We thank the referees for their time reviewing the manuscript and for providing thoughtful comments. In the main comments (RC1 1-2 and the first paragraph of RC2) the referees ask about (1) the impacts of parameters to workflow outputs (ASVs, abundances) and diazotroph community composition, and (2) how the workflow and its outputs compare to previously published pipelines and results, in particular from the published 19 studies. These topics were addressed with two new analyses that now appear in Appendices B and C:

B. Read trimming method effects on workflow outputs (pages 33-36)

C. Comparison of communities from the workflow to previous studies (pages 36-38)

We reply to specific comments in RC1 and RC2 below but first respond to (1) and (2).

Referee text is black; our responses, blue; and manuscript text, red. All line numbers, figures, and tables are for the revised manuscript unless indicated with the subscript *orig* in which case they refer to the ESSD preprint. (We do not reference line numbers in Highlight_changes_in_revised_essd-2024-163.pdf.)

(1) Workflow parameters

The quality of the raw sequencing data is critical to the ASVs and abundances output by the workflow. The DADA2 *nifH* pipeline exposes key parameters related to data quality, especially truncLen and truncQ which specify two fundamentally different ways to trim low quality bases off the tails of the reads (Table 2; approaches described in lines 263-265 and in more detail in lines 750-762 in Appendix B). Moreover, read trimming affects downstream steps including the discarding of low quality reads (maxEE in Table 2), denoising, and merging (minOverlap and maxMismatch in Table 2). Therefore, the trimming method could substantially impact workflow outputs. Appendix B (pages 33-36) investigated this by running the workflow using truncLen and truncQ on the 21 studies. For most studies using either approach resulted in very similar diazotroph communities when resolved to *nifH* clusters (Fig. B1), a classification often used in the 19 previously published studies.

However, using truncLen retained substantially more of the data (Table B1, and Fig. R1 below) which noticeably shifted *nifH* clusters for three studies (BentzonTilia_2015, Hallstrom_2022, and Shiozaki_2020; Fig. B1 and lines 779-788 in Appendix B). Therefore, a new version of the *nifH* ASV database is now available where every study was run through the workflow using truncLen. (The original, submitted database used truncQ for all but five studies [lines 426-431$_{orig}$,Table 2$_{orig}$]). For each study in the new database, the script estimateTrimLengths.R was used to determine lengths that would maximize the percentages of reads retained by DADA2 (lines 266-272). The new *nifH* ASV database includes more samples (944 vs. 865), reads (43.0 vs. 34.4 million), and ASVs (9383 vs. 7909) than the original. However, these gains did not substantially impact the patterns described in the analysis of the DNA dataset (section 3.3 starting at line 569). This is not surprising because the total reads from dominant *nifH* clusters in each sample were usually similar using either trimming approach (Fig. R1).

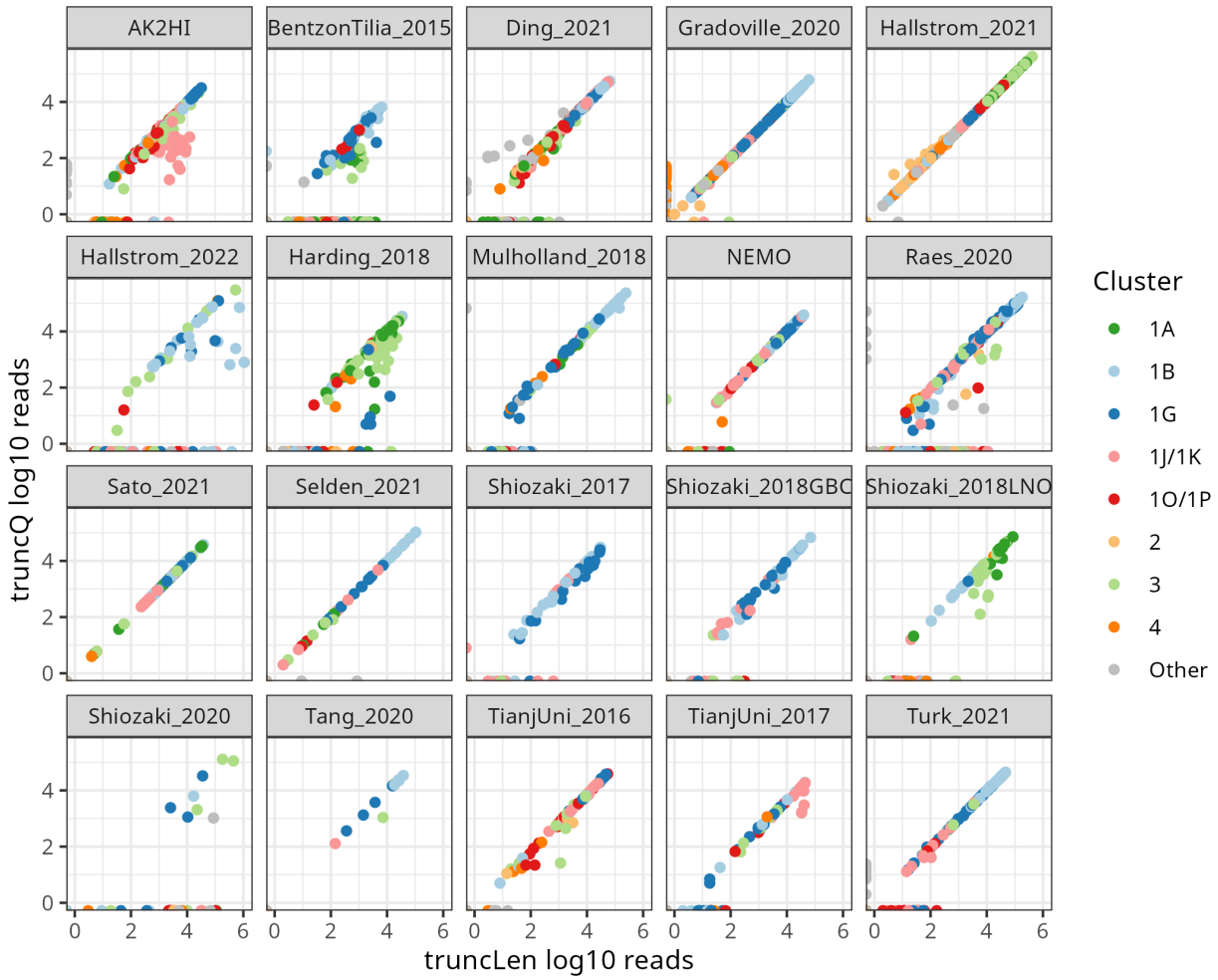

Figure R1. *nifH* clusters in each sample in the original (truncQ) versus new (truncLen) *nifH* ASV database. Each point represents a *nifH* cluster in one sample. Clusters for most samples are near the line y=x which indicates similar abundances using either trimming approach. Using truncLen sometimes resulted in higher abundances (x>y) or the detection of less abundant clusters that were missed using truncQ (x>0 and y=0). In the original database, Gradoville_2020, Hallstrom_2021, Sato_2021, and Selden_2021 were processed using truncLen.

(2) Comparison to previous pipelines and published results

Our workflow (the DADA2 *nifH* pipeline and post-pipeline stages) is the only software specifically designed to process Illumina-sequenced *nifH* amplicon data into ASVs and then perform rigorous quality checks, annotation, and colocalization with environmental data. For the 19 previously published studies, custom scripts were not available for comparison, and methods did not fully describe parameters needed to approximate the original results. Furthermore, the standard practice is to upload only the raw sequencing data to the NCBI Short Read Archive, and in most studies the representative sequences and abundance tables were not provided.

These factors limited options for comparing to previous pipelines or results. Since 13 of the 19 published studies were OTU analyses (Table C1), running all studies though an OTU pipeline was a reasonable alternative, and provided a more systematic approach given the original studies used a variety of tools. The NifMAP OTU pipeline (Angel et al., 2018) was selected because it shares some features with the workflow. Both use CART (Frank et al., 2016) for *nifH* cluster annotation, and both distinguish *nifH* from orthologs that are often mistaken for *nifH* (lines 815-823 in Appendix C). For most studies the workflow and NifMAP identified the same dominant *nifH* clusters at similar relative abundances (Fig. B1; lines 825-828 in Appendix C). For two studies the workflow detected additional clusters missed by NifMAP likely because the workflow retained much higher percentages of the reads (Fig. B1; lines 830-844 in Appendix C). Far fewer ASVs were detected than OTUs consistent with known limitations of OTUs (Table C1 and lines 850-852 in Appendix C) that have shifted researchers to ASV-based analysis.

Please note that:

- Three studies had name corrections: TianjUni_2016 is now Wu_2021; TianjUni_2017 is now Wu_2019; and Turk_2021 is now TurkKubo_2021.
- The submitted manuscript said that FilterAuids discards samples with ≤1 K reads. This should have been 500 reads and has been fixed in Fig. 1, the Table 3 header and lines 328 and 446.

Referee #1 comments RC1

The document by Morendo et al. reported a standardized pipeline to process microbial molecular data for cross-study comparison and reveling the distribution pattern of marine nitrogen-fixing organisms. Overall, the value of this study should be recognized due to nitrogen-fixers' ecological role in the ocean, and their sensitivity to future climate change (i.e. temperature), and the study is well suited for prompt publication on ESSD. At present I suggest the authors consider a few issues to improve the clarity of the manuscript.

REPLY: Thank you so much for seeing the potential for our software and database to contribute to marine nitrogen fixation research.

(1) The purpose of establishing such a novel processing approach was to allow cross-study comparisons. It would be helpful to present some improved aspects of this new pipeline compared to previous ones; especially for general marine biologists, some of the following questions can be better addressed: What would be changed if higher or lower retention of reads occurred (Fig. 3)? Could interpretation of marine diazotrophic distribution be somewhat different if higher or lower % retention of reads at each stage of the post-pipeline workflow (Fig. 4)? The previously reported pipelines may be tuned to suit some specific goals, could the new pipeline be used to suit those goals?

REPLY: We appreciate the comment and agree that comparing to previous pipelines would help readers. As explained above, the NifMAP OTU pipeline seemed the best option for a "previous

[pipeline]" in the absence of original sequence data and scripts (Appendix C). Moreover, none of the methods for the 19 previous studies appeared to tune their parameters to specific goals.

Usually communities obtained from our *nifH* workflow and NifMAP were similar at the taxonomic resolution of *nifH* clusters (Fig. B1; lines 825-828 in Appendix C). Several exceptions were likely explained by the workflow retaining more reads than NifMAP for studies with lower quality sequencing data (lines 830-844 in Appendix C).

For parameter impacts to the workflow (Appendix B) we focussed on the read trimming method. This seemed the most interpretable and feasible way to address the comment given the large parameter space (even for the handful of parameters in Table 2), interactions among DADA2 steps, and dependence of ASVs on sequencing data quality. Figure R1 above suggests that read counts obtained in each sample for each *nifH* cluster were usually similar using either of two very different trimming methods. Figure B1 shows this in a simpler way with all samples in a study pooled. Therefore, additional tuning of Table 2 parameters that are "downstream" to DADA2 filterAndTrim probably would not have a meaningful impact on results. The revised manuscript suggests (lines 766-767 in Appendix B):

> If both methods produce similar ASVs and abundances, additional parameter tuning is unlikely to impact the analysis meaningfully.

(2) Similar to (1), I think it may be useful to compare the relative abundance of each *nifH* cluster in previous studies using customized pipelines, with the present study (shown in Fig. 7b). And the authors could explore the advantage/disadvantage of the new pipeline, say more or less sensitive to particular *nifH* clusters.

REPLY: This was a great suggestion and addressed in Appendix C. We checked whether communities as represented in the *nifH* ASV database were similar to proxy communities for the original publications. Because the original studies often used OTU pipelines and *nifH* cluster annotation, we used the NifMAP OTU pipeline (which includes CART for *nifH* clusters) to produce proxy communities. Usually the proxies and their counterparts in the *nifH* ASV database were similar (when samples were pooled, shown in Fig. B1). Though our workflow detected additional, less abundant clusters in Hallstrom_2022 (line 831 in Appendix C).

We did not emphasize advantages or disadvantages of our workflow compared to NifMAP. ASVs have replaced OTUs in amplicon analyses and advantages of ASVs vs. OTUs have been described in detail elsewhere (Callahan et al., 2017).

(3) Dose the temperature-dependent distribution of *nifH* clusters (Fig. 8b) include only surface samples? Were deep samples, e.g. below 75 m (shown in Fig. 6b) be excluded? I think it more reasonable to plot SST vs. surface samples, say top 10 m. If all samples from surface to deep were included, can the authors explain the reasons?

REPLY: Originally all DNA samples were used to analyze the effects of SST on *nifH* cluster abundance and generate Figure 8 a and b to determine if a simple and commonly collected measurement would be able to provide a good predictor of the community. But we appreciate you pointing out that restricting depths to the photic zone is more relevant. We have now restricted this to only the photic samples (samples with depths less than 50 m for coastal, 100 m for open ocean [lines 508-509]) and the revised manuscript and figure now reflect this. This change had minimal effects on the distribution of these *nifH* clusters with respect to absolute latitude or SST.

(4) Please provide details of acquiring DNA and cDNA datasets (in Fig. 6) in the methods section preferably, and discuss the value of these datasets, their relative advantages and/or disadvantages, etc.

REPLY: Published datasets were acquired using GrabSeqs at line 137-138 (or line 136-138$_{orig}$)

> Datasets were downloaded from the National Center for Biotechnology Information (NCBI) Sequencing Read Archive (SRA) using the GrabSeqs tool (Taylor et al., 2020) by specifying the study's NCBI project accession.

GrabSeqs is the best way to obtain the data because it is simple to use, robust to download errors, and consistent if one wants to download multiple datasets.

If the question is on how samples were collected and processed, we only describe this for the two new datasets AK2HI and NEMO in section 2.2.2 (line 161-179$_{orig}$) since prior to this publication this information was not available publically. Sample collection and processing for the 19 published sets can be obtained using the publication DOIs in Table 1.

As to the value of the datasets, we included all published datasets that we could process and the two new datasets (described in section 2.2.1 and Table A1). The *nifH* ASV database becomes more valuable as global coverage of diazotroph communities expands (Fig. 2 and this interactive Google Map). Although cDNA datasets were somewhat limited, they were included because they provide insight into which organisms are actively expressing *nifH* at the time of sampling, which is a proxy for active N_2 fixation.

(5) In figure 7, why is it necessary to show % total reads (in panel a) and % relative abundance (in panel b) together? Most of the studies have similar abundance vs. reads of respective clusters, but the “Shiozaki_2020” shows obvious difference, e.g. higher % reads of “cluster 3” (light purple) while higher abundance of “others” (deep purple). Can the authors provide more explanations?

REPLY: Percent of composition plots like these often do not specify if they are using % total reads or relative abundance, therefore we considered presenting both. However, due to the confusion raised by both reviewers, we have removed the % total reads portion from the manuscript and Figure 7. Thank you for your feedback.

Citation: https://doi.org/10.5194/essd-2024-163-RC1

Referee #2 comments RC2

Morando and coauthors developed a bioinformatic workflow to analyze nifH gene amplicon sequences. This workflow was applied to analyze nifH amplicon datasets compiled from 21 studies and to build a nifH ASV database along with physical and chemical parameters extracted from CMAP. The workflow and nifH ASV database can facilitate comparison of marine diazotrophs diversity and biogeography across studies. The manuscript is well-written. My major comments are 1) In addition to the variations in software pipelines and parameters used to analyze nifH sequences by different studies, sample collection, DNA/RNA extraction and PCR conditions vary across studies, which makes cross-study comparisons challenging. Could you also provide some guidelines or best practice for these procedures? 2) how much difference it makes in the resulting diazotroph diversity comparing the new workflow to the sequence analysis procedures used in previous studies? It may be good to show an example in terms of the retained reads, identified ASVs, and relative abundance of different diazotrophs comparing this new workflow and the previous studies.

REPLY: Thank you very much for your feedback. Please note that we supplement the reply below at the top of this document.

1) As discussed above, impacts of software pipelines and parameters are addressed in Appendices B and C. As for sample collection and processing, although we agree this is a very important consideration in interpreting data, providing detailed guidelines and best practices is beyond the scope of this work, except for section 3.4 Limits and caveats to interpreting nifH amplicon data. Amplification biases and contamination inherent to *nifH* amplicon data (using nifH1-4 primers) are mentioned along with relevant references to help guide users on these issues (lines 662-668), as are complementary approaches for targeting specific diazotroph groups (qPCR, ddPCR, FISH; lines 682-683). For sample collection and extraction, methods from previous publications for the relevant ocean region (open ocean vs. coastal, tropical vs. polar) are a good place to start (DOIs in Table 1).
2) Thank you for raising this important question. Please see the discussion above that describes Appendix C, and in particular Fig. C1 for relative abundances and Table C1 for ASV versus OTU total abundances and distinct sequences.

Below are some minor comments.

REPLY: Thank you. We have fixed all of the items below.

Line 10: Marine nitrogen (N2) fixation…

REPLY: Fixed (line 10): "Marine dinitrogen (N_2) fixation…"

Line 17: diazotroph diversity, biogeography, and …

REPLY: Fixed (line 17) to use suggested order for consistency with line 13.

Line 47: Benavides et al. reference year is missing.

REPLY: This is a 2021 paper. Fixed (line 48 and 891).

Line 337: reference database (DB)

REPLY: "(DB)" was removed (line 351) and was replaced in the Table A1 header with "database."

Line 381: only removing 94 samples out of total xx samples

REPLY: Fixed at line 396: "108 cDNA samples out of 944 total samples"

Figure 3. Could you show the proportion reads retained as percentage as what you did in Table 4 and Figure 4?

REPLY: Fixed. The y axis label is now "% of initial reads retained"

Figure 4 caption: Study-specific loss of reads? Why not showing the subplots in alphabetical order to be consistent with other figures and tables across the study?

REPLY: Thank you! The title now indicates the "loss" of reads rather than "retention" and studies are now ordered alphabetically for easier comparison to other figures.

Figure 5. are these stacked bars? Or Northern and Southern Hemisphere bars overlapping?

REPLY: Yes, bars are stacked and the caption now mentions this (line 540).

Figure 6. Please double-check the sampling locations. Atlantic is missing.

REPLY: Thank you! Atlantic samples now are correctly labeled, and Shiozaki_2018LNO was corrected to be Artic.

Figure 7. Could you clarify the difference between % total reads and % relative abundance?

REPLY: The % total reads panel has been removed.

Line 597: nifH cluster 1E is not shown in the figure 7 legend.

REPLY: Cluster 1E is lumped in with 'other' clusters because it was usually rare and its bar would have been too thin to see the color. However, 1E was abundant in a single study. We thought that this was worth pointing out in the main text (lines 608-609).

Citation: https://doi.org/10.5194/essd-2024-163-RC2

References in this response

Angel, R., Nepel M., Panhölzl C., Schmidt H., Herbold C. W., Eichorst S. A., and Woebken D.: Evaluation of Primers Targeting the Diazotroph Functional Gene and Development of NifMAP - A Bioinformatics Pipeline for Analyzing nifH Amplicon Data, Front Microbiol., 9, 703, doi: 10.3389/fmicb.2018.00703, 2018.

Callahan, B. J., McMurdie P. J., and Holmes S. P.: Exact sequence variants should replace operational taxonomic units in marker-gene data analysis, ISME J, 11, 2639-2643, doi: 10.1038/ismej.2017.119, 2017.

Frank I.E., Turk-Kubo K. A., and Zehr J. P.: Rapid annotation of nifH gene sequences using classification and regression trees facilitates environmental functional gene analysis, Env Microbiol Rep., 8, 905-916, doi: 10.1111/1758-2229.12455, 2016.