Referee comment

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

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.

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.

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.

***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).

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).

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?

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.

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.

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: (qP**RC**) should be (qP**CR**)

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?

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.