3	46.07	2152.23	383.28	46.07	1153.82	55.61
11.4	21.39	20241.08	1360.48	21.39	18049.39	1381.87
11.5	12.35	12213.64	770.56	12.35	8402.08	1165.03
11.6	5.84	36343.33	1861.80	5.84	28713.60	5077.47
12.1	127.48	13.34	4.36	127.48	0.00	0.00

12.2	84.81	114.12	7.08	84.81	0.00	0.00
12.3	43.39	24045.90	1552.06	43.39	4784.42	327.22
12.4	19.16	15179.65	2300.00	19.16	5259.23	47.69
12.5	10.49	10852.23	2110.81	10.49	6513.01	1419.06
12.6	5.75	2990.12	366.52	5.75	2011.18	386.35
14.1	165.17	56.86	0.00	165.17	0.00	0.00
14.2	110.71	107.76	23.75	110.71	0.00	0.00
14.3	56.22	89.37	11.96	56.22	0.00	0.00
14.4	26.26	70.62	8.46	26.26	0.00	0.00
14.5	15.30	86.11	33.82	15.30	0.00	0.00
14.6	6.01	72.03	12.37	6.01	0.00	0.00
15.1	174.16	14.62	9.37	174.16	0.00	0.00
15.2	115.72	15.80	3.85	115.72	0.00	0.00
15.3	58.03	13.93	0.11	58.03	0.00	0.00
15.4	26.58	65.20	11.48	26.58	0.00	0.00
15.5	16.81	73.24	9.42	16.81	0.00	0.00
15.6	5.44	67.07	1.25	5.44	0.00	0.00

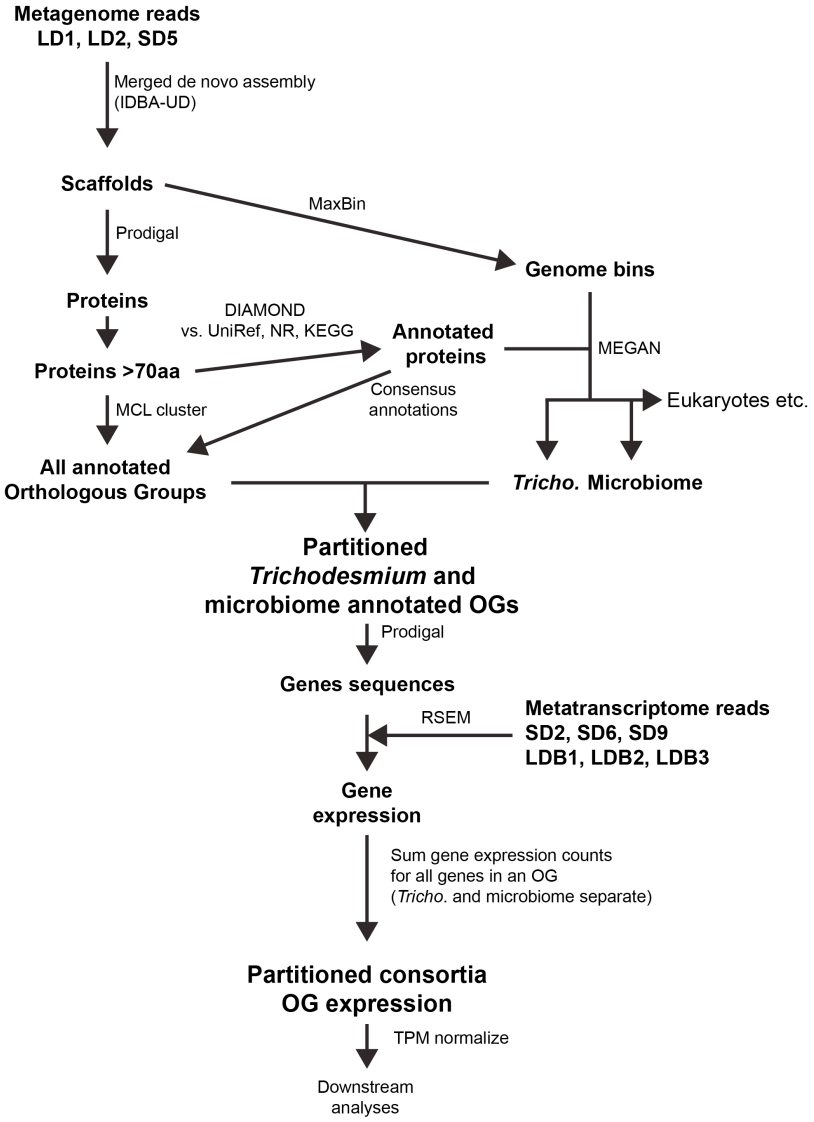
Supplemental Table 3. Table of gene expression counts (tags per million, TPM) summed across orthologous groups (OGs) and partitioned between the *Trichodesmium* and microbiome fractions.

Supplemental Table 4. The sequence of the assembled scaffold containing the *ppm* cassette from the *Trichodesmium* metagenome assembly. Note: full sequence will be included at re-submission.

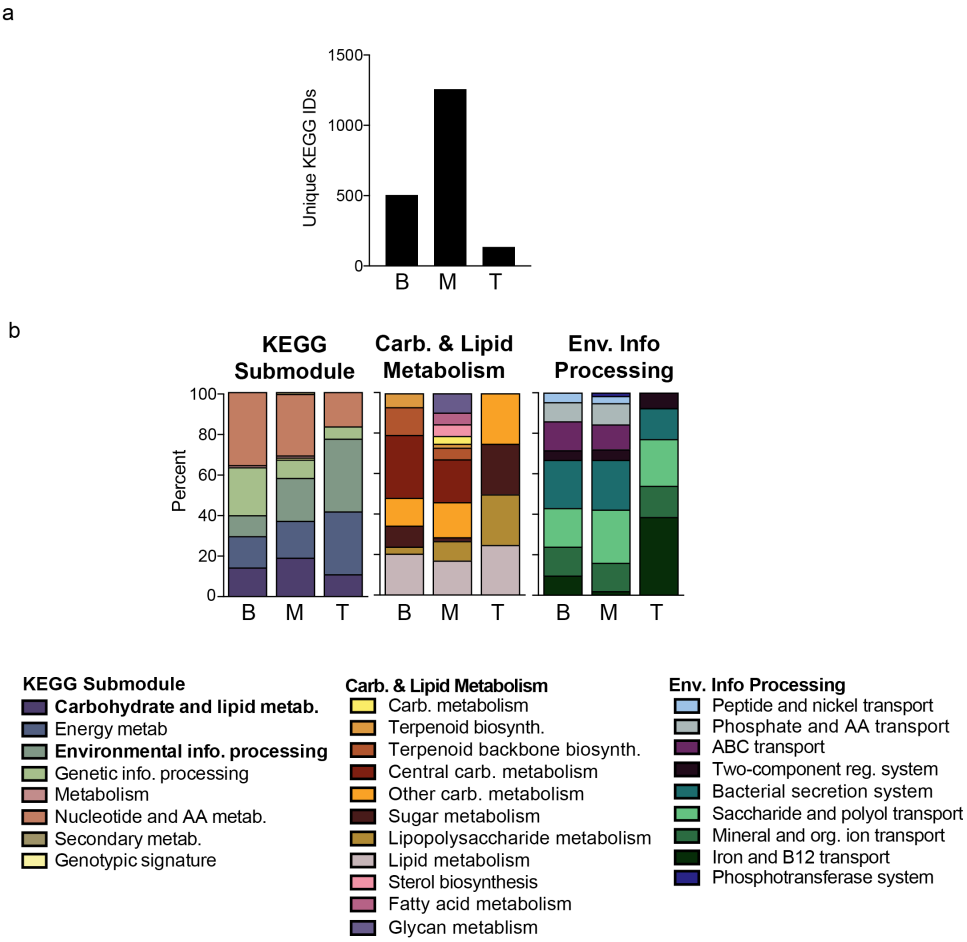
>scaffold_3391
ATTTTGTAACCCTTTTTTCTGCTATCCCTTTGCTCCATCTCCTGACTCGTGACTCCTGTACCTCGCCCTA
CTGACGGCTCCCCTACCCGATCAAAAAACTTAATAGTGGCTATTGATACCAAATGGGTTCTATAGAGTTG
GAAGTTGTTGAATGAATTTTTGGTAGTCCATTTTGGTCTGTTGGTTCTTTGGGGTTATTGTAGAGGTTGG
TTGAGTGGGGGGCGATCGTTTTTGTGGGGAAATGGGGAATAATTAGAGGTGTTAAAGTTTTCTACTCCCT
TCCCGGAGTAATACATTCAGTATAATTATAATTATGTCCGTTAACAAGATAAGTTTATGCTCAGAAAAAC
AACCCAATTGAAGAACTTACTGCAGTCAGAAAAATTGGAATTTTTGATGGAAGCTCATAACGGCTTGAGT
GCAAAAATTGTAGAAGAAGCAGGCTTTAAGGGCATTTGGGGGAGCGGTCTTTCAATTTCGGCAGCAATGG
GTGTCAGGGATAATAATGAAGCTAGCTGGACTCAGATCTTAGAAATTCTGGAGTTTATGTCTGATGCTAC
CTCAATTCCTATTTTACTAGATGGTGACACGGGTTATGGTAACTTTAATAACTTACGACGTTTGGTAAAA
AAACTAGAGCAGCGTGGTGTGGCAGGAGTTTGTATTGAAGATAAACAGTTTCCGAAAAAAAAATAGTTTTA
TTGACGGTCATACTCAACCACTAGCTGATATAGAAGAGTTTTGTGGCAAAATCAAAGCAGCAAAAGATGC
TCAAAAGGATAATGATTTTGTAGTAATTTCTCGTGTAGAAGCTTTTATTGCCGGTTGGGGTTTGTCAGAA
GCACTAAAACGGGCAGAAGCATATTATCATTCCGGAGCTGACGGAATTTTGATTCACAGTTCTCTAAGAG
TTCCAAATGAAATTTTGGCATTTAAACAAGATTGGGGCGATCGTTGTCCTGTAATTATTGTGCCGACTAA
GTATTATACTACTCCCACTCAAGTATTTAGAGATTACAAGTTCTCAATTGCTATTTGGGCCAATCAAACT
CTAAGGGCGGCGATCACAGAAATGCAGAAAATAGTTAATAAATTATCGCGACAAGAAAATTTACTGGATG
CAGAAGAGTCAATTATTCCTGTTTCTGAGGTTTTTAGACTTCAAGGTTGCCTTGAGTTACAAGAAGCAGA
AAAACTTTATTTACCCCAAAATAACAAAAAAATAGCAAGTTTATTGTTAGCAGCATCTAGAGGTATAGAA
TTAGGAACTTTTACTAAAGAAAAACCTAAATGTATGTTGACTTTAAAAAGAAAGCCGATTTTAGGACAAA
TAATAGCTATTCATCATGAAATAGGAATAAAAAATATTAGTGTTGTTCGCGGTTATAAAAAAGAAGCTAT
TAATTTAGCCAACATTAAATATGTAGATAACGATGAATATGAGTCAACGGGGGAAATATTTTCCCTCTAT
TTAGGCTTGAAAAAAATCATTCAAGATAATGAAAATCAAGACCTAGTAATTGGCTATGGTGATGTGTTAT
TGAAAAAGTATATACTGCAACTCCTCTTAGAAAATAATTATGACTTGGCAGTAATAGTTGATAGCAACTG
GCATAAATGCTCGAATCAAGTTAGGGCAGATTATACTAATTGCTCTGTACCTAATTCTAAACGAGCATTT
TATCAAGATGTTGGCTTACACCAAATAGACGCAAATATGCCAGAAGAAAAAATTTGTGGAATGTGGACTG
GTTTACTGAAAGTTTCAAGTAACGTGCAAAGACCATTATTAGATGCTCTTGAAAAGTTACTTCTACAGCT
GGAAATCAAAAGTAGCGGTAGAATGCCAAACTTGATAAATGAGTTAATTGCTTTAGGATATCCAGTTAAT
GTGGTTTACATTACAGGAGATTGGTTAGATGTAGATGAAGTAGAAGATATGATCAAGGCGGGAACTTTTT
AAATGATTAAAGCAGAAAACTTTTTTGAAGTTGCAGGCAAACTAGGATTTGGCTTATATACAGGGGTTCC
CTGTTCTTACCTGAAGTCTTTTATAAACTATGTAATTGATTCTCCTGACTTGCGGTATGTAGGTGCTACA
AATGAGGGAAATGCGGTAGCTATCGCCTCTGGTGCAGAATTAGCCGGAGTTAGAAGTGTGGTTATGTTGC
AGAATTCTGGTTTGGGTAATGGGGTTAACCCGCTAACATCTCTGAATCAAACCTTTAAAATACCAATTTT

Supplemental Figures

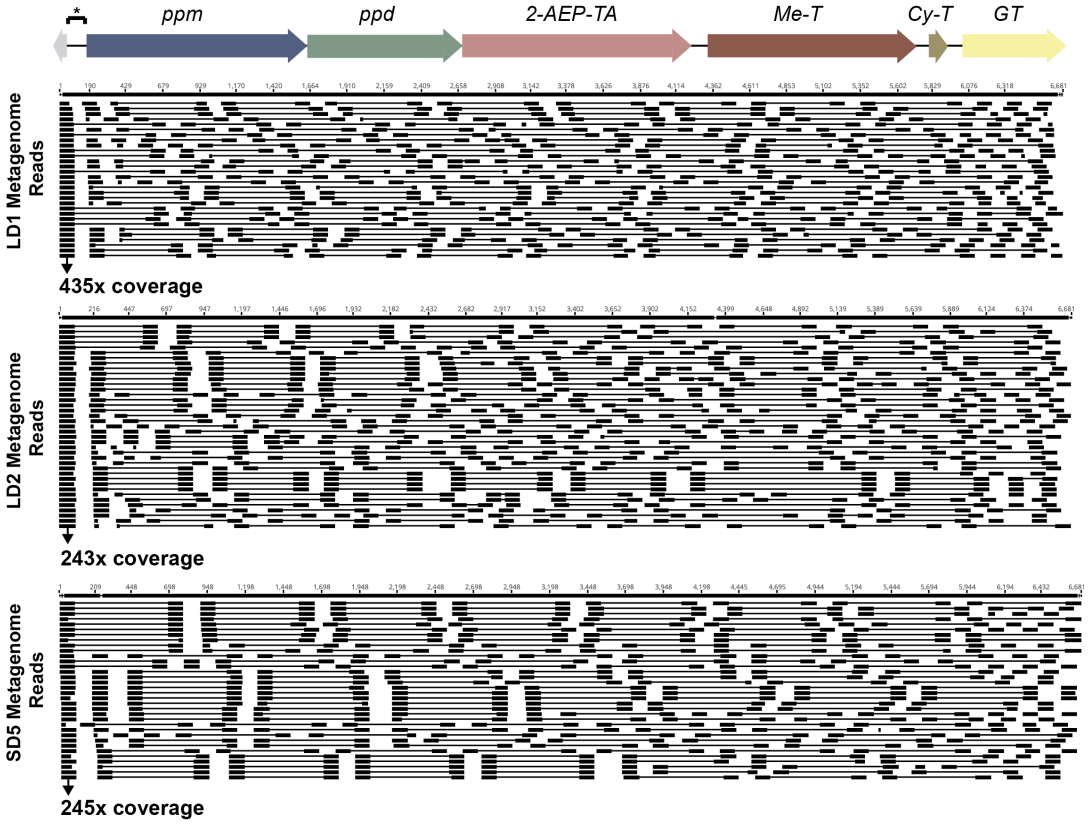
Supplemental Figure 1. Diagram detailing approach used to assemble and partition the metagenome between *Trichodesmium* and the microbiome, annotate and cluster protein coding sequences into orthologous groups (OGs), and obtain gene expression values from metatranscriptomes. Eukaryote identified sequences as well as phototrophs other than *Trichodesmium* were excluded from downstream analysis after the genome binning and analysis step. TPM, transcripts per million.



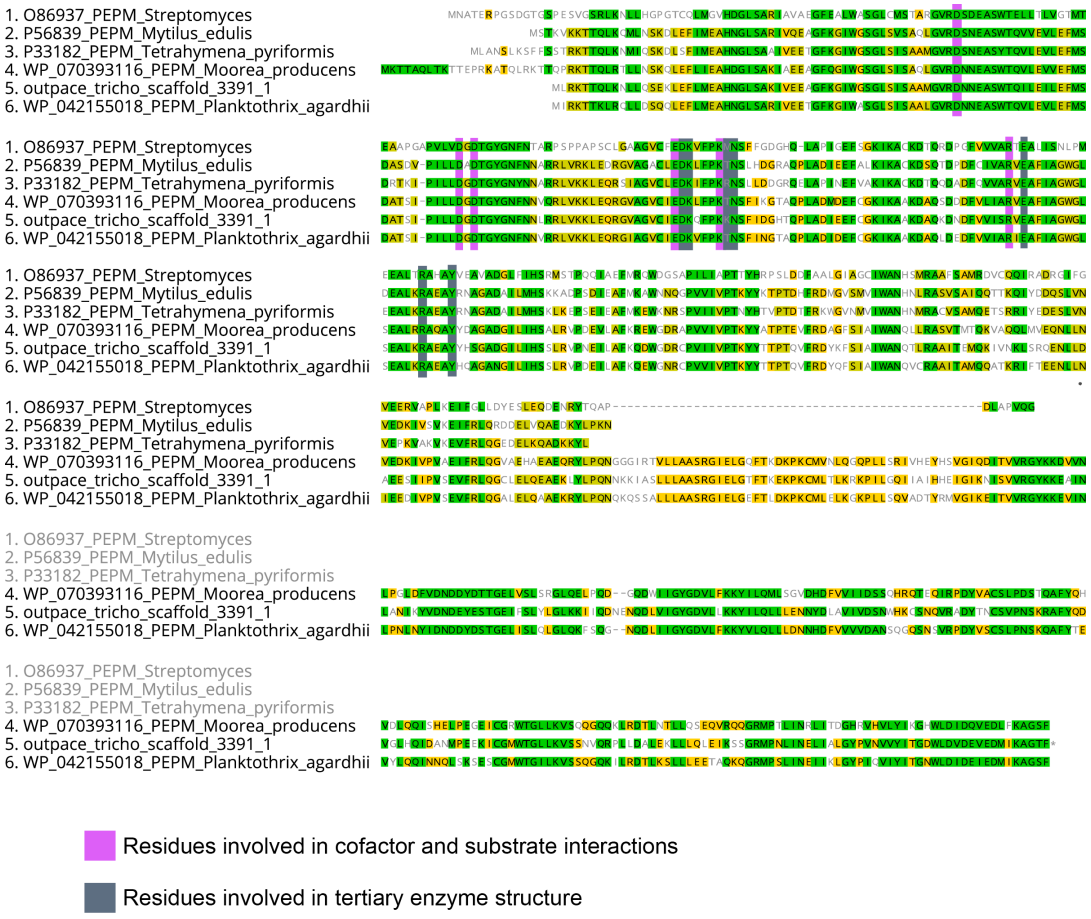
Supplemental Figure 2. Distribution and annotations of KEGG functional annotations of OGs found uniquely in the *Trichodesmium* (T) or microbiome (M) or those composed of both *Trichodesmium* and microbiome proteins (B). (a) Total number of OGs in each category. (b) A functional breakdown of these annotations at the KEGG module level and detailed annotations from within two KEGG module categories.



Supplemental Figure 3. Metagenome reads aligned to the scaffold containing the *ppm* cassette. Thick black rectangles depict ~100 bp reads, connected by thin black lines to their mate paired read.



Supplemental Figure 4. Amino acid alignment of the Ppm protein recovered from a *Trichodesmium* metagenome bin against experimentally verified Ppm sequences in other organisms. Genes highlighted in green denote regions with 100% amino acid identity across all sequences. Shades of yellow denote conservation across the majority of amino acids in the column. Un-highlighted amino acids indicate divergent residues or regions with little conservation. Purple and grey highlighted columns denote residues that were previously determined to be important to the structure or activity of this enzyme (Chen et al., 2006).



Supplemental Figure 5. Phylogenetic tree showing the placement of the *Trichodesmium* Ppm protein (red), a microbiome proteins from this study's metagenome assembly that is similar to Ppm but lacking conservation at key residues (green), as well as homologous proteins from a previously assembled North Atlantic *Trichodesmium* metagenome assembly (blue) (Frischkorn et al., 2017), along with homologous sequences obtained from the NCBI nr database (black). The tree was generated with FastTree using the default settings (Price et al., 2010). Numbers at the branch labels indicate FastTree support percentages for the sequences in that branch.