Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original Paper

Sequence analysis

PPIT: an R package for inferring microbial taxonomy from *nifH* sequences

Bennett J. Kapili* and Anne E. Dekas*

Department of Earth System Science, Stanford University, Stanford, CA 94305, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on April 15 2020; revised on July 22 2020; accepted on January 31 2021

Abstract

Motivation: Linking microbial community members to their ecological functions is a central goal of environmental microbiology. When assigned taxonomy, amplicon sequences of metabolic marker genes can suggest such links, thereby offering an overview of the phylogenetic structure underpinning particular ecosystem functions. However, inferring microbial taxonomy from metabolic marker gene sequences remains a challenge, particularly for the frequently sequenced nitrogen fixation marker gene, nitrogenase reductase (*nifH*). Horizontal gene transfer in recent *nifH* evolutionary history can confound taxonomic inferences drawn from the pairwise identity methods used in existing software. Other methods for inferring taxonomy are not standardized and require manual inspection that is difficult to scale.

Results: We present Phylogenetic Placement for Inferring Taxonomy (PPIT), an R package that infers microbial taxonomy from *nifH* amplicons using both phylogenetic and sequence identity approaches. After users place query sequences on a reference *nifH* gene tree provided by PPIT (n = 6317 full-length *nifH* sequences), PPIT searches the phylogenetic neighborhood of each query sequence and attempts to infer microbial taxonomy. An inference is drawn only if references in the phylogenetic neighborhood are: (1) taxonomically consistent and (2) share sufficient pairwise identity with the query, thereby avoiding erroneous inferences due to known horizontal gene transfer events. We find that PPIT returns a higher proportion of correct taxonomic inferences than BLAST-based approaches at the cost of fewer total inferences. We demonstrate PPIT on deep-sea sediment and find that *Deltaproteobacteria* are the most abundant potential diazotrophs. Using this dataset we show that emending PPIT inferences based on visual inspection of query sequence placement can achieve taxonomic inferences for nearly all sequences in a query set. We additionally discuss how users can apply PPIT to the analysis of other marker genes.

Availability: PPIT is freely available to non-commercial users at https://github.com/bkapili/ppit. Installation includes a vignette that demonstrates package use and reproduces the *nifH* amplicon analysis discussed here. The raw *nifH* amplicon sequence data have been deposited in the GenBank, EMBL, and DDBJ databases under BioProject number PRJEB37167.

Contact: kapili@stanford.edu and dekas@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Nucleotide sequencing is a fundamental technique for studying microbial communities. While metagenomic approaches provide simultaneous insight into the taxonomic identity and metabolic potential of organisms, PCR amplicon sequencing maintains utility in offering greater sequencing depth for targeted genetic loci and as a cost-effective strategy for studies containing many samples. In addition to targeting the 16S rRNA gene for snapshots of community structures (Pace *et al.*, 1986; Ward *et al.*, 1990), amplicon sequencing projects can target metabolic marker genes to survey community metabolic potential (Kirshtein *et al.*, 1991;

Ohkuma *et al.*, 1995; Rotthauwe *et al.*, 1997; Braker *et al.*, 1998; Cottrell and Cary, 1999). When marker genes are linked to their source organisms, amplicon sequencing can further suggest connections between ecological functions and taxonomic identities. However, accurately linking marker gene amplicons to their source organisms remains a significant challenge.

Linking sequence to source organism is both particularly desirable and challenging for nitrogenase reductase (*nifH*), which is targeted to study nitrogen fixation (*i.e.*, the reduction of N_2 to bioavailable NH_3). *nifH* is one of the most frequently sequenced marker genes (n = 81 889 sequences in GenBank as of October 2019) and can be remarkably diverse in environmental samples (Izquierdo and Nüsslein, 2006; Fernández-Méndez *et al.*, 2016; Kapili *et al.*, 2020). Recent software packages have addressed the need to link *nifH* sequences to source organism taxonomy using pairwise sequence identity methods, particularly using query sequences' top BLAST hits (*i.e.*, lowest E-value match) to infer identity (Gaby *et al.*, 2018; Angel *et al.*, 2018). However, the occurrence of horizontal gene transfer in *nifH* evolutionary history (Raymond *et al.*, 2004) can confound taxonomic assignments for recently transferred sequences. Existing programs cannot address inference errors from horizontal inheritance because their current implementations base taxonomic inferences on only one reference sequence.

Another common approach for inferring *nifH* source organism taxonomy – observing the position of novel sequences on a reference gene tree (Farnelid *et al.*, 2011; Bertics *et al.*, 2013; Collavino *et al.*, 2014; Igai *et al.*, 2016; Wang *et al.*, 2016) – can allow investigators to evaluate evidence of horizontal inheritance for query sequences. Although this technique addresses the shortcomings inherent to pairwise identity approaches, it currently lacks standardization, uses *ad hoc* reference trees not optimized for detecting horizontal gene transfer events, and requires manual inspection that is difficult to scale. With advancements to sequencing technologies (Singer *et al.*, 2019) and bioinformatics programs (Callahan *et al.*, 2016; Amir *et al.*, 2017) enabling greater recovery of environmental sequence diversity, the need for accurate, high-throughput taxonomic inferencing for *nifH* sequences continues to grow.

To address these issues, we present Phylogenetic Placement for Inferring Taxonomy (PPIT), an R software package that combines phylogenetic and sequence identity approaches to infer the source organism taxonomy of *nifH* sequences. PPIT reads the output from SEPP (Mirarab *et al.*, 2012), a phylogenetic placement program that resolves the phylogeny of short nucleotide sequences more accurately than the construction of *de novo* trees (Janssen *et al.*, 2018), and progressively searches each query sequence's local phylogenetic neighborhood after placement on a reference *nifH* tree. The taxonomic identity of the *nifH* source organism is then only inferred if nearby reference sequences have consistent taxonomic classifications and share sufficient pairwise identity with the query sequence. PPIT standardizes and automates the process of interpreting *nifH* phylogenetic placement to draw taxonomic inferences, which supports analytical reproducibility and reduces the demand for manual inspection.

We include in the package a reference *nifH* tree that contains nearly all the full-length *nifH* sequences currently available in GenBank (n = 6317), as well as the necessary supporting files for users to place query sequences on the reference tree using SEPP. The output from PPIT is a data frame of taxonomic inferences that can be easily supplied to a phyloseq object as the accompanying taxonomy table to query sequences. Thus, PPIT leverages cumulative sequencing efforts to infer taxonomic identity and is built around existing software to support reproducible bioinformatic workflows. Here, we detail PPIT's concept, compare the accuracy of PPIT to pairwise identity methods, use PPIT to analyze *nifH* amplicons generated from deep-sea sediment, and discuss the applicability of PPIT to other environmental samples and metabolic marker genes.

2 Concept and methods

2.1 PPIT concept and implementation

PPIT is based on the analysis of *nifH* amplicons presented in Kapili *et al.*, 2020 with some modifications. We provide a schematic of the PPIT workflow in Figure 1. Prior to analysis, users must run SEPP to insert their query sequences into the curated reference alignment and tree that are included in the package. PPIT requires seven inputs, four of which are user-supplied for *nifH* analyses: (1) the reference *nifH* tree and (2) alignment containing inserted query sequences, (3) the list of query sequence names, and (4) the type of sequences supplied (*i.e.,* partial- or full-length). The other inputs, which include (5) references' taxonomic and gene location information (*i.e.*, if the *nifH* sequence was located on a chromosome, plasmid, or undetermined), (6) empirically-determined taxonomic rank cutoffs, and (7) operational phylogenetic neighborhood are either included with installation or calculated.

The algorithm is composed of three steps: (1) defining each query sequence's operational phylogenetic neighborhood, (2) evaluating both the taxonomic consistency in its neighborhood and pairwise identity to the nearest reference sequence, and (3) optimizing the phylogenetic neighborhood.

In Step 1, the operational phylogenetic neighborhood is set, which is defined here as the maximum summed branch length (*i.e.*, patristic distance) allowed between a query sequence and a reference sequence for the reference to be used during inferencing. Setting an upper bound to this distance is necessary to avoid overexploring tree topology during Step 2, which otherwise could result in mischaracterizing vertical inheritance as horizontal inheritance (see Results and discussion).This parameter is initially set to a value less than that expected to be used, and is later optimized (Step 3).

In Step 2, the taxonomic and gene location information for the references within each query sequence's operational phylogenetic neighborhood are collected. If the nearest reference is a suspected *nifH* homolog, the query is marked as a suspected homolog and no taxonomic inference is drawn (see Section 2.3 for discussion of suspected homologs). Otherwise, the gathered references for each query are placed into taxonomic rank subsets based on their patristic distances to the query. PPIT uses empirically-defined patristic distance rank cutoffs below which references are expected to be from the same genus, family, order, class, or phylum. The algorithm begins with evaluating the genus reference subset. If all collected references share the same genus and if the pairwise sequence identity with the nearest reference is greater than the genus percent identity threshold, then the query sequence's identity is inferred as the references' taxonomy (domain through genus). If no references are within the genus patristic distance cutoff, the collected references are taxonomically inconsistent, or the query is insufficiently similar to the nearest reference, then the family reference subset is analyzed at the family rank and similarity evaluated using the family percent identity cutoff. The process repeats until the phylum rank, at which point if an inference remains unmade, then no inference for the query sequence is drawn. Inferencing proceeds from genus to phylum because otherwise we would implicitly require no horizontal gene transfer to have occurred in a phylum on average, including ancestral events that preceded the rise of a class, order, *etc*. Query sequences are flagged for potential horizon-

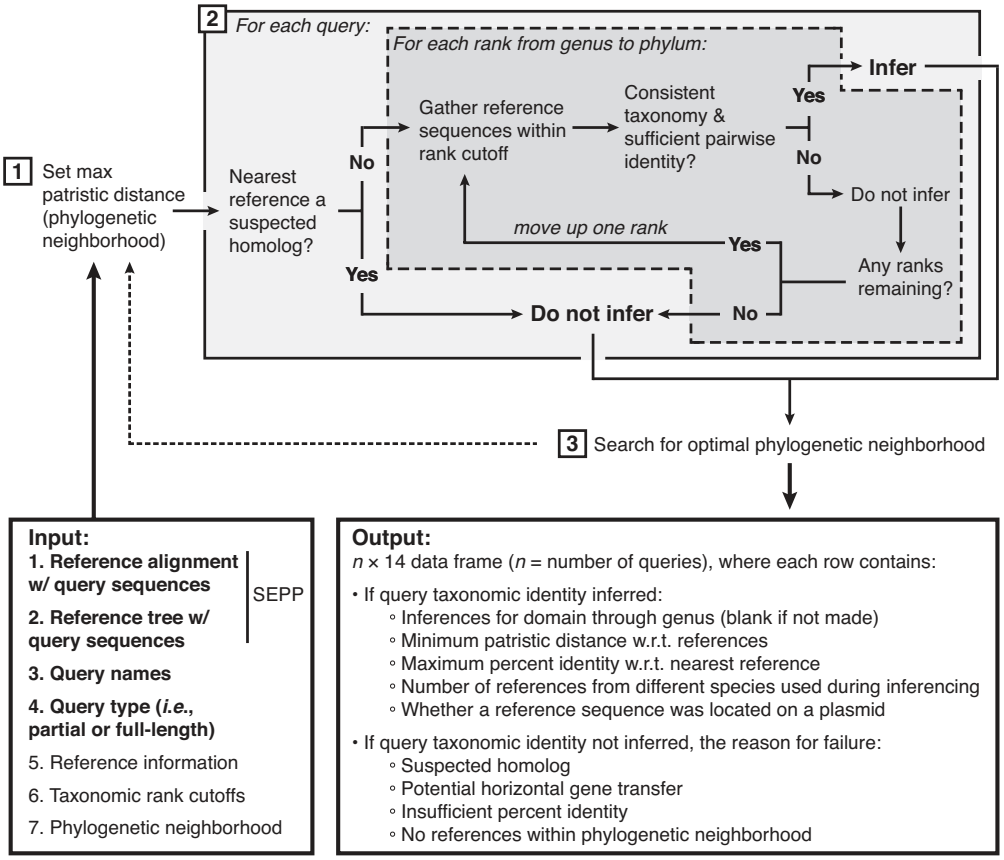


Figure 1 Conceptual diagram of Phylogenetic Placement for Inferring Taxonomy (PPIT) workflow, including program inputs and outputs. Bolded inputs denote those that are user-supplied and non-bolded inputs denote those that are included with installation for *nifH* analyses. Boxed numbers represent steps described in Section 2.1. Step 1, set operational phylogenetic neighborhood. Step 2, infer identity if the nearest reference is not a suspected homolog and the gathered references are taxonomically consistent and share sufficient pairwise identity with the query sequence. Step 3, search for phylogenetic neighborhood value that returns the maximum number of inferences at the phylum rank. Dashed line indicates iterations during phylogenetic neighborhood optimization.

tal gene transfer when taxonomic inconsistencies remain at the phylum rank for the phylum reference subset.

PPIT's approach for inferring taxonomic identity also helps prevent the propagation of erroneous taxonomic assignments from sequences in the reference database (Bagheri *et al.*, 2020) to query sequences. Since PPIT draws taxonomic inferences based on a collection of references rather than a single reference – and requires these references to have consistent taxonomies – if a reference sequence contains an incorrect taxonomic assignment that conflicts with the taxonomy of the other collected references, then no inference for the query will be drawn.

In Step 3, the operational phylogenetic neighborhood is optimized for the set of query sequences. The optimal value is defined here as the value that maximizes the number of query sequences for which inferences are drawn at the phylum rank. The optimal value of the operational phylogenetic neighborhood depends on the query set and can be lower than the family, order, class, and phylum patristic distance cutoffs, but cannot be lower than the genus cutoff. This requires that, at minimum, PPIT uses all reference sequences within the genus patristic distance cutoff during inferencing. It is important to note that, while the inferred identity for a given query will not change based on the other sequences in the query set, the taxonomic depth to which the source organism is inferred may change.

PPIT output consists of an $n \times 14$ data frame (n = number of query sequences) where each row contains a query's taxonomic inferences for domain through genus (blank if not made), as well as the minimum patristic distance to a reference and the pairwise percent identity with respect to that reference. The output also contains the number of references from different species used to infer identity for each query and indicates if one of the reference sequences was located on a plasmid. If the taxonomic identity for a given query was not inferred, the reason for failure is provided (*i.e.*, suspected homolog, potential horizontal gene transfer, no references within phylogenetic neighborhood, or insufficient percent identity). The data frame PPIT provides can be easily reformatted and supplied to a phyloseq object as the accompanying taxonomy table to the query sequences (McMurdie and Holmes, 2013).

2.2 Database curation

We curated a reference *nifH* database using ARBitrator, which identifies putative NifH sequences in GenBank's non-redundant (nr) sequence database using a refined sequence similarity-based approach (Heller *et al.*, 2014). We supplemented the default query NifH sequences with sequences from each phylum in the Genome Taxonomy Database (r86) containing at least one representative with NifH adjacent to NifDKENB. We collected these sequences via blastp search using a NifH sequence from *Azotobacter vinelandii* CA6 (accession no. WP_012698831.1) to search bacterial phyla and *Methanococcus maripaludis* (accession no. WP_011170797.1) to search archaeal phyla. In total, we used thirty-eight

reference NifH sequences (n = 27 bacterial phyla, n = 1 archaeal phylum) and one CfbC sequence – formerly Group-IV NifH (Raymond *et al.*, 2004) – for the ARBitrator search (October 2019).

We obtained 43 357 protein sequences, for which the corresponding nucleotide sequences were downloaded using eutils. We then filtered the sequences to remove duplicate nucleotide records and those containing ambiguous bases. For each of the remaining sequences, we obtained the source organism, full taxonomy (as listed on NCBI), and genetic location (*i.e.*, chromosome, plasmid, or undetermined) using eutils. We identified those flanked by start and stop codons (Genetic Code 11), adjusting for sequence direction, to create a subset of full-length sequences. We additionally filtered these full-length records to remove those >1400 bp in length (shortest length of *nifEH* gene fusion) or <553 bp (2 standard deviations below average sequence length after removal of *nifEH* fusions). We aligned the translated sequences using MAFFT (FFT-NS-1, v.7.427) (Katoh and Standley, 2013) and manually removed poorly aligned sequences (n = 8876 remaining).

2.3 Reference tree construction

We aligned all filtered full-length sequences using MAFFT (FFT-NS-1) and inferred an initial phylogenetic tree using FastTree (WAG, Gamma20; v.2.1.11) (Price *et al.*, 2010). Upon tree visualization, we identified an initial set of suspected NifH homologs based on the placements of the reference NifH sequences and homolog used for ARBitrator query. Excluding the set of suspected homologs, we determined the rank order of remaining NifH sequences that maximized the total length of the tree after addition to the ARBitrator query set (Supplementary Figure 1). Determining this rank order identifies which sequences cover the greatest sequence diversity. We aligned the top 736 NifH sequences and 14 suspected homologs that represented the major divergent homolog clades on the initial tree using MAFFT-DASH (G-INS-I, v.7.427), which incorporates protein structural information to increase alignment accuracy (Rozewicki *et al.*, 2019). We then aligned the remaining amino acid sequences to this seed alignment using MAFFT (G-INS-I) and converted the alignment to a nucleotide alignment using PAL2NAL (v.14; Suyama *et al.*, 2006) and the corresponding *nifH* nucleotide sequences. On the CIPRES Science Gateway (Miller *et al.*, 2010), we ran ModelTestNG (v.0.1.5) to determine the optimal evolutionary model given the nucleotide alignment, and then executed RAxML (v.8.2.12; Stamatakis, 2014) and IQ-TREE (v.1.6.10; Nguyen *et al.*, 2015) ten times each using the GTR+G+I (4 rate categories) model with 100 bootstrap replicates. We selected the gene tree estimate with the highest log-likelihood score as the *nifH* reference tree (RAxML).

To identify suspected *nifH* homologs in the final reference tree, we visualized the tree using Interactive Tree of Life (v.4) (Letunic and Bork, 2019) and defined the *nifH* crown group as that containing all ARBitrator reference *nifH* queries, as well as the immediately basal clades in which at least one *nifH* is adjacent to *nifDKENB* (Supplementary Figure 2). The sequences not within the *nifH* crown group appear highly diverged from *nifH* (Supplementary Figure 2), suggesting that their protein products are not involved in nitrogen fixation. Some of these sequences include the *nifH* paralog *cfbC* (Group IV), which was recently shown to be involved in methanogenesis and methanotrophy (Zheng *et al.*, 2016; Moore *et al.*, 2017). Although there are no apparent *chlL*/*bchL*/*bchX* sequences (Group V) on the reference tree, which are involved in chlorophyll and bacteriochlorophyll biosynthesis (Fujita *et al.*, 1992; Fujita and Bauer, 2000), PPIT successfully identified a set of diverse partial-length *bchL* sequences as *nifH* homologs when used as queries (Supplementary Table 1). Given the phylogenetic diversity of sequences outside the *nifH* crown group, we consider it possible that the reference tree also contains novel *nifH* homologs. We refer to the collection of these divergent sequences as suspected *nifH* homologs, agnostic to both function and evolutionary relationship to *nifH*.

2.4 Calculation of rank cutoffs

To minimize bias towards overrepresented species or species with multiple similar copies of *nifH* during the calculation of rank cutoffs, we first clustered sequences in each assembly at 95% similarity and selected one strain to represent each species in the curated full-length *nifH* database (n = 1620 *nifH* sequences; 1347 unique taxa). We chose type strains when available, otherwise we chose strains with the contig containing the most remaining *nifH* sequences, or lastly strains with the longest *nifH*-containing contig. When multiple taxa were missing classifications for ranks between class and species, but were identical in the provided classifications, we kept only one taxon using the criteria described above.

We excluded *nifH* sequences on contigs <10 kb to minimize erroneous taxonomic comparisons due to misidentified contigs. We set the threshold to 10 kb because previous benchmarking of the widely used metagenomic binning software CONCOCT suggests the pipeline has a species-level binning error probability of <0.05 for contigs >10 kb (Alneberg *et al.*, 2014). This error probability therefore represents the lower limit to taxonomic misclassification during taxonomic profiling for at least one software approach. These sequences are also excluded during the taxonomic inferencing step.

We calculated pairwise percent identity and patristic distance rank cutoffs from phylum to genus using an approach similar to that described in Yarza *et al.*, 2014. Briefly, we calculated global matrices of pairwise percent identities (including gaps) and patristic distances based on the nucleotide alignment and reference tree, respectively. In the calculation of each rank's cutoff, we only recorded comparisons between taxa that shared the same taxonomic identity at that rank but differed in the immediate subrank (*e.g.*, for phylum cutoff, comparisons between two sequences from the same phylum but different classes). We removed groups containing fewer than three comparisons and those identified as patristic distance outliers (median ± 2.5 MAD) to exclude taxonomic groups in which the presence of numerous horizontally-inherited *nifH* sequences skewed the group's comparisons. Estimating cutoffs using only comparisons between vertically-inherited *nifH* sequences is necessary to avoid relaxed pairwise percent identity and patristic distance thresholds. For each taxonomic classification at each rank, we then calculated the first quartile for pairwise percent identity and third quartile for patristic distance based on the recorded comparisons. We used the Hodges-Lehmann estimator based on the remaining pairwise percent identity and patristic distance quartiles for each rank as the rank cutoffs (Figure 2). Basing the cutoffs on the selected quartiles provides flexibility if the relationships in novel lineages deviate from those in the existing dataset, and results in cutoffs that are lower relative to other estimates for *nifH* (*cf.* Gaby *et al.*, 2018).

2.5 Error analysis

To benchmark package performance against other common approaches, we performed a 10-fold cross-validation analysis comparing inferences derived from PPIT, blastn, and blastp using partial-length *nifH* sequences. We created partial-length *nifH* sequences representative of those produced during amplicon sequencing by extracting the region targeted by the *nifH* primers described in Mehta *et al.*, 2003 for each sequence in the reference alignment. These partial-length sequences correspond to

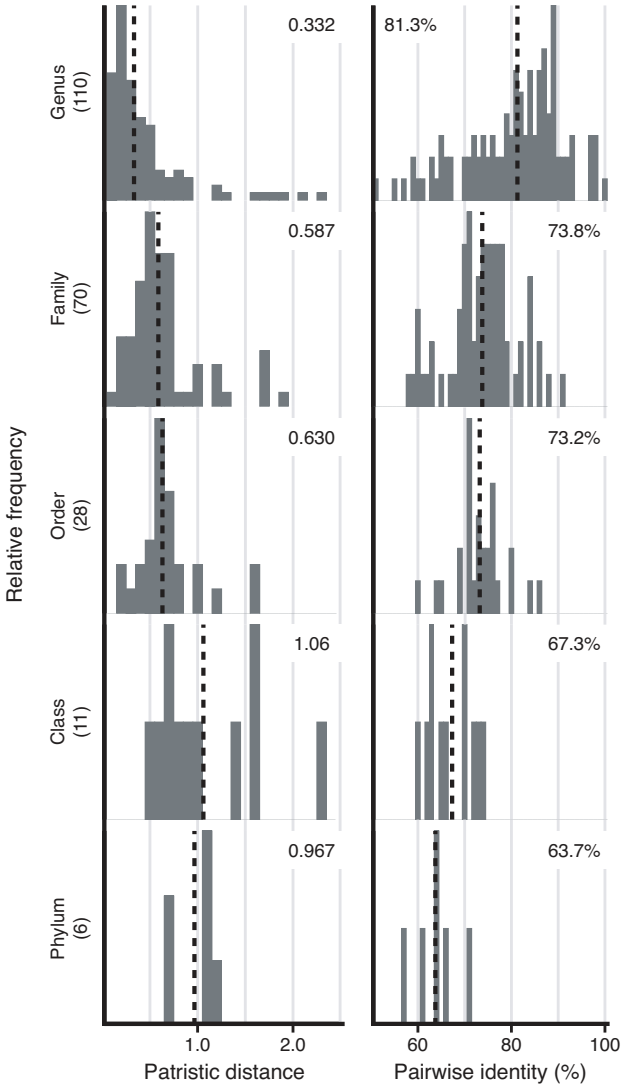


Figure 2 *nifH* patristic distance and pairwise percent identity rank cutoffs used during inferencing. Each histogram contains the distribution of third quartiles for patristic distance and first quartiles for pairwise percent identity based on comparisons between *nifH* sequences from taxa at their lowest shared rank. Each histogram is normalized to its maximum frequency. Total number of taxonomic groups represented at each rank reported in parentheses (*e.g.*, number of unique phyla, number of unique classes, *etc*.). Dashed lines represent set pseudomedians, which are used as the rank cutoffs and reported on each histogram.

the most frequently targeted *nifH* region for amplicon sequencing (Gaby and Buckley, 2012).

For taxonomic inferencing using PPIT, we conducted sequence alignment, tree estimation, calculation of rank cutoffs, and query placement as previously described, except only one run of RAxML was executed for each train set. We clustered the *nifH* sequences in each test set for each species at 5% pairwise identity based on the full nucleotide alignment used for total gene tree estimation (n = 360 – 391 *nifH* seqs. per test set). For inferencing using blastn and blastp, we performed BLAST searches locally using rBLAST (v.0.99.2) on the *nifH* database described in Section 2.2 (containing both partial- and full-length sequences; n = 81 889). We then inferred source organism taxonomy using the match with the lowest E-value. For each fold, the sequences in the test set were removed from the database prior to BLAST.

We analyzed inference accuracy at each rank as a function of the maximum pairwise percent identity between a query in the test set and a reference in the train set. To estimate the probability that a given inference is correct, we assigned correct inferences a value of 1 and incorrect inferences a value of 0, neglecting inferences at ranks for which the query's identity was unassigned in GenBank. To estimate the probability that an inference will be made for a given query, we assigned all inferences a value of 1, regardless of whether they were correct, and unmade inferences a value of 0. We then combined values from each fold into a single dataset and fit LOESS curves to the results from each test fold using first-degree polynomials. We optimized the proportion of inferences per window using the bias-corrected Akaike Information Criterion as implemented in fANCOVA (v.0.5-1), but manually increased the span for PPIT's phylum and class inference probability to remove local fitting errors (Supplementary Figure 3). We interpret the LOESS curves as estimates of inference accuracy.

2.6 Field demonstration

We analyzed *nifH* sequences from a deep-sea sediment sample using PPIT to demonstrate its application on an environmental dataset. Briefly, we collected a sediment core at 3535 m water depth offshore San Francisco, CA, USA using a multicorer on board the R/V *Oceanus* in March 2017. We sectioned the core on board and stored aliquots at -80ºC until DNA extraction in the laboratory using an RNeasy PowerSoil DNA elution kit (Qiagen, cat. no. 12867-25) after RNA extraction using an RNeasy PowerSoil Total RNA kit (Qiagen, cat. no. 12866-25). Here, we present an analysis of the DNA extracted from the 0 – 2.5 cm below seafloor (cmbsf) horizon.

We prepared *nifH* amplicons for sequencing with primers from Mehta *et al.*, 2003 following the two-step PCR protocol described in Kapili *et al.*, 2020 without pooling duplicate reactions. We sent the samples to the UC Davis DNA Technologies Core Facility (Davis, CA, USA) for 2 × 250 bp sequencing on an Illumina MiSeq platform. For a summary of sequencing statistics, including negative control and *nifH* mock community results, see Supplementary Table 2.

We trimmed primer sequences from demultiplexed samples using cutadapt (v.1.18; Martin, 2011), then quality-filtered reads (including chimera removal) and inferred amplicon sequence variants (ASVs) using DADA2 (v.1.12.1; Callahan *et al.*, 2016). We removed ASVs that either did not align to the *nifH* target region or were less than 320 bp or greater than 367 bp in length. We then inserted the remaining ASVs (n = 1245) into the reference alignment and tree described in Section 2.3 using SEPP (v.4.3.5). We analyzed the ASVs using PPIT (v.1.2.0) on a 2016 MacBook Pro, which took 67 mins for phylogenetic neighborhood optimization and 6.2 mins for final taxonomic inference. We then ran PPIT again using the phylum patristic distance cutoff as the optimal phylogenetic neighborhood to attempt identity inference for ASVs that did not have a reference within the calculated optimal distance (n = 46 ASVs).

We additionally analyzed previously published *nifH* Illumina MiSeq data generated from soil samples (Tu *et al.*, 2016). We downloaded the data from NCBI's Sequence Read Archive (BioProject number PRJNA308872) and processed the raw reads using the same protocol as previously described, except removing sequences not between 300 and 325 bp in length to account for a different *nifH* amplicon length.

3 Results and discussion

3.1 PPIT returns a higher proportion of correct inferences than blastn and blastps

PPIT produced the highest proportion of correct taxonomic inferences in comparison to blastn and blastp at each rank for all tested percent identities (Figure 3). For a *nifH* query sequence that shares the minimum pairwise identity with a reference expected for sequences from the same genus (*i.e.*, 81% nucleotide pairwise identity; Figure 2), the probability of PPIT correctly inferring the source organism's phylum is 0.93 while for blastn it is 0.78 (Figure 3). PPIT increasingly outperforms blastn and blastp with respect to the proportion of correct inferences as pairwise identity between the query and closest reference sequence decreases. Inference accuracy when query sequences are dissimilar to existing references is particularly important to studies with environmental samples because metagenomics-based estimates suggest that the majority of microorganisms in marine sediments, soils, the terrestrial subsurface, seawater, and freshwater are phylogenetically dissimilar to cultured organisms (Lloyd *et al.*, 2018). At 72% nucleotide pairwise identity – the lowest pairwise identity at which blastn returned a match – PPIT correctly infers query phylum with 0.77 probability while blastn correctly infers with 0.53 probability (Figure 3). Since other existing software for inferring the source organism identity of *nifH* sequences relies on blastn (Gaby *et al.*, 2018) or blastp (Angel *et al.*, 2018), PPIT is therefore the preferred tool for analyzing a variety of environmental samples.

3.2 PPIT returns fewer total inferences than blastn and blastp

Due to its conservative approach, PPIT draws fewer taxonomic inferences for a given query set than both blastn and blastp (Figure 3). At 81% nucleotide pairwise identity, the probability of PPIT drawing an inference for a given query is 0.67 while the probability of blastn returning a match, and therefore drawing an inference, is 0.92 (Figure 3). For inferences at the phylum and class rank, the disparity in the number of returned inferences increases as pairwise identity between the query and closest reference decreases (Figure 3).

Our results also show that, for ranks class through genus, PPIT's inference probability either locally or globally decreases with increasing identity to a reference sequence (Figure 3). We believe this trend is due to the combined effects of the strict inference criteria, overestimated patristic distance cutoffs for some lineages, and the relatively short terminal branch lengths when queries are similar to reference sequences. PPIT does not infer identity if there is at least one taxonomic inconsistency in the gathered reference subset, which is defined using the patristic distance cutoffs. However, these cutoffs likely are overestimates for some taxonomic lineages. Therefore, the reference subsets for query sequences from these lineages may include sequences from distant taxa along a shared line of vertical descent. This overexploration of local tree topology can result in no inference being drawn due to mischaracterizing vertical inheritance as horizontal inheritance. The issue is exacerbated as sequences increase in similarity to references because the query sequences' terminal branch lengths tend to decrease, permitting greater exploration of the tree topology. The problem of overexploring tree topology is particularly relevant to sequences from sparsely sampled taxonomic groups, such as the *Acidobacteria*, *Chloroflexi*, and *Plancto-*

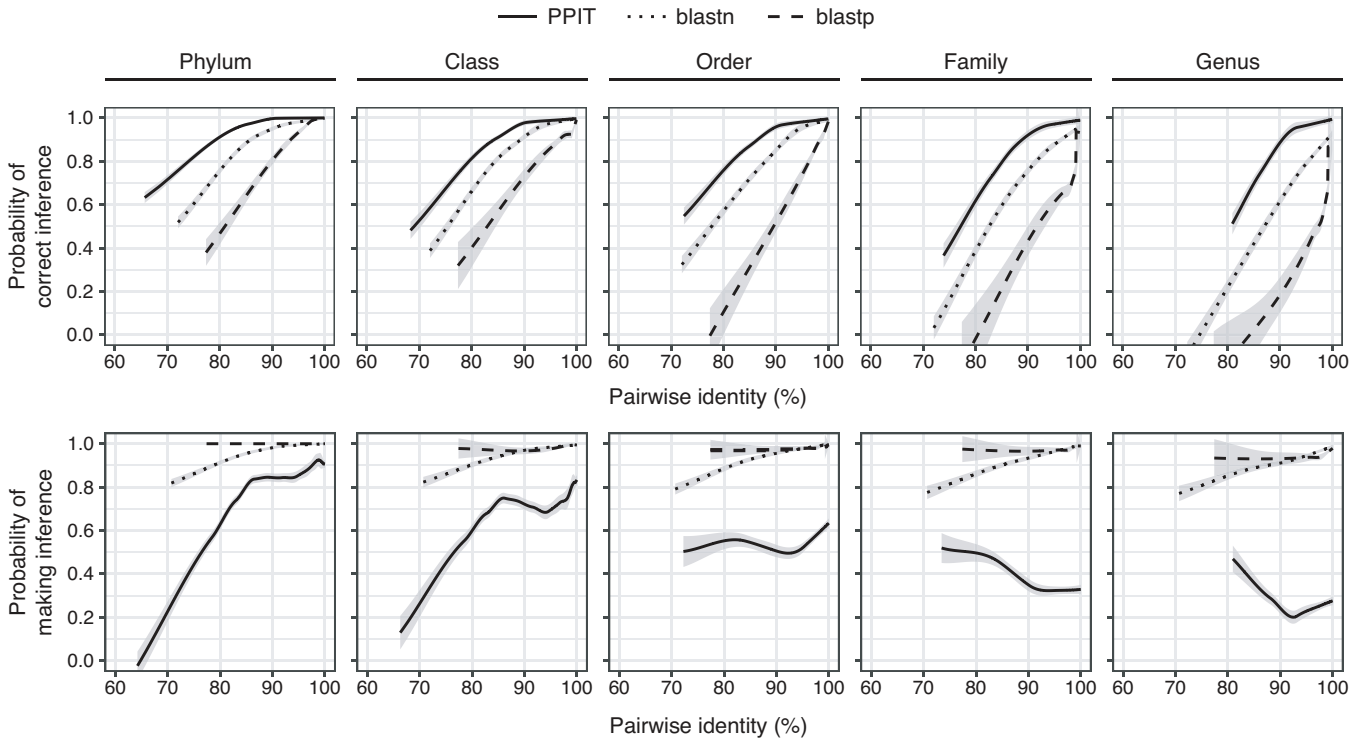


Figure 3 Comparison of taxonomic inference accuracy between PPIT, blastn, and blastp at each rank. First row shows probability that an inference is correct; second row shows probability that an inference is provided for a given query. Lines represent LOESS curves fit to binary accuracy evaluations (1 = inference correct or made, 0 = inference incorrect or not made) with shaded grey regions showing 95% confidence intervals. Probabilities reported as functions of the maximum pairwise percent identity between a query sequence and a reference sequence. Partial-length *nifH* sequences used as queries.

mycetes (Supplementary Table 3). Branches leading to undersampled taxa tend to have underestimated lengths (Fitch and Bruschi, 1987; Heath *et al.*, 2008), which in turn increases the probability of overexploring the local topology for queries if they are placed nearby.

Additionally, some references do not contain taxonomic assignments at all ranks (min. 0.54% of references [missing phylum], max 7.3% [missing class]). Taxonomic inferences based solely on these references will therefore have gaps and contribute to lower inference probabilities. This phenomenon explains the observed dip in inference probability for blastp (Figure 3). However, PPIT does not consider missing taxonomic information as a taxonomic inconsistency. For example, if all the collected references for a given query are taxonomically consistent but one reference contains information at a rank that is missing for the others (or vice versa), an inference would still be made and would include the taxonomic information at the rank missing in the other references.

Lastly, issues inherent to NCBI taxonomy, particularly conflicts between taxonomy and organism phylogeny, may contribute to the fewer inferences PPIT provides. For example, NCBI taxonomy currently classifies *Betaproteobacteria* and *Gammaproteobacteria* as separate classes within the *Proteobacteria*, although recent efforts to improve phylogenetic consistency across taxonomy suggest the *Betaproteobacteria* are more accurately considered an order of the *Gammaproteobacteria* (Parks *et al.*, 2018). To account for this revision, we have modified PPIT to consider the *Betaproteobacteria* an order of the *Gammaproteobacteria* during taxonomic consistency evaluation. Similar manual adjustments will likely be required to future PPIT reference databases until the NCBI taxonomy adopts the recommendations in the Genome Taxonomy Database.

3.3 Field test suggests abundance of potential Deltaproteobacteria diazotrophs

PPIT inferred the taxonomic identity for 44 of the 59 recovered *nifH* ASVs (Table 1). *Deltaproteobacteria* compose most of the inferred *nifH* source organism assemblage, accounting for 36 ASVs and 78% of reads (Table 1). The inferred source organisms are split amongst the *Desulfobacterales* (n = 8) and *Desulfuromonadales* (n = 28), both orders of which were previously implicated as active nitrogen fixers in marine sediments using stable isotope tracer and molecular analyses (Kapili *et al.*, 2020). The phylogenetic placement of the inferred *Desulfuromonadales nifH* ASVs shows that they are most similar to *nifH* sequences from *Desulfuromonadales* organisms either isolated (Liesack and Finster, 1994; Holmes *et al.*, 2004) or sequenced (Barnum *et al.*, 2018) from estuarine sediments near Boston Harbor, MA, USA, Aarhus Bay, Denmark or San Francisco Bay, CA, USA (Figure 4A). See Supplementary Figure 4 for their placement in the context of the entire reference tree.

Furthermore, the recovered *nifH* ASVs share high sequence identity with each of the references in the clade (Figure 4B). The similarity between the *nifH* ASVs recovered here and the reference *nifH* sequences from geographically distant samples suggest that lineages phylogenetically similar to these potential *Desulfuromonadales* diazotrophs are widespread in coastal marine sediments.

In addition to *Proteobacteria*, PPIT inferred *nifH* source organisms from the *Lentisphaerae*, *Kiritimatiellaeota*, and *Planctomycetes* (Table 1). These results demonstrate PPIT's ability to draw inferences for poorly represented taxonomic groups, as only two sequences from the *Lentisphaerae* and *Kiritimatiellaeota*, and ten from the *Planctomycetes*, were present in the reference tree at the time of analysis.

PPIT identified 90% of the sequences in the field test sample as suspected *nifH* homologs (Supplementary Table 2). To demonstrate PPIT use on samples containing fewer *nifH* homologs, we analyzed previously published *nifH* Illumina MiSeq data generated from soil samples (n = 21 samples) from Niwot Ridge, Colorado, USA (Tu *et al.*, 2016) (Supplementary Figure 5). PPIT identified 90% of the ASVs as *nifH* (n = 2921 out of 3260 ASVs), and inferred the identity for an average of 41% (±10%; min. 25%, max. 59%) of the *nifH* ASVs in each sample, with taxonomic inferences spanning 6 phyla. On average, these ASVs composed 60% (±25%; min. 11%, max 96%) of *nifH* reads in each sample (Supplementary Figure 5). The most common reason for failure to infer taxonomic identity was insufficient pairwise percent identity between queries and reference sequences and potential horizontal gene transfer.

Table 1. Inferred taxonomy of *nifH* source organisms in a deep-sea sediment sample using PPIT (n = 59 ASVs). Taxonomy for all lineages reported at the phylum rank except for *Proteobacteria*, which are reported at the class rank. Average pairwise percent identity reported with ±1 standard deviation.

	Number of ASVs	Relative abundance (%)	Avg. id[a] (%)
Deltaproteobacteria	36	78.5	86.0 ± 3.0
Kiritimatiellaeota[b]	4	2.3	82.0 ± 2.2
Gammaproteobacteria	2	1.0	77.4, 81.2[c]
Lentisphaerae	1	0.3	78.3
Planctomycetes	1	0.1	80.6
No inference	15[d]	17.9	76.9 ± 4.5

[a] With respect to nearest reference on the reference tree.
[b] Visual inspection of ASV placement suggests *Kiritimatiellaeota* and/or *Verrucomicrobia* source organisms.
[c] Percent identities for both ASVs shown.
[d] Visual inspection suggests 14 ASVs were erroneously flagged for horizontal gene transfer.

3.4 Recommendations for visually inspecting PPIT results

Visually inspecting the field test ASVs' placements on the input reference tree (*i.e.*, inspecting the SEPP output tree) suggests vertical inheritance for 14 of the 15 query sequences that PPIT conservatively flagged for potential horizontal gene transfer (Figures 5 and 6). For example, ten of these ASVs were placed among reference *nifH* sequences from uncultured *Gammaproteobacteria*, including two species of the proposed genus *Candidatus* Thiodiazotropha, and were flagged for potential horizontal gene transfer at the class rank (Figure 5A). However, PPIT was unable to evaluate taxonomic consistency at ranks lower than class because either no references were within the patristic distance rank cutoff (genus) or no references contained taxonomic assignments at the necessary ranks (family and order). Although PPIT could not infer source organism identity due to references' incomplete taxonomic assignments, our manual inspection reveals that these 10 *nifH* ASVs form a clade with *nifH* sequences from multiple different *Gammaproteobacteria* (Figure 5A). We therefore infer these 10 *nifH* ASVs to be from *Gammaproteobacteria* source organisms because their placement is consistent with vertical inheritance from a *Gammaproteobacteria* ancestor. Visual inspection revealed three additional ASVs that PPIT flagged for horizontal inheritance – one placed with the *Deltaproteobacteria* and two with another *Gammaproteobacteria* clade – although their placements also

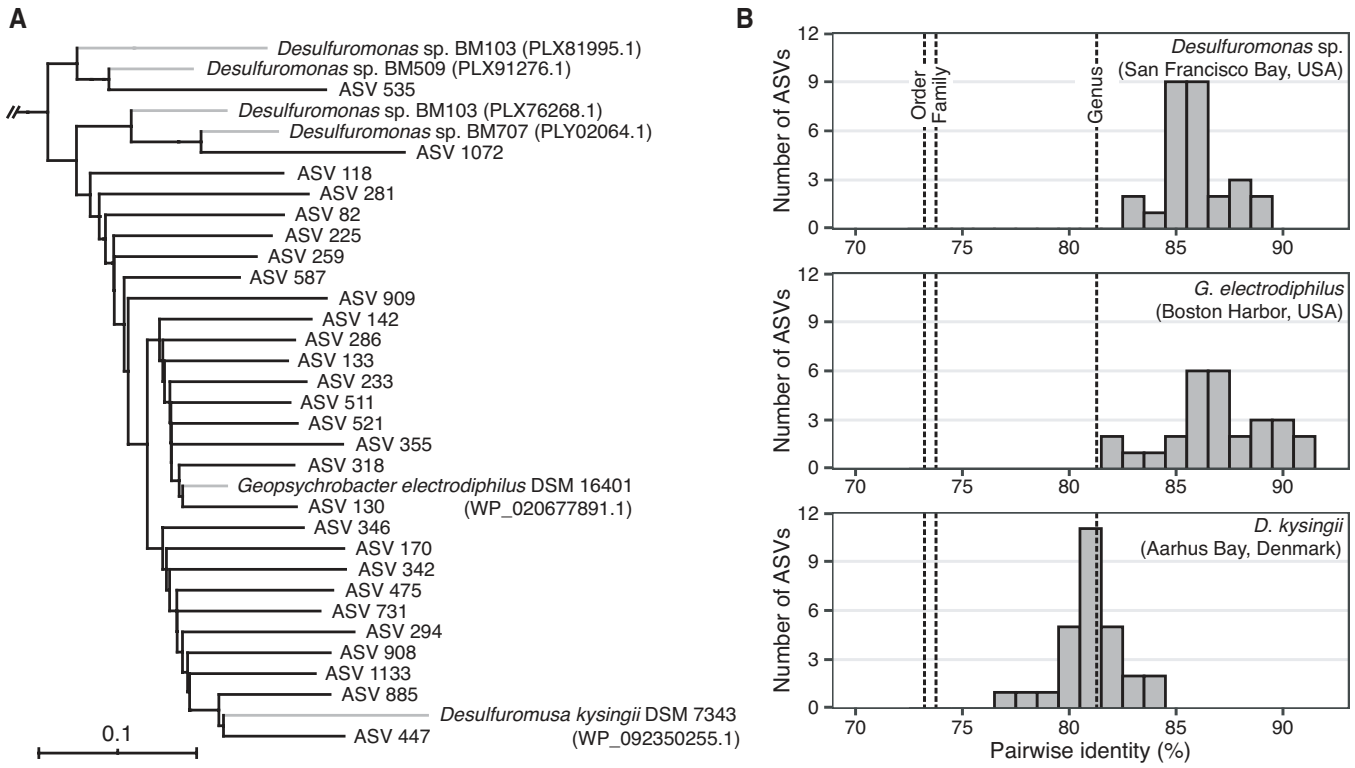


Figure 4 Phylogenetic characterization and sequence novelty of *nifH* ASVs from inferred *Desulfuromonadales* source organisms (n = 28 ASVs). **A,** Phylogenetic placement of *nifH* ASVs. GenBank accession numbers for each reference sequence are reported in parentheses. Branches leading to reference *nifH* sequences are colored grey. Scale bar shows the expected number of substitutions per site. **B,** Histograms of pairwise percent identity between inferred *Desulfuromonadales nifH* ASVs and *nifH* sequences from *Desulfuromonas* sp., *G. electrodiphilus*, and *D. kysingii*. For comparisons between ASVs and the four *nifH* sequences from *Desulfuromonas* sp., only the maximum pairwise percent identity of the four comparisons is reported. Dashed lines indicate order, family, and genus pairwise percent identity rank cutoffs.

appeared consistent with vertical inheritance (Figure 5B,C). The single *nifH* ASV for which taxonomic identity was not inferred did not have a phylogenetic placement that was clearly interpretable (Supplementary Figure 6).

Inspecting the placement of query sequences also reveals some taxonomic inferences that are in partial conflict with reference tree topology. The four ASVs from inferred *Kiritimatiellaeota* source organisms were placed in a clade consisting of *Spirochaetes* (n = 1 reference), *Verrucomicrobia* (n = 1), and *Kiritimatiellaeota* (n = 2). However, another *nifH* ASV was placed among the four inferred *Kiritimatiellaeota nifH* ASVs, but was instead flagged for potential horizontal gene transfer because the *Verrucomicrobia* reference *nifH* sequence was within the genus rank cutoff in addition to the *Kiritimatiellaeota* reference sequences. Visual inspection does not provide a biologically meaningful distinction between the five ASVs' placements (Figure 6). Since the *Kiritimatiellaeota* were classified as *Verrucomicrobia* subdivision 5 until recently (Spring *et al.*, 2016) and currently contain only two *nifH* sequences, we consider it likely that these reference *nifH* sequences were vertically inherited but appear to be horizontally inherited due to poor taxon sampling. To reconcile the conflicting taxonomic inferences, we interpret that all five *nifH* ASVs are from members of the *Kiritimatiellaeota* and/or *Verrucomicrobia*, leaving open the possibility that either one or both phyla are represented in the dataset.

Our manual analysis provided an additional 14 taxonomic inferences, allowing us to infer the taxonomic identity for 58 of the 59 *nifH* ASVs in total. We recommend that users manually inspect the phylogenetic placement of ASVs that PPIT either: (1) flags for potential horizontal gene transfer or (2) infers to belong to poorly sampled phyla. We note that Case 1 may be particularly relevant to samples containing many *Euryarchaeota nifH* sequences (*e.g.*, methane seeps, wetlands) due to challenges stemming from a probable inter-domain, ancient horizontal gene transfer event with the *Clostridia* (Doolittle, 2000; Boyd *et al.*, 2011) (see Supplementary Text 1 and Supplementary Figure 7). To support users in addressing Case 2, we provide the number of reference sequences in each phylum in Supplementary Table 3. Overall, emending PPIT results based on visual inspection helps address the fewer number of taxonomic inferences PPIT provides relative to BLAST-based approaches, and as illustrated here, can result in drawing taxonomic inferences for nearly all the *nifH* sequences in a query set.

3.5 PPIT applicability to other marker genes

PPIT can be applied to other marker genes, particularly genes with many full-length sequences available, such that the estimation of informative reference trees is possible, and/or genes with an evolutionary history of horizontal gene transfer. For example, the depth of publicly-available *amoA* (Pester *et al*., 2012) and *mcrA* (Speth and Orphan, 2018) sequences, used as marker genes for ammonia oxidation and methanogenesis/methanotrophy, respectively, makes them suitable targets for analysis with PPIT. Additionally, PPIT may help address erroneous taxonomic inferences for *nirS* and *nirK* amplicon sequences, used as marker genes for denitrification, due to the prevalence of horizontal gene transfer in the evolutionary history of both genes (Heylen *et al*., 2006).

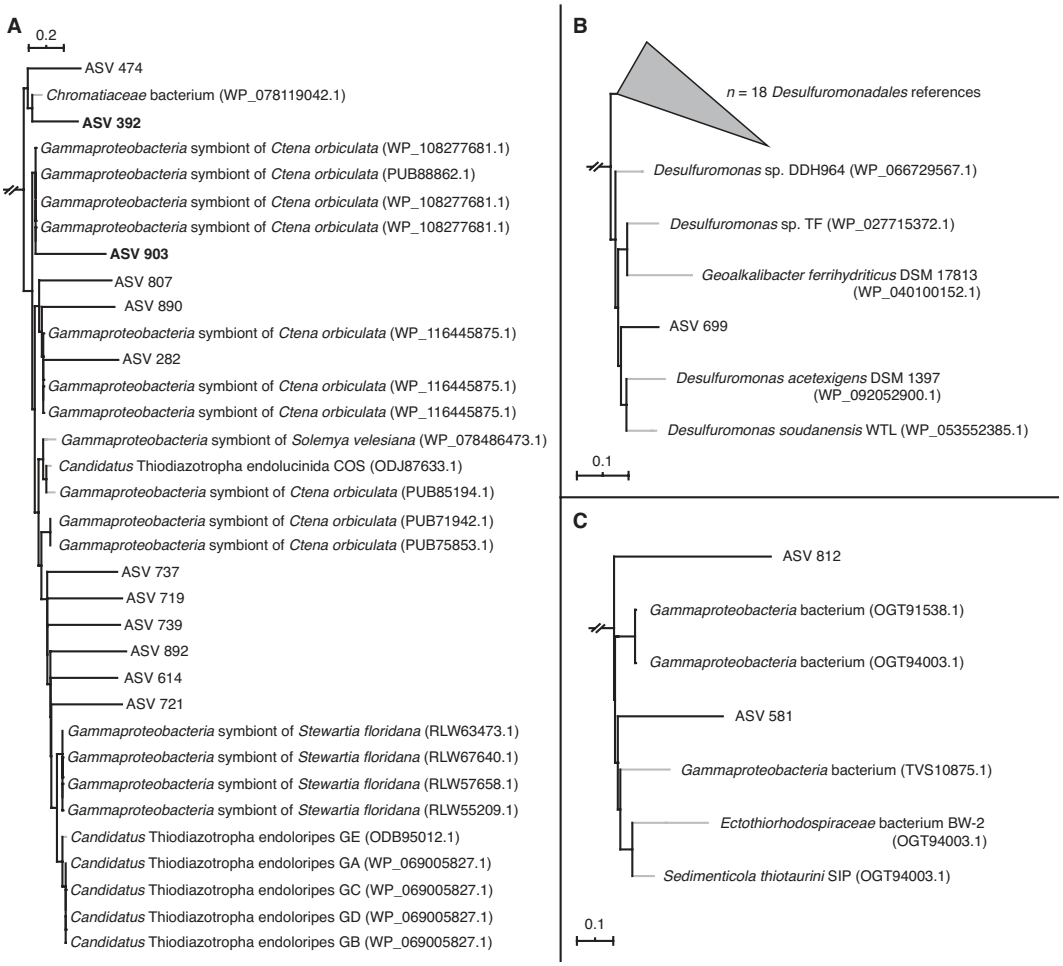


Figure 5 Phylogenetic placement of *nifH* ASVs flagged for horizontal gene transfer. Panels show *nifH* ASVs placed within the **(A)** *Gammaproteobacteria*, **(B)** *Deltaproteobacteria*, and **(C)** a separate *Gammaproteobacteria* clade. GenBank accession numbers for each reference sequence are reported in parentheses. Branches leading to reference *nifH* sequences are colored grey. Scale bar shows the expected number of substitutions per site. Note that PPIT inferred the taxonomic identity for the bold ASVs in Panel A (ASV 392 and 903).

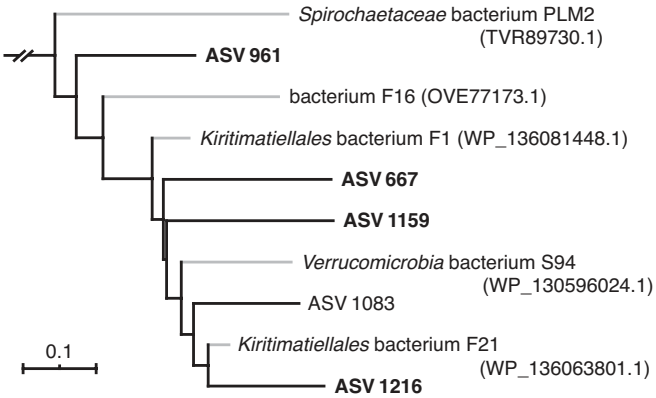


Figure 6 Phylogenetic placement of inferred *Kiritimatiellaeota nifH* ASVs (bold) and one ASV flagged for potential horizontal gene transfer (ASV 1083). GenBank accession numbers for each reference sequence are reported in parentheses. Branches leading to reference *nifH* sequences are colored grey. Scale bar shows the expected number of substitutions per site.

To adapt PPIT to the analysis of other genes, users need to supply the appropriate reference alignment, gene tree estimate, accompanying taxonomic information, and taxonomic rank cutoffs (see Sections 2.2 – 2.4).

4 Conclusions

We present PPIT to address the need for accurate, high-throughput taxonomy inferencing of *nifH* source organisms. We show that PPIT returns a higher proportion of correct taxonomic inferences than BLAST-based approaches at each taxonomic rank at the cost of fewer total inferences. However, we show that visual inspection of query sequence placements can recover the difference in taxonomic inferences. Furthermore, as the depth of the reference *nifH* database increases, we expect PPIT accuracy and inference rate to increase. We demonstrate PPIT on *nifH* amplicons from deep-sea sediment and, combined with visual inspection of results, were able to draw taxonomic inferences for 58 of the 59 *nifH* sequences detected, including inferences from sparsely sampled phyla. We therefore recommend PPIT over alternative inferencing approaches for most environmental studies based on its higher accuracy and higher throughput. PPIT is readily integrated into current bioinformatic workflows, and allows users to substitute the provided *nifH*

resources with resources specific to other genes. PPIT is therefore a tool broadly applicable to the analysis of metabolic marker gene sequences.

Acknowledgements

We thank all members of the Dekas Laboratory for valuable discussions and feedback. We also thank Hanon McShea for assistance with computing resources and two anonymous reviewers whose comments improved the quality of this manuscript.

Funding

This work was supported by the National Science Foundation (OCE-1634297 to AED and a Graduate Research Fellowship to BJK).

Conflict of Interest: none declared.

References

Alneberg,J. *et al.* (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods*, **11**, 1144–1146.

Amir,A. *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *Am. Soc. Microbiol.*, **2**, 1–7.

Angel,R. *et al.* (2018) Evaluation of primers targeting the diazotroph functional gene and development of NifMAP – a bioinformatics pipeline for analyzing *nifH* amplicon data. **9**, 1–15.

Bagheri,H. *et al.* (2020) Detecting and correcting misclassified sequences in the large-scale public databases. *Bioinformatics*, **36**, 4699–4705.

Barnum,T.P. *et al.* (2018) Genome-resolved metagenomics identifies genetic mobility, metabolic interactions, and unexpected diversity in perchlorate-reducing communities. *ISME J.*, **12**, 1568–1581.

Bertics,V.J. *et al.* (2013) Occurrence of benthic microbial nitrogen fixation coupled to sulfate reduction in the seasonally hypoxic Eckernförde Bay, Baltic Sea. *Biogeosciences*, **10**, 1243–1258.

Boyd,E.S. *et al.* (2011) A late methanogen origin for molybdenum-dependent nitrogenase. *Geobiology*, **9**, 221–232.

Braker,G. *et al.* (1998) Development of PCR primer systems for amplification of nitrite reductase genes (*nirK* and *nirS*) to detect denitrifying bacteria in environmental samples. *Appl. Environ. Microbiol.*, **64**, 3769–3775.

Callahan,B.J. *et al.* (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**, 581–583.

Collavino,M.M. *et al.* (2014) *nifH* pyrosequencing reveals the potential for location-specific soil chemistry to influence N_2-fixing community dynamics. *Environ. Microbiol.*, **16**, 3211–3223.

Cottrell,M.T. and Cary,S.C. (1999) Diversity of dissimilatory bisulfite reductase genes of bacteria associated with the deep-sea hydrothermal vent polychaete annelid *Alvinella pompejana*. *Appl. Environ. Microbiol.*, **65**, 1127–1132.

Doolittle,R.F. (2000) Searching for the common ancestor. *Res. Microbiol.*, **151**, 85–89.

Farnelid,H. *et al.* (2011) Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One*, **6**, e19223.

Fernández-Méndez,M. *et al.* (2016) 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.

Fitch,W.M. and Bruschi,M. (1987) The evolution of prokaryotic ferredoxins – with a general method correcting for unobserved substitutions in less branched lineages. *Mol. Biol. Evol.*, **4**, 381–394.

Fujita,Y. *et al.* (1992) The *nifH*-like (*frxC*) gene is involved in the biosynthesis of chlorophyll in the filamentous cyanobacterium *Plectonema boryanum*. *Plant Cell Physiol.*, **33**, 81–92.

Fujita,Y. and Bauer,C.E. (2000) Reconstitution of light-independent protochlorophyllide reductase from purified Bchl and BchN-BchB subunits: In vitro confirmation of nitrogenase-like features of a bacteriochlorophyll biosynthesis enzyme. *J. Biol. Chem.*, **275**, 23583–23588.

Gaby,J.C. *et al.* (2018) Diazotroph community characterization via a high-throughput *nifH* amplicon sequencing and analysis pipeline. **84**, 1–16.

Gaby,J.C. and Buckley,D.H. (2012) A comprehensive evaluation of PCR primers to amplify the *nifH* gene of nitrogenase. *PLoS One*, **7**, e42149.

Heath,T.A. *et al.* (2008) Taxon sampling and the accuracy of phylogenetic analyses. *J. Syst. Evol.*, **46**, 239–257.

Heller,P. *et al.* (2014) ARBitrator: a software pipeline for on-demand retrieval of auto-curated *nifH* sequences from GenBank. *Bioinformatics*, **30**, 2883–2890.

Heylen,K. *et al.* (2006) The incidence of *nirS* and *nirK* and their genetic heterogeneity in cultivated denitrifiers. *Environ. Microbiol.*, **8**, 2012–2021.

Holmes,D.E. *et al.* (2004) Potential role of a novel psychrotolerant member of the family *Geobacteraceae*, *Geopsychrobacter electrodiphilus* gen. nov., sp. nov., in electricity production by a marine sediment fuel cell. *Appl. Environ. Microbiol.*, **70**, 6023–6030.

Igai,K. *et al.* (2016) Nitrogen fixation and *nifH* diversity in human gut microbiota. *Sci. Rep.*, **6**, 1–11.

Izquierdo,J.A. and Nüsslein,K. (2006) Distribution of extensive *nifH* gene diversity across physical soil microenvironments. *Microb. Ecol.*, **51**, 441–452.

Janssen,S. *et al.* (2018) Phylogenetic placement of exact amplicon sequences improves associations with clinical information. *mSystems*, **3**, e00021-18.

Kapili,B.J. *et al.* (2020) Evidence for phylogenetically and catabolically diverse active diazotrophs in deep-sea sediment. *ISME J.*, **14**, 971–983.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Kirshtein,J.D. *et al.* (1991) Amplification, cloning, and sequencing of a *nifH* segment from aquatic microorganisms and natural communities. *Appl. Environ. Microbiol.*, **57**, 2645–2650.

Letunic,I. and Bork,P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, **47**, W256–W259.

Liesack,W. and Finster,K. (1994) Phylogenetic analysis of five strains of gram-negative, obligately anaerobic, sulfur-reducing bacteria and description of *Desulfuromusa* gen. nov., including *Desulfuromusa kysingii* sp. nov., *Desulfuromusa bakii* sp. nov., and *Desulfuromusa succinoxidans* sp. nov. *Int. J. Syst. Bacteriol.*, **44**, 753–758.

Lloyd,K.G. *et al.* (2018) Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *mSystems*, **3**, e00055-18.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.

McMurdie,P.J. and Holmes,S. (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, **8**, e61217.

Mehta,M.P. *et al.* (2003) Phylogenetic diversity of nitrogenase (*nifH*) genes in deep-sea and hydrothermal vent environments of the Juan de Fuca Ridge.

Appl. Environ. Microbiol., **69**, 960–970.

Miller,M.A. *et al.* (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gatew. Comput. Environ. Work. GCE 2010*, 1–8.

Mirarab,S. *et al.* (2012) SEPP: SATé-Enabled Phylogenetic Placement. *Pacific Symp. Biocomput.*, **17**, 247–258.

Moore,S.J. *et al.* (2017) Elucidation of the biosynthesis of the methane catalyst coenzyme F430. *Nature*, **543**, 78–82.

Nguyen,L.T. *et al.* (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

Ohkuma,M. *et al.* (1995) Phylogeny of symbiotic methanogens in the gut of the termite *Reticulitermes speratus*. *FEMS Microbiol. Lett.*, **134**, 45–50.

Pace,N.R. *et al.* (1986) The analysis of natural microbial populations by ribosomal RNA sequences. In, Marshall,K.C. (ed), *Advances in Microbial Ecology*. Springer, Boston, pp. 1–55.

Parks,D.H. *et al.* (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996.

Pester,M. *et al.* (2012) *amoA*-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of *amoA* genes from soils of four different geographic regions. *Environ. Microbiol.*, **14**, 525–539.

Price,M.N. *et al.* (2010) FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Raymond,J. *et al.* (2004) The natural history of nitrogen fixation. *Mol. Biol. Evol.*, **21**, 541–554.

Rotthauwe,J.H. *et al.* (1997) The ammonia monooxygenase structural gene *amoA* as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. *Appl. Environ. Microbiol.*, **63**, 4704–4712.

Rozewicki,J. *et al.* (2019) MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.*, **47**, W5–W10.

Singer,G.A.C. *et al.* (2019) Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Sci. Rep.*, **9**, 1–12.

Speth,D.R. and Orphan,V.J. (2018) Metabolic marker gene mining provides insight in global *mcrA* diversity and, coupled with targeted genome reconstruction, sheds further light on metabolic potential of the *Methanomassiliicoccales*. *PeerJ*, **2018**, 10.7717/peerj.5614.

Spring,S. *et al.* (2016) Characterization of the first cultured representative of *Verrucomicrobia* subdivision 5 indicates the proposal of a novel phylum. *ISME J.*, **10**, 2801–2816.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Suyama,M. *et al.* (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, 609–612.

Tu,Q. *et al.* (2016) Biogeographic patterns of soil diazotrophic communities across six forests in the North America. *Mol. Ecol.*, **25**, 2937–2948.

Wang,J. *et al.* (2016) Molecular ecology of *nifH* genes and transcripts along a chronosequence in revegetated areas of the Tengger Desert. *Microb. Ecol.*, **71**, 150–163.

Ward,D.M. *et al.* (1990) 16S rRNA sequences reveal numerous uncultured inhabitants in a well-studied natural community. *Nature*, **345**, 63–65.

Yarza,P. *et al.* (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Publ. Gr.*, **12**, 635–645.

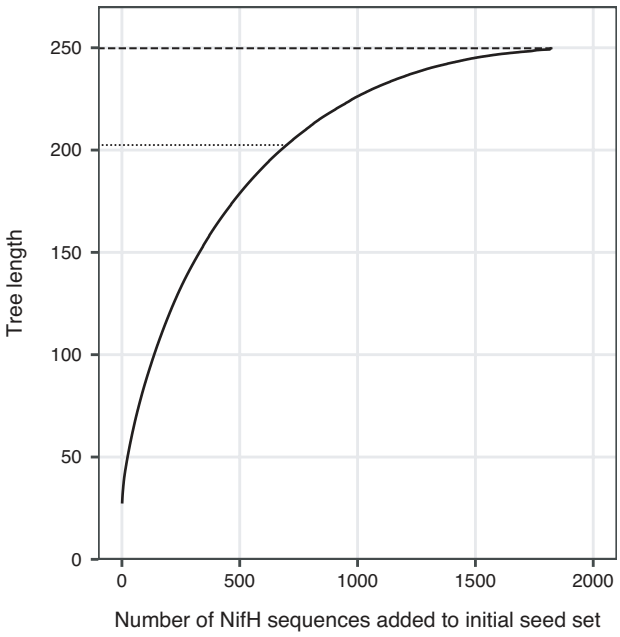
Zheng,K. *et al.* (2016) The biosynthetic pathway of coenzyme F430 in methanogenic and methanotrophic archaea. *Science*, **354**, 339–342.

Supplementary Information

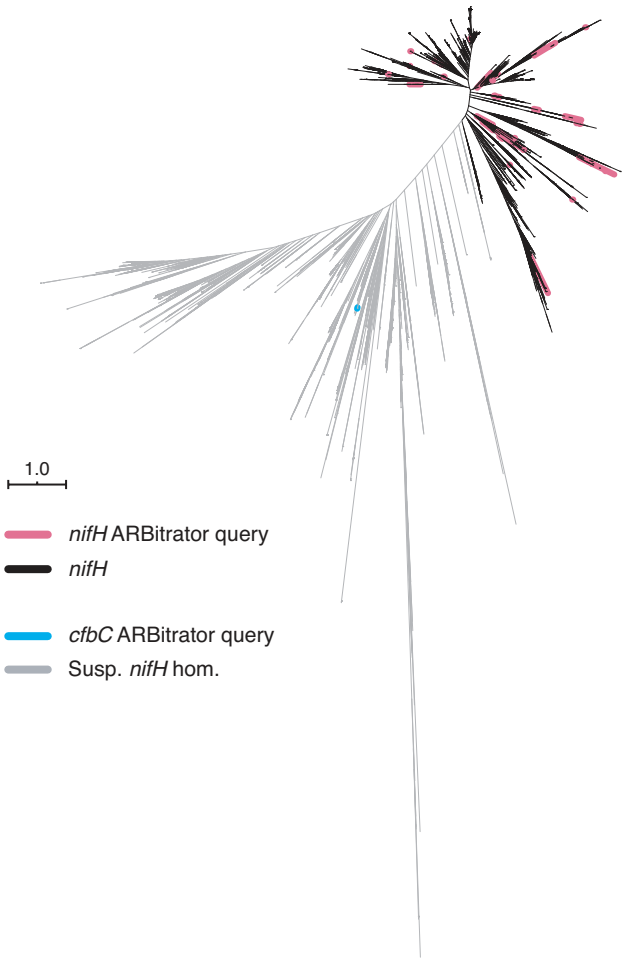
PPIT: an R package for inferring microbial taxonomy from *nifH* sequences

Bennett J. Kapili* & Anne E. Dekas*
Department of Earth System Science, Stanford University, Stanford, CA 94305, USA

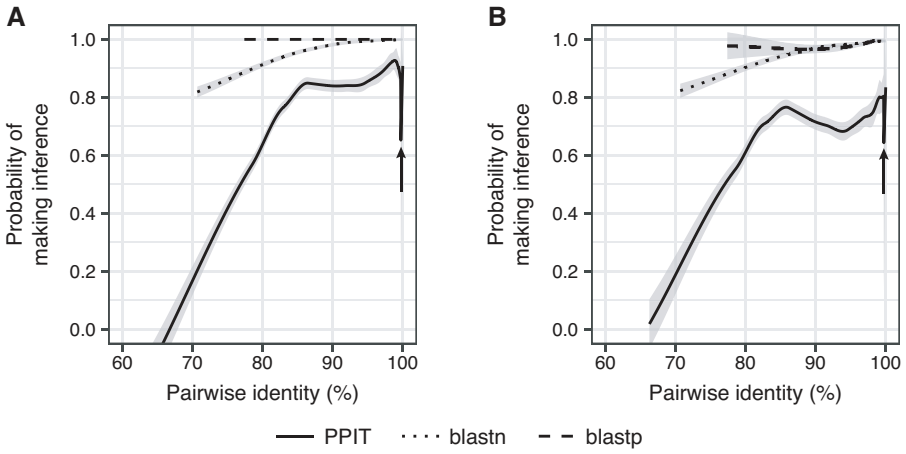
Email: kapili@stanford.edu, dekas@stanford.edu
To whom correspondence should be addressed.



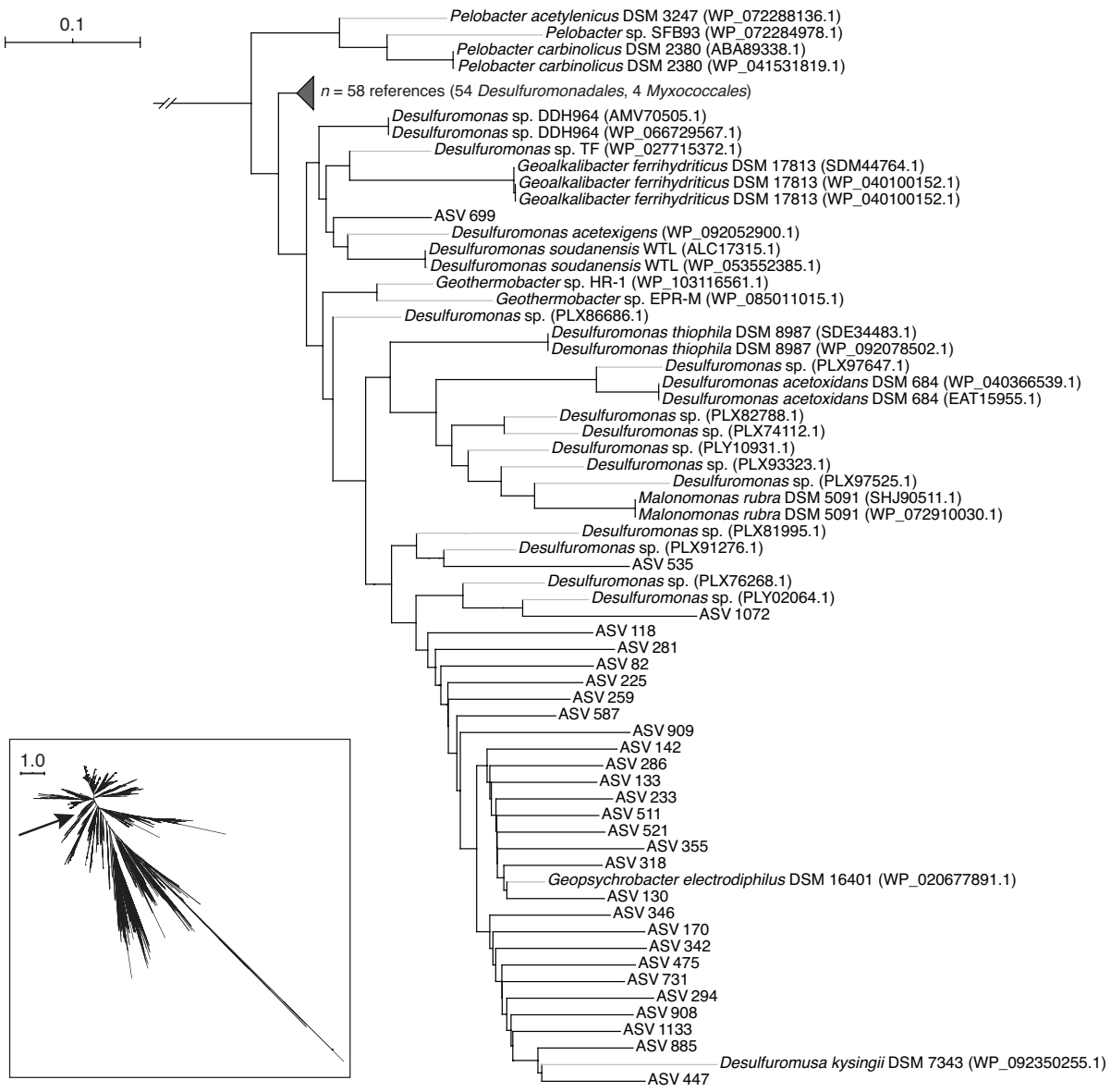
Supplementary Figure 1 Summed edge length of NifH subtree as a function of sequences added to initial seed set (n = 37). Horizontal dashed line: summed edge length of total initial NifH tree (n = 6292 sequences). Horizontal dotted line: summed edge length of initial NifH tree containing only the NifH sequences selected for MAFFT-DASH alignment (n = 736).



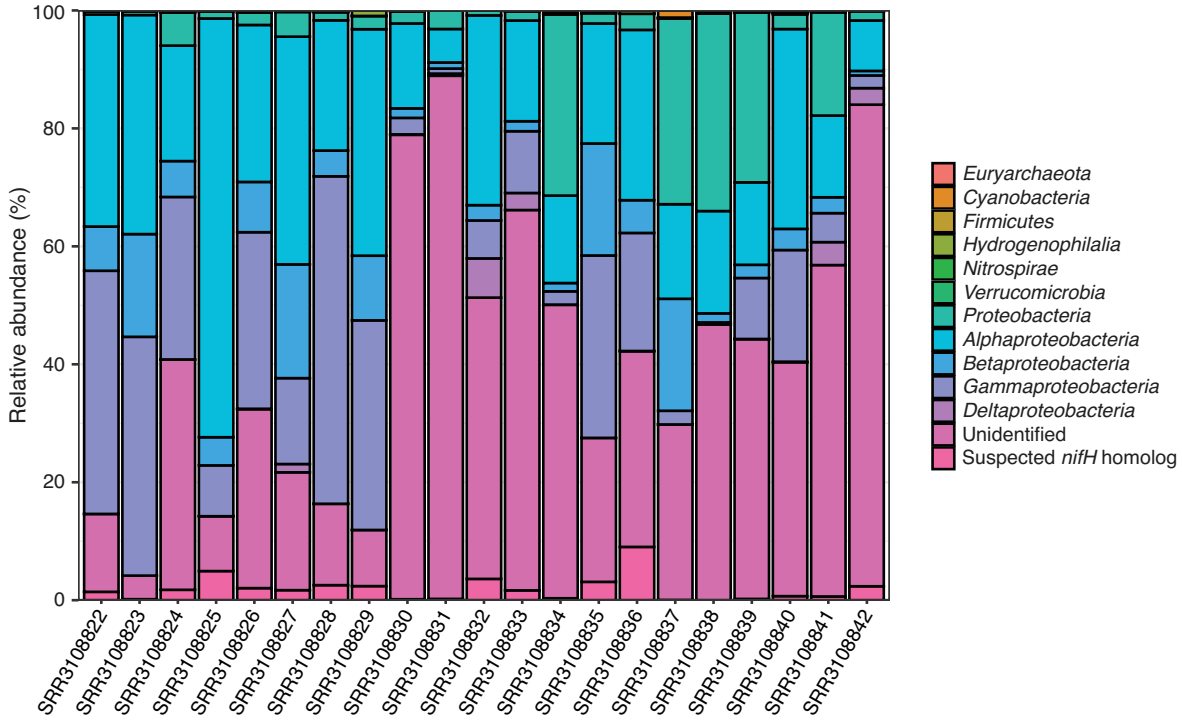
Supplementary Figure 2 Identification of suspected *nifH* homologs. Scale bar shows the expected number of substitutions per site.



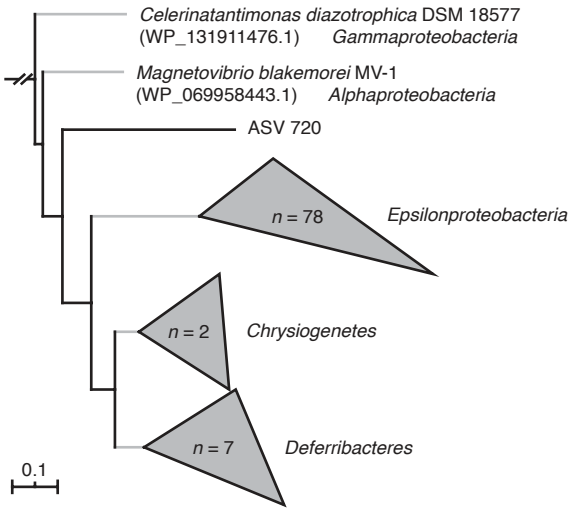
Supplementary Figure 3 LOESS curve fitting errors during estimation of PPIT inference probability when the span is optimized using the bias-corrected Akaike Information Criterion at the **(A)** phylum and **(B)** class ranks. Arrows point to fitting errors. To smooth the fitted curve, the phylum span was manually adjusted from 0.33 to 0.50; class span was manually adjusted from 0.41 to 0.50.



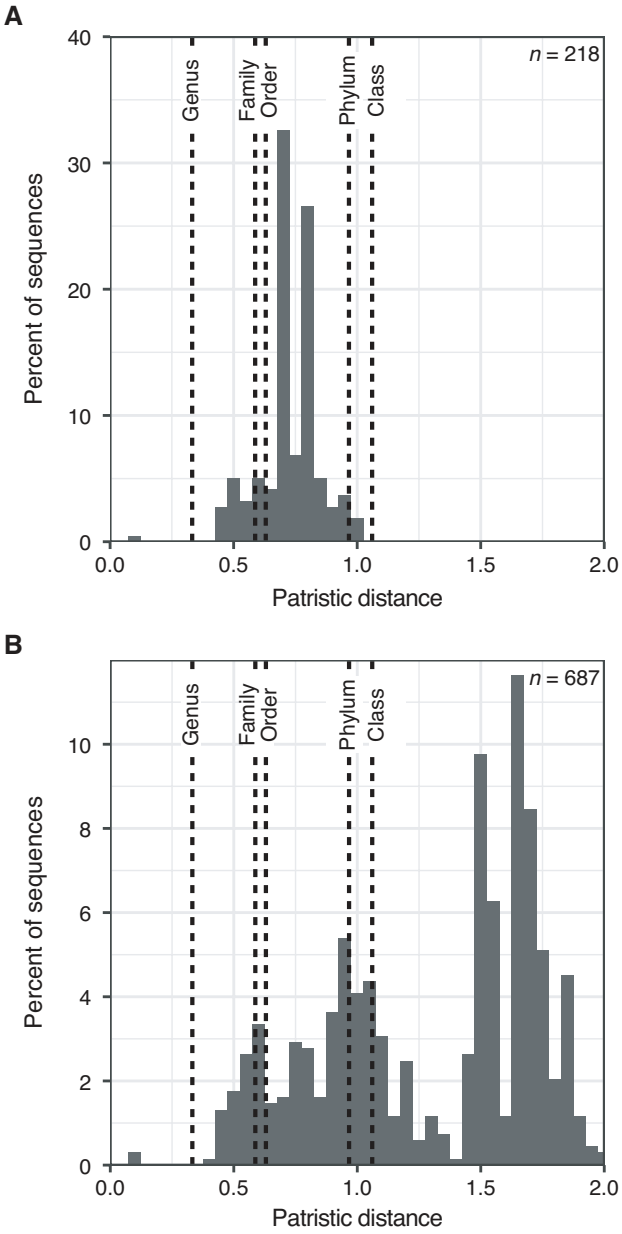
Supplementary Figure 4 Phylogenetic placement of *nifH* ASVs from inferred *Desulfuromonadales* source organisms. Branches leading to reference *nifH* sequences are colored grey; tip labels include GenBank accession numbers in parentheses. Inset shows full *nifH* reference tree (homologs collapsed into grey wedge) with arrow pointing to the clade containing the inferred *Desulfuromonadales nifH* ASVs. Scale bar shows expected number of nucleotide substitutions per site.



Supplementary Figure 5 Relative abundance of previously published (Tu *et al.*, 2016) *nifH* amplicon data generated from Niwot Ridge, Colorado, USA soil samples. Sequence Read Archive run numbers corresponding to each sample are reported on the x-axis. Taxa are reported at the phylum rank, except for the *Proteobacteria*, which are reported at the class rank when possible. *nifH* ASVs inferred to belong to the *Proteobacteria*, but for which the class could not be inferred, are shown as *Proteobacteria*.



Supplementary Figure 6 Visual inspection of the query sequence (ASV 720) for which taxonomic identity was not inferred. GenBank accession numbers for each reference sequence are reported in parentheses. Branches leading to reference *nifH* sequences are colored grey. Scale bar shows the expected number of substitutions per site.



Supplementary Figure 7 Minimum patristic distances between reference *Euryarchaeota* and *Clostridia nifH* sequences. **A,** Minimum patristic distance to a reference *Clostridia nifH* sequence for each reference *Euryarchaeota nifH* sequence (*n* = 218). **B,** Minimum patristic distance to a reference *Euryarchaeota nifH* sequence for each reference *Clostridia nifH* sequence (*n* = 687). Vertical dashed lines represent patristic distance taxonomic rank cutoffs.

Supplementary Text 1

If we construct a histogram displaying the patristic distances between the reference *Euryarchaeota* (n = 218) and *Clostridia nifH* (n = 687) sequences on the reference tree, then we can explore a "worst-case-scenario" for inferring the taxonomy of query sequences from either of these two lineages (Supplementary Figure 7). Examining the existing tree is equivalent to examining a tree with query sequences that have terminal branch lengths of 0, which occurs when query sequences are identical to reference sequences. Analysis of the histograms reveals that 0.5% of *Euryarchaeota* references contain at least one reference *Clostridia nifH* within the genus cutoff, 12.4% within the family cutoff, and 17.9% within the order cutoff (Supplementary Figure 7A). All *Euryarchaeota* references contain at least one reference *Clostridia nifH* within the class cutoff and would therefore be flagged for horizontal gene transfer if an inference could not be made at the genus, family, or order ranks (Supplementary Figure 7A). The value of visually inspecting the phylogenetic placement of query sequences flagged for horizontal gene transfer is further underscored since inferencing power reduces to zero for the *Euryarchaeota* if an inference is not able to be made by the order rank.

Potential query sequences that are identical to existing *Clostridia* references, however, fare better. Only 0.3% of *Clostridia* references contain at least one reference *Euryarchaeota nifH* within the genus cutoff, 6.8% within the family cutoff, and 9.8% within the order cutoff (Supplementary Figure 7B). Similar to the *Euryarchaeota*, 36.2% of reference *Clostridia nifH* sequences contain a *Euryarchaeota nifH* sequence within the class cutoff and 28.2% contain one within the phylum cutoff (Supplementary Figure 7B). Under the worst-case-scenario, PPIT remains capable of inferring the taxonomy for query sequences from the *Euryarchaeota* and *Clostridia* despite an inter-domain, albeit ancient, horizontal gene transfer event.

Supplementary Table 1 Nucleotide accession numbers and coding sequence start/stop positions of diverse *bchL* sequences (*n* = 18). Sequences identified using a similar approach to how the ARBitrator *nifH* query set was gathered. Partial sequences created using the same method that was used for creating partial *nifH* sequences.

Nucleotide accession	CDS start	CDS stop
NC_016025.1	102584	103462
LGEI01000001.1	275528	274635
NZ_JAAXMP010000003.1	367812	366946
NZ_GL501404.1	172500	173321
NZ_NKFP01000006.1	2844089	2844952
NZ_ANCI01000003.1	4181932	4181066
NC_022600.1	4148557	4147730
NZ_CP017675.1	2741729	2742583
NZ_CAIY01000027.1	61572	62474
JAAHGA010000056.1	46105	45643
NZ_JRFE01000050.1	20067	20945
NZ_JH980292.1	744116	743250
NZ_WBXO01000004.1	231259	232092
NZ_CP011454.1	1570572	1571468
JAAUUL010000242.1	1925	2824
SACE01000163.1	2184	1288
NHKM02000072.1	1956	2846
NZ_LJHQ01000063.1	35793	36722

Supplementary Table 2 Summary of read filtering statistics and replicate similarity.

	3500 m, 0 – 2.5 cmbsf		Mock community			Negativ
	Rep. 1	Rep. 2	Rep. 1	Rep. 2	Rep. 3	
No. raw reads	22506	21299	12494	15887	13381	77
No. filtered reads	13774	12748	9593	12147	9938	6
No. *nifH*[a] reads	1148	1059	6213	14193	6656	1
No. homolog[a] reads	12626	11689	3380	4611	3282	5
No. mock ASVs[b]	–	–	12	12	12	–
No. unexpected ASVs[c]	–	–	2	2	1	6
No. unexpected reads	–	–	4	7	3	6
Bray-Curtis similarity	0.88		0.91 ± 0.05			–

[a]As determined by PPIT.

[b]Out of 12 (8 *nifH*, 4 homologs; even community).

[c]ASVs not matching a sequence that was included in the mock community. One chimera was share among all three replicates.

Supplementary Table 3 Number of *nifH* references in each phylum used for taxonomic inferencing in PPIT (v.1.2.0).

Phylum	Num. references
Euryarchaeota	197
Acidobacteria	1
Actinobacteria	40
Aquificae	5
Bacteroidetes	52
Candidatus Dadabacteria	1
Candidatus Margulisbacteria	4
Chlorobi	34
Chloroflexi	9
Chrysiogenetes	2
Cyanobacteria	412
Deferribacteres	7
Elusimicrobia	1
Fibrobacteres	29
Firmicutes	940
Fusobacteria	1
Kiritimatiellaeota	2
Lentisphaerae	2
Nitrospirae	22
Planctomycetes	10
Proteobacteria	4083
Spirochaetes	28
Thermodesulfobacteria	4
Verrucomicrobia	31