Global biogeography of N_2-fixing microbes: *nifH* amplicon database and analytics workflow

Michael Morando[1*], Jonathan Magasin[1*], Shunyan Cheung[1,2], Matthew M. Mills[3], Jonathan P. Zehr[1], Kendra A. Turk-Kubo[1]

[1]Ocean Sciences Department, University of California, Santa Cruz, Santa Cruz, 95064, United States
[2]Institute of Marine Biology and Center of Excellence for the Oceans, National Taiwan Ocean University, Keelung, Taiwan
[3]Earth System Science, Stanford University, Stanford, 94305, United States
* equal contributions

Correspondence to: Kendra A. Turk-Kubo (kturk@ucsc.edu)

Abstract. Marine ~~nitrogen (N~~dinitrogen (N_2) fixation is a globally significant biogeochemical process carried out by a specialized group of prokaryotes (diazotrophs), yet our understanding of their ecology is constantly evolving. Although marine ~~dinitrogen (N_2)~~ N_2 fixation is often ascribed to cyanobacterial diazotrophs, indirect evidence suggests that non-cyanobacterial diazotrophs (NCDs) might also be important. One widely used approach for understanding diazotroph diversity and biogeography is polymerase chain reaction (PCR)-amplification of a portion of the *nifH* gene, which encodes a structural component of the N_2-fixing enzyme complex, nitrogenase. An array of bioinformatic tools exists to process *nifH* amplicon data, however, the lack of standardized practices has hindered cross-study comparisons. This has led to a missed opportunity to more thoroughly assess diazotroph ~~biogeography, diversity~~diversity, biogeography, and their potential contributions to the marine N cycle. To address these knowledge gaps a bioinformatic workflow was designed that standardizes the processing of *nifH* amplicon datasets originating from high-throughput sequencing (HTS). Multiple datasets are efficiently and consistently processed with a specialized DADA2 pipeline to identify amplicon sequence variants (ASVs). A series of customizable post-pipeline stages then detect and discard spurious *nifH* sequences and annotate the subsequent quality-filtered *nifH* ASVs using multiple reference databases and classification approaches. This newly developed workflow was used to reprocess nearly all publicly available *nifH* amplicon HTS datasets from marine studies, and to generate a comprehensive *nifH* ASV database containing ~~7909~~9383 ASVs aggregated from 21 studies that represent the diazotrophic populations in the global ocean. For each sample, the database includes physical and chemical metadata obtained from the Simons Collaborative Marine Atlas Project (CMAP). Here we demonstrate the utility of this database for revealing global biogeographical patterns of prominent diazotroph groups and highlight the influence of sea surface temperature. The workflow and *nifH* ASV database provide a robust framework for studying marine N_2 fixation and diazotrophic diversity captured by *nifH* amplicon HTS. Future datasets that target understudied ocean regions can be added easily, and users can tune parameters and studies included for their specific focus. The workflow and database are available,

respectively, in GitHub (https://github.com/jdmagasin/nifH-ASV-workflow; Morando et al., 2024c) and Figshare (https://doi.org/10.6084/m9.figshare.23795943.~~v1~~v2; Morando et al., 2024a).

1 Introduction

Dinitrogen (N_2) fixation, the reduction of N_2 into bioavailable NH_3 is a source of new nitrogen (N) in the oceans and can support as much as 70 % of new primary production in N-limited oligotrophic gyres (Jickells et al., 2017). Over millennia, N_2 fixation may balance the loss of N from the marine system through denitrification and annamox (Zehr and Capone, 2020). N_2 fixation was thought to be performed exclusively by prokaryotes, yet it was recently demonstrated that the marine haptophyte alga, *Braarudosphaera bigelowii*, contains a cyanobacterially-derived organelle specialized for N_2 fixation (Coale et al., 2024). Noting this exception, microorganisms able to fix N_2 (diazotrophs), are broadly characterized into two main groups, cyanobacterial diazotrophs (those phylogenetically related to cyanobacteria) and non-cyanobacterial diazotrophs (NCDs). Historically, cyanobacterial diazotrophs have been considered the most important contributors to marine N_2 fixation (Villareal, 1994; Capone et al., 2005). NCDs, first detected by Zehr et al. (1998), have since been demonstrated to be ubiquitous in pelagic marine waters, and are generally thought to be putative chemoheterotrophs with a highly diverse lineage that includes the massive phylum Proteobacteria as well as Firmicutes, Actinobacteria, and Chloroflexi (Turk-Kubo et al., 2022). However, their contribution of fixed N and their role in the global ocean is not well-understood (Moisander et al., 2017).

Diazotrophs are often present at low abundances relative to other members of ocean microbiomes, which makes them challenging to study (Moisander et al., 2017; Benavides et al., 2021). Distinctive pigments and morphologies that enable some cyanobacterial diazotrophs to be identified by microscopy are lacking in many diazotrophs (Carpenter and Capone, 1983; Carpenter and Foster, 2002), including NCDs. Furthermore, many marine diazotrophs are uncultivated, which has required the use of cultivation-independent approaches such as PCR and quantitative PCR (qPCR) (Luo et al., 2012; Shao and Luo, 2022; Turk-Kubo et al., 2022). The *nifH* gene encodes the identical subunits of the Fe protein of nitrogenase, the enzyme that catalyzes the N_2 fixation reaction, and contains both highly conserved and variable regions enabling its use as a phylogenetic marker and as a proxy for N_2-fixing potential in marine ecosystems globally (Gaby and Buckley, 2011).

Although the importance of marine N_2 fixation is well-established, knowledge gaps remain, and discoveries continue to be made (Zehr and Capone, 2020). For example, high-throughput sequencing (HTS) of *nifH* amplicons is expanding our knowledge of diazotroph biogeography and activity and has revealed surprising new diversity. However, HTS studies often utilize different or custom software pipelines and parameters, rendering direct comparisons between studies difficult. Additionally, many studies do not address the full breadth of diazotrophic diversity because they focus on cyanobacterial diazotrophs while providing only a superficial analysis of the NCDs present. The resulting lack of information on NCD *in* *situ* distributions limits our understanding of diazotroph ecology and N_2 fixation as well as our ability to predict how these populations will respond, e.g., trait-based ecological models, to a continually changing ocean.

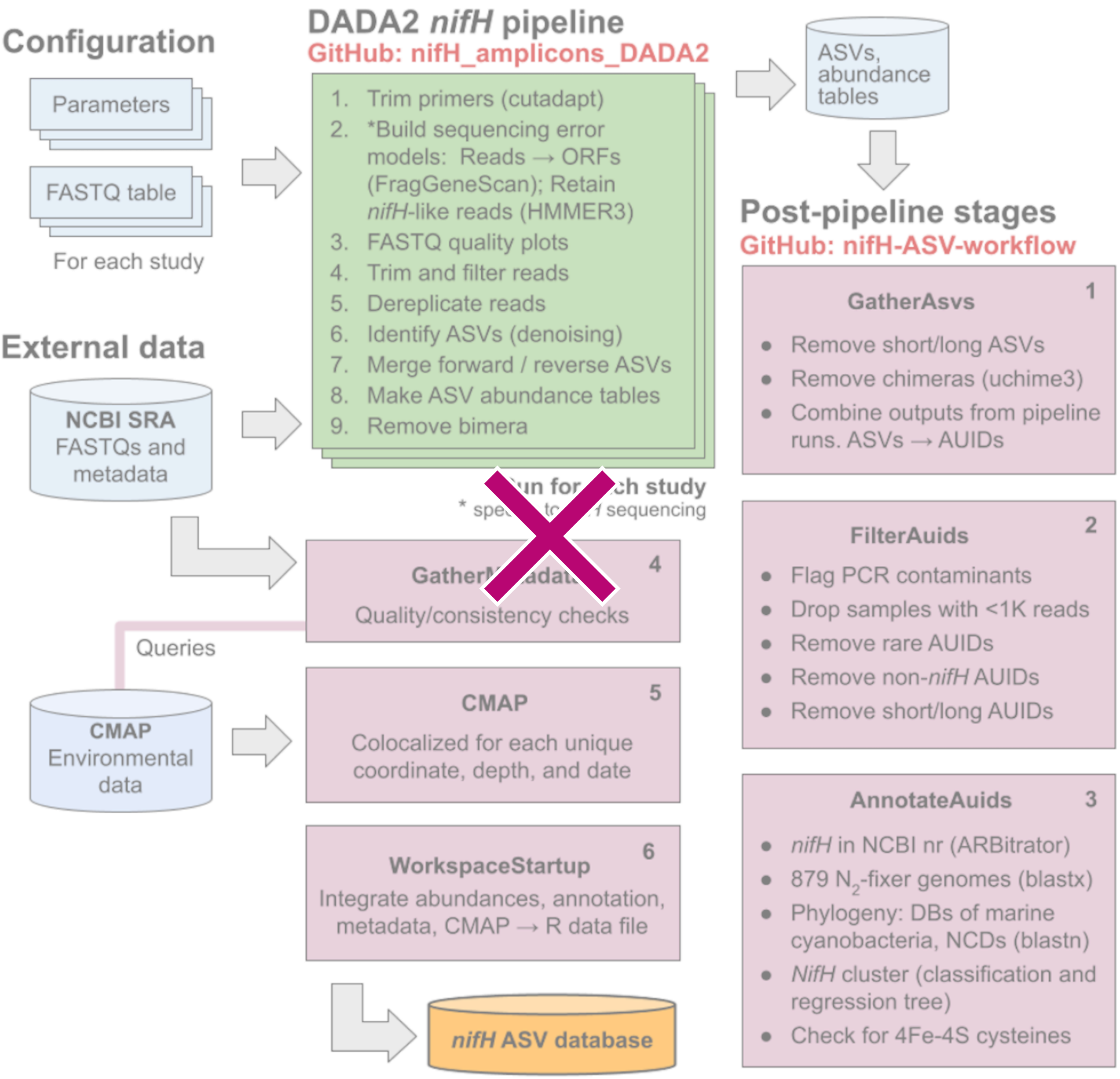
To address these issues, we compiled published *nifH* amplicon HTS datasets along with two new datasets. Twenty-one studies were reprocessed by our newly developed software workflow, which streamlines the integration of multiple, large amplicon datasets for reproducible analyses. The workflow identifies amplicon sequence variants (ASVs) using a pipeline developed around DADA2 (Callahan et al., 2016) — the DADA2 *nifH* pipeline — and then executes rigorous post-pipeline stages to: remove spurious *nifH* ASVs; annotate the remaining quality-filtered ASVs using multiple reference databases and classification approaches; and obtain *in situ* and modeled environmental data for each sample from the Simons Collaborative Marine Atlas Project (CMAP; https://simonscmap.com). Although created to support research into N_2 fixation (*nifH*), the complete workflow (ASV pipeline followed by the post-pipeline stages) can be adapted for use with other amplicon datasets, including other functional genes or taxonomic markers (16S rRNA genes), with some simple modifications.

In addition to the workflow, our efforts resulted in the construction of a comprehensive database of *nifH* ASVs with contextual metadata that will be a community resource for marine diazotroph investigations, enhancing comparability between previous and future *nifH* amplicon datasets. The *nifH* ASV database is available in Figshare (https://doi.org/10.6084/m9.figshare.23795943.~~v1~~v2; Morando et al., 2024a). The entire workflow required to produce the *nifH* ASV database is available in two GitHub repositories, the DADA2 *nifH* pipeline (https://github.com/jdmagasin/nifH_amplicons_DADA2; Morando et al., 2024b), and the post-pipeline stages (https://github.com/jdmagasin/nifH-ASV-workflow; Morando et al., 2024c).

2 Data and Methods

2.1 Overview of *nifH* amplicon workflow and *nifH* ASV database generation

The full workflow is comprised of two parts: 1) the DADA2 *nifH* pipeline; and 2) a series of post-pipeline stages (Fig. 1).


Configuration
Parameters
FASTQ table
For each study
External data
NCBI SRA FASTQs and metadata
CMAP Environmental data
Queries
DADA2 nifH pipeline
GitHub: nifH_amplicons_DADA2
1. Trim primers (cutadapt)
2. *Build sequencing error models: Reads → ORFs (FragGeneScan); Retain nifH-like reads (HMMER3)
3. FASTQ quality plots
4. Trim and filter reads
5. Dereplicate reads
6. Identify ASVs (denoising)
7. Merge forward / reverse ASVs
8. Make ASV abundance tables
9. Remove bimera
Run for each study
* specific to nifH sequencing
GatherMetadata 4
Quality/consistency checks
CMAP 5
Colocalized for each unique coordinate, depth, and date
WorkspaceStartup 6
Integrate abundances, annotation, metadata, CMAP → R data file
nifH ASV database
ASVs, abundance tables
Post-pipeline stages
GitHub: nifH-ASV-workflow
GatherAsvs 1
Remove short/long ASVs
Remove chimeras (uchime3)
Combine outputs from pipeline runs. ASVs → AUIDs
FilterAuids 2
Flag PCR contaminants
Drop samples with <1K reads
Remove rare AUIDs
Remove non-nifH AUIDs
Remove short/long AUIDs
AnnotateAuids 3
nifH in NCBI nr (ARBitrator)
879 N2-fixer genomes (blastx)
Phylogeny: DBs of marine cyanobacteria, NCDs (blastn)
NifH cluster (classification and regression tree)
Check for 4Fe-4S cysteines

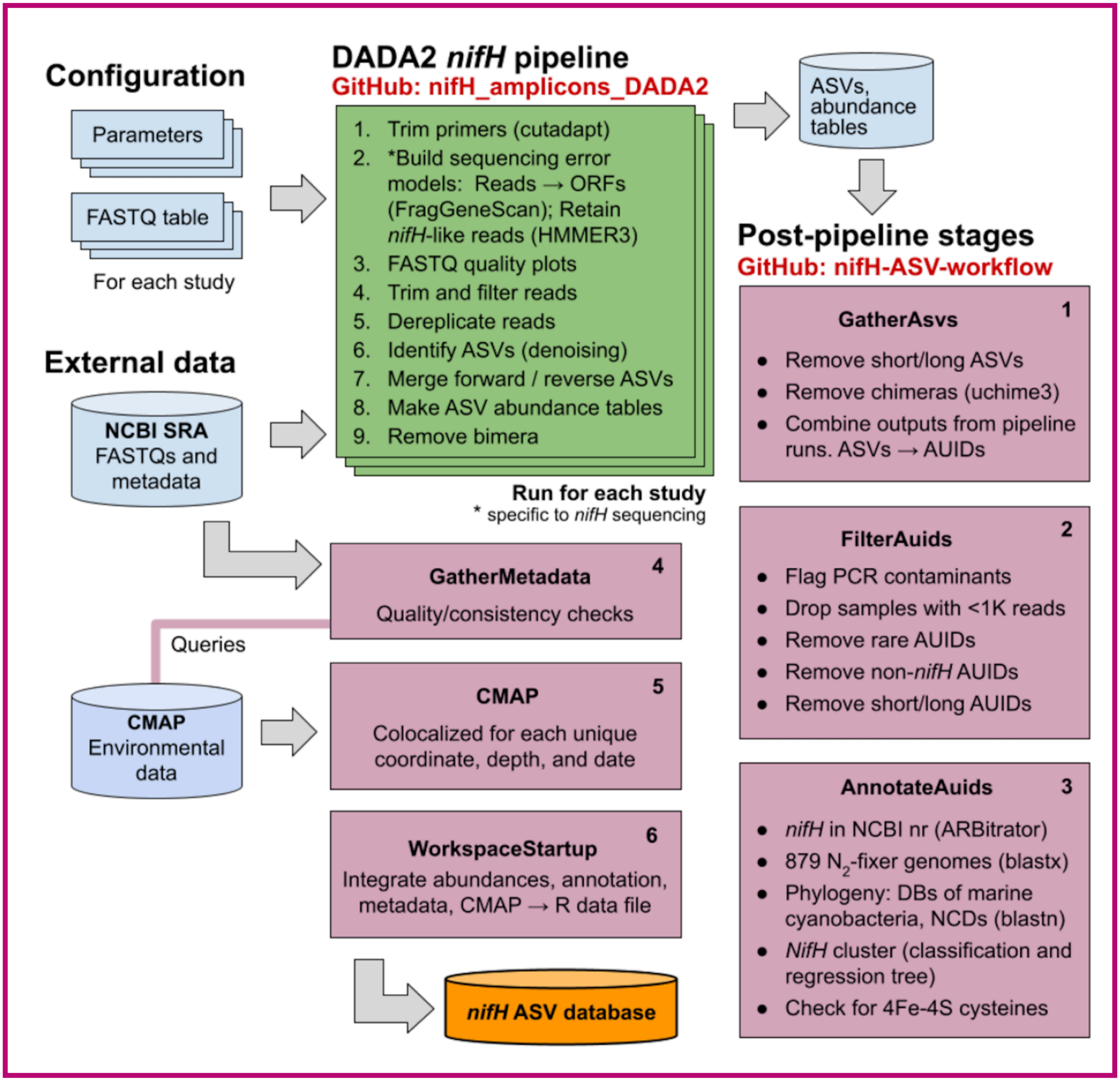

Configuration
Parameters
FASTQ table
For each study
External data
NCBI SRA FASTQs and metadata
DADA2 nifH pipeline
GitHub: nifH_amplicons_DADA2
1. Trim primers (cutadapt)
2. *Build sequencing error models: Reads → ORFs (FragGeneScan); Retain nifH-like reads (HMMER3)
3. FASTQ quality plots
4. Trim and filter reads
5. Dereplicate reads
6. Identify ASVs (denoising)
7. Merge forward / reverse ASVs
8. Make ASV abundance tables
9. Remove bimera
Run for each study
* specific to nifH sequencing
ASVs, abundance tables
Post-pipeline stages
GitHub: nifH-ASV-workflow
GatherAsvs 1
Remove short/long ASVs
Remove chimeras (uchime3)
Combine outputs from pipeline runs. ASVs → AUIDs
FilterAuids 2
Flag PCR contaminants
Drop samples with <1K reads
Remove rare AUIDs
Remove non-nifH AUIDs
Remove short/long AUIDs
AnnotateAuids 3
nifH in NCBI nr (ARBitrator)
879 N_2-fixer genomes (blastx)
Phylogeny: DBs of marine cyanobacteria, NCDs (blastn)
NifH cluster (classification and regression tree)
Check for 4Fe-4S cysteines
GatherMetadata 4
Quality/consistency checks
Queries
CMAP Environmental data
CMAP 5
Colocalized for each unique coordinate, depth, and date
WorkspaceStartup 6
Integrate abundances, annotation, metadata, CMAP → R data file
nifH ASV database

Figure 1: Schematic of the *nifH* amplicon data workflow. Data from all studies that met our criteria (Sect. 2.2) were downloaded from the NCBI Sequence Read Archive (SRA) and processed separately through the DADA2 *nifH* pipeline (green; Sect. 2.3.2), generally using identical parameters. ASV sequences and abundance tables from all studies were then combined and processed through each stage of the post-pipeline workflow (purple, Sect. 2.3.3) by executing the Makefile associated with each stage. Post-pipeline stages quality-filtered and then annotated the ASVs by reference to several *nifH* databases (DBs), and downloaded CMAP environmental data matched to the date, coordinates, and depth of each amplicon dataset. The main output of the entire workflow (pipeline and post-pipeline) is the *nifH* ASV database, which is available in Figshare (https://doi.org/10.6084/m9.figshare.23795943.~~v1~~v2; Morando et al., 2024a). The workflow is

maintained in two GitHub repositories, one for the DADA2 *nifH* pipeline (https://github.com/jdmagasin/nifH_amplicons_DADA2; Morando et al., 2024b) and one for the post-pipeline stages (https://github.com/jdmagasin/nifH-ASV-workflow; Morando et al., 2024c).

Required inputs for the pipeline are raw *nifH* amplicon sequencing reads and sample collection metadata (at minimum the latitude and longitude, depth and sample collection date and time) used to acquire environmental metadata from CMAP. Criteria for including publicly available datasets are detailed in Section 2.2.1.

The DADA2 software package is frequently used for processing 16/18S rRNA gene amplicon sequencing data due to its ability to remove base calling errors (“denoising”) and thereby infer error-free ASVs (Callahan et al., 2016). We have developed a customizable pipeline to improve the error models utilized by DADA2 by training them only on reads in a dataset that are valid *nifH* sequences (not PCR artifacts). The DADA2 pipeline runs from the command line in a Unix-like shell, moving through nine steps (Fig. 1 DADA2 *nifH* pipeline) described in Section 2.3.2 for each study independently. After the DADA2 pipeline is completed, outputs from all studies are integrated and refined by the six post-pipeline stages of the workflow, which perform additional quality filtering (e.g., size- and abundance-based selection), identify and remove spurious sequences (e.g., potential contaminants and non-target sequences), and annotate the ASVs (Fig. 1 Post-pipeline stages). By considering ASVs from all studies simultaneously, the workflow considers rare ASVs that might be discarded as irrelevant in a single-study analysis. Workflow stages are executed manually by running their associated Makefiles and Snakefiles within a Unix-like shell.

The workflow generates the final data product published in this work, the *nifH* ASV database, which includes ASV sequences, abundance and annotation tables, sample collection metadata, and sample environmental data from CMAP (Fig. 1). The database is available in Figshare (https://doi.org/10.6084/m9.figshare.23795943.~~v1~~v2; Morando et al., 2024a) as a set of tables (comma-separated value files) and an ASV FASTA file. However, these are also provided within an R data file, workspace.RData, in the WorkspaceStartup directory in the workflow GitHub repository, for users who wish to analyze, curate, or customize the database using R packages for ecological analysis. All documentation, scripts, and data needed to run the workflow and produce the *nifH* ASV database are provided in the workflow GitHub repository (https://github.com/jdmagasin/nifH-ASV-workflow; Morando et al., 2024c). This includes pre-generated pipeline results for each of the 21 studies as well as the pipeline parameters files.

In summary, the workflow facilitates the systematic and reproducible exploration of *nifH*-based diversity within microbial communities and was applied to available *nifH* amplicon data to generate a globally distributed *nifH* ASV database. Together the workflow and *nifH* ASV database will serve as valuable community resources, fostering future investigations while ensuring comparability between previous and forthcoming studies. In the following sections, detailed descriptions of each stage of the workflow are provided.

2.2 Compilation of *nifH* amplicon studies

2.2.1 Published studies

We compiled all publicly available *nifH* amplicon HTS data that were generated using the nifH1-4 primers (Zani, 1999; Zehr and ~~Mcreynolds~~McReynolds, 1989) and subsequently sequenced on the Illumina MiSeq/HiSeq platform totaling ~~19~~21 studies (Table 1). Limiting the scope to investigations that used the same amplification primers enabled a more tractable comparison across studies by different research groups that employed varying approaches to sample collection and preparation for sequencing by different centers. Datasets were downloaded ~~directly~~ 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. Each dataset obtained included paired-end sequencing reads (in FASTQ files) and a table with the collection metadata for each sample. Some datasets could not be retrieved directly from the SRA and were obtained ~~directly~~ from the authors (Table A1). Note that we did not include studies where data was generated from experimental perturbations or particle enrichments (Table A1). Data were last accessed from NCBI SRA on 17 April 2024.

Table 1: Information on the studies compiled to generate the *nifH* ASV database. All compiled studies and associated information. This includes the study ID used to refer to each dataset, the number of samples, NCBI BioProject accession, a reference to each publication and its corresponding DOI.

Study ID	Samples	NCBI BioProject	Reference	DOI
AK2HI	43	PRJNA1062410	This study	n/a
BentzonTilia_2015	56	PRJNA239310	Bentzon-Tilia et al., 2015	10.1038/ismej.2014.119
Ding_2021	32	SUB7406573	Ding et al., 2021	10.3390/biology10060555
Gradoville_2020_G1	111	PRJNA530276	Gradoville et al., 2020	10.1002/lno.11423
Gradoville_2020_G2	56	PRJNA530276	Gradoville et al., 2020	10.1002/lno.11423
Hallstrom_2021	82	PRJNA656687	Hallstrøm et al., 2022b	10.1002/lno.11997
Hallstrom_2022	83	PRJNA756869	Hallstrøm et al., 2022a	10.1007/s10533-022-00940-w
Harding_2018	91	PRJNA476143	Harding et al., 2018	10.1073/pnas.1813658115
Mulholland_2018	29	PRJNA841982	Mulholland et al., 2019	10.1029/2018GB006130
NEMO	56	PRJNA1062391	This study	n/a
Raes_2020	121	PRJNA385736	Raes et al., 2020	10.3389/fmars.2020.00389
Sato_2021	28	PRJDB10819	Sato et al., 2021	10.1029/2020JC017071
Selden_2021	10	PRJNA683637	Selden et al., 2021	10.1002/lno.11727
Shiozaki_2017	22	PRJDB5199	Shiozaki et al., 2017	10.1002/2017GB005681
Shiozaki_2018GBC	20	PRJDB6603	Shiozaki et al., 2018b	10.1029/2017GB005869

Shiozaki_2018LNO	20	PRJDB5679	Shiozaki et al., 2018a	10.1002/lno.10933
Shiozaki_2020	14	PRJDB9222	Shiozaki et al., 2020	10.1038/s41561-020-00651-7
Tang_2020	6	PRJNA554315	Tang et al., 2020	10.1038/s41396-020-0703-6
~~TianjUni_2016~~Turk Kubo_2021	~~14~~136	~~PRJNA637983~~PRJNA695866	~~Wu~~Turk-Kubo et al., 2021	~~10.1007/s10021-021-00702-z~~10.1038/s43705-021-00039-7
~~TianjUni_2017~~Wu_2019	18	PRJNA438304	Wu et al., 2019	10.1007/s00248-019-01355-1
~~Turk_2021~~Wu_2021	~~136~~14	~~PRJNA695866~~PRJNA637983	~~Turk-Kubo~~Wu et al., 2021	~~10.1038/s43705-021-00039-7~~10.1007/s10021-021-00702-z

Sample quality was validated prior to processing through the DADA2 *nifH* pipeline. Samples were discarded if they did not contain unmerged pairs of forward and reverse reads with properly oriented primer sequences (Table A1). There were two exceptions, studies by Shiozaki et al. (2017) and Shiozaki et al. (2018b), that used mixed-orientation sequence libraries and required preprocessing. The reads in each of these studies were partitioned by whether they captured the coding or template strand of *nifH*, determined by primer orientation. Because HTS sequence quality generally degrades from 5' to 3', the partitioned data were run separately through the pipeline to preserve their sequencing error profiles for DADA2. The ASVs from the misoriented reads (e.g. forward reads with template sequence) were then reverse-complemented and combined with the properly oriented ASVs into a single ASV abundance table and FASTA file. Table 1 and Table A1 provide information for obtaining the raw FASTQ files for all samples evaluated for the *nifH* ASV database including information regarding studies excluded from the database.

2.2.2 Unpublished *nifH* amplicon datasets

Additional *nifH* gene HTS datasets were included from DNA samples collected on two cruises in the North Pacific. One was a transect cruise across the Eastern North Pacific (NEMO; R/V New Horizon, August 2014; Shilova et al., 2017), and the other was a transect cruise from Alaska to Hawaii (AK2HI; R/V Kilo Moana, September 2017). Euphotic zone samples were collected from Niskin bottles deployed on a CTD-rosette (NEMO) or from the underway water system (5 m; AK2HI). NEMO samples (2-4 L) were filtered through 0.2 µm and 3 µm pore-size filters (in series), while AK2HI samples (ca. 2 L) were filtered through 0.2 µm pore-size filters using gentle peristaltic pumping. Filters were dried, flash frozen and stored at -80ºC until processing. DNA was extracted using a modified DNeasy Plant Kit (Qiagen, Germantown, MD) protocol, described in detail in Moisander et al. (2008), with on-column washing steps automated by a QIAcube (Qiagen).

Partial *nifH* DNA sequences were PCR-amplified using the nifH1-4 primers in a nested *nifH* PCR assay (Zani, 1999; Zehr and ~~Mcreynolds~~McReynolds, 1989) according to details in Cabello et al. (2020). All samples were amplified in duplicate and pooled prior to sequencing. A targeted amplicon sequencing approach was used to create barcoded libraries as described in Green et al. (2015), using 5’ common sequence linkers (Moonsamy et al., 2013) on second round primers, nifH1 and nifH2. Sequence libraries were prepared at the DNA Service Facility at the University of Illinois at Chicago, and multiplexed amplicons were bidirectionally sequenced (2 × 300 bp) using the Illumina MiSeq platform at the W.M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign. Samples were multiplexed to achieve ca. 40,000 high quality paired reads per sample. The AK2HI and NEMO datasets can be found in the SRA (BioProjects PRJNA1062410 and PRJNA1062391, respectively).

2.2.3 Sample collection data and co-localized CMAP environmental data

Sample collection data (e.g. coordinates, depth, date) and environmental data provide essential context for the interpretation of diazotroph ‘omics datasets. Large-scale multivariate analyses depend on properly formatted, complete, and ideally quality checked metadata from consistently collected and analyzed measurements. However, accessibility to this information is often limited (especially environmental data) for datasets published across multiple decades. Therefore, we first obtained sample collection metadata from the SRA, and corrected or flagged errors and inconsistencies in the GatherMetadata stage of our post-pipeline workflow (described below), to ensure consistency and completeness. For each sample, the geographic coordinates, depth, and collection date (at local noon) from the SRA were used to query the Simons Collaborative Marine Atlas Project on 24 March 2023 (CMAP; https://simonscmap.com/; Ashkezari et al., 2021) for co-localized environmental data using a custom script (query_CMAP.py) in the CMAP stage of the workflow (Fig. 1). CMAP is an open-source data portal designed for retrieving, visualizing, and analyzing diverse ocean datasets including research cruise-based and autonomous measurements of biological, chemical, and physical properties, multi-decadal global satellite products, and output from global-scale biogeochemical models. For each sample a mixture of ~~102~~100 satellite derived and modeled environmental variables from the CMAP repository were obtained. These, along with the SRA collection data, are included in our database. Aggregated metadata for all samples are summarized in Supplementary Table 1 but a detailed description of environmental metadata can be found at the CMAP website (https://simonscmap.com/catalog). Metadata are available in the *nifH* ASV database (metaTab.csv for sample metadata and cmapTab.csv for environmental data).

2.3 Automated workflow for processing datasets with the DADA2 *nifH* pipeline

2.3.1 Installation of the DADA2 *nifH* pipeline and the post-pipeline workflow

The workflow (Fig. 1) comprises two software projects installed from separate GitHub repositories, nifH_amplicons_DADA2 which ~~comprises~~contains the ASV pipeline and ancillary scripts, and nifH-ASV-workflow which integrates pipeline results for all datasets with annotation and CMAP environmental data to produce the data deliverable of the present work, the *nifH* ASV database. Installation requires cloning the nifH_amplicons_DADA2 repository (https://github.com/jdmagasin/nifH_amplicons_DADA2; Morando et al., 2024b) to a local machine and then downloading several external software packages using miniconda3. Detailed installation instructions are available from the GitHub homepage, as well as a small tutorial to verify the installation on a small *nifH* amplicon dataset and introduce the two main pipeline commands (organizeFastqs.R and run_DADA2_pipeline.sh). Altogether the installation and example take 30–40 min.

After installing the ASV pipeline, installation of the nifH-ASV-workflow proceeds similarly: Clone the GitHub repository (https://github.com/jdmagasin/nifH-ASV-workflow**;** Morando et al., 2024c) and then download a few additional packages with miniconda3 (~10 min to complete). For each study, the nifH-ASV-workflow includes the pipeline outputs (ASVs and abundance tables) which were used to create the *nifH* ASV database. Pipeline parameters and FASTQ input tables for each study are also provided for users who instead wish to rerun the pipeline starting from FASTQs downloaded from the SRA. Because the nifH-ASV-workflow includes data and parameters specific to the studies used in this work, it has a separate GitHub repository from the pipeline. However, we emphasize that together they comprise the *nifH* amplicon workflow in Fig. 1.

Adding a new dataset to the workflow can be summarized in four steps: (1) Start a Unix-like shell that includes the required software (by “activating” a minconda3 environment called nifH_ASV_workflow). (2) Generate ASVs for the new dataset by running it through the pipeline, likely multiple times to tune parameters (Table 2). Output can be placed in the Data directory alongside other studies used in this work, and SRA metadata must be added to Data/StudyMetadata. (3) Include the new ASVs in the workflow by appending rows to the table GatherASVs/asvs.noChimera.fasta_table.tsv, which has file paths to all ASV abundance tables. (4) For each stage shown in Fig. 1, run the associated Makefile or Snakefile from the Unix-like shell by executing "make" or "snakemake -c1 --use-conda", respectively. Documentation resides within each Makefile or Snakefile. Input tables from the post-pipeline workflow also have embedded documentation.

Table 2. Parameters for controlling the DADA2 *nifH* pipeline. Default values can be overridden in the text file that is passed to run_DADA2_pipeline.sh. Parameters for "Read trimming" and "Error models" are used in steps 1 and 2 of the pipeline (Fig. 1). The remaining parameters are important for controlling how DADA2 trims and quality filters the reads, and merges forward and reverse sequences to create ASVs.

DADA2 *nifH* pipeline step	**Parameter name**	**Default value**	**Description**	**Studies with non-default parameters**
Read Trimming Remove **primers** with cutadapt	forward	TGYGAYCCN AARGCNGA	Forward primer 5' to 3'. Default is nifH2 (Zehr and ~~Mcreynolds~~McReynolds, 1989).	None
	reverse	ADNGCCATC ATYTCNCC	Reverse primer 5' to 3'. Default is nifH1 (Zehr and ~~Mcreynolds~~McReynolds, 1989).	None

	allowMissingPrimers	FALSE	If TRUE, retain read pairs even if primers are absent, e.g. if trimmed reads were uploaded to NCBI SRA.	Ding et al., 2021
Error Models	skipNifHErrorModels	FALSE	By default, use only *nifH*-like reads to train error models. If TRUE, use a random sample of all reads.	None
	NifH_minBits	150	Train error models using reads that align to PFAM00142 at ≥ the specified bit score. The trusted cut off in PFAM00142 (25 bits) is always used to filter reads, then NifH_minBits. If set to 0, only the trusted cut off is used.	Set to 0 for most studies. Exceptions that used 100 bits were: Bentzon-Tilia et al., 2015; Gradoville et al., 2020; Shiozaki et al., 2018a; Turk-Kubo et al., 2021.
	NifH_minLen	33	Train error models using reads with ORFs that align with ≥ this many residues to PFAM00142.	None
DADA2 filterAndTrim~~(~~ ~~)~~	id.field	NA	Specify number of ID field if reads do not follow the CASAVA format. Forwarded to filterAndTrim(). If set, usually to 1.	Ding et al., 2021; Wu et al., 2021; Wu et al., 2019; Mulholland et al., 2019; Raes et al., 2020; Tang et al., 2020; Selden et al., 2021; Hallstrøm et al., 2022b; Hallstrøm et al., 2022a
	~~truncQ~~maxEE.fwd	~~2~~Inf	Forwarded to filterAndTrim().	All studies set to ~~16 unless used truncLen~~2.
	~~maxEE.fwd~~maxEE.rev	Inf		All studies set to ~~2~~4.
	~~maxEE.rev~~minLen	~~Inf~~20	Forwarded to filterAndTrim().	~~All studies set to 4.~~None
	~~minLen~~truncLen.fwd	~~20~~0	Forwarded to filterAndTrim().	~~None~~Ancillary script estimateTrimLengths.R determined optimal lengths.
	~~truncLen.fwd~~truncLen.rev	0		
	~~truncLen.rev~~truncQ	~~0~~2	Forwarded to filterAndTrim()~~.~~	~~(See truncLen.fwd.)~~All studies used truncLen.
	useOnlyR1Reads	FALSE	If TRUE, only use R1 reads (and do not call mergePairs()). Used if R2 reads are very low quality.	None
DADA2 mergePairs~~()~~	minOverlap	12	Forwarded to mergePairs().	None
	maxMismatch	0	Forwarded to mergePairs().	All studies set to 1.
	justConcatenate	FALSE	Forwarded to mergePairs().	None

2.3.2 DADA2 *nifH* pipeline

To encourage reproducible outputs and usage by non-programmers, the DADA2 pipeline (GitHub repository: nifH_amplicons_DADA2) is controlled by a plain text parameters file (Table 2) and a descriptive table of input samples (the "FASTQ map"). Since a study might include samples with vastly different diazotroph communities and relative abundances, potentially impacting ASV inferences by DADA2, the FASTQ map for a study enables samples to be partitioned into "processing groups" that are each run separately through DADA2. For example, in the present work processing groups usually partitioned the samples in a study by the unique combinations of collection station or date, nucleic acid type (DNA or RNA), size fraction, and collection depth. Pipeline outputs for each processing group are stored in a directory hierarchy with levels that follow the processing group definition. Partitioning datasets into processing groups greatly improves the overall speed of DADA2 and simplifies subsequent analyses that compare ASVs detected in different kinds of samples (e.g., detected versus transcriptionally active diazotrophs, or presence across different stations, depths, and/or size fractions). For generating the *nifH* ASV database, studies that met selection criteria (Sect. 2.2.1 and Table 1) were run through the pipeline using the study-specific FASTQ maps and parameters available in the Data directory of the nifH-ASV-workflow GitHub repository.

The DADA2 pipeline runs from the command line in a Unix-like shell, moving through 9 main steps (Fig. 1 DADA2 *nifH* pipeline): (1) trim reads of primers using cutadapt (Martin, 2011); (2) build sequencing error models; (3) make FASTQ quality plots; (4) trim and filter reads based on quality; (5) dereplicate; (6) denoise (ASV inference); (7) merge forward and reverse sequences; (8) make the ASV abundance table; and (9) remove bimera (Callahan et al., 2016 for steps 2 through 9). These steps will be familiar to DADA2 users, except that for step 2 the error models are trained only on *nifH*-like reads (discussed below). To run the pipeline on other functional genes, the parameters file would need to be edited to disable *nifH*-based error models and to include the expected primers. We again note that the DADA2 pipeline is distinct from the post-pipeline workflow stages which are specific to this work, but together they comprise the workflow in Fig. 1.

DADA2 parameters impact the ASV sequences identified, and the number of reads used. Thus, exploring parameters is essential for checking the robustness of ASVs (particularly rare ones) and their relative abundances. The ~~DADA2 pipeline supports the optimization of parameters (Table 2). For example, one~~method and parameters used to trim the reads are especially important because most pipeline steps occur after filterAndTrim (Fig. 1). Two methods are supported: One can trim each read based on its quality degradation (truncQ parameter to the DADA2 filterAndTrim function; Table 2) or all reads at the same position determined by inspecting sample FASTQ quality plots~~.~~ (truncLen parameter; Table 2, and comparison of methods in Appendix B). The latter approach can be labor-intensive and unsystematic for studies with tens to hundreds of samples. To address this the ancillary script estimateTrimLengths.R can be used to determine trimming lengths that will maximize the percentage of reads that make it through the pipeline. For each FASTQ file in a study, the script chooses 1 K read pairs at random and removes the primers. Then the read pairs are trimmed using every combination of lengths over a window (from 55—85 % of the median read length in 15 bp steps) and successful merges (with ≥12 bp overlapping and ≤2 mismatches) are counted. The counts are averaged across all samples (weighting by sequencing depths) and the top ten combinations of forward and reverse trimming lengths are reported in a table, with estimates for the percentages of reads retained and the mean errors per read to help choose the maxEE parameters (Table 2).

The pipeline allows one to rerun DADA2 steps 3~~-~~–9, with outputs saved in separate, date-stamped directories. ~~Read trimming~~Primer removal and error models (steps 1~~-~~–2) are unlikely to benefit much from parameter tuning, so the pipeline reuses outputs from those steps. Log files and diagnostic plots created by the pipeline are intended to facilitate parameter evaluation as well to capture statistics to support publication. Moreover, logs and other pipeline outputs are consistently formatted across pipeline runs, which enables scripts to aggregate and analyze results across datasets such as in our workflow.

Step 1 consisted only of ~~read trimming~~primer removal using cutadapt (Martin, 2011). Raw reads were ~~trimmed and retained only when read pairs for which~~retained only if the forward (nifH2) and reverse (nifH1) primers were both found on the R1 and R2 reads, respectively. DADA2 sequencing error models were built at step 2 using only the reads predicted to be *nifH*, rather than a subsample of all reads as in typical use of DADA2. Reads likely to encode *nifH* were identified as follows: FragGeneScan (version 1.31, (Rho et al., 2010)) was used to predict open reading frames (ORFs) on R1 reads which were then aligned to the nitrogenase PFAM model (PF00142.20) using HMMer3 (hmmsearch version 3.3.2; hmmer.org). ORFs with >33 residues and a bit score that exceeded the trusted cut-off encoded in the model (25.0 bits) were retained. Prefiltering the reads aims to reduce effects of PCR artifacts on the error models. For some studies this approach resulted in increases (~3–10 %) in the total percentage of reads retained in ASVs, and fewer total ASVs, compared to using error models based on a subsample of all reads. Adapting the pipeline to a different marker gene would only require substituting an appropriate PFAM model, or disabling step 2 (by setting skipNifHErrorModels to TRUE; Table 2), which forces the pipeline to make error models by subsampling from all reads. At step 4, DADA2 filterAndTrim() ~~truncated reads at the first base with PHRED score ≤16 and~~trimmed forward and reverse reads using the lengths suggested by estimateTrimLengths.R and then discarded read pairs that had excessive errors (>2 for R1 reads, >4 for R2 reads) or were <20 bp. ~~The PHRED quality cut off, which corresponds to a 2.5 % base call error rate, was complemented by conservative parameters~~Conservative parameters were used for merging sequences: At most 1 base pair was allowed to mismatch in the forward and reverse sequence overlap of minimally 12 bp (stage 7). Dereplicating (step 5) and denoising, ASV calling (step 6), generating an abundance table (step 8), and bimera detection (step 9), were all performed with default DADA2 parameters. ~~Data sets~~Datasets that passed pre-processing steps (Table 1) were run through the DADA2 pipeline using mostly identical parameters ~~(~~except for the trimming lengths (truncLen.fwd and truncLen.rev in Table 2).

2.3.3 Post-pipeline stages

The workflow post-pipeline stages (GitHub repository: nifH-ASV-workflow) combine the pipeline outputs, conduct further quality control steps, co-locate the samples with environmental data from the CMAP data portal, and annotate the ASVs (Fig. 1 Post-pipeline stages). Key outputs from the post-pipeline are: a unified FASTA with all the unique ASVs detected across all the studies (i.e. all samples); tables of ASV total counts and relative abundances in all studies; multiple annotations for each ASV by comparison to several *nifH* reference databases; and CMAP environmental data for each sample. These outputs comprise the *nifH* ASV database, and are all available within an R image file (workspace.RData) generated by the workflow which is included in the nifH-ASV-workflow repository. Provision as an R image will make the outputs immediately accessible to many researchers who prefer R due to its extensive packages for ecological analysis. The *nifH* ASV database is also available on Figshare (https://doi.org/10.6084/m9.figshare.23795943.~~v1~~v2; Morando et al., 2024a). The remainder of this section describes each of the post-pipeline stages.

The GatherAsvs stage aggregated ASV sequences and abundances across all DADA2 pipeline runs (i.e. from all samples and studies). First, ASVs were filtered based on length. Chimera sequences were then removed using UCHIME3 denovo (Edgar, ~~2016~~2016a) via VSEARCH (Rognes et al., 2016). Chimera sequences were identified within each sample, but the final classification was based on majority vote (chimera or not) across the samples in the processing group. Second, the GatherAsvs stage combined the non-chimeric ASVs from all studies into a single abundance table and FASTA file. Since each study is run independently through the DADA2 pipeline, ASV identifiers are not consistent across studies. Therefore, each unique ASV sequence was renamed with a new unique identifier of the form AUID.*i*, where AUID stands for **A**SV **U**niversal **ID**entifier. The scripts used to rename the ASVs (assignAUIDs2ASVs.R) and to create the new abundance table (makeAUIDCountTable.R) are available at the nifH_amplicons_DADA2 GitHub repository (in scripts.ancillary/ASVs_to_AUIDs). The script assignAUIDs2ASVs.R optionally takes an AUID reference FASTA so that AUIDs can be preserved as new datasets are added to future versions of the *nifH* ASV database.

Both rare and potential non-*nifH* sequences were assessed on the unified AUID tables in the next stage, FilterAuids (Fig. 1). First, possible contaminants were identified by the Makefile invocation of check_nifH_contaminants.sh, provided as an ancillary script in the pipeline GitHub repository. In brief, check_nifH_contaminants.sh first translated all ASVs into amino acid sequences using FragGeneScan (Rho et al., 2010), which were then compared using *blastp* to 26 contaminants known from previous *nifH* amplicon studies (Zehr et al., 2003; Goto et al., 2005; Farnelid et al., 2009; Turk et al., 2011). ASVs that aligned at >96 % amino acid identity to known contaminants were flagged. Next FilterAuids removed samples with ≤~~1000~~500 reads, and rare ASVs, defined as those that did not have at least one read in at least two samples or ≥1000 reads in one sample.

Next, the ancillary script, classifyNifH.sh, was employed to identify and remove non-*nifH*-like sequences. The script utilized *blastx* to search each ASV against ~44 K positive and ~15 K negative examples of NifH protein sequences that were found in NCBI GenBank by ARBitrator (run on April 28, 2020; Heller et al., 2014). ASVs were classified based on the relative quality of their best hits in the two databases, similar to the "superiority" check in ARBitrator. An ASV was classified as positive if the E-value of its best positive hit was ≥10 times smaller than the E-value for the best negative hit, and vice versa for negative classifications. ASVs failing to meet these criteria were classified as ‘uncertain’. The *blastx* searches used the same effective sizes for the two databases (-dbsize 1000000), so that E-values could be compared, and retained up to 10 hits (-max_target_seqs 10).

The FilterAuids stage of the workflow exclusively discarded ASVs with negative classifications. “Uncertain” ASVs were retained as potential *nifH* sequences not in GenBank. In the last stage, FilterAuids excluded ASVs with lengths that fell outside 281–359 nucleotides, a size range which in our experience encompasses the majority of valid *nifH* amplicon sequences generated by nested PCR with nifH1–4 primers.

For each AUID in the *nifH* ASV database, we provide taxonomical annotations using several different approaches, encompassed by the AnnotateAuids stage (Fig. 1) and accessible through ancillary scripts in the GitHub repository (in scripts.ancillary/Annotation). The script blastxGenome879.sh enables a protein level comparison via *blastx* against a database of 879 sequenced diazotroph genomes (“genome879”, https://www.jzehrlab.com/~~nifh~~nifH). Here, the closest cultivated relative for each AUID was determined by smallest E-value among alignments with ≥50 % amino acid identity and ≥90 % query sequence coverage. Cautious interpretation is suggested because the reference ~~DB~~ database is small and contains only cultivable taxa. Similarly, the top nucleotide match of each AUID was identified by E-value within alignments possessing ≥70 % nt identity and ≥90 % query sequence coverage obtained by *blastn* against a curated database of *nifH* sequences (July 2017 *nifH* database, https://~~wwwzehr.pmc.ucsc.edu/nifH_Database_Public/~~www.jzehrlab.com/nifH) by executing the blastnARB2017.sh script. Additionally, *nifH* cluster annotations were assigned to each ASV using the classification and regression tree (CART) method of Frank et al. (2016). This approach was implemented as part of a custom tool that predicted ORFs for the ASVs with FragGeneScan, then performed a multiple sequence alignment on the ORFs, and then applied the CART classifier. The tool is available as the ancillary script assignNifHclustersToNuclSeqs.sh.

The Makefile created and searched against two "phylotype" databases, one containing 223 *nifH* sequences from prominent marine diazotrophs including NCDs (Turk-Kubo et al., 2022) and another with 44 UCYN-A *nifH* oligotype sequences (Turk-Kubo et al., 2017). These databases were searched using *blastn* with the effective database size of the ARB2017 database (-dbsize set to ~29 million bases) to enable E-value comparisons across all three searches. For each ASV, we provide phylotype annotations based on the top hit by E-value if the alignment had ≥97 % nt identity and covered ≥70 % of the ASV. Finally, ORFs for all ASVs were searched for highly conserved residues which are thought to coordinate the 4Fe-4S cluster in NifH, specifically for paired cysteines shortly followed by AMP residues (described in Schlessman et al. 1998). This simple check, performed by the script check_CCAMP.R, was intended to complement the reference-based annotations above. Presence of cysteines and AMP could be used to retain ASVs that have no close reference. Absence could be used to flag ASVs that, despite high similarity to a reference sequence, might not represent functional *nifH* (e.g. due to frameshifts).

Since the annotation scripts provided multiple taxonomic identifications for most of the AUIDs, a primary taxonomic ID was assigned for each AUID using the script make_primary_taxon_id.py. If a phylotype annotation (e.g., Gamma A) was assigned, this became the primary taxonomic ID; otherwise, cultivated diazotrophs from genome879 were used (e.g., "*Pseudomonas stutzeri*"). Finally, when neither a phylotype nor a cultivated diazotroph could be determined, the *nifH* cluster (e.g. "unknown 1G") was used. AUIDs without an assigned *nifH* cluster or taxonomic rank below domain were removed from the final *nifH* ASV database unless paired cysteines and AMP were detected. This final data filtration step occurred in the WorkspaceStartup stage described below.

The CMAP stage was managed by a Snakefile that called the script query_cmap.py to query the CMAP data portal for co-localized environmental data (Fig. 1). The script was passed the main output from the GatherMetadata stage, metadata.cmap.tsv, a table of the collection coordinates, dates at local noon, and depths from all the samples. GatherMetadata reported any samples with missing metadata and ensured standardized formats for the required query fields. Additionally, query_cmap.py validated fields prior to querying CMAP. It should be noted that the precision of values obtained from CMAP depend on floating point arithmetic, not the significant digits of the underlying measurement or model. Therefore, prior to an analysis requiring high precision for specific CMAP variables, it is recommended to consult the original producer of the data to determine the significant digits.

The last stage of the workflow, WorkspaceStartup, filtered out AUIDs that had no annotation and then generated the final *nifH* ASV database, which is comprised of AUID abundance tables (counts and relative), AUID annotations, sample metadata and corresponding environmental data. These data are provided as text files (.csv and FASTA) within a single compressed file (.tgz) that is available in Figshare (https://doi.org/10.6084/m9.figshare.23795943.~~v1~~v2; Morando et al., 2024a) as well as within the workflow GitHub repository within an R image file (workspace.RData).

2.4 Diazotroph biogeography from DNA dataset of the *nifH* ASV database

The DNA dataset, a custom version of the *nifH* ASV database restricted to DNA samples (representing a majority of the database, only removing ~~94~~108 cDNA samples out of 944 total samples), was created to showcase the utility of the workflow. Additional data reduction steps were conducted, averaging replicates and samples from the same location but different size fractions, to enable comparisons between different sampling methodologies.

3 Results and Discussion

3.1 Generation of the marine *nifH* ASV database

All publicly available marine *nifH* amplicon HTS data from studies that met our criteria, including two new studies, were compiled in the present investigation (see Sect. 2.2 and Table A1). Altogether 982 samples from 21 studies, comprising a total of 87.7 million reads (Table 3), were processed through the entire workflow, i.e., the DADA2 *nifH* pipeline (Sect. 2.2.2) as well as the post-pipeline stages (Sect. 2.2.3). The *nifH* ASV database, i.e., the ASV sequences, abundances, and annotations, as well as sample collection and CMAP environmental data, was generated from the ~~865~~944 samples, ~~7909~~9383 ASVs, and ~~34.4~~43.0 million reads that were retained by this workflow (Figs. 1 and 2 and Table 3). To our knowledge it is the only global database for marine diazotrophs detected using *nifH* HTS amplicon sequencing, with comprehensive, standardized ancillary data (Fig. 2 and Supplementary Table 1).

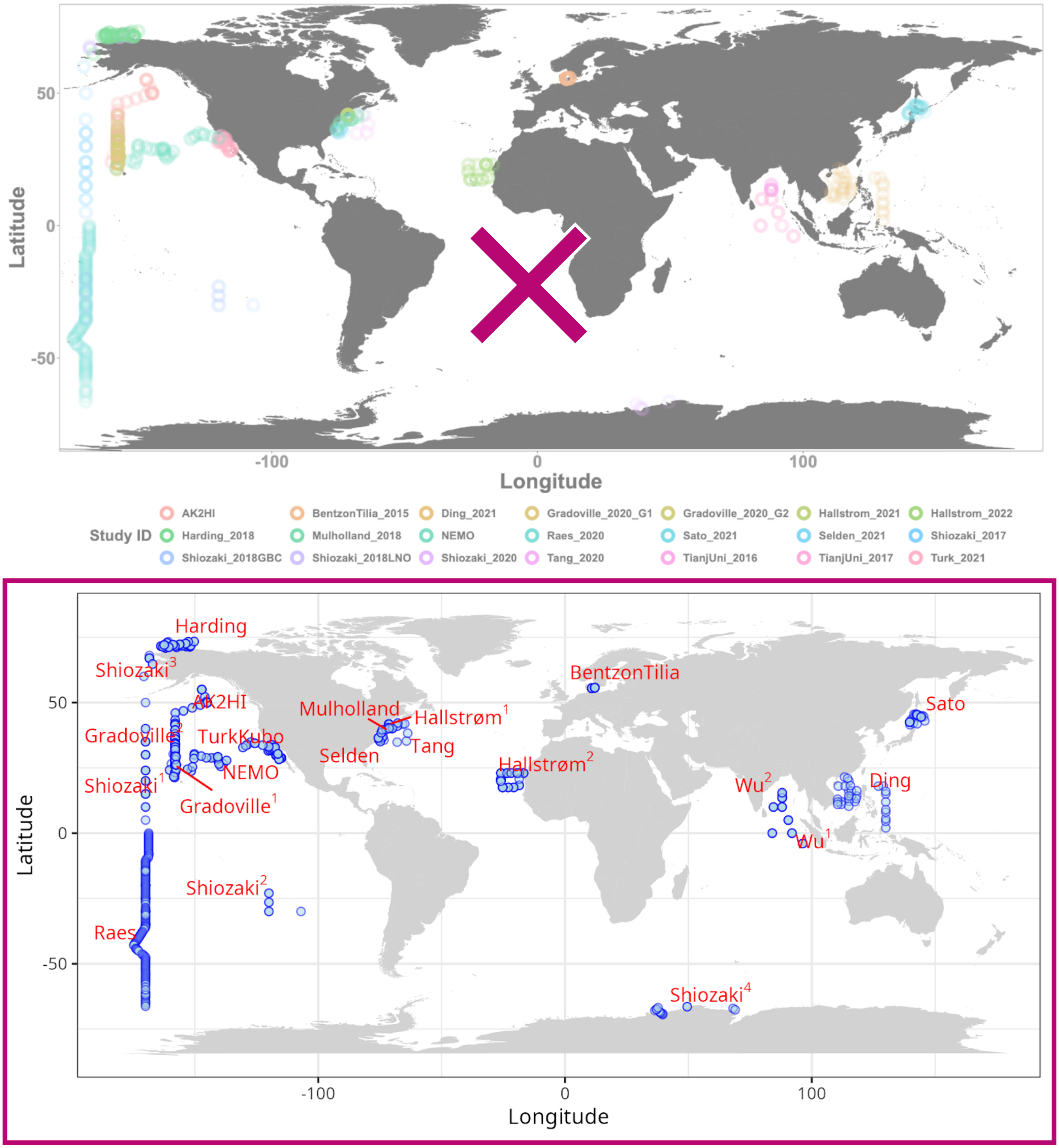


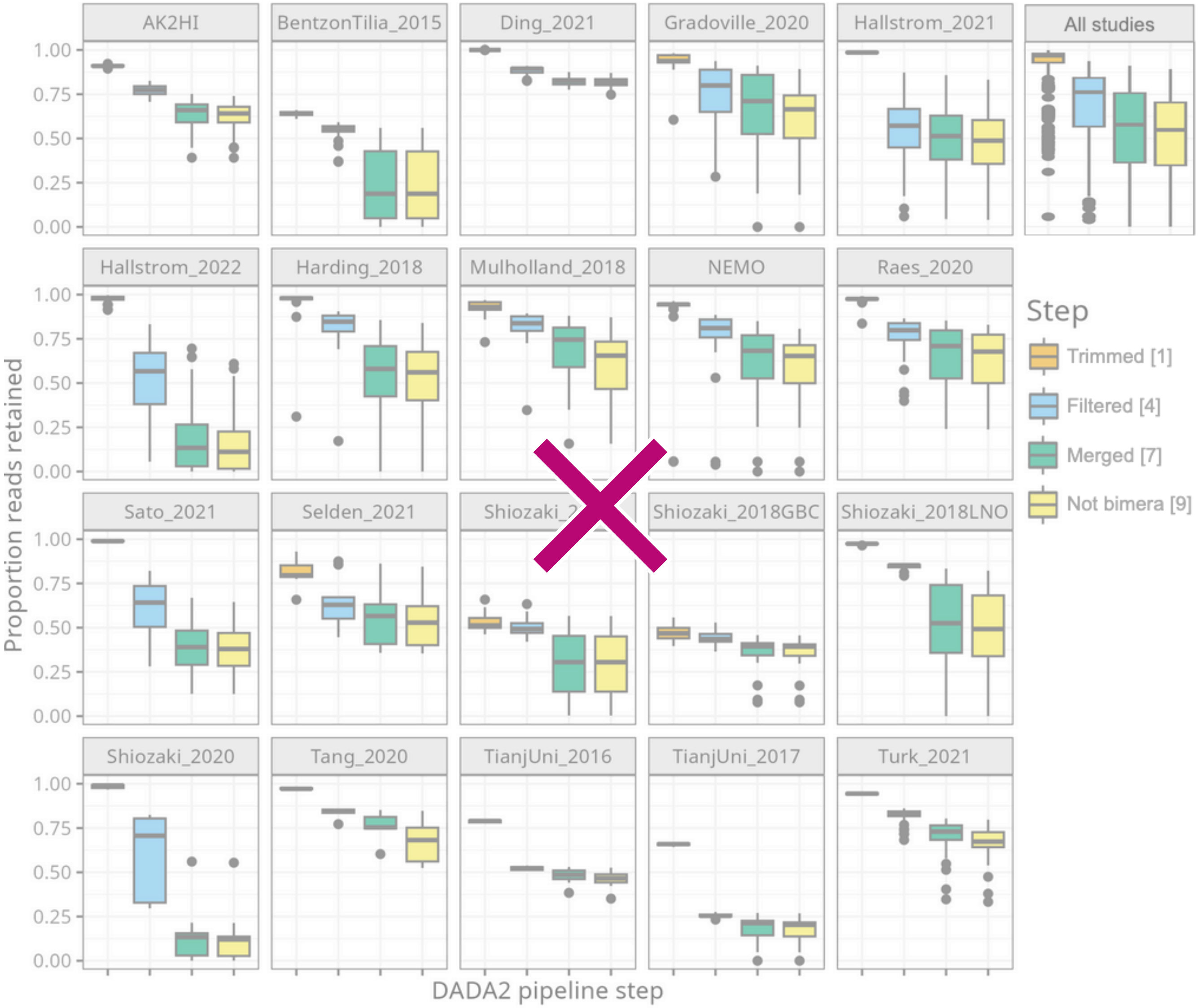
Figure 2: Global sampling distribution of the *nifH* ASV database. World map of sampling locations for the datasets compiled and processed to construct the *nifH* ASV database. Abbreviated study IDs are used with superscripts ordered by publication year for Shiozaki

(2017, 2018GBC, 2018LNO, and 2020), Hallstrøm (2021 and 2022), and Wu (2019 and 2021). For Gradoville the superscripts indicate Gradients cruises 1 and 2. See Table 1 for the citation source linked to each study ID.

Table 3: Summary of the full *nifH* workflow. The number of samples, ASVs, and reads retained through the entire workflow (the DADA2 *nifH* pipeline and major post-pipeline stages) to create the *nifH* ASV database. The vast majority ASVs that were removed by GatherAsvs fell outside 200–450 nt. WorkspaceStartup removed ASVs with no annotation and samples that had zero reads after ASV filtering.

	Initial	**DADA2 pipeline**	**Gather Asvs**	**FilterAuids**				**Workspace Startup**
				~~<1K~~≤500 reads in sample	**rare**	**non-NifH**	**length**	
Samples	982	982	982	~~894~~951	~~890~~951	~~890~~951	~~890~~951	~~865~~944
ASVs	n/a	~~177,935~~152,915	~~97,205~~139,355	~~97,172~~139,334	~~13,774~~18,193	~~12,479~~16,253	~~9,416~~11,915	~~7,909~~9,383
Reads (millions)	87.7	~~43.3~~48.7	~~38.7~~48.4	~~38.6~~48.4	~~36.4~~45.5	~~36.0~~45.0	~~35.1~~43.8	~~34.4~~43.0

Interestingly, studies were affected differently by each step of the DADA2 *nifH* pipeline (Fig. 3 and Table 4). There were major losses of reads during ASV merging, with several studies retaining <~~25~~40 % of their total reads by the end of the pipeline (i.e., ~~BentzonTilia_2015,~~ Hallstrom_2022, ~~Shiozaki_2020~~Sato_2021, and ~~TianjUni_2016~~Shiozaki_2020), though on average about ~~half~~60 % of the reads were retained across studies (Fig. 3 and Table 4).


AK2HI
BentzonTilia_2015
Ding_2021
Gradoville_2020
Hallstrom_2021
All studies
Hallstrom_2022
Harding_2018
Mulholland_2018
NEMO
Raes_2020
Sato_2021
Selden_2021
Shiozaki_2
Shiozaki_2018GBC
Shiozaki_2018LNO
Shiozaki_2020
Tang_2020
TianjUni_2016
TianjUni_2017
Turk_2021
1.00
0.75
0.50
0.25
0.00
Proportion reads retained
DADA2 pipeline step
Step
Trimmed [1]
Filtered [4]
Merged [7]
Not bimera [9]

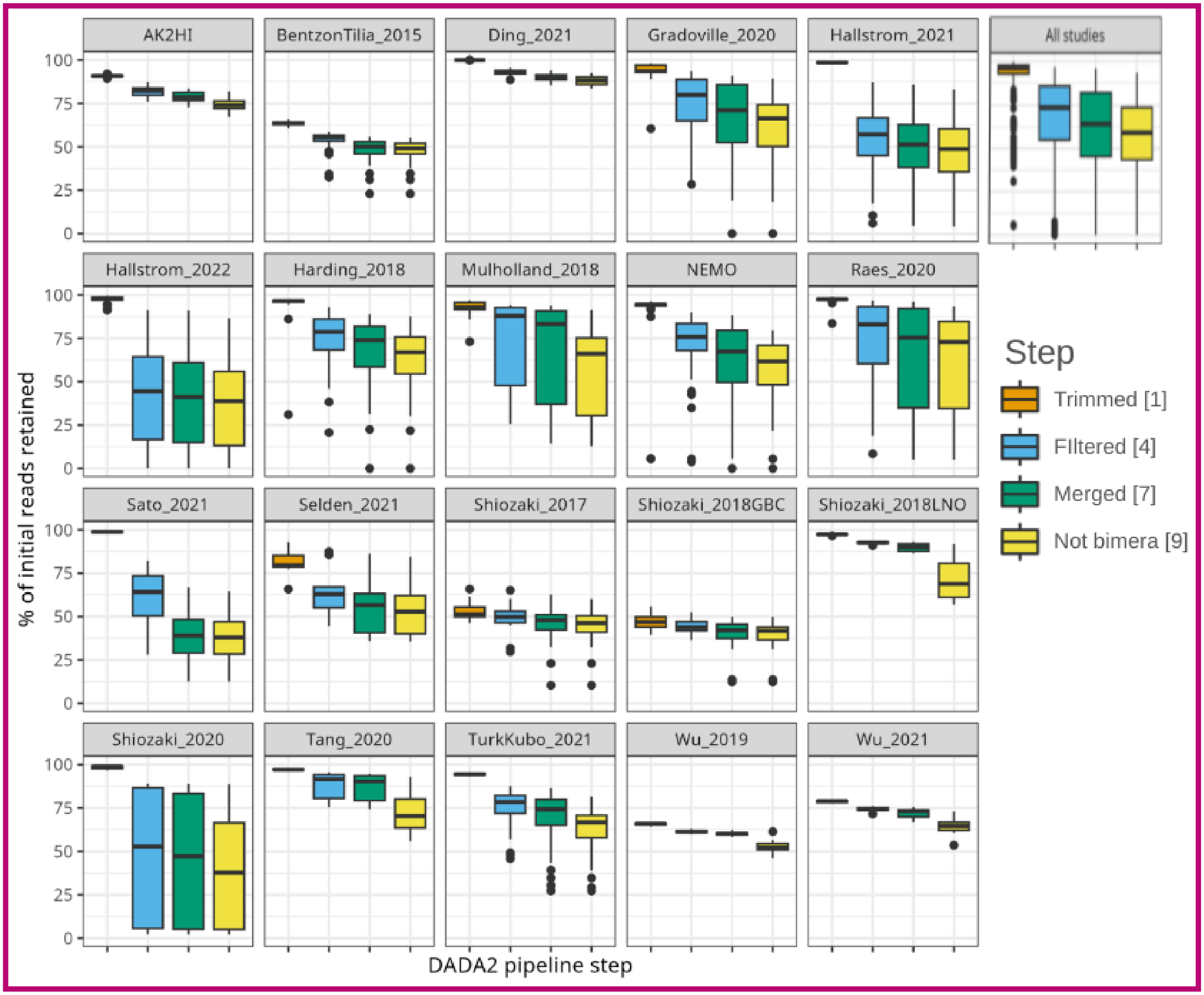


Figure 3: Study-specific retention of reads at each stage of the pipeline. The proportion of total reads in each sample that are retained at the completion of each step of the DADA2 *nifH* pipeline. Each box shows the distribution for samples in the indicated study (using Study IDs in Table 1), or for all samples together (top right). Proportions for Shiozaki_2017 and Shiozaki_2018GBC reflect that approximately half the amplicons were not in the orientation expected by the pipeline (see text). Numbers in the legend indicate pipeline steps in ~~Fig.~~Figure 1.

Table 4: Quality filtering by the DADA2 *nifH* pipeline. For each study ID are shown the mean numbers of reads retained per sample at the end of each stage of the DADA2 *nifH* pipeline, as well as the mean percentage of reads retained. Statistics in the bottom three rows pool all samples. Initial, Trimmed^4, Filtered^4, and Merged^7 and non-Bimera^9 and their superscripts are specific to the pipeline steps in ~~Fig.~~Figure 1. At each step (column) the calculations include only the samples that have >0 reads.

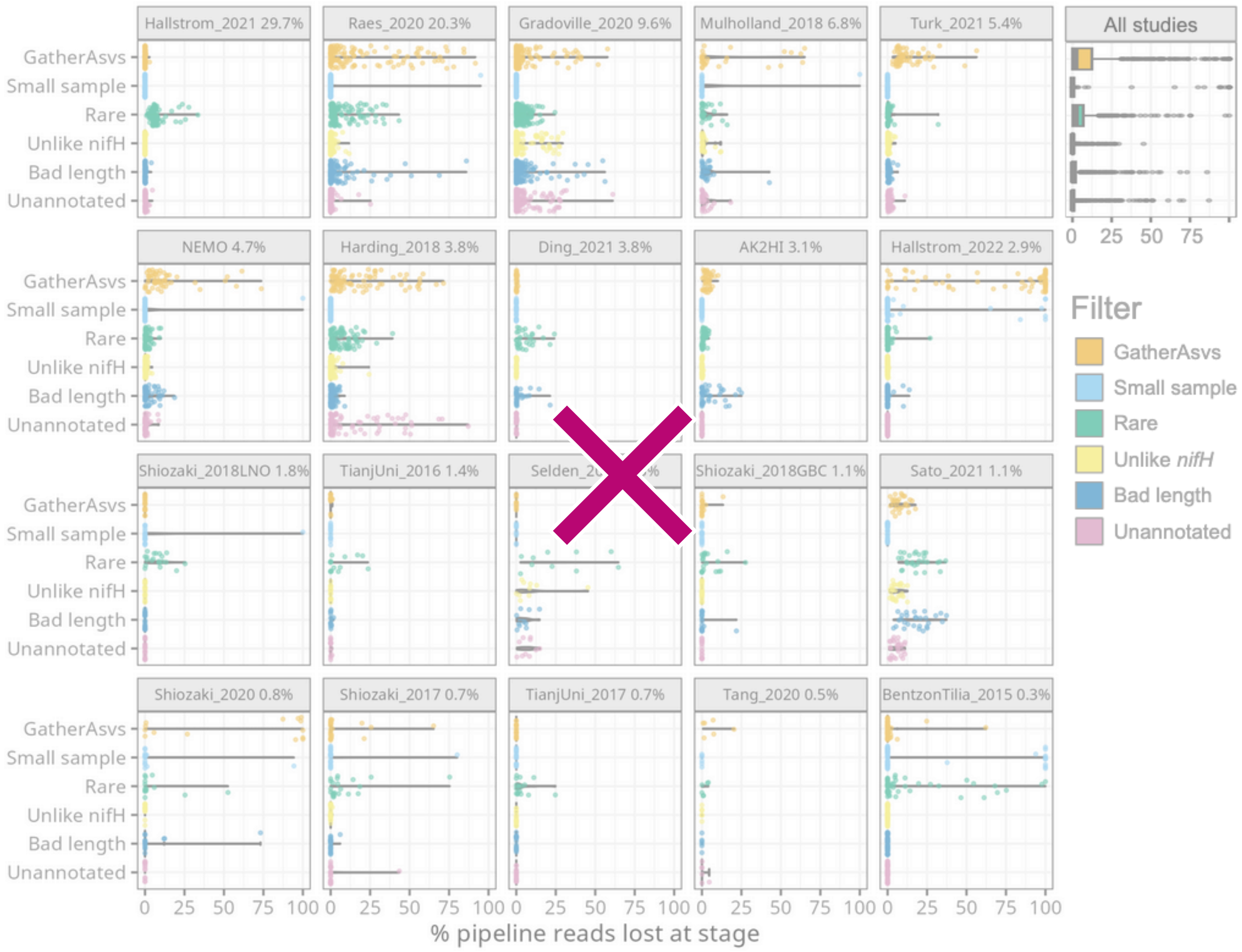
Study	Initial	Trimmed^4	Filtered^4	Merged^9	Non-bimera^9	Retained (%)

AK2HI		4.5E+04	4.1E+04	~~3.5E~~3.7E+04	~~2.8E~~3.6E+04	~~2.8E~~3.3E+04	~~62~~74.1
BentzonTilia_2015		8.2E+03	~~5.3E~~5.2E+03	4.6E+03	~~2.2E~~4.1E+03	~~2.1E~~4.1E+03	~~26~~48.1
Ding_2021		5.6E+04	5.6E+04	~~4.8E~~5.2E+04	~~4.5E~~5.0E+04	~~4.5E~~4.9E+04	~~82~~88.1
Gradoville_2020		4.0E+04	3.8E+04	2.9E+04	2.6E+04	2.4E+04	~~61~~60.3
Hallstrom_2021		2.5E+05	2.5E+05	1.5E+05	1.4E+05	1.4E+05	~~49~~48.7
Hallstrom_2022		2.0E+05	1.9E+05	~~1.0E+05~~7.5E+04	~~5.4E~~7.4E+04	~~4.6E~~6.6E+04	~~19~~36.3
Harding_2018		4.2E+04	4.1E+04	~~3.5E~~3.1E+04	~~2.4E~~2.9E+04	~~2.3E~~2.6E+04	~~54~~63.2
Mulholland_2018		1.8E+05	1.6E+05	~~1.5E~~1.3E+05	1.2E+05	~~1.1E~~1.0E+05	~~61~~58.5
NEMO		5.7E+04	5.4E+04	~~4.6E~~4.2E+04	~~3.7E~~3.6E+04	~~3.5E~~3.3E+04	~~60~~57.1
Raes_2020		9.3E+04	9.1E+04	~~7.4E~~7.7E+04	~~6.6E~~6.9E+04	~~6.3E~~6.5E+04	~~63~~61.0
Sato_2021		7.5E+04	7.4E+04	4.5E+04	2.9E+04	2.9E+04	~~39~~38.8
Selden_2021		1.5E+05	1.2E+05	9.2E+04	8.2E+04	8.0E+04	~~55~~54.7
Shiozaki_2017		1.8E+04	9.3E+03	8.9E+03	~~5.8E~~8.4E+03	~~5.8E~~8.2E+03	~~28~~44.1
Shiozaki_2018GBC		2.4E+04	1.1E+04	1.1E+04	~~9.2E+03~~1.0E+04	~~9.1E~~9.8E+03	~~35~~38.6
Shiozaki_2018LNO		6.7E+04	6.5E+04	~~5.6E~~6.2E+04	~~3.5E~~6.0E+04	~~3.3E~~4.8E+04	~~49~~71.5
Shiozaki_2020		2.5E+05	2.5E+05	~~1.9E~~1.8E+05	~~3.4E+04~~1.8E+05	~~3.3E+04~~1.4E+05	~~12~~39.1
Tang_2020		4.7E+04	4.6E+04	~~3.9E~~4.1E+04	~~3.5E~~4.0E+04	~~3.2E~~3.4E+04	~~67~~72.4
~~TianjUni_2016~~TurkKubo_2021		~~8.0E~~5.5E+04	~~6.3E~~5.2E+04	4.2E+04	~~3.9E~~4.0E+04	~~3.7E~~3.6E+04	~~46~~63.2
~~TianjUni_2017~~Wu_2019		8.0E+04	5.3E+04	~~2.0E~~4.9E+04	~~1.5E~~4.8E+04	~~1.4E~~4.2E+04	~~18~~52.9
~~Turk_2021~~Wu_2021		~~5.5E~~8.0E+04	~~5.2E~~6.3E+04	~~4.6E~~6.0E+04	~~4.0E~~5.8E+04	~~3.7E~~5.2E+04	~~66~~64.4
All samples	**mean**	8.9E+04	8.5E+04	~~6.1E~~5.8E+04	~~4.8E~~5.4E+04	~~4.5E~~4.9E+04	~~52~~56.9

and **studies**	**median**	5.1E+04	4.8E+04	~~3.8E~~3.7E+04	~~2.9E~~3.2E+04	~~2.8E~~3.0E+04	~~56~~59.0
	sum	8.8E+07	8.4E+07	~~5.9E~~5.7E+07	~~4.6E~~5.3E+07	~~4.3E~~4.8E+07	~~49~~60.0

~~Switching the trimming approach from one based on individual read quality profiles (using truncQ in Table 3) to fixed-length trimming based on overall quality profiles of the forward and reverse reads (using truncLen.fwd and truncLen.rev in Table 2) resulted in more reads being retained for some studies (Sato et al., 2021; Selden et al., 2021; Hallstrøm et al., 2022b; Gradoville et al., 2020). However, fixed-length trimming would have required the selection of trim lengths based on visual, qualitative assessments of hundreds of FASTQ quality plots which is difficult to accomplish in a systematic manner. For consistency we preferred to use nearly identical parameters for most studies (Table 3). ¶~~

Post-pipeline stages of the workflow further refined the data (detailed in Methods) (Fig. 4). First, GatherAsvs identified and removed ~~112~~163 chimeras using uchime3 denovo (distinct from the bimera filtering done by the pipeline), and then removed ~~81~~8.7 K ASVs that were far outside expected *nifH* lengths (200–450 nt). AUIDs were assigned to the remaining ~~97~~139 K unique non-chimeric ASVs (comprising ~~38.7~~48.4 million total reads; Tables 3 and 5). The ~~GatherAsvs length filter had by far the largest impact of any post-pipeline quality filtering, removing 10 % of the reads from the pipeline. Next, FilterAuids dropped four poorly sequenced samples (7 K total reads), as they would likely misrepresent their diazotrophic communities, and then removed 83 K rare ASVs (2.3 million~~FilterAuids stage had the largest impacts on retained data. Thirty-one samples with ≤500 reads were removed because they would likely misrepresent their diazotrophic communities. The FilterAuids rarity check had the greatest reduction to pipeline outputs (121 K ASVs removed and 6.0 % of reads), followed by the length filter (4 K ASVs and 2.7 % of reads; Tables 3 and 5).


Hallstrom_2021 29.7%
Raes_2020 20.3%
Gradoville_2020 9.6%
Mulholland_2018 6.8%
Turk_2021 5.4%
All studies
NEMO 4.7%
Harding_2018 3.8%
Ding_2021 3.8%
AK2HI 3.1%
Hallstrom_2022 2.9%
Shiozaki_2018LNO 1.8%
TianjUni_2016 1.4%
Selden_20
Shiozaki_2018GBC 1.1%
Sato_2021 1.1%
Shiozaki_2020 0.8%
Shiozaki_2017 0.7%
TianjUni_2017 0.7%
Tang_2020 0.5%
BentzonTilia_2015 0.3%
GatherAsvs
Small sample
Rare
Unlike nifH
Bad length
Unannotated
0
25
50
75
100
% pipeline reads lost at stage
Filter
GatherAsvs
Small sample
Rare
Unlike *nifH*
Bad length
Unannotated

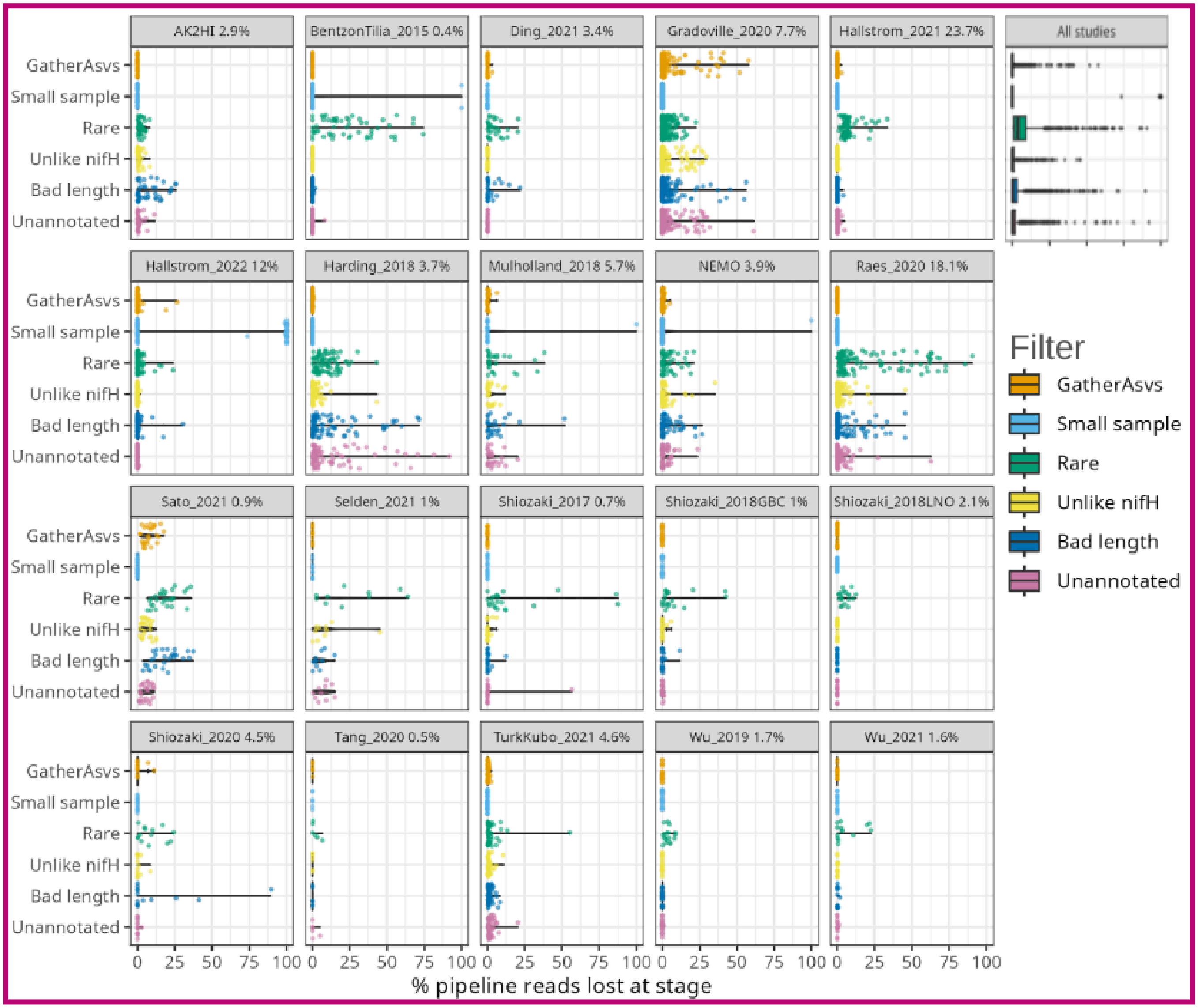


Figure 4: Study-specific ~~retention~~loss of reads at each stage of the post-pipeline workflow. For each study the violin plots show how many reads from the pipeline were removed by GatherAsvs due to length, the four filtering steps of FilterAuids, or WorkspaceStartup due to the ASV having no annotation (shown in Fig. 1). Losses for all samples combined are shown in the box plot (top right). ~~Studies are ordered by contribution~~Bracketed numbers after each study ID indicate the percentage of reads contributed to the *nifH* ASV database, e.g. ~~29.7~~23.7 % of all the reads in the database were from Hallstrom_2021.

Table 5. Quality filtering by the post-pipeline workflow. For each study are shown the mean numbers of reads per sample that were output by the DADA2 *nifH* pipeline and retained by the GatherAsvs, FilterAuids, and WorkspaceStartup stages of the post-pipeline workflow. The Retained (%) column has the mean percentages of reads retained per sample (relative to column DADA2 pipeline values).

Additionally, the last three rows show the overall means, medians, and sums of reads across all samples and studies. Superscripts correspond to stage numbers in Fig. 1 Post-pipeline stages. The GatherAsvs[1] column mainly reflects length filtering (200–450 nt), and the WorkspaceStartup[6] column reflects discarding of ASVs that had no annotation. At each stage (column) the calculations include only the samples that have >0 reads.

¶

~~Study ID¶~~		~~DADA2 pipeline¶~~	~~Gather Asvs[1]¶~~	~~FilterAuids[2]¶~~			~~Workspace Startup[6]¶~~	~~Retained (%)¶~~
				~~Rare¶~~	~~Non-NifH¶~~	~~Length¶~~		
~~AK2HI¶~~		~~2.8E+04¶~~	~~2.7E+04¶~~	~~2.7E+04¶~~	~~2.7E+04¶~~	~~2.5E+04¶~~	~~2.5E+04¶~~	~~90¶~~
~~BentzonTilia_2015¶~~		~~2.1E+03¶~~	~~2.1E+03¶~~	~~2.6E+03¶~~	~~2.6E+03¶~~	~~2.6E+03¶~~	~~2.6E+03¶~~	~~85¶~~
~~Ding_2021¶~~		~~4.5E+04¶~~	~~4.5E+04¶~~	~~4.2E+04¶~~	~~4.2E+04¶~~	~~4.1E+04¶~~	~~4.1E+04¶~~	~~91¶~~
~~Gradoville_2020¶~~		~~2.4E+04¶~~	~~2.3E+04¶~~	~~2.2E+04¶~~	~~2.1E+04¶~~	~~2.1E+04¶~~	~~2.0E+04¶~~	~~80¶~~
~~Hallstrom_2021¶~~		~~1.4E+05¶~~	~~1.4E+05¶~~	~~1.3E+05¶~~	~~1.3E+05¶~~	~~1.2E+05¶~~	~~1.2E+05¶~~	~~92¶~~
~~Hallstrom_2022¶~~		~~4.6E+04¶~~	~~2.6E+04¶~~	~~3.8E+04¶~~	~~3.8E+04¶~~	~~3.4E+04¶~~	~~3.4E+04¶~~	~~50¶~~
~~Harding_2018¶~~		~~2.3E+04¶~~	~~1.9E+04¶~~	~~1.8E+04¶~~	~~1.7E+04¶~~	~~1.7E+04¶~~	~~1.5E+04¶~~	~~64¶~~
~~Mulholland_2018¶~~		~~1.1E+05¶~~	~~9.3E+04¶~~	~~9.3E+04¶~~	~~9.1E+04¶~~	~~8.7E+04¶~~	~~8.4E+04¶~~	~~72¶~~
~~NEMO¶~~		~~3.5E+04¶~~	~~3.1E+04¶~~	~~3.1E+04¶~~	~~3.1E+04¶~~	~~3.0E+04¶~~	~~3.0E+04¶~~	~~80¶~~
~~Raes_2020¶~~		~~6.3E+04¶~~	~~5.8E+04¶~~	~~5.6E+04¶~~	~~5.6E+04¶~~	~~5.6E+04¶~~	~~6.0E+04¶~~	~~76¶~~
~~Sato_2021¶~~		~~2.9E+04¶~~	~~2.7E+04¶~~	~~2.1E+04¶~~	~~2.0E+04¶~~	~~1.5E+04¶~~	~~1.4E+04¶~~	~~43¶~~
~~Selden_2021¶~~		~~8.0E+04¶~~	~~8.0E+04¶~~	~~6.0E+04¶~~	~~5.2E+04¶~~	~~4.9E+04¶~~	~~4.4E+04¶~~	~~52¶~~
~~Shiozaki_2017¶~~		~~1.2E+04¶~~	~~1.2E+04¶~~	~~1.1E+04¶~~	~~1.1E+04¶~~	~~1.1E+04¶~~	~~1.1E+04¶~~	~~83¶~~
~~Shiozaki_2018GBC¶~~		~~2.0E+04¶~~	~~2.0E+04¶~~	~~2.0E+04¶~~	~~2.0E+04¶~~	~~2.0E+04¶~~	~~2.0E+04¶~~	~~93¶~~
~~Shiozaki_2018LNO¶~~		~~3.3E+04¶~~	~~3.3E+04¶~~	~~3.3E+04¶~~	~~3.3E+04¶~~	~~3.3E+04¶~~	~~3.3E+04¶~~	~~92¶~~
~~Shiozaki_2020¶~~		~~3.3E+04¶~~	~~2.8E+04¶~~	~~4.2E+04¶~~	~~4.2E+04¶~~	~~5.7E+04¶~~	~~5.7E+04¶~~	~~61¶~~
~~Tang_2020¶~~		~~3.2E+04¶~~	~~3.0E+04¶~~	~~2.9E+04¶~~	~~2.9E+04¶~~	~~2.9E+04¶~~	~~2.9E+04¶~~	~~91¶~~
~~TianjUni_2016¶~~		~~3.7E+04¶~~	~~3.7E+04¶~~	~~3.5E+04¶~~	~~3.5E+04¶~~	~~3.5E+04¶~~	~~3.5E+04¶~~	~~93¶~~
~~TianjUni_2017¶~~		~~1.4E+04¶~~	~~1.4E+04¶~~	~~1.4E+04¶~~	~~1.4E+04¶~~	~~1.4E+04¶~~	~~1.4E+04¶~~	~~96¶~~
~~Turk_2021¶~~		~~3.7E+04¶~~	~~3.3E+04¶~~	~~3.3E+04¶~~	~~3.2E+04¶~~	~~3.2E+04¶~~	~~3.2E+04¶~~	~~83¶~~
~~All samples and studies¶~~	~~mean¶~~	~~4.5E+04¶~~	~~4.2E+04¶~~	~~4.1E+04¶~~	~~4.0E+04¶~~	~~4.0E+04¶~~	~~4.0E+04¶~~	~~79¶~~
	~~median¶~~	~~2.8E+04¶~~	~~2.6E+04¶~~	~~2.6E+04¶~~	~~2.6E+04¶~~	~~2.5E+04¶~~	~~2.6E+04¶~~	~~90¶~~
	~~sum¶~~	~~4.3E+07¶~~	~~3.9E+07¶~~	~~3.6E+07¶~~	~~3.6E+07¶~~	~~3.5E+07¶~~	~~3.4E+07¶~~	~~79¶~~

Study ID	DADA2 pipeline	Gather Asvs[1]	FilterAuids[2]				Workspace Startup[6]	Retained (%)
			Small	Rare	Non-NifH	Length		
AK2HI	3.3E+04	3.3E+04	3.3E+04	3.3E+04	3.2E+04	3.0E+04	2.9E+04	89.2
BentzonTilia_2015	4.1E+03	4.1E+03	4.0E+03	3.1E+03	3.1E+03	3.1E+03	3.0E+03	72.8
Ding_2021	4.9E+04	4.9E+04	4.9E+04	4.6E+04	4.6E+04	4.5E+04	4.5E+04	92.2
Gradoville_2020	2.4E+04	2.3E+04	2.3E+04	2.2E+04	2.1E+04	2.1E+04	2.0E+04	82.6

Hallstrom_2021		1.4E+05	1.4E+05	1.4E+05	1.3E+05	1.3E+05	1.2E+05	1.2E+05	92.2
Hallstrom_2022		6.6E+04	6.5E+04	6.5E+04	6.4E+04	6.4E+04	6.2E+04	6.2E+04	68.1
Harding_2018		2.6E+04	2.6E+04	2.6E+04	2.4E+04	2.3E+04	2.0E+04	1.7E+04	75.6
Mulholland_2018		1.0E+05	1.0E+05	1.0E+05	9.5E+04	9.3E+04	8.8E+04	8.4E+04	80.0
NEMO		3.3E+04	3.3E+04	3.3E+04	3.2E+04	3.2E+04	3.0E+04	3.0E+04	84.2
Raes_2020		6.5E+04	6.5E+04	6.5E+04	6.1E+04	6.1E+04	6.0E+04	5.9E+04	75.3
Sato_2021		2.9E+04	2.7E+04	2.7E+04	2.2E+04	2.0E+04	1.5E+04	1.4E+04	49.2
Selden_2021		8.0E+04	8.0E+04	8.0E+04	6.0E+04	5.2E+04	4.9E+04	4.5E+04	59.0
Shiozaki_2017		1.6E+04	1.6E+04	1.6E+04	1.5E+04	1.5E+04	1.4E+04	1.4E+04	82.5
Shiozaki_2018GBC		2.2E+04	2.2E+04	2.2E+04	2.1E+04	2.1E+04	2.1E+04	2.1E+04	90.4
Shiozaki_2018LNO		4.8E+04	4.8E+04	4.8E+04	4.6E+04	4.6E+04	4.6E+04	4.6E+04	95.0
Shiozaki_2020		1.4E+05	1.4E+05	1.4E+05	1.4E+05	1.4E+05	1.4E+05	1.4E+05	76.6
Tang_2020		3.4E+04	3.4E+04	3.4E+04	3.3E+04	3.3E+04	3.3E+04	3.3E+04	97.9
TurkKubo_2021		3.6E+04	3.5E+04	3.5E+04	3.5E+04	3.5E+04	3.4E+04	3.3E+04	94.1
Wu_2019		4.2E+04	4.2E+04	4.2E+04	4.1E+04	4.1E+04	4.1E+04	4.1E+04	96.3
Wu_2021		5.2E+04	5.2E+04	5.2E+04	4.8E+04	4.8E+04	4.8E+04	4.8E+04	93.2
All samples and studies	**mean**	5.0E+04	4.9E+04	4.9E+04	4.6E+04	4.6E+04	4.5E+04	4.4E+04	80.9
	median	3.0E+04	3.0E+04	3.0E+04	2.9E+04	2.8E+04	2.7E+04	2.6E+04	93.0
	sum	4.9E+07	4.8E+07	4.8E+07	4.6E+07	4.5E+07	4.4E+07	4.3E+07	90.0

Finally, ASVs were removed if they were classified as non-*nifH*, based on a strong alignment to sequences in NCBI nr that ARBitrator (Heller et al., 2014) classified as non-*nifH*. Specifically, an ASV was classified as non-*nifH* if the ratio of E-values for its best ~~negative and positive~~positive and negative hits, among sequences classified by ARBitrator, was >10. A total of ~~96,095~~137,366 of the ~~97,205~~139,355 non-chimera ASVs had database hits which resulted in ~~40,448~~50,233 positive, ~~12,977~~20,528 negative, and ~~42,670~~66,605 uncertain classifications. This approach was used to leverage ARBitrator's high specificity for detecting *nifH* as well as to enable users to identify ASVs that have high percent identity matches to sequences in GenBank. An alternative approach would have been to classify the ASVs based on their alignments to HMMs for NifH versus NifH-like proteins (e.g. protochlorophyllide reductase), used by the NifMAP pipeline for *nifH* operational taxonomic units (Angel et al., 2018). Finally, FilterAuids removed ASVs with lengths outside 281–359 nt, a total of ~~974 K reads and 3063 ASVs~~4338 ASVs comprising 1.2 million reads (Figs. 1, 4 and Tables 3 and 5). After FilterAUIDs, the total number of samples in the dataset was reduced from 982 to ~~890~~951 and the number of ASVs from ~~97,205 to 9416~~139,355 to 11,915.

FilterAuids also flagged a total of ~~2000~~2342 ASVs as possible PCR contaminants. Although we opted to flag, not remove, these ASVs, the workflow can be easily altered to remove contaminants. Most studies contained low levels of contamination (≤1 %) based on our criteria. However, several studies were flagged with ~9–~~30~~29 % of their reads being similar to known contaminants. Identifying potential contaminants is challenging given their numerous sources, study specific nature (Zehr et al., 2003), and lack of control sequence data from blanks.

Next, AnnotateAuids assigned annotations using our three *nifH* reference databases and CART (Fig. 1). In total ~~7931~~9406 of the ~~9416~~11,915 quality filtered ASVs were annotated, usually with multiple references (Fig. A1). Most (~~7926~~9322 ASVs) had hits to both genome879 and ARB2017, likely because the 879 sequenced diazotrophs had *nifH* homologs in GenBank that were found by ARBitrator. Fewer ASVs had hits to the databases that targeted UCYN-A oligos (~~102~~217 ASVs) and other marine diazotrophs (~~645~~938 ASVs; ~~96~~211 ASVs also had UCYN-A hits). Most ASVs (~~7905~~9380 total) were assigned to NifH clusters 1–4 by CART (respectively, ~~4100; 79; 3607~~4923; 101; 4205; and ~~109~~151 ASVs), including five ASVs that had no hits to our databases. The majority of ASVs (~~7749~~9257 total) had open reading frames (ORFs) that contained paired cysteines and AMP which might coordinate the 4Fe-4S cluster, and all ~~7749~~9257 also had ~~annotations~~annotation from the reference databases or CART. A few ASVs had annotations but lacked residues to coordinate 4Fe-4S: ~~23~~29 ORFs lacked the paired cysteines and another ~~159~~120 ORFs had paired cysteines but not AMP, usually due to a substitution for M. The last step of AnnotateAuids assigned primary IDs (described above) to ~~7908~~9383 ASVs. ~~In~~All of them were retained in the final stage of the post-pipeline workflow, WorkspaceStartup ~~retained these 7908 ASVs. One ASV, which had no phylogroup but did have paired cysteines and AMP, was also retained. In total the *nifH* ASV database had 7909 ASVs comprising 34.4 million reads (Table 3~~(below).

In the CMAP stage, sample collection metadata (date, latitude, longitude, and depth) were used to download CMAP environmental data (~~102~~100 variables) for each sample in the *nifH* ASV database (Fig. 1). The CMAP data will enable analyses of potential factors that influence the global distribution of the diazotrophic community. Aggregated metadata for all samples are available in the *nifH* ASV database (metaTab.csv for sample metadata and cmapTab.csv for environmental data).

The last stage of the post-pipeline workflow is WorkspaceStartup, which generates the *nifH* ASV database (Fig. 1). ASVs with no annotation are removed as well as samples with zero total reads due to ASV filtering steps. The *nifH* ASV database consisted of 21 studies, ~~865~~944 samples, ~~7909~~9383 AVS and ~~34.4~~43.0 million total reads (Tables 3 and 5). The database is heavily biased toward euphotic zone DNA samples, with euphotic heuristically defined as follows: Samples were classified as coastal (< 200 km from a major landmass) or open ocean. Euphotic samples were then identified as those collected above a depth cut off, 50 m for coastal samples and 100 m for open ocean. Samples obtained from DNA (n=~~768~~836) far exceeded those from RNA (n=~~94~~108) extracts. Likewise, a majority of the samples were from the euphotic zone (~~789~~861 compared to

~~73~~83 from the aphotic zone). The database also includes replicate samples (n=~~256~~286) and size fractionated samples (n=~~142~~170).

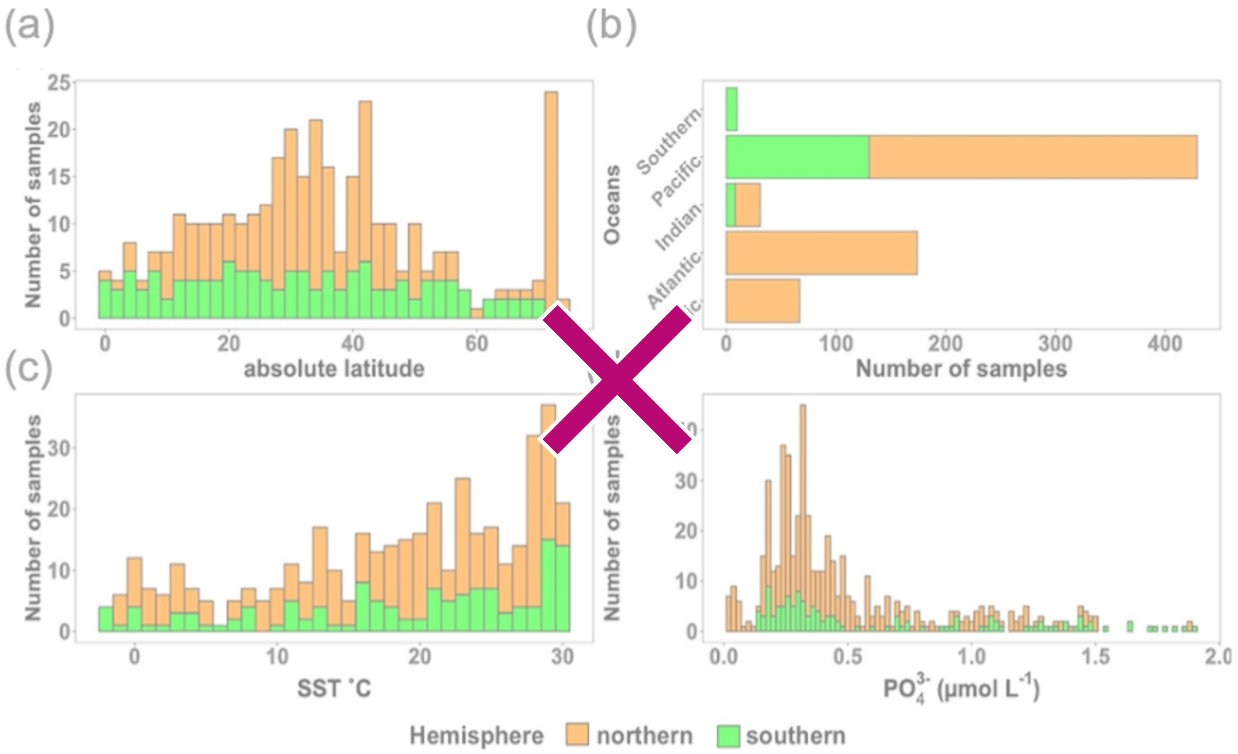
3.2 Global *nifH* ASV database

3.2.1. Comparison to an OTU database

New studies with Illumina amplicon data have mainly used DADA2 (Callahan et al., 2016) and other methods that distinguish fine-scale variation from sequencing errors (Eren et al., 2014; Edgar, 2016b; Amir et al., 2017). Earlier studies, including 13 of the 19 previously published studies in the *nifH* ASV database (Table C1), used *de novo* operational taxonomic units (OTUs) which were obtained by clustering the sequences at 97 % nucleotide identity. OTUs masked sequencing errors as well as fine-scale variation and had other disadvantages compared to ASV approaches (Callahan et al., 2017). Although cross-study comparisons ideally will use the same pipeline for all the studies—the motivation for our workflow—previously published results should be considered. Therefore, for each study in the *nifH* ASV database, diazotroph communities were compared to versions generated using the NifMAP OTU pipeline (Appendix C). The ASV and OTU communities mainly had similar *nifH* clusters, except for several studies where the workflow retained substantially more sequencing reads (Fig. C1, Table C1).

3.2.2. Sample Distribution

Investigations of N_2 fixation and diazotrophic communities have focused on specific ocean regions and this is reflected by the uneven global distribution of *nifH* amplicon datasets in the *nifH* ASV database (Figs. 2, 5a, and 5b). There is an outsized influence of the northern hemisphere, especially in the Pacific Ocean where most of the database samples were located (~~429~~439) and ~~69.7~~68.3 % of these samples originated from the northern hemisphere (Figs. 2, 5a, 5b, and 6). Ten studies are found within the Pacific, with several containing >50 samples (Figs. 2 and 6). Notably, Raes_2020 (Raes et al., 2020) is the largest dataset stretching from the equator to the Southern Ocean, making up almost the entirety of the southern hemisphere Pacific samples (Figs. 2 and 6). Two new studies carried out in the North Pacific constitute the only previously unpublished data of the *nifH* ASV database (Table 1). AK2HI was a latitudinal transect from Alaska (U.S.) to Hawaii (U.S.) and NEMO was a longitudinal transect across the Eastern North Pacific from San Diego, CA (U.S.) to Hawaii (U.S.) (Fig. 2; Sect. 2.2.2). The amplicon data compiled for the *nifH* ASV database was primarily generated from DNA, with most RNA samples deriving from Atlantic Ocean studies and no contribution from RNA samples in the Arctic or Indian Oceans (Fig. 6).



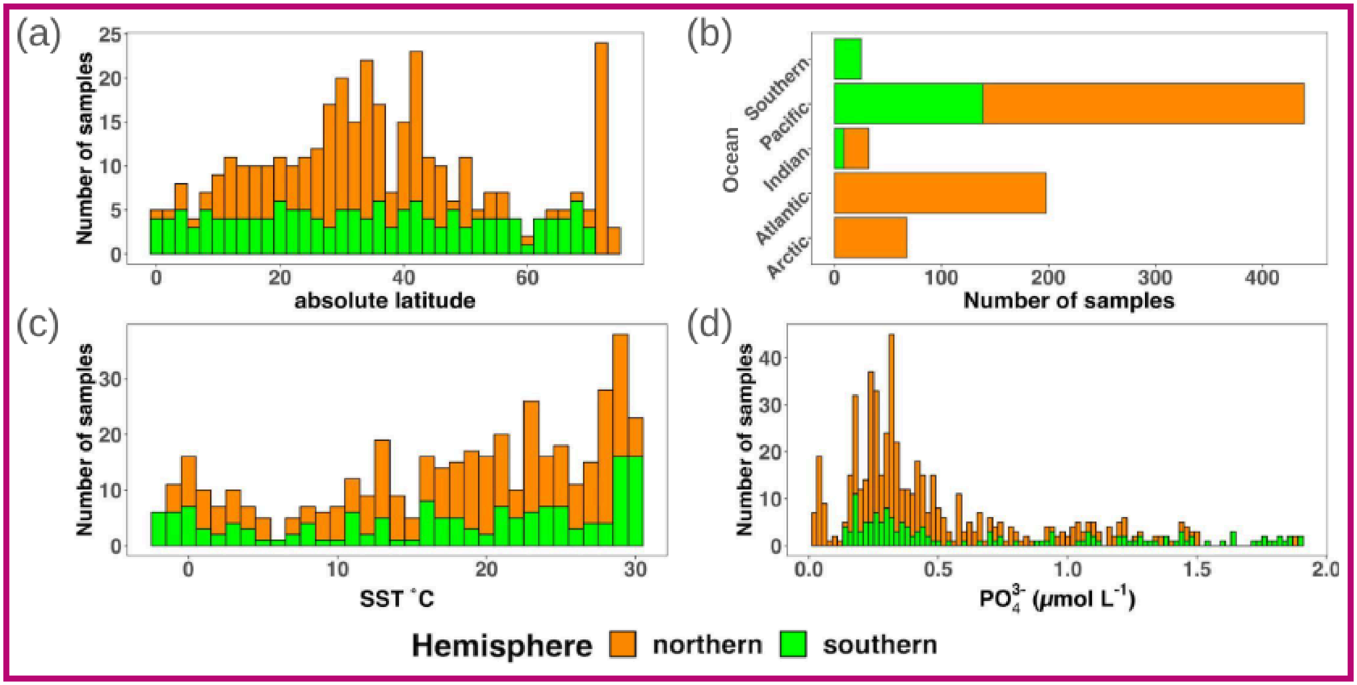
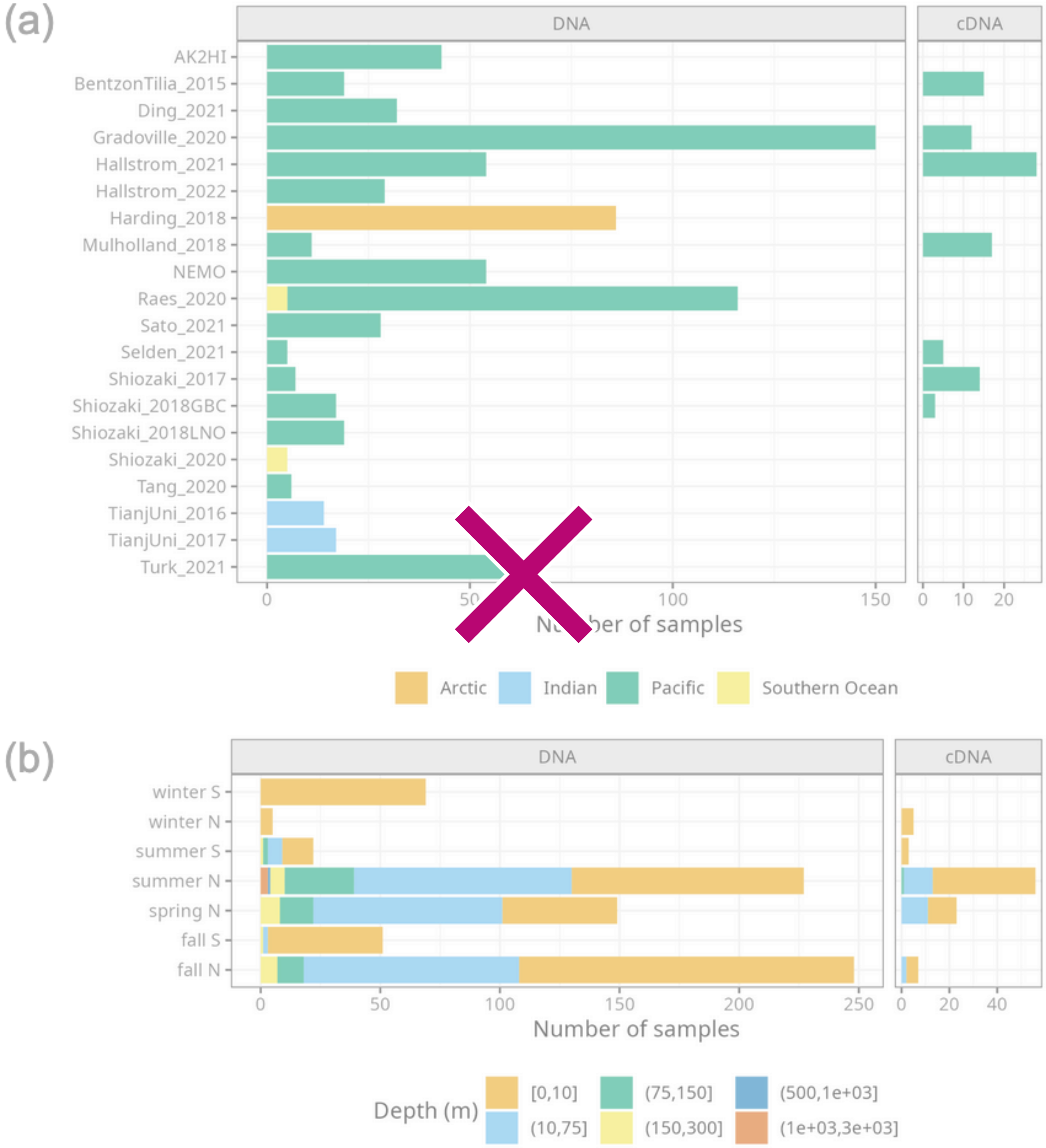


Figure 5. Location, temperature, and phosphate distributions of the *nifH* ASV database. The number of samples from the *nifH* ASV database by (a) absolute latitude, (b) the world's oceans, (c) sea surface temperature (SST, ˚C) and (d) Pisces-derived PO_4^{3-} (µmol L^{-1}). Environmental data, (c) and (d), were retrieved from the CMAP data portal.

¶

All bars are stacked.


(a)
DNA
cDNA
AK2HI
BentzonTilia_2015
Ding_2021
Gradoville_2020
Hallstrom_2021
Hallstrom_2022
Harding_2018
Mulholland_2018
NEMO
Raes_2020
Sato_2021
Selden_2021
Shiozaki_2017
Shiozaki_2018GBC
Shiozaki_2018LNO
Shiozaki_2020
Tang_2020
TianjUni_2016
TianjUni_2017
Turk_2021
0
50
100
150
0
10
20
Number of samples
Arctic
Indian
Pacific
Southern Ocean
(b)
DNA
cDNA
winter S
winter N
summer S
summer N
spring N
fall S
fall N
0
50
100
150
200
250
0
20
40
Number of samples
Depth (m)
[0,10]
(10,75]
(75,150]
(150,300]
(500,1e+03]
(1e+03,3e+03]

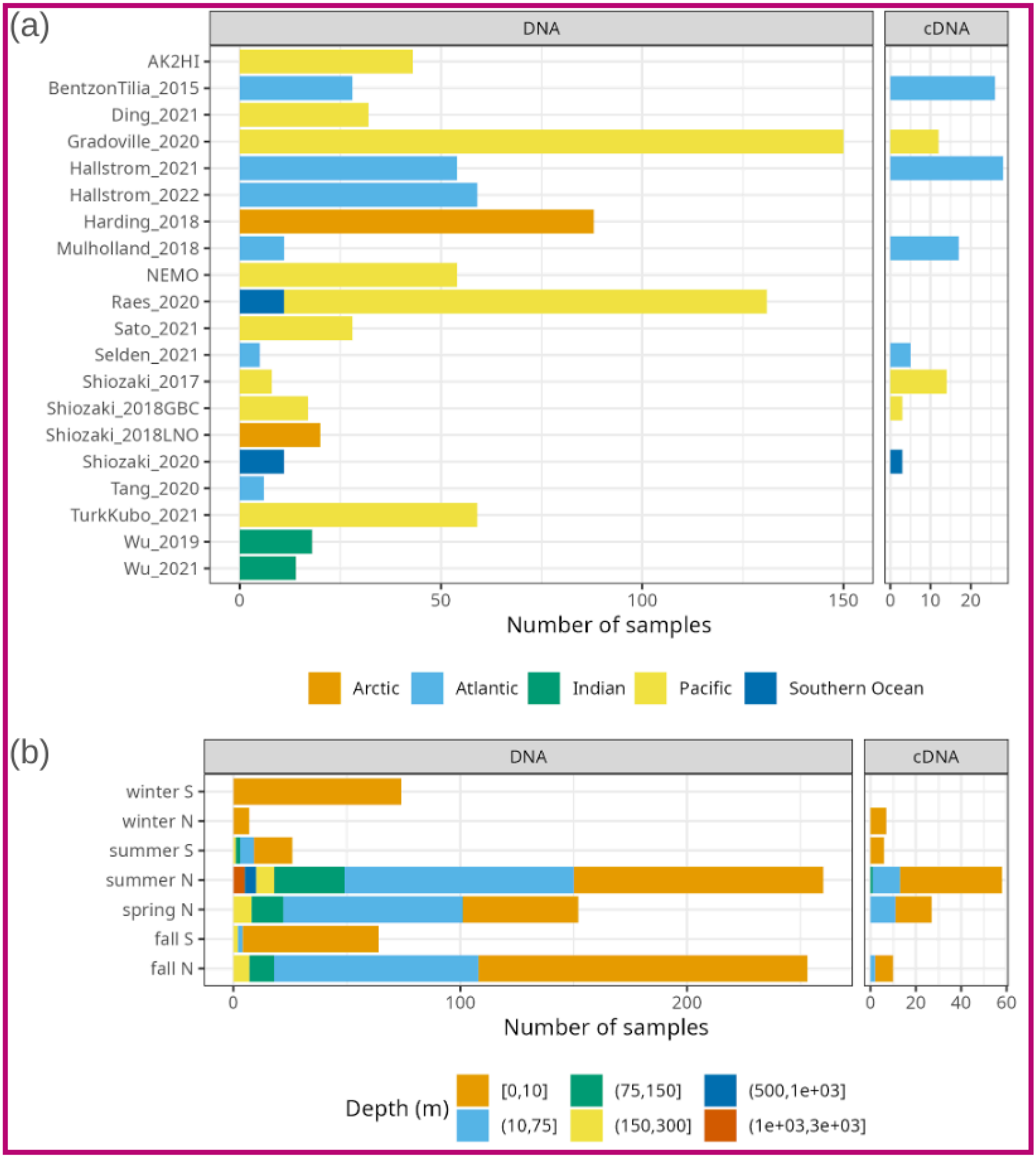


Figure 6. Samples in the *nifH* ASV database by collection location, season, and amplicon type. The number of samples from each study are shown by ocean and study (a), and by the collection season, hemisphere, and depth (b). For both panels the amplicon type (DNA or cDNA) is shown, but *x* axis scales differ between (a) and (b). See Table 1 for citations for the studies in (a). ~~For (b) there were no samples collected between 500 – 1000 m.~~

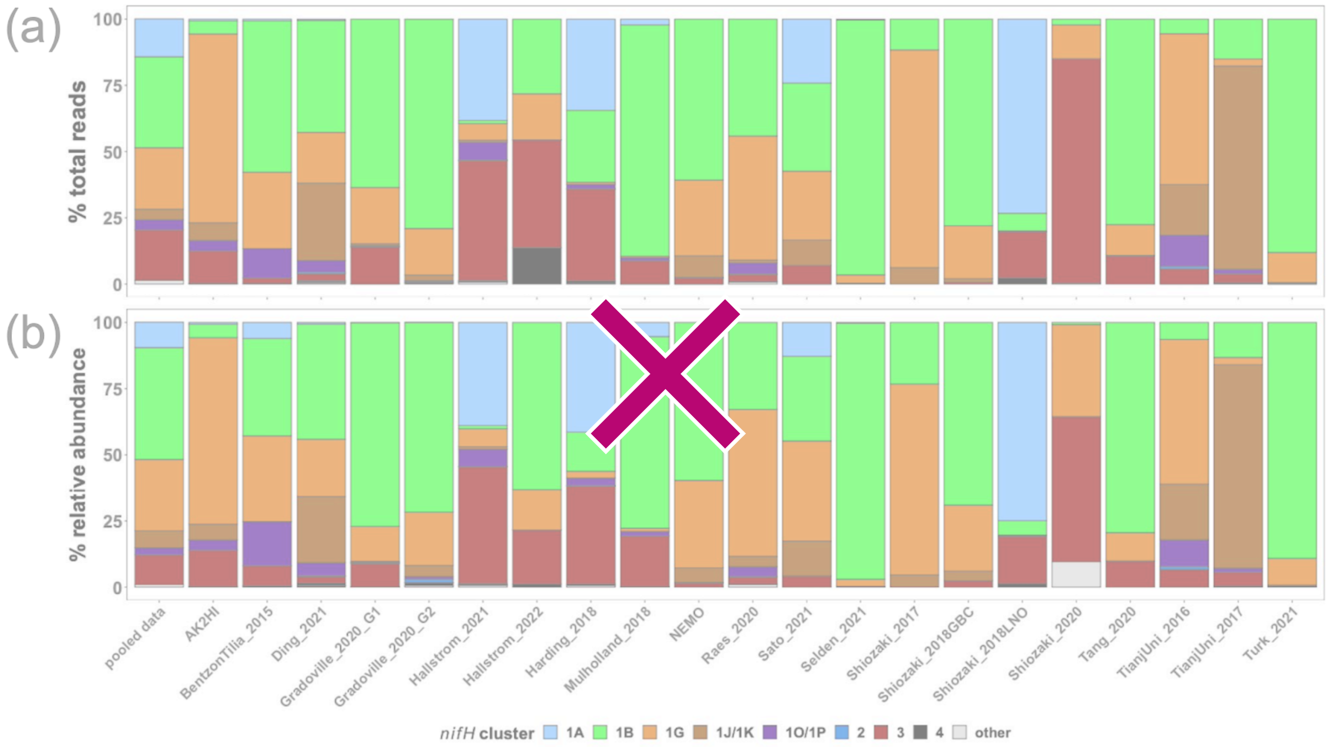
Under-sampled regions include the Eastern South Pacific (n=6) and the Western Indian Ocean (n=0) (Figs. 2, 5a, and 6a). Only two studies originated from the Indian Ocean, a unique environment with intense weather and shifting circulation patterns that include monsoon seasons and upwelling conditions that will require much greater sampling coverage to capture diazotroph biogeography. No South Atlantic samples were found during compilation that met the criteria for inclusion in the *nifH* ASV database, though there are several studies from this region (Table A1). Most Atlantic Ocean samples were coastal and from the North Atlantic. Thus, the Atlantic subtropical gyres, which are known to host diverse diazotrophs (Langlois et al., 2005), are underrepresented by *nifH* amplicon data (Fig. 2).

Tropical and subtropical regions, often associated with high temperatures and low nutrients, are highly represented in the database (Figs. 2 and 5a). This likely influenced the ranges of environmental variables with most samples in the database originating from locations with SST above 15 ˚C and PO_4^{3-} below 0.5 µmol L^{-1} (Figs. 5c and 5d). Northern hemisphere samples were collected in all seasons, though fewer from the winter. In contrast, most southern hemisphere samples were collected in the winter and fall (Fig. 6b). While most DNA samples are from the euphotic zone (Fig. 6b), cDNA samples are almost exclusively from the euphotic zone, and mainly from the northern hemisphere during the spring and summer (Fig. 6b), indicating an incomplete picture of diazotroph activity.

The disproportionate spatial and seasonal coverage between hemispheres in the *nifH* ASV database mirrors collection biases in other N_2 fixation metrics including: N_2 fixation rate measurements; diazotroph cell counts; and *nifH* qPCR data, which are heavily sourced from the North Atlantic (Shao et al., 2023) or, when targeting NCDs, also the North Pacific (Turk-Kubo et al., 2022). These biases underscore the need for future work in understudied regions and seasons.

3.3 Study-specific patterns in global diazotroph assemblages in the DNA dataset

To demonstrate how the *nifH* ASV database can be used, a subset of the data was created that comprised ~~of~~ all DNA samples (~~89.1~~88.8 % of the total dataset; Fig. 7) and referred to herein as the “DNA dataset~~”~~.” Samples derived from cDNA (n=~~94~~108; Fig. 6) were removed. Replicate samples (n=~~256~~286) or those with multiple size fractions (n=~~142~~170) were combined by averaging across replicates or size fractions. This reduced the number of DNA samples to ~~711~~762 and the total number of reads in the count table to ~~30.0~~36.6 million from ~~34.4~~43.0 million.


(a)
% total reads
100
75
50
25
0
(b)
% relative abundance
100
75
50
25
0
pooled data
AK2HI
BentzonTilia_2015
Ding_2021
Gradoville_2020_G1
Gradoville_2020_G2
Hallstrom_2021
Hallstrom_2022
Harding_2018
Mulholland_2018
NEMO
Raes_2020
Sato_2021
Selden_2021
Shiozaki_2017
Shiozaki_2018GBC
Shiozaki_2018LNO
Shiozaki_2020
Tang_2020
TianjUni_2016
TianjUni_2017
Turk_2021
nifH cluster
1A
1B
1G
1J/1K
1O/1P
2
3
4
other

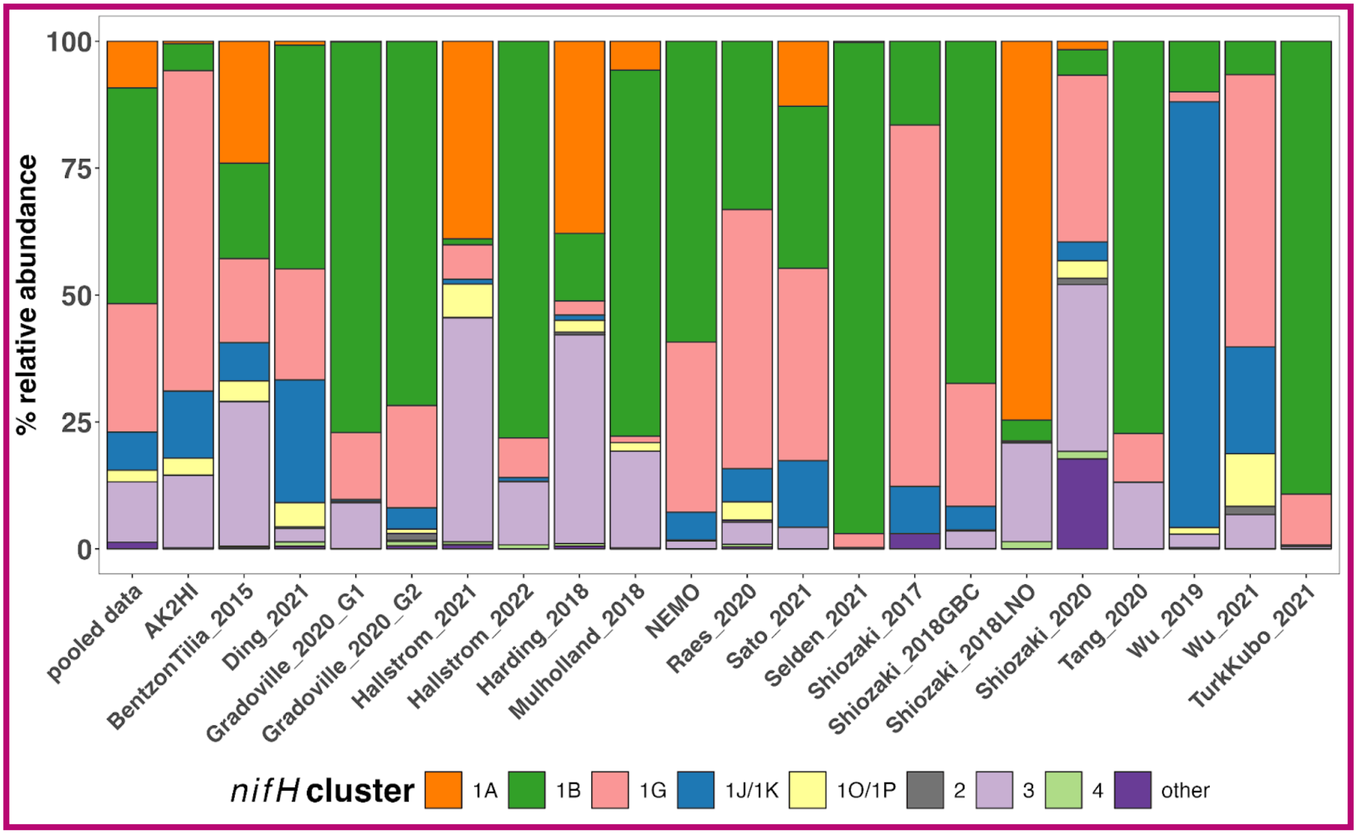


Figure 7. Study-specific diazotroph assemblage patterns in the DNA dataset. The percentage of ~~(a) total reads and (b)~~ relative abundance over the DNA dataset for each major *nifH* cluster. The first column ~~of each panel~~ ('pooled data') uses all the compiled data while each subsequent column only uses data from the indicated study. Colors represent different *nifH* subclusters; 'other' are the remaining *nifH* clusters.

As demonstrated in a previous global analysis of diazotroph assemblages (Farnelid et al., 2011), cyanobacterial sequences (cluster 1B) dominate the samples, making up ~~34 % and~~ 42 % of the total ~~reads and relative abundance, respectively~~relative abundance (Fig. 7). Though photosynthetic cyanobacteria would be expected to thrive in euphotic waters, NCDs are also widespread in the ocean surface (Langlois et al., 2005; Delmont et al., 2018; Delmont et al., 2022; Pierella Karlusich et al., 2021; Turk-Kubo et al., 2022). Indeed, among the NCDs, γ-proteobacteria (*nifH* cluster 1G) were the most prevalent, comprising ~~ca. 23 % of total reads and 27 % of~~ 27 % of the total relative abundance, while δ-proteobacteria (clusters 1A and 3) accounted for ~~33~~21 % of ~~total reads and 21 % of~~the total relative abundance of the DNA dataset (Fig. 7). Less prominent clusters 1J/1K (α- and β-proteobacteria) and 1O/1P (γ-/β-proteobacteria and Deferribacteres) were ~~ca. 4 % and 6 % of the reads and~~ 4 % and 3 % of the relative abundance, respectively. The remaining ASVs comprised <1.5 % of the total ~~reads and~~ relative ~~abundances~~abundance and came from clusters associated with nitrogenases that do not use iron (e.g. cluster 2) or that are uncharacterized (cluster 4) (Fig. 7).

Cluster 1B (cyanobacteria) were generally high in individual studies across the *nifH* DNA dataset, comprising ≥25 % of the ~~relative abundance~~ community in two-thirds of the studies (Fig. 7), which is the highest of any cluster. Studies carried out in polar regions (Harding_2018, Shiozaki_2018LNO, Shiozaki_2020) and the Indian Ocean (~~TianjUni_2016 and TianjUni_2017~~Wu_2019 and Wu_2021) were distinct from this pattern, with low relative abundances of cluster 1B. Instead, Arctic studies had high relative abundances of cluster 1A and 3 (both primarily comprised of δ-proteobacteria) and while clusters 1J/1K (α- and β-proteobacteria) and 1O/1P (γ-/β-proteobacteria and Deferribacteres) were the ~~predominate~~predominant groups in the Indian Ocean.

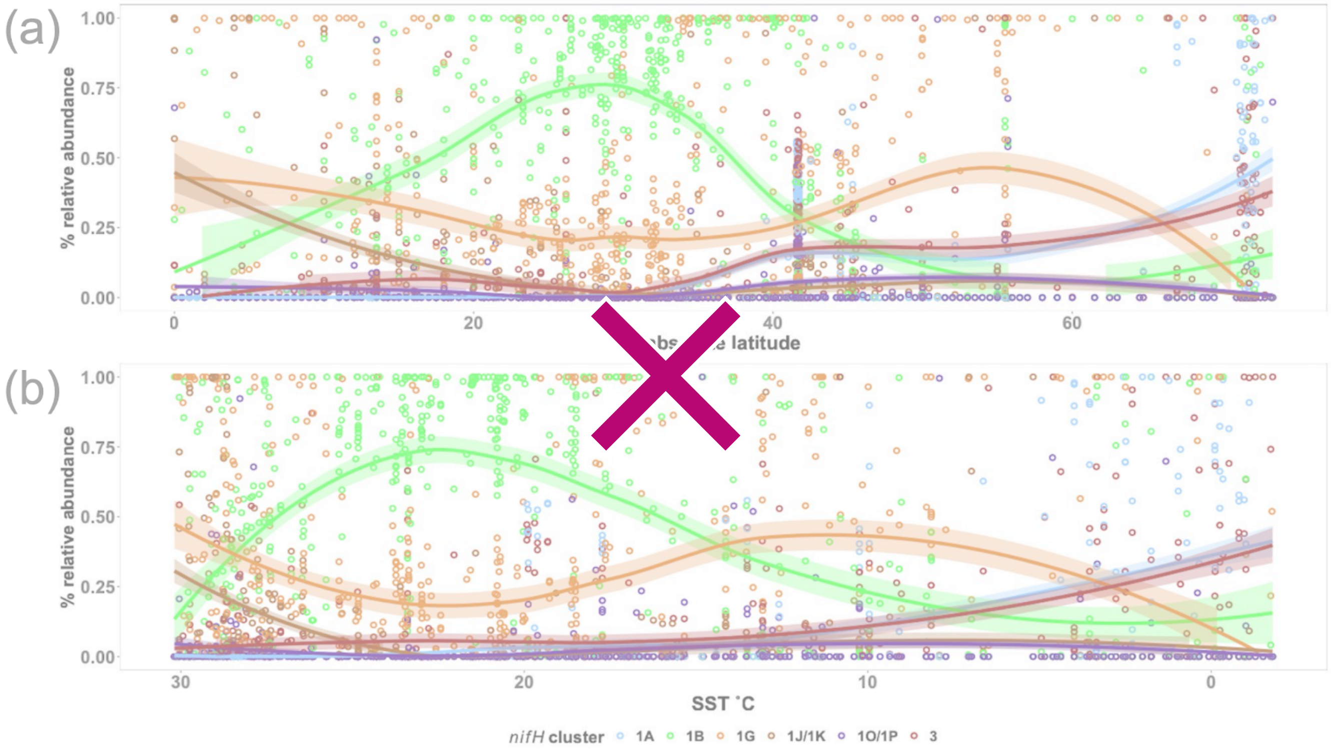
The second most abundant group was the cluster 1G (γ-proteobacteria), making up ca. 25 % of the total ~~reads~~relative abundance across the DNA dataset, with study-specific relative abundances greater than 25 % in eight out of 21 studies (Fig. 7). Members of this group were often found at high relative abundances in Pacific Ocean studies (AK2HI, NEMO, Raes_2020, Sato_2021, Shiozaki_2017), as well as in other ocean regions including the Atlantic (BentzonTilla_2015), Indian (~~TianjUni_2016~~Wu_2021) and Southern Ocean (Shiozaki_2020). The notable exception is in Arctic studies~~,~~ (Harding_2018, Shiozaki_2018LNO) where cluster 1G was almost absent (Fig. 7).

In several studies, including BentzonTillia_2015, Hallstrom_2021, Mulholland_2018, Selden_2021, Tang_2020, and Hallstrom_2022, diazotroph assemblages had high relative abundances of putative δ-proteobacteria (clusters 1A and 3), reflecting possibly a coastal/shelf or upwelling signature (Figs. 2 and 7). The only study with samples primarily from the Southern Ocean (Shiozaki_2020) was also the only study with a large portion of *nifH* cluster 1E (*Bacillota*).

3.3.2 Emerging patterns in global diazotroph assemblages across the DNA dataset

The *nifH* ASV database enables new analyses of global diazotroph biogeography in the context of environmental parameters, through co-localization with satellite and model outputs publicly available through CMAP (Ashkezari et al., 2021). To demonstrate the utility of the *nifH* ASV database, we present here patterns in relative abundances of *nifH* clusters across absolute latitude and SST in the DNA dataset. Cosmopolitan distributions were evident for γ-proteobacterial (1G) and cyanobacterial diazotrophs (1B; Fig. 8a), corroborating and extending previous findings (Farnelid et al., 2011; Shao and Luo, 2022; Halm et al., 2012; Fernandez et al., 2011; Löscher et al., 2014; Cheung et al., 2016). At low to mid latitudes, γ-proteobacterial (1G) diazotrophs generally had high relative abundances and were often the dominant taxa when present. However, they declined within the gyre regions, ranging between ~25–50 % of the population when present, while cyanobacterial diazotrophs (1B) increased and became dominant in the subtropical gyres (Fig. 8a). Notably, cluster 1G diazotrophs reached high relative abundances in each transitional zone, before mainly disappearing at latitudes above 56^{o} (Fig. 8a). However, as mentioned previously, sampling bias likely plays a large role at these higher latitudes where the number of studies and samples are sparse (Figs. 2 and 5).

Clusters 1B and 1G were both detected over the full range of SST (approximately -2–30 ˚C) but peaks in their relative abundances occurred in distinct SST ranges. Cyanobacterial diazotrophs had multiple peaks in relative abundance in waters >18 ˚C underscoring their dominance in tropical gyre regions (Fig. 8b). The 1G cluster also spanned the entire temperature spectrum but had notably higher presence and relative abundance above SSTs of 8 ˚C and 11 ˚C, respectively (Fig. 8b). The overlap between 1G and 1B has been reported previously, however the factors controlling this are unknown (Moisander et al., 2014; Shiozaki et al., 2017; Shiozaki et al., 2018b; Liu et al., 2020; Tang et al., 2020; Messer et al., 2015).


(a)
% relative abundance
1.00
0.75
0.50
0.25
0.00
0
20
40
60
absolute latitude
(b)
% relative abundance
1.00
0.75
0.50
0.25
0.00
30
20
10
0
SST ˚C
nifH cluster
1A
1B
1G
1J/1K
1O/1P
3

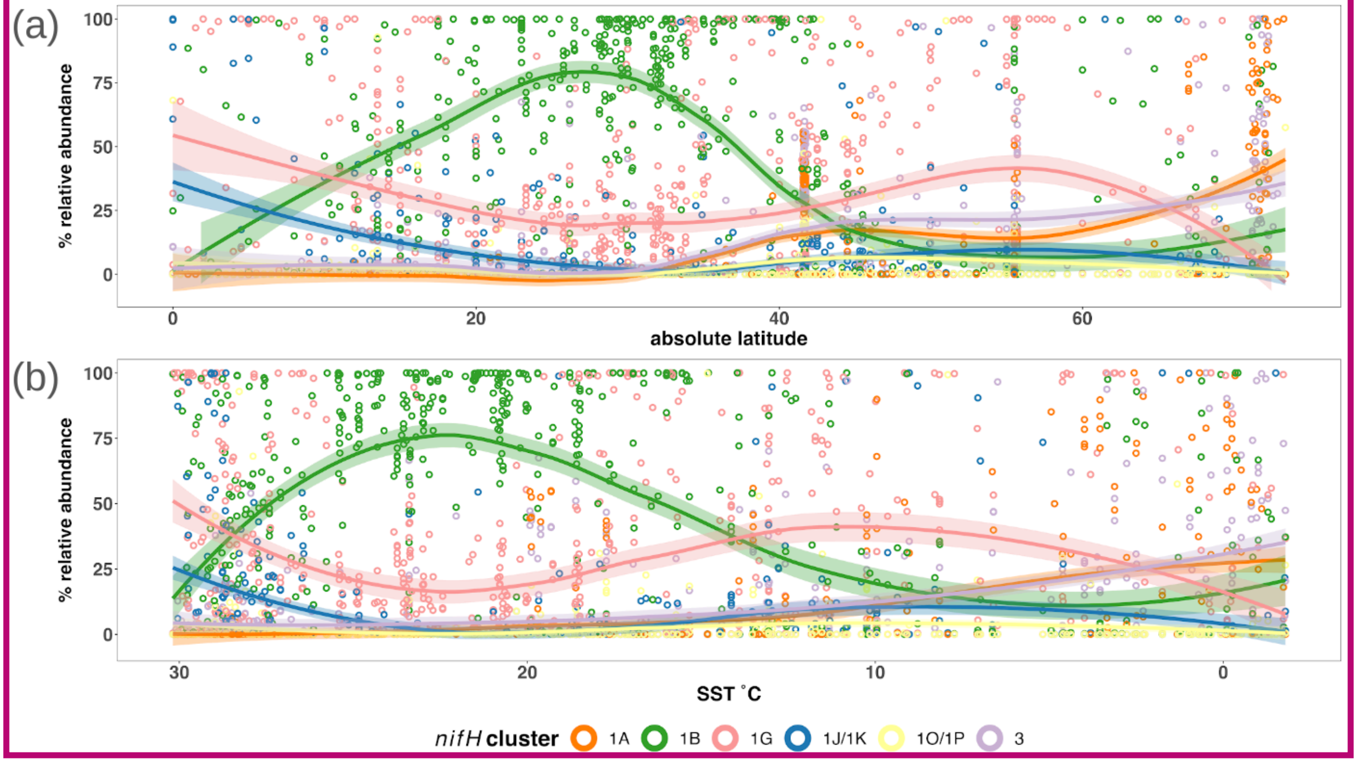

(a)
% relative abundance
100
75
50
25
0
0
20
40
60
absolute latitude
(b)
% relative abundance
100
75
50
25
0
30
20
10
0
SST ˚C
nifH cluster
1A
1B
1G
1J/1K
1O/1P
3

Figure 8: ~~Global~~Influence of SST on the global distribution of major *nifH* clusters in the photic zone of the DNA dataset. The relative abundance of *nifH* genes for each major *nifH* cluster from every photic zone sample compiled in the DNA dataset versus (a) absolute ~~latitudinal~~latitude and (b) SST. Smoothing averages (lines) were calculated using local polynomial regression fitting (LOESS) with 95 % confidence intervals (translucent colored areas). Each color represents a different *nifH* cluster. SST in (b) is from warmest to ~~coolest~~coldest temperatures to show that trends are similar to those in (a).

δ-proteobacterial diazotrophs (clusters 1A and 3) were generally found in cooler, higher latitude waters. Notably, both clusters 1A and 3 were mainly found below ~10˚C (Fig. 8b). δ-proteobacteria associated with cluster 1A were generally found at latitudes >32˚ and reached maximum relative abundances near the poles, including in the Beaufort Sea, the highest latitude region surveyed (72˚; Figs. 2, 5, and 8a). The vast majority of cluster 1A δ-proteobacteria were found at SST ≤5 ˚C (Fig. 8b). Though cluster 3 and 1A distributions were similar, cluster 3 showed broader spatial and temperature ranges, with consistent but low relative abundances in the subtropics and tropics (Fig. 8).

In contrast, the relative abundances of cluster 1J/1K and 1O/1P diazotrophs declined as SST decreased and latitude increased, becoming rare at higher latitudes (Fig 8). The highest relative abundances for these clusters were observed near the equator, and in some cases, comprised 100% of the diazotroph assemblage in high SST, tropical samples. These patterns suggest that temperature was an important factor controlling the narrow SST band (≥26 ˚C) clusters 1J/1K and 1O/1P occupied, establishing them as the *nifH* clusters with the smallest geographic range in the *nifH* ASV database (Fig. 8).

3.4 Limits and caveats to interpreting *nifH* amplicon data

The PCR amplification of the *nifH* gene and its transcripts has been vital in advancing the knowledge of diazotroph ecology due to its high sensitivity, detecting diazotrophs at abundances that are often orders of magnitude lower than other marine microbes. This approach has facilitated the discovery of many novel diazotrophs, and provided the first evidence of the widespread distribution of unicellular diazotrophs throughout the open oceans (~~Falcon~~Falcón et al., 2004; ~~Falcon~~Falcón et al., 2002; Zehr et al., 1998; Zehr et al., 2001). Advances in HTS technologies have revealed diverse diazotrophic assemblages, including the ubiquitously distributed NCDs (Turk-Kubo et al., 2014; Shiozaki et al., 2017; Raes et al., 2020). These discoveries have fostered a new perspective of global diazotrophic ecology (Zehr and Capone, 2020), improved our models of diazotrophic distributions and global N fixation rates (Tang et al., 2019) and will continue to drive new research questions.

However, interpreting *nifH* PCR-based data requires the consideration of several important caveats. Diazotrophs constitute a small fraction of the total microbial community, and thus often require numerous PCR cycles in conjunction with nested PCR for detection. Increasing the number of cycles can exacerbate known amplification biases (Turk et al., 2011) and increase the likelihood of detecting contaminant sequences (Zehr et al., 2003). Strategies to mitigate and assess contamination exist, e.g., by employing ultrafiltration of reagents and including blanks at different stages of the sampling and sequencing process (Bostrom et al., 2007; Farnelid et al., 2011; Blais et al., 2012; Moisander et al., 2014; Langlois et al., 2015; Fernandez-Mendez et al., 2016; Cheung et al., 2021), but such strategies have not been universally adopted. Additionally, relative abundances of PCR amplicons cannot easily be related to absolute abundances. For example, the relative abundance of a taxon can change even if its absolute abundance remains constant, or the relative abundance can remain constant despite changes in the total assemblage size. Moreover, the complexity of the diazotroph assemblage can, if the HTS sequencing depth is insufficient, cause rare ASVs to go undetected, or have relative abundances which are too low to interpret.

Primary objectives in studying marine diazotrophic populations include understanding the contribution of each group to N_2 fixation, the factors influencing their activity, and their global distributions. The relative abundances of *nifH* genes and transcripts estimated by the workflow can point to potentially significant contributors to N_2 fixation rates. Yet, the presence of *nifH* genes or transcripts does not always correlate with N_2 fixation rates (e.g. (Gradoville et al., 2017)). This underscores the need for cell-specific rates to better constrain N_2 fixation, the assemblages driving given rates, and the taxa-specific regulatory factors of N_2 fixation to better constrain global biogeochemical modeling.

Various methods are available to target specific diazotroph taxa over space and time (e.g. qPCR/ddPCR, fluorescent in situ hybridization (FISH)-based methods). Universal PCR assays, e.g., those used in the studies compiled here (nifH1-4), are an important complement because they better capture the overall diversity of the diazotrophic assemblage. Unlike primers designed for specific sequences, universal primers can amplify unknown or ambiguous sequences, enabling the discovery of genetic diversity. This includes microdiversity, where sequences show subtle variations from known ones, or even identifying entirely novel taxa. Primers specific to novel sequences can then be developed for use in the mentioned quantitative methods, enabling experiments to characterize the growth, activity, and controlling factors/dynamics of putative diazotrophs growth.

Tools like RT-qPCR, where transcript abundances are assessed directly, or FISH-based methods where single-cells are identified for cell-specific analysis, provide complementary perspectives into the activities of putative diazotrophs. Enumerating diazotrophs using techniques like these can help standardize the relative abundances associated with amplicon sequencing via matching taxa across each method. By assessing diversity and abundance simultaneously, major players can potentially be identified and monitored.

Through genome reconstruction, `omics studies can enhance the characterization of putative diazotroph amplicon sequences by providing a robust suite of associated genetic data, e.g., taxonomic, phylogenetic, and metabolic. Previous studies have led to the assembly of dozens of diazotrophic genomes (Delmont et al., 2022; Delmont et al., 2018). However, `omics methods often require massive amounts of data to detect rare community members, and linking genes of interest to other genomic information, e.g., taxonomy, remains quite difficult. Gene-specific models are also required to retrieve diazotrophic information and these models can benefit greatly from the high quality diazotrophic sequences of the *nifH* ASV database. In summary, the complementary perspectives afforded by the methods just described should all be used to obtain robust insights into diazotrophic assemblages.

4 Data availability

The *nifH* ASV database is freely available in Figshare (https://doi.org/10.6084/m9.figshare.23795943.~~v1~~v2; Morando et al., ~~2024~~2024a). HTS datasets for the 21 studies in the database can be obtained from the NCBI Sequence Read Archive using the NCBI BioProject accessions in Table 1.

5 Code availability

The workflow used to generate the *nifH* ASV database is freely available in two GitHub repositories, one for the DADA2 *nifH* pipeline (https://github.com/jdmagasin/nifH_amplicons_DADA2; Morando et all., 2024b) and one for the post-pipeline stages (https://github.com/jdmagasin/nifH-ASV-workflow; Morando et al., 2024c).

6 Conclusions

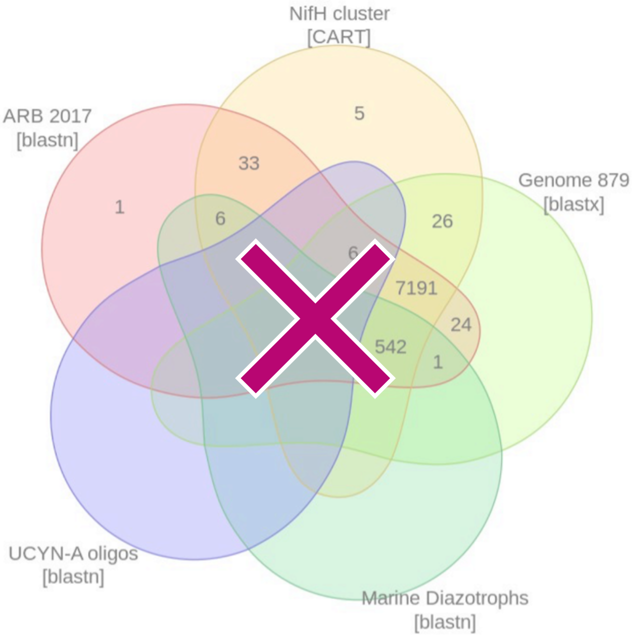
The workflow and *nifH* ASV database represent a significant step towards a unified framework that facilitates cross-study comparisons of marine diazotroph diversity and biogeography. Furthermore, they could guide future research, including cruise planning, e.g., focusing more on the southern hemisphere and areas outside of the tropics, and molecular assay development, e.g., assays to characterize NCDs for single-cell activity rates.

To demonstrate the utility of our framework, the DNA dataset was used to identify potentially important ASVs and diazotrophic groups, establishing global biogeographic patterns from this aggregated amplicon data. Cyanobacteria were the dominant diazotrophic group, but cumulatively the NCDs made up more than half of the total data. Distinct latitudinal patterns were seen among these major diazotrophic groups, with NCDs (clusters 1G, 1J/K, 1O/1P, 1A, and 3) having a greater contribution to relative abundances near the equator and at higher latitudes, while cyanobacteria (1B) comprised a majority of the diazotroph assemblage in the subtropics. SST appeared to restrict and differentiate the biogeography of clusters 1J/1K and 1O/1P (warm tropics/subtropics) from clusters 3 and 1A (cool, high latitude waters), but did not play as large of a role for the biogeography of clusters 1B and 1G.

We provide the workflow and database for future investigations into the ecological factors driving global diazotrophic biogeography and responses to a changing climate. Ultimately, we hope that insights derived from the use of our framework will inform global biogeochemical models and improve predictions of future assemblages.

Appendix A:

Figures:



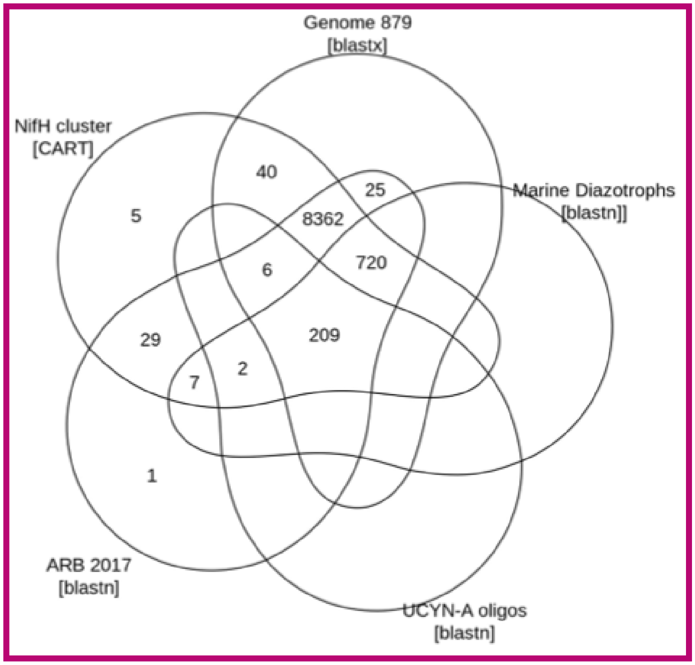


Figure A1. ASV annotations. The Venn diagram summarizes annotations assigned to ~~7931~~9406 ASVs during the AnnotateAuids stage of the workflow (Fig. 1). Numbers indicate how many ASVs received each type of annotation. Of the ~~9416~~11,915 ASVs from the preceding workflow stage, FilterAuids, only the ~~7931~~9406 ASVs shown received annotations.

Tables:

Table A1. Compiled *nifH* amplicon studies. Information on all studies compiled to generate the *nifH* ASV database, as well as studies that were not ultimately included and the reasons for this. The table provides the study ID used to refer to each dataset, the NCBI BioProject accession, the number of samples, and the DOI of the publication in which the dataset became public.¶

~~Study ID¶~~	~~NCBI BioProject¶~~	~~Samples¶~~	~~Publication DOI¶~~	~~In *nifH* ASV DB?¶~~	~~Reason excluded¶~~
~~AK2HI¶~~	~~PRJNA1062410¶~~	~~43¶~~	~~This study¶~~	~~y¶~~	¶
~~NEMO¶~~	~~PRJNA1062391¶~~	~~56¶~~	~~This study¶~~	~~y¶~~	¶
~~Cabello_2020¶~~	~~PRJNA605009¶~~	~~75¶~~	~~10.1111/jpy.13045-20-043¶~~	~~n¶~~	~~Time series samples¶~~
~~Harding_2018¶~~	~~PRJNA476143¶~~	~~91¶~~	~~10.1073/pnas.1813658115¶~~	~~y¶~~	¶
~~Turk_2021¶~~	~~PRJNA695866¶~~	~~136¶~~	~~10.1038/s43705-021-00039-7¶~~	~~y¶~~	¶
~~Gradoville_2020_G1¶~~	~~PRJNA530276¶~~	~~111¶~~	~~10.1002/lno.11423¶~~	~~y¶~~	¶
~~Gradoville_2020_G2¶~~	~~PRJNA530276¶~~	~~56¶~~	~~10.1002/lno.11423¶~~	~~y¶~~	¶
~~Turk-Kubo_2015¶~~	~~PRJNA300416¶~~	~~11¶~~	~~10.5194/bg-12-7435-2015¶~~	~~n¶~~	~~Mesocosm samples¶~~
~~Farnelid 2019¶~~	~~PRJNA392595¶~~	~~155¶~~	~~10.1002/2017GB005681¶~~	~~n¶~~	¶
~~Shiozaki_2017¶~~	~~PRJDB5199¶~~	~~22¶~~	~~10.1002/lno.10933¶~~	~~y¶~~	¶
~~Shiozaki_2018LNO¶~~	~~PRJDB5679¶~~	~~20¶~~	~~10.1038/s41561-020-00651-7¶~~	~~y¶~~	¶
~~Shiozaki_2020¶~~	~~PRJDB9222¶~~	~~14¶~~	~~10.1029/2017GB005869¶~~	~~y¶~~	¶
~~Shiozaki_2018GBC¶~~	~~PRJDB6603¶~~	~~20¶~~	~~10.3389/fmicb.2018.00797¶~~	~~y¶~~	¶
~~Li_2018¶~~	~~PRJNA434503¶~~	~~16¶~~	~~10.1002/lno.10542¶~~	~~n¶~~	~~Issues merging reads¶~~
~~Gradoville_2017¶~~	~~PRJNA328516¶~~	~~49¶~~	~~10.1038/ismej.2014.119¶~~	~~y¶~~	¶
~~BentzonTilia_2015¶~~	~~PRJNA239310¶~~	~~56¶~~	~~10.3389/fmicb.2017.01122¶~~	~~y¶~~	¶
~~Gradoville 2017 Frontiers¶~~	~~PRJNA358796¶~~	~~45¶~~	~~10.1038/srep27858¶~~	~~n¶~~	~~Perturbation experiments¶~~
~~Rahav 2016¶~~	~~n/a¶~~	~~n/a¶~~	~~10.1038/s41396-018-0050-z¶~~	~~n¶~~	~~Samples were sorted prior to sequencing¶~~
~~Gerikas Ribeiro 2018¶~~	~~PRJNA377956¶~~	~~55¶~~	~~10.1038/nmicrobiol.2016.163¶~~	~~n¶~~	~~Samples contained very few sequences¶~~
~~MartinezPerez_2016¶~~	~~PRJNA326820¶~~	~~27¶~~	~~10.1029/2020JC017071¶~~	~~y¶~~	¶
~~Sato_2021¶~~	~~PRJDB10819¶~~	~~28¶~~	~~10.1002/lno.11727¶~~	~~y¶~~	¶
~~Selden_2021¶~~	~~PRJNA683637¶~~	~~10¶~~	~~10.1029/2018GB006130¶~~	~~y¶~~	¶
~~Mulholland_2018¶~~	~~PRJNA841982¶~~	~~29¶~~	~~10.1038/s41598-019-39586-4¶~~	~~y¶~~	¶
~~MoreiraCoello_2019¶~~	~~PRJNA473903¶~~	~~24¶~~	~~10.1007/s10021-021-00702-z¶~~	~~y¶~~	¶
~~TianjUni_2016¶~~	~~PRJNA637983¶~~	~~14¶~~	~~10.1007/s00248-019-01355-1¶~~	~~y¶~~	¶
~~TianjUni_2017¶~~	~~PRJNA438304¶~~	~~18¶~~	~~10.1002/lno.11997¶~~	~~y¶~~	¶
~~Hallstrom_2021¶~~	~~PRJNA656687¶~~	~~82¶~~	~~10.1007/s10533-022-00940-w¶~~	~~y¶~~	¶
~~Hallstrom_2022¶~~	~~PRJNA756869¶~~	~~83¶~~	~~10.3389/fmars.2020.00389¶~~	~~y¶~~	¶
~~Raes_2020¶~~	~~PRJNA385736¶~~	~~121¶~~	~~10.1038/s41396-020-0703-6¶~~	~~y¶~~	¶
~~Tang_2020¶~~	~~PRJNA554315¶~~	~~6¶~~	~~10.3390/biology10060555¶~~	~~y¶~~	¶
~~Ding_2021¶~~	~~SUB7406573¶~~	~~32¶~~	~~10.1007/s13131-019-1513-4¶~~	~~y¶~~	¶

*: Data were obtained from authors, not the SRA.

Study ID	**Samples**	**NCBI BioProject**	**Reference**	**DOI**	**In *nifH* ASV database?**
AK2HI	43	PRJNA1062410	This study	n/a	Yes
BentzonTilia_2015	56	PRJNA239310	Bentzon-Tilia et al., 2015	10.1038/ismej.2014.119	Yes
Cabello 2020	75	PRJNA605009	Cabello et al., 2020	10.1111/jpy.13045	No. Time series samples
Ding_2021	32	SUB7406573	Ding et al., 2021	10.3390/biology10060555	Yes

Farnelid 2019	155	PRJNA392595	Farnelid et al., 2019	10.1038/s41396-018-0259-x	No. Particle enrichment samples
Gérikas Ribeiro 2018	55	PRJNA377956	Gérikas Ribeiro et al., 2018	10.1038/s41396-018-0050-z	No. Samples had very few sequences
Gradoville 2017 Frontiers	45	PRJNA358796	Gradoville et al., 2017	10.3389/fmicb.2017.01122	No. Perturbation experiments
Gradoville_2020_G1	111	PRJNA530276	Gradoville et al., 2020	10.1002/lno.11423	Yes
Gradoville_2020_G2	56	PRJNA530276	Gradoville et al., 2020	10.1002/lno.11423	Yes
Hallstrom_2021	82	PRJNA656687	Hallstrøm et al., 2022b	10.1002/lno.11997	Yes
Hallstrom_2022	83	PRJNA756869	Hallstrøm et al., 2022a	10.1007/s10533-022-00940-w	Yes
Harding_2018	91	PRJNA476143	Harding et al., 2018	10.1073/pnas.1813658115	Yes
Li 2018	16	PRJNA434503	Li et al., 2018	10.3389/fmicb.2018.00797	No. Issues merging reads
Mulholland_2018	29	PRJNA841982	Mulholland et al., 2019	10.1029/2018GB006130	Yes
NEMO	56	PRJNA1062391	This study	n/a	Yes
Raes_2020	121	PRJNA385736	Raes et al., 2020	10.3389/fmars.2020.00389	Yes
Rahav 2016	n/a	n/a	Rahav et al., 2016	10.1038/srep27858	No. Samples sorted prior to sequencing
Sato_2021	28	PRJDB10819	Sato et al., 2021	10.1029/2020JC017071	Yes
Selden_2021	10	PRJNA683637	Selden et al., 2021	10.1002/lno.11727	Yes
Shiozaki_2017*	22	PRJDB5199	Shiozaki et al., 2017	10.1002/2017GB005681	Yes
Shiozaki_2018GBC*	20	PRJDB6603	Shiozaki et al., 2018b	10.1029/2017GB005869	Yes
Shiozaki_2018LNO	20	PRJDB5679	Shiozaki et al., 2018a	10.1002/lno.10933	Yes
Shiozaki_2020	14	PRJDB9222	Shiozaki et al., 2020	10.1038/s41561-020-00651-7	Yes
Tang_2020	6	PRJNA554315	Tang et al., 2020	10.1038/s41396-020-0703-6	Yes
Turk-Kubo 2015	11	PRJNA300416	Turk-Kubo et al., 2015	10.5194/bg-12-7435-2015	No. Mesocosm samples
TurkKubo_2021	136	PRJNA695866	Turk-Kubo et al., 2021	10.1038/s43705-021-00039-7	Yes
Wu_2019	18	PRJNA438304	Wu et al., 2019	10.1007/s00248-019-01355-1	Yes
Wu_2021*	14	PRJNA637983	Wu et al., 2021	10.1007/s10021-021-00702-z	Yes

Appendix B: Read trimming method effects on workflow outputs

It is well-established that error rates increase with the number of PCR cycles during Illumina sequencing (Manley et al., 2016). DADA2 trims the reads to remove the low-quality tails, an important early step that impacts the proportion of sequences retained during quality-filtering and merging, as well as the ASVs detected (Fig. 1). Usually sequencing quality plots are inspected to identify a trimming length that will on average cut the reads before quality declines significantly. However, inspecting tens to hundreds of quality plots (depending on the study size) is laborious and unsystematic. For the present work, the pipeline ancillary script estimateTrimLengths.R was used to efficiently identify lengths that maximized the percentages of reads retained for each study (Section 2.3.2). The optimized lengths appeared in the parameter files as truncLen.fwd and truncLen.rev used by DADA2 filterAndTrim (Table 2).

An alternative to fixed-length trimming is to trim each read based on its individual quality profile, at the first position where the estimated sequencing error rate exceeds a threshold specified in the truncQ parameter to filterAndTrim (Table 2). This approach might reduce mismatches in the overlapping regions during the merge step and thus retain more read pairs. However, spurious low-quality bases could cause overly aggressive trimming, and picking a threshold that allows most sequences to overlap is not straightforward.

The quality of the raw sequencing data is a critical factor in the generation of the final ASV table. When analyzing a new dataset, testing both the fixed-length (truncLen) and quality-based (truncQ) trimming methods is suggested because they are fundamentally different and filterAndTrim impacts all downstream DADA2 steps. If both methods produce similar ASVs and abundances, additional parameter tuning is unlikely to impact the analysis meaningfully.

To illustrate how the trimming approach can impact workflow outputs, a version of the *nifH* ASV database was generated as shown in Figure 1 except that reads were trimmed at the first position where the estimated error rate was >2.5 % (truncQ = 16 in Table 2). This threshold typically produces forward and reverse ASVs of sufficient length to overlap without mismatches. The truncQ version of the database had substantially fewer samples, reads, and ASVs (Table B1), partly because truncQ appeared more affected by low quality reads (discussed below). Only 1783 ASVs out of 9383 in the *nifH* ASV database were detected by both trimming methods, but they comprised 88.3 % of the total reads in the database (Table B1). The 7600 ASVs (16.7 % of reads) that were found only using truncLen had mainly low abundances and were detected mainly in one to several samples. Although truncQ was less sensitive to rare ASVs, for most studies the relative abundances of *nifH* groups were similar using either trimming approach (Fig. B1).

There were three exceptions where sequencing quality issues caused substantial differences in the results from truncQ and truncLen, BentzonTilia_2015, Hallstrom_2022, and Shiozaki_2020. Using either trimming method, all three studies lost high percentages of reads during filterAndTrim (Fig. 3; losses using truncQ were comparable). This indicates that sequencing errors remained after trimming (>2 errors in the trimmed forward reads and >4 in the reverse; maxEE in Table 2). However, the subsequent losses during mergePairs were much higher using truncQ (vs. truncLen), respectively 58 % (10 %), 61 % (5 %), and 72 % (6 %) of reads. This suggests that trimming with truncQ=16 more frequently produced reads that failed to overlap during the merge step. For these three studies the workflow discarded many samples due to having ≤500 reads, but more with truncQ (vs. truncLen), respectively n=54 (34); 59 (29); and 14 (5) samples discarded. These three exceptions suggest that truncLen-based trimming can retain substantially more reads and samples for FASTQs with lower quality reads, which could impact relative abundances (Fig. B1).

Figures:

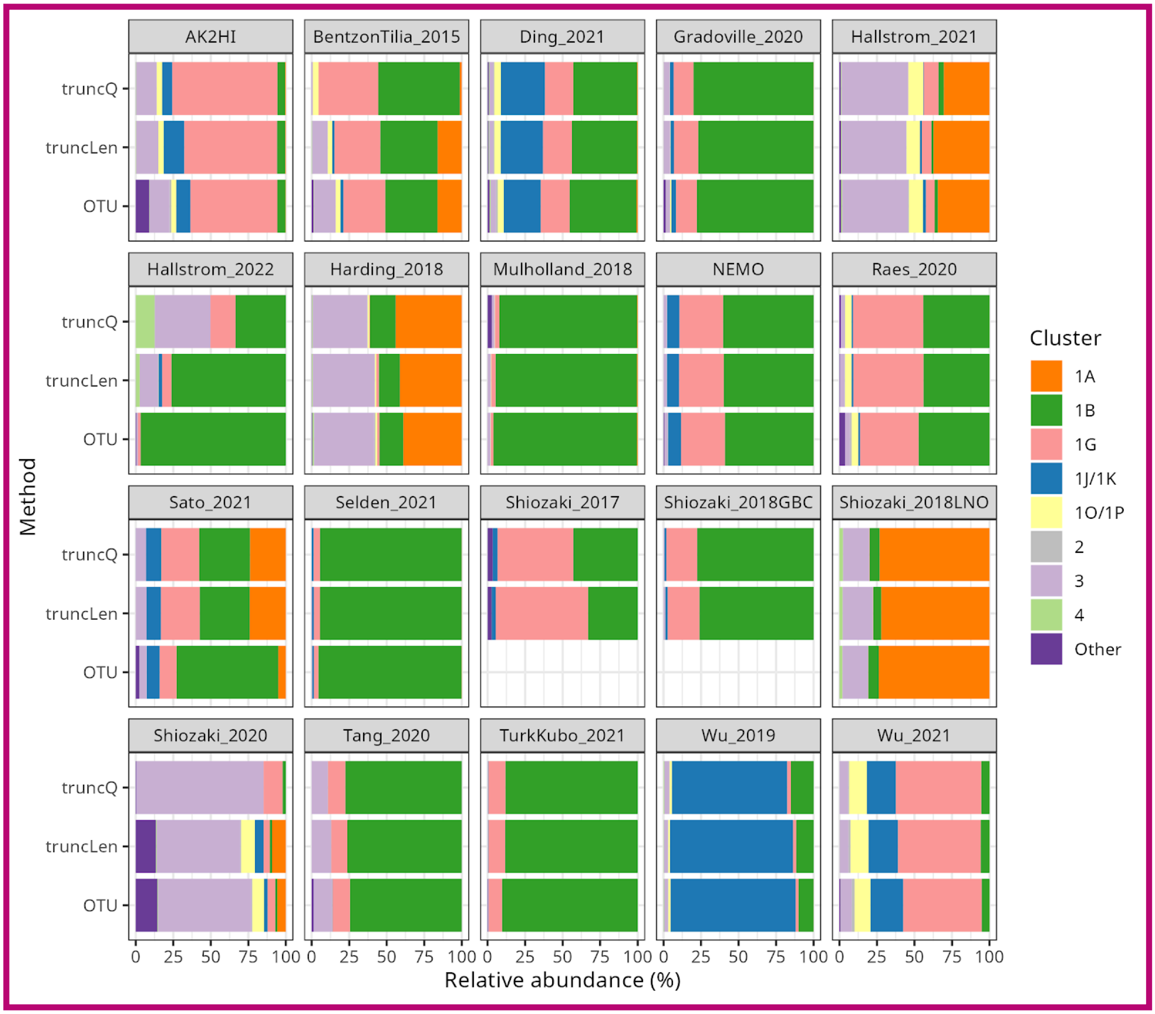


Figure B1. Relative abundances using different DADA2 trimming methods and the NifMAP OTU pipeline. *nifH* cluster relative abundances are shown for each study when processed using the NifMAP OTU pipeline (Angel et al., 2018) or by the *nifH* workflow using two methods for trimming reads, quality-based (truncQ) or fixed-length (truncLen). ASV or OTU abundances for the samples in a study were pooled to calculate the relative abundances shown. The three results for each study were calculated using only the samples that were retained by both runs of the *nifH* workflow. Shiozaki_2017 and Shiozaki_2018GBC used mixed-orientation sequencing libraries and could not be processed by NifMAP.

Tables:

Table B1. Impact of read trimming method on workflow outputs. The table compares the *nifH* ASV database, generated using fixed-length read trimming (truncLen for DADA2 filterAndTrim), to an alternative database for which reads were trimmed at the first nucleotide where the error rate was >2.5 % (truncQ=16). No other pipeline or post-pipeline parameters were changed.

	truncLen	truncQ	% decrease

Samples	944	847	10.4
ASVs	9383	1997	78.7
Reads	43.0E+6	26.3E+6	38.9

Appendix C: Comparison of communities from the workflow to previous studies

Prior to DADA2 (Callhan et al. 2016) and other approaches that distinguish fine-scale variation from sequencing errors (Eren et al. 2014, Edgar 2016b, Amir et al. 2017), most amplicon studies—for 16S rRNA as well as functional genes—processed their sequencing data into operational taxonomic units (OTUs). Usually this meant *de novo* clustering the amplicon sequences at 97 % nucleotide identity and using a representative sequence from each of the OTUs (clusters) for subsequent analyses. For 16S rRNA genes, it is known that PCR artifacts and sequencing errors can inflate the number of OTUs and cause diversity to be overestimated (Quince et al., 2009; Eren et al., 2013). For *nifH* amplicon data, these issues have been mitigated in previously published OTU analyses by analyzing broad diazotroph groups (Table C1).

To demonstrate whether communities derived from the workflow differ substantially from those previously published, a comparison was made between the results from the *nifH* workflow and another *nifH* pipeline, NifMAP (Angel et al. 2018). NifMAP is an OTU pipeline that uses hidden Markov models in an attempt to distinguish true *nifH* sequences from orthologs often mistaken for *nifH*. NifMAP was used to generate proxies for most of the 21 studies since complete OTU sequences and abundances were not available for the 19 original studies. Using NifMAP for all studies was more systematic than trying to reproduce the original results which depended on different software and methods for quality filtering. Additionally, the workflow and NifMAP both use CART (Frank et al. 2016) to identify *nifH* clusters enabling the cross-comparison of major *nifH* groups. Both also distinguish *nifH* from orthologs, the workflow using classifyNifH.sh described in section 2.3.3). Only the samples that were processed by both the workflow and NifMAP were compared (n=902).

The main result was that similar diazotroph communities were detected by the *nifH* workflow and NifMAP (Fig. B1). For every study they agreed on the two most abundant *nifH* subclusters, usually with ≤3 % difference between the relative abundances from the workflow and NifMAP. These results suggest that comparisons between new and previously published *nifH* amplicon studies are possible, especially if both use similarly broad taxonomic levels, e.g., *nifH* subclusters.

However, for two studies there were clear differences between the *nifH* workflow and NifMAP that speak to the utility of the workflow. For Hallstrom_2022 the workflow detected additional *nifH* subclusters, mainly 3 and 1G, and for Sato_20201 the workflow detected 1G and 1A at much higher levels (Fig. B1). These compositional differences likely stemmed from vastly greater numbers of reads retained by the workflow compared to NifMAP (1034 % and 264 % more reads, respectively for the two studies; Table C1). The NifMAP logs revealed that poor read quality caused NifMAP to discard the majority of reads in the first two steps. Only 10% of the Hallstrom_2022 reads could be merged, the lowest of any study (median 78 %, range 10–94 %), and 56 % of the reads from Sato_2021. The merged reads were short for both Hallstrom_2022 (mean 174 nt) and Sato_2021 (198 nt) in comparison to all studies (median of 307 nt). NifMAP then discarded, respectively, 66 % and 58 % of the merged reads due to lengths < 200 nt. In comparison, for Hallstrom_2022 the workflow discarded most reads during DADA2 filterAndTrim (using truncLen) due to sequencing errors but discarded few reads during mergePairs (Fig. 3 and Table 4). This suggests that DADA2 denoising worked very well for this dataset because the forward and reverse ASVs were allowed at most one mismatch in their overlapping region (Table 2). In contrast, Sato_2021 had substantial losses of reads during both filterAndTrim and mergePairs (Fig. 3 and Table 4). Together these results indicate that the *nifH* workflow can potentially retain more reads than NifMAP, particularly when data quality is low, with noticeable impacts on community composition.

Although community compositions from the workflow and NifMAP were mainly similar (Fig. B1), the workflow tended to retain more of the sequencing reads (Table C1). For 9 of the 18 studies analyzed by both the workflow and NifMAP, there was <10 % difference in the number of reads retained into final sequences (ASVs or OTUs; Table C1). However, 6 of the other 9 studies had more reads retained by the workflow (14–1034 %) and 3 had more reads retained by NifMAP (10–23 %). Although the workflow retained more reads, usually there were fewer ASVs than OTUs despite compression from clustering at 97 % nucleotide identity (Table C1). This is consistent with the known limitations of OTUs mentioned earlier, errors and overestimated diversity.

Tables:

Table C1. Summary of the total reads and final sequences obtained by the workflow (ASVs) and NifMAP (OTUs) applied to the same samples. A total of 902 of 944 samples in the *nifH* ASV database were compared. This excludes 42 samples from Shiozaki_2017 and Shiozaki_2018GBC that used mixed-orientation sequencing libraries and could not be processed by NifMAP. The Change (%) column is relative to reads in OTUs. OTUs in column 6 count clusters (97 % nucleotide identity). *: The original publication analyzed OTUs.

		Reads (K)			Sequences	
Study ID	**Samples compared**	**In OTUs**	**In ASVs**	**Change (%)**	**OTUs**	**ASVs**
AK2HI	43	1319	1259	4.6	987	283
BentzonTilia_2015*	54	220	171	22.6	1043	352
Ding_2021*	32	1358	1446	-6.5	1362	435
Gradoville_2020 (G1,G2)*	162	3200	3304	-3.3	642	333
Hallstrom_2021	82	4531	10,216	-125.5	14,606	6403
Hallstrom_2022*	59	455	5155	-1033.8	91	165

Harding_2018*	88	1384	1579	-14.1	1715	842
Mulholland_2018	28	2527	2439	3.5	1706	549
NEMO	54	1830	1665	9.0	591	177
Raes_2020	131	7668	7793	-1.6	1421	395
Sato_2021	28	106	388	-264.1	141	169
Selden_2021	10	405	445	-9.9	217	60
Shiozaki_2018LNO*	20	618	913	-47.8	929	283
Shiozaki_2020	14	946	1935	-104.7	1664	123
Tang_2020*	6	229	196	14.2	235	35
TurkKubo_2021*	59	2011	1976	1.8	305	74
Wu_2019*	18	801	734	8.3	504	102
Wu_2021*	14	749	674	10.0	1315	180

Author Contributions

KTK and MM designed the study with input from SC and MMM. JM created and optimized the DADA2 pipeline for *nifH* amplicon analyses. JM and MM developed the post-pipeline workflow. MM and JM compiled the database, retrieved environmental data from CMAP, and analyzed the database. MM, JM and KTK wrote the manuscript with input from MMM, SC, and JPZ.

Competing Interests

No competing interest is declared.

Acknowledgements

We gratefully acknowledge Mohammad Ashkezari and the Simons CMAP team, Stefan Green (Rush University) and the DNA genomics core at University of Illinois at Chicago, Irina Shilova, Julie Robidart and Grace Reed for NEMO sampling and sample processing, and Angelicque White (University of Hawaii, Manoa) and Mary R. Gradoville (Columbia River Inter-Tribal Fish Commission) for AK2HI sampling and sample processing. We would like to thank the authors who directly provided access to sequences: Takuhei Shiozaki (Shiozaki_2017 and Shiozaki_2018GBC) and Jun Sun, Changling Ding, and Chao Wu (~~TianjUni_2016~~Wu_2021). This work was supported by grants from the National Science Foundation to KTK

(OCE-2023498) and the Simons Foundation to JPZ (Simons Collaboration on Ocean Processes and Ecology, Award ID 724220).

References

Amir, A. McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., and Knight, R.: Deblur rapidly resolves single-nucleotide community sequence patterns, mSystems, 2, 10.1128/msystems.00191-16, 2017.

Angel, R., Nepel, M., Panholzl, 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, 10.3389/fmicb.2018.00703, 2018.

Ashkezari, M. D., Hagen, N. R., Denholtz, M., Neang, A., Burns, T. C., Morales, R. L., Lee, C. P., Hill, C. N., and Armbrust, E. V.: Simons Collaborative Marine Atlas Project (Simons CMAP): An open-source portal to share, visualize, and analyze ocean data, Limnol. Oceanogr.: Methods, 19, 488-496, 2021.

Benavides, M., Conradt, L., Bonnet, S., Berman-Frank, I., Barrillon, S., Petrenko, A., and Dogliolii, A.: Fine-scale sampling unveils diazotroph patchiness in the South Pacific Ocean, ISME Communications, 1, 3, 2021.

Bentzon-Tilia, M., Traving, S. J., Mantikci, M., Knudsen-Leerbeck, H., Hansen, J. L. S., Markager, S., and Riemann, L.: Significant N_2 fixation by heterotrophs, photoheterotrophs and heterocystous cyanobacteria in two temperate estuaries, ISME J, 9, 273-285, 2015.

Blais, M., Tremblay, J. É., Jungblut, A. D., Gagnon, J., Martin, J., Thaler, M., and Lovejoy, C.: Nitrogen fixation and identification of potential diazotrophs in the Canadian Arctic, Global Biogeochem. Cy., 26, GB3022, 10.1029/2011gb004096, 2012.

Bostrom, K. H., Riemann, L., Kuhl, M., and Hagstrom, A.: Isolation and gene quantification of heterotrophic N_2-fixing bacterioplankton in the Baltic Sea, Environ. Microbiol., 9, 152-164, doi:10.1111/j.1462-2920.2006.01124.x, 2007.

Cabello, A. M., Turk-Kubo, K. A., Hayashi, K., Jacobs, L., Kudela, R. M., and Zehr, J. P.: Unexpected presence of the nitrogen-fixing symbiotic cyanobacterium UCYN-A in Monterey Bay, California, J Phycol, 56, 1521-1533, 10.1111/jpy.13045, 2020.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P.: DADA2: High-resolution sample inference from Illumina amplicon data, Nat Methods, 13, 581-583, 10.1038/nmeth.3869, 2016.

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, 12, 2639–2643, 2017.

Capone, D. G., Burns, J. A., Montoya, J. P., Subramaniam, A., Mahaffey, C., Gunderson, T., Michaels, A. F., and Carpenter, E. J.: Nitrogen fixation by Trichodesmium spp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean, Global Biogeochem. Cy., 19, GB2024: 2021-2017, 2005.

Carpenter, E. J. and Capone, D. G.: Nitrogen in the marine environment, Academic Press, New York, 900 pp.1983.

Carpenter, E. J. and Foster, R. A.: Marine symbioses, in: Cyanobacteria in Symbiosis, edited by: Rai, A. N., Bergman, B., and Rasmussen, U., Kluwer Academic Publishers, The Netherlands, 11-18, 2002.

Cheung, S., Xia, X., Guo, C., and Liu, H.: Diazotroph community structure in the deep oxygen minimum zone of the Costa Rica Dome, J Plankton Res, 38, 380-391, 2016.

Cheung, S., Zehr, J. P., Xia, X., Tsurumoto, C., Endo, H., Nakaoka, S. I., Mak, W., Suzuki, K., and Liu, H.: Gamma4: a genetically versatile Gammaproteobacterial *nifH* phylotype that is widely distributed in the North Pacific Ocean, Environ Microbiol, 23, 4246-4259, 10.1111/1462-2920.15604, 2021.

Coale, T. H., Loconte, V., Turk-Kubo, K. A., Vanslembrouck, B., Mak, W. K. E., Cheung, S., Ekman, A., Chen, J. H., Hagino, K., Takano, Y., Nishimura, T., Adachi, M., Le Gros, M., Larabell, C., and Zehr, J. P.: Nitrogen-fixing organelle in a marine alga, Science, 384, 217-222, 10.1126/science.adk1075, 2024.

Delmont, T. O., ~~Karlusich, J. J. P., Veseli, I., Fuessel, J., Eren, A. M., Foster, R. A., Bowler, C., Wincker, P., and Pelletier, E.: Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean, ISME J, 16, 927-936, 2022.¶~~

~~Delmont, T. O., Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. T., Rappé, M. S., MacLellan, S. L., Lücker, S., and Eren, A. M.: Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes, Nature microbiology, 3, 804-813, 2018.¶~~

~~Ding, C., Wu, C., Li, L., Pujari, L., Zhang, G., and Sun, J.: Comparison of Diazotrophic Composition and Distribution in the South China Sea and the Western Pacific Ocean, Biology (Basel), 10, 10.3390/biology10060555, 2021.¶~~

~~Edgar, R.: UCHIME2: improved chimera prediction for amplicon sequencing, BioRxiv, doi.org/10.1101/074252, 2016.¶~~

~~Falcon~~Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. T., Rappé, M. S., MacLellan, S. L., Lücker, S., and Eren, A. M.: Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes, Nature microbiology, 3, 804-813, 2018.

Delmont, T. O., Karlusich, J. J. P., Veseli, I., Fuessel, J., Eren, A. M., Foster, R. A., Bowler, C., Wincker, P., and Pelletier, E.: Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean, ISME J, 16, 927-936, 2022.

Ding, C., Wu, C., Li, L., Pujari, L., Zhang, G., and Sun, J.: Comparison of Diazotrophic Composition and Distribution in the South China Sea and the Western Pacific Ocean, Biology (Basel), 10, 10.3390/biology10060555, 2021.

Edgar, R.: UCHIME2: improved chimera prediction for amplicon sequencing, BioRxiv, doi.org/10.1101/074252, 2016a.

Edgar, R.: UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing, BioRxiv, doi.org/10.1101/081257, 2016b.

Eren, A. M., Vineis, J. H., Morrison, H. G., and Sogin, M. L.: A filtering method to generate high quality short reads using Illumina paired-end technology, PLOS ONE, 8, 6, e66643, 10.1371/journal.pone.0066643, 2013.

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H. and Sogin, M. L.: Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences, ISME J, 9, 4, 968-979, doi.org/10.1038/ismej.2014.195, 2014.

Falcón, L., Cipriano, F., Chistoserdov, A., and Carpenter, E.: Diversity of diazotrophic unicellular cyanobacteria in the tropical North Atlantic Ocean, Appl Environ Microbiol, 68, 5760, 2002.

~~Falcon~~Falcón, L., Carpenter, E., Cipriano, F., Bergman, B., and Capone, D.: N_2 fixation by unicellular bacterioplankton from the Atlantic and Pacific Oceans: phylogeny and in situ rates, Appl Environ Microbiol, 70, 765-770, 2004.

Farnelid, H., Oberg, T., and Riemann, L.: Identity and dynamics of putative N_2-fixing picoplankton in the Baltic Sea proper suggest complex patterns of regulation, Environmental Microbiology Reports, 1, 145-154, 10.1111/j.1758-2229.2009.00021.x, 2009.

Farnelid, H., Andersson, A. F., Bertilsson, S., Al-Soud, W. A., Hansen, L. H., Sørensen, S., Steward, G. F., Hagström, A., and Riemann, L.: Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria, PLOS ONE, 6, e19223, 10.1371/journal.pone.0019223, 2011.

Fernandez, C., Farias, L., and Ulloa, O.: Nitrogen fixation in denitrified marine waters, PLOS ONE, 6, e20539, 10.1371/journal.pone.0020539, 2011.

Fernandez-Mendez, M., Turk-Kubo, K. A., Buttigieg, P. L., Rapp, J. Z., Krumpen, T., Zehr, J. P., and Boetius, A.: Diazotroph Diversity in the Sea Ice, Melt Ponds, and Surface Waters of the Eurasian Basin of the Central Arctic Ocean, Front Microbiol, 7, 1-18, 10.3389/fmicb.2016.01884, 2016.

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, 2016.

Gaby, J. C. and Buckley, D. H.: A global census of nitrogenase diversity, Environ Microbiol, 13, 1790-1799, 10.1111/j.1462-2920.2011.02488.x, 2011.

Goto, M., Ando, S., Hachisuka, Y., and Yoneyama, T.: Contamination of diverse nifH and nifH-like DNA into commercial PCR primers, FEMS Microbiol Lett, 246, 33-38, 10.1016/j.femsle.2005.03.042, 2005.

Gradoville, M. R., Bombar, D., Crump, B. C., Letelier, R. M., Zehr, J. P., and White, A. E.: Diversity and activity of nitrogen-fixing communities across ocean basins, Limnol Oceanogr, 62, 1895-1909, 2017.

Gradoville, M. R., Farnelid, H., White, A. E., Turk-Kubo, K. A., Stewart, B., Ribalet, F., Ferrón, S., Pinedo-Gonzalez, P., Armbrust, E. V., Karl, D. M., John, S., and Zehr, J. P.: Latitudinal constraints on the abundance and activity of the cyanobacterium UCYN-A and other marine diazotrophs in the North Pacific, Limnol Oceanogr, 65, 1858-1875, 10.1002/lno.11423, 2020.

Green, S. J., Venkatramanan, R., and Naqib, A.: Deconstructing the polymerase chain reaction: Understanding and correcting bias associated with primer degeneracies and primer-template mismatches, PLOS ONE, 10, e0128122, doi:10.1371/journal.pone.0128122, 2015.

Hallstrøm, S., Benavides, M., Salamon, E. R., Arístegui, J., and Riemann, L.: Activity and distribution of diazotrophic communities across the Cape Verde Frontal Zone in the Northeast Atlantic Ocean, Biogeochem, 1-19, 2022a.

Hallstrøm, S., Benavides, M., Salamon, E. R., Evans, C. W., Potts, L. J., Granger, J., Tobias, C. R., Moisander, P. H., and Riemann, L.: Pelagic N_2 fixation dominated by sediment diazotrophic communities in a shallow temperate estuary, Limnol Oceanogr, 67, 364-378, 2022b.

Halm, H., Lam, P., Ferdelman, T. G., Lavik, G., Dittmar, T., LaRoche, J., D'Hondt, S., and Kuypers, M. M.: Heterotrophic organisms dominate nitrogen fixation in the South Pacific Gyre, ISME J, 6, 1238-1249, 10.1038/ismej.2011.182, 2012.

Harding, K., Turk-Kubo, K. A., Sipler, R. E., Mills, M. M., Bronk, D. A., and Zehr, J. P.: Symbiotic unicellular cyanobacteria fix nitrogen in the Arctic Ocean, Proc Natl Acad Sci U S A, 115, 13371-13375, 10.1073/pnas.1813658115, 2018.

Heller, P., Tripp, H. J., Turk-Kubo, K., and Zehr, J. P.: ARBitrator: a software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank, Bioinformatics, 10.1093/bioinformatics/btu417, 2014.

Jickells, T., Buitenhuis, E., Altieri, K., Baker, A., Capone, D., Duce, R., Dentener, F., Fennel, K., Kanakidou, M., and LaRoche, J.: A reevaluation of the magnitude and impacts of anthropogenic atmospheric nitrogen inputs on the ocean, Global Biogeochem. Cy., 31, 289-305, 2017.

Langlois, R. J., LaRoche, J., and Raab, P. A.: Diazotrophic diversity and distribution in the tropical and subtropical Atlantic Ocean, Appl Environ Microbiol, 71, 7910-7919, 10.1128/AEM.71.12.7910-7919.2005, 2005.

Langlois, R., Großkopf, T., Mills, M., Takeda, S., and LaRoche, J.: Widespread distribution and expression of Gamma A (UMB), an uncultured, diazotrophic, γ-proteobacterial nifH phylotype, PLOS ONE, 10, e0128912, 2015.

~~Langlois, R. J., LaRoche, J., and Raab, P. A.: Diazotrophic diversity and distribution in the tropical and subtropical Atlantic Ocean, Appl Environ Microbiol, 71, 7910-7919, 10.1128/AEM.71.12.7910-7919.2005, 2005.¶~~

Liu, J., Zhou, L., Li, J., Lin, Y., Ke, Z., Zhao, C., Liu, H., Jiang, X., He, Y., and Tan, Y.: Effect of mesoscale eddies on diazotroph community structure and nitrogen fixation rates in the South China Sea, Regional Studies in Marine Science, 35, 101106, 2020.

Löscher, C. R., Großkopf, T., Desai, F. D., Gill, D., Schunck, H., Croot, P. L., Schlosser, C., Neulinger, S. C., Pinnow, N., and Lavik, G.: Facets of diazotrophy in the oxygen minimum zone waters off Peru, ISME J, 8, 2180-2192, 2014.

Luo, Y. W., Doney, S. C., Anderson, L. A., Benavides, M., Berman-Frank, I., Bode, A., Bonnet, S., Boström, K. H., Böttjer, D., Capone, D. G., Carpenter, E. J., Chen, Y. L., Church, M. J., Dore, J. E., Falcón, L. I., Fernández, A., Foster, R. A., Furuya, K., Gómez, F., Gundersen, K., Hynes, A. M., Karl, D. M., Kitajima, S., Langlois, R. J., LaRoche, J., Letelier, R. M., Marañón, E., McGillicuddy, D. J., Moisander, P. H., Moore, C. M., Mouriño-Carballido, B., Mulholland, M. R., Needoba, J. A., Orcutt, K. M., Poulton, A. J., Rahav, E., Raimbault, P., Rees, A. P., Riemann, L., Shiozaki, T., Subramaniam, A., Tyrrell, T., Turk-Kubo, K. A., Varela, M., Villareal, T. A., Webb, E. A., White, A. E., Wu, J., and Zehr, J. P.: Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates, Earth System Science Data, 4, 47-73, 10.5194/essd-4-47-2012, 2012.

Manley, L. J., Ma, D., and Levine, S. S.: Monitoring error rates In Illumina sequencing, J Biomol Tech, 27, 4, 125-128, 10.7171/jbt.16-2704-002, 2016.

Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet, 17, 10-12, 2011.

Messer, L. F., Mahaffey, C., Robinson, C. M., Jeffries, T. C., Baker, K. G., Isaksson, J. B., Ostrowski, M., Doblin, M. A., Brown, M. V., and Seymour, J. R.: High levels of heterogeneity in diazotroph diversity and activity within a putative hotspot for marine nitrogen fixation, ISME J, 1499-1513, 2015.

Moisander, P. H., Beinart, R. A., Voss, M., and Zehr, J. P.: Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon, ISME J, 2, 954-967, 10.1038/ismej.2008.51, 2008.

Moisander, P. H., Serros, T., Paerl, R. W., Beinart, R. A., and Zehr, J. P.: Gammaproteobacterial diazotrophs and nifH gene expression in surface waters of the South Pacific Ocean, ISME J, 8, 1962-1973, 10.1038/ismej.2014.49, 2014.

Moisander, P. H., Benavides, M., Bonnet, S., Berman-Frank, I., White, A. E., and Riemann, L.: Chasing after non-cyanobacterial nitrogen fixation in marine pelagic environments, Front Microbiol, 8, 1736, 2017.

Moonsamy, P. V., Williams, T., Bonella, P., Holcomb, C. L., Hoglund, B. N., Hillman, G., Goodridge, D., Turenchalk, G. S., Blake, L. A., Daigle, D. A., Simen, B. B., Hamilton, A., May, A. P., and Erlich, H. A.: High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array System for simplified amplicon library preparation, Tissue Antigens, 81, 141-149, 10.1111/tan.12071, 2013.

Morando, M., Magasin, J., Cheung, S., Mills, M. M., Zehr, J. P., and Turk-Kubo, K. A.: nifH ASV database in Global biogeography of N_2-fixing microbes: nifH amplicon database and analytics workflow, Figshare [dataset], https://doi.org/10.6084/m9.figshare.23795943.~~v1, 2024~~v2, 2024a.

Morando, M., Magasin, J., Cheung, S., Mills, M. M., Zehr, J. P., and Turk-Kubo, K. A.: DADA2 nifH pipeline in Global biogeography of N_2-fixing microbes: nifH amplicon database and analytics workflow, GitHub [code], https://github.com/jdmagasin/nifH_amplicons_DADA2, 2024b.

Morando, M., Magasin, J., Cheung, S., Mills, M. M., Zehr, J. P., and Turk-Kubo, K. A.: nifH ASV workflow in Global biogeography of N_2-fixing microbes: nifH amplicon database and analytics workflow, GitHub [code], https://github.com/jdmagasin/nifH-ASV-workflow, 2024c.

Mulholland, M. R., Bernhardt, P. W., Widner, B. N., Selden, C. R., Chappell, P. D., Clayton, S., Mannino, A., and Hyde, K.: High Rates of N_2 Fixation in Temperate, Western North Atlantic Coastal Waters Expand the Realm of Marine Diazotrophy, Global Biogeochem. Cy., 33, 826-840, 10.1029/2018gb006130, 2019.

Pierella Karlusich, J. J., Pelletier, E., Lombard, F., Carsique, M., Dvorak, E., Colin, S., Picheral, M., Cornejo-Castillo, F. M., Acinas, S. G., Pepperkok, R., Karsenti, E., de Vargas, C., Wincker, P., Bowler, C., and Foster, R. A.: Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods, Nat Commun, 12, 1-18, 10.1038/s41467-021-24299-y, 2021.

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., Read, L. F., and Sloan, W. T.: Accurate determination of microbial diversity from 454 pyrosequencing data, Nat Methods, 6, 639–641, doi.org/10.1038/nmeth.1361, 2009.

Raes, E. J., Van de Kamp, J., Bodrossy, L., Fong, A. A., Riekenberg, J., Holmes, B. H., Erler, D. V., Eyre, B. D., Weil, S. S., and Waite, A. M.: N_2 fixation and new insights into nitrification from the ice-edge to the equator in the South Pacific Ocean, Frontiers in Marine Science, 7, 1-20, 2020.

Rho, M., Tang, H., and Ye, Y.: FragGeneScan: predicting genes in short and error-prone reads, Nucleic Acids Res, 38, e191, 10.1093/nar/gkq747, 2010.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahe, F.: VSEARCH: a versatile open source tool for metagenomics, PeerJ, 4, e2584, 10.7717/peerj.2584, 2016.

Sato, T., Shiozaki, T., Taniuchi, Y., Kasai, H., and Takahashi, K.: Nitrogen Fixation and Diazotroph Community in the Subarctic Sea of Japan and Sea of Okhotsk, Journal of Geophysical Research: Oceans, 126, e2020JC017071, 2021.

Schlessman, J. L., Woo, D., Joshua-Tor, L., Howard, J. B., and Rees, D. C.: Conformational variability in structures of the nitrogenase iron proteins from Azotobacter vinelandii and Clostridium pasteurianum, J Mol Biol, 280, 4, 669-685, 1998.

Selden, C. R., Chappell, P. D., Clayton, S., Macías-Tapia, A., Bernhardt, P. W., and Mulholland, M. R.: A coastal N_2 fixation hotspot at the Cape Hatteras front: Elucidating spatial heterogeneity in diazotroph activity via supervised machine learning, Limnol Oceanogr, 66, 1832-1849, 2021.

Shao, Z. and Luo, Y. W.: Controlling factors on the global distribution of a representative marine heterotrophic diazotroph phylotype (Gamma A), Biogeosciences, 19, 2939-2952, 2022.

Shao, Z., Xu, Y., Wang, H., Luo, W., Wang, L., Huang, Y., Agawin, N. S. R., Ahmed, A., Benavides, M., Bentzon-Tilia, M., and Berman-Frank, I.: Global Oceanic Diazotroph Database Version 2 and Elevated Estimate of Global N_2 Fixation, Earth System Science Data, 15, 2023.

Shilova, I., Mills, M., Robidart, J., Turk-Kubo, K., Björkman, K., Kolber, Z., Rapp, I., van Dijken, G., Church, M., and Arrigo, K.: Differential effects of nitrate, ammonium, and urea as N sources for microbial communities in the North Pacific Ocean, Limnol Oceanogr, 62, 2550-2574, 2017.

Shiozaki, T., Fujiwara, A., Inomura, K., Hirose, Y., Hashihama, F., and Harada, N.: Biological nitrogen fixation detected under Antarctic sea ice, Nature Geoscience, 13, 729–732, 2020.

Shiozaki, T., ~~Fujiwara, A., Ijichi, M., Harada, N., Nishino, S., Nishi, S., Nagata, T~~Bombar, D., Riemann, L., Hashihama, F., Takeda, S., Yamaguchi, T., Ehama, M., Hamasaki, K., and ~~Hamasaki~~Furuya, K.: ~~Diazotroph community structure and the role of nitrogen fixation in the nitrogen cycle in the Chukchi Sea (western Arctic Ocean), Limnol Oceanogr, 63, 2191-2205, 10.1002/lno.10933, 2018a.¶~~

~~Shiozaki, T., Bombar, D., Riemann, L., Hashihama, F., Takeda, S., Yamaguchi, T., Ehama, M., Hamasaki, K~~Basin scale variability of active diazotrophs and nitrogen fixation in the North Pacific, from the tropics to the subarctic Bering Sea, Global Biogeochem. Cy., 31, 996-1009, 10.1002/2017gb005681, 2017.

Shiozaki, T., Fujiwara, A., Ijichi, M., Harada, N., Nishino, S., Nishi, S., Nagata, T., and ~~Furuya~~Hamasaki, K.: ~~Basin scale variability of active diazotrophs and nitrogen fixation in the North Pacific, from the tropics to the subarctic Bering Sea, Global Biogeochem. Cy., 31, 996-1009, 10.1002/2017gb005681, 2017~~Diazotroph community structure and the role of nitrogen fixation in the nitrogen cycle in the Chukchi Sea (western Arctic Ocean), Limnol Oceanogr, 63, 2191-2205, 10.1002/lno.10933, 2018a.

Shiozaki, T., Bombar, D., Riemann, L., Sato, M., Hashihama, F., Kodama, T., Tanita, I., Takeda, S., Saito, H., Hamasaki, K., and Furuya, K.: Linkage between dinitrogen fixation and primary production in the oligotrophic South Pacific Ocean, Global Biogeochem. Cy., 32, 1028-1044, 2018b.

Tang, W., Li, Z., and Cassar, N.: Machine learning estimates of global marine nitrogen fixation, Journal of Geophysical Research: Biogeosciences, 124, 717-730, 2019.

Tang, W., Cerdan-Garcia, E., Berthelot, H., Polyviou, D., Wang, S., Baylay, A., Whitby, H., Planquette, H., Mowlem, M., Robidart, J., and Cassar, N.: New insights into the distributions of nitrogen fixation and diazotrophs revealed by high-resolution sensing and sampling methods, ISME J, 14, 2514-2526, 10.1038/s41396-020-0703-6, 2020.

Taylor, L. J., Abbas, A., and Bushman, F. D.: grabseqs: Simple downloading of reads and metadata from multiple next-generation sequencing data repositories, Bioinformatics, doi.org/10.1093/bioinformatics/btaa167, 2020.

Turk, K., Rees, A. P., Zehr, J. P., Pereira, N., Swift, P., Shelley, R., Lohan, M., Woodward, E. M. S., and Gilbert, J.: Nitrogen fixation and nitrogenase (nifH) expression in tropical waters of the eastern North Atlantic, ISME J, 5, 1201-1212, 10.1038/ismej.2010.205, 2011.

Turk-Kubo, K. A., Karamchandani, M., Capone, D. G., and Zehr, J. P.: The paradox of marine heterotrophic nitrogen fixation: abundances of heterotrophic diazotrophs do not account for nitrogen fixation rates in the Eastern Tropical South Pacific, Environ Microbiol, 16, 3095-3114, 10.1111/1462-2920.12346, 2014.

Turk-Kubo, K. A., Farnelid, H. M., Shilova, I. N., Henke, B., and Zehr, J. P.: Distinct ecological niches of marine symbiotic N_2-fixing cyanobacterium *Candidatus* Atelocyanobacterium thalassa sublineages, J Phycol, 53, 451-461, 10.1111/jpy.12505, 2017.

Turk-Kubo, K. A., ~~Gradoville~~Mills, M. ~~R., Cheung, S., Cornejo-Castillo, F., Harding, K. J., Morando, M., Mills, M., and Zehr, J. P.: Non-cyanobacterial diazotrophs: Global diversity, distribution, ecophysiology, and activity in marine waters, FEMS Microbiol Rev, 10.1093/femsre/fuac046, 2022~~M., Arrigo, K. R., van Dijken, G., Henke, B. A., Stewart, B., Wilson, S. T., and Zehr, J. P.: UCYN-A/haptophyte symbioses dominate N_2 fixation in the Southern California Current System, ISME Communications, 1, 1-13, 2021.

Turk-Kubo, K. A., ~~Mills~~Gradoville, M. ~~M., Arrigo, K. R., van Dijken, G., Henke, B. A., Stewart, B., Wilson, S. T., and Zehr, J. P.: UCYN-A/haptophyte symbioses dominate N_2 fixation in the Southern California Current System, ISME Communications, 1, 1-13, 2021~~R., Cheung, S., Cornejo-Castillo, F., Harding, K. J., Morando, M., Mills, M., and Zehr, J. P.: Non-cyanobacterial diazotrophs: Global diversity, distribution, ecophysiology, and activity in marine waters, FEMS Microbiol Rev, 10.1093/femsre/fuac046, 2022.

Villareal, T. A.: Widespread occurrence of the *Hemiaulus*-cyanobacterial symbiosis in the southwest North-Atlantic Ocean, Bulletin of Marine Science, 54, 1-7, 1994.

Wu, C., Kan, J., Liu, H., Pujari, L., Guo, C., Wang, X., and Sun, J.: Heterotrophic Bacteria Dominate the Diazotrophic Community in the Eastern Indian Ocean (EIO) during Pre-Southwest Monsoon, Microb Ecol, 78, 804-819, 10.1007/s00248-019-01355-1, 2019.

Wu, C., Sun, J., Liu, H., Xu, W., Zhang, G., Lu, H., and Guo, Y.: Evidence of the Significant Contribution of Heterotrophic Diazotrophs to Nitrogen Fixation in the Eastern Indian Ocean During Pre-Southwest Monsoon Period, Ecosyst, 25, 1066-1083, 2021.

Zani, S.: Application of a nested reverse transcriptase polymerase chain reaction assay for the detection of nifH expression in Lake George, New York, M. S. Thesis, Rensselaer Polytechnic Institute, 1999.

Zehr, J. P. and Capone, D. G.: Changing perspectives in marine nitrogen fixation, Science, 368, eaay9514, 10.1126/science.aay9514, 2020.

Zehr, J. and McReynolds, L.: Use of degenerate oligonucleotides for amplification of the nifH gene from the marine cyanobacterium Trichodesmium thiebautii, Appl Environ Microbiol, 55, 2522-2526, 1989.

Zehr, J., Mellon, M., and Zani, S.: New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (*nifH*) genes, Appl. Environ. Microbiol, 64, 3444-3450, 1998.

Zehr, J. P~~. and Capone, D. G.: Changing perspectives in marine nitrogen fixation, Science, 368, eaay9514, 10.1126/science.aay9514, 2020.¶~~

~~Zehr, J. P., Crumbliss, L. L., Church, M. J., Omoregie, E. O., and Jenkins, B. D.: Nitrogenase genes in PCR and RT-PCR reagents: implications for studies of diversity of functional genes, Biotechniques, 35, 996-1005, 2003.¶~~

~~Zehr, J. P.~~, Waterbury, J. B., Turner, P. J., Montoya, J. P., Omoregie, E., Steward, G. F., Hansen, A., and Karl, D. M.: Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean, Nature, 412, 635-638, 2001.

Zehr, J. P., Crumbliss, L. L., Church, M. J., Omoregie, E. O., and Jenkins, B. D.: Nitrogenase genes in PCR and RT-PCR reagents: implications for studies of diversity of functional genes, Biotechniques, 35, 996-1005, 2003.