Rapid annotation of *nifH* gene sequences using Classification and Regression Trees (CART) facilitates environmental functional gene analysis

Ildiko E. Frank, Kendra A. Turk-Kubo, Jonathan P. Zehr

Affiliation: Department of Ocean Sciences, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA

Corresponding author: Ildiko Frank; email: ildiko1frank@gmail.com; cell: 650-796-3509; mail: 790 Esplanada Way, Stanford, CA 94305

Running title: Rapid *nifH* annotation

Summary

The n*ifH* gene is a widely used molecular proxy for studying nitrogen fixation. Phylogenetic classification of *nifH* gene sequences is an essential step in diazotroph community analysis that requires a fast automated solution due to increasing size of environmental sequence libraries and increasing yield of *nifH* sequences from high-throughput technologies. We present a novel approach to rapidly classify *nifH* amino acid sequences into well-defined phylogenetic clusters that provides a common platform for comparative analysis across studies. Phylogenetic group membership can be accurately predicted with decision tree-type statistical models that identify and utilize signature residues in the amino acid sequences. Our classification models were trained and evaluated with a publicly available and manually curated *nifH* gene database containing cluster annotations. Model-independent sequence sets from diverse ecosystems were used for further assessment of the models' prediction accuracy. We demonstrate the utility of this novel sequence binning approach in a comparative study where joint treatment of diazotroph assemblages from a wide range of habitats identified habitat-specific and widely-distributed diazotrophs and revealed a marine – terrestrial distinction in community composition. Our rapid and automated phylogenetic cluster assignment circumvents extensive phylogenetic analysis of *nifH* sequences; hence, it saves substantial time and resources in nitrogen fixation studies.

Introduction

Biological nitrogen fixation is a prokaryote-driven biogeochemical process that sustains the trophic web in nitrogen limited habitats including vast areas of the ocean (Vitousek and Howarth, 1991), where it is linked to atmospheric carbon dioxide fixation and carbon export from surface waters (Falkowski, 1997). Since the microbial majority is recalcitrant to cultivation (Rappe and Giovannoni, 2003), and biogeochemical interactions cannot be investigated in a laboratory setting (DeLong, 2009), cultivation-independent molecular surveys are indispensable in assessing microbial diversity and metabolic complexity (Zehr et al., 2009). The small subunit ribosomal RNA gene sequence is a universal phylogenetic marker (Lane et al., 1985); however, it does not provide information on nitrogen fixation capabilities. Therefore, the *nifH* gene coding for the Fe protein of the nitrogenase enzyme was proposed as molecular proxy for nitrogen fixation potential (Zehr and McReynolds, 1989), which led to the recognition of high diversity of nitrogen-fixing microbes (diazotrophs) (Zehr et al., 1995, Ueda et al., 1995) and the discovery of a widely distributed marine nitrogen-fixing organism with an unusual physiology (Zehr, 2011). Several curated *nifH* gene databases are available to the public (Cole et al., 2009, Heller et al., 2014, Gaby and Buckley, 2014). Our publicly available *nifH* database at www.jzehrlab.com has been a valuable resource facilitating numerous investigations of nitrogen-fixing assemblages; examples cover marine environments (Bombar et al., 2011, Bonnet et al., 2013, Farnelid et al., 2011, Fong et al., 2008, Halm et al., 2012, Hamersley et al., 2011, Moisander et al., 2007, Moisander et al., 2008, Rahav et al., 2013, Turk et al., 2011, Zehr et al., 2007), terrestrial environments (Desai et al., 2013, Duc et al., 2009, Furnkranz et al., 2008, Steward et al., 2004), and host symbionts (Desai and Brune, 2012, Lema et al., 2012, Mohamed et al., 2008, Yamada et al., 2007).

Diazotroph community composition analysis requires classifying *nifH* sequences into annotated taxonomic groups. Despite some disagreement (Raymond et al., 2004, Gaby and Buckley, 2011), a phylogenetic division of four main clusters (Chien and Zinder, 1994) is widely used in publications. Clusters IV and/ or V are irrelevant for nitrogen fixation studies, since they are *nifH*-like genes not involved in the fixation of atmospheric nitrogen. Cluster I is composed mainly of Cyanobacteria, α-, β-, γ- and δ-Proteobacteria, Firmicutes, and Actinobacteria (Zehr et al., 2003). Sequences from prokaryotes with alternative nitrogenase enzymes (Betancourt et al., 2008) and from methanogenic Archaea (Chien et al., 2000) form cluster II. The distantly related sequences in cluster III come predominantly from anaerobic organisms, such as *Chlorobium*, *Desulfovibrio*, *Clostridium*, and *Acetobacterium* genera.

In order to analyze diazotroph diversity in any given environment, a finer-level sequence grouping in diazotroph community analyses is commonly employed. This may be accomplished either by merging sequences into more manageable but a priori unknown number of groups, called operational taxonomic units (OTUs), or by classifying sequences into subclusters, intra-cluster branches of the *nifH* phylogenetic tree (Zehr et al., 2003). Cluster I contains well-defined subclusters with high phylogenetic similarity to 16S rRNA gene tree topology, e.g. subcluster 1A contains δ-Proteobacteria, subcluster 1B is comprised of Cyanobacteria, etc. (Zehr et al., 2003). OTUs can be calculated with distance-based hierarchical clustering (Schloss et al., 2009), or with fast clustering algorithms suited for large data sets, for example, CD-HIT (Li and Godzik, 2006) or UCLUST (Edgar, 2010). In contrast to the phylogenetically established subclusters, the resulting OTU groups are study specific and not comparable across ecosystems. Subclusters provide a common platform, but labeling newly acquired sequences currently necessitates a

63 time-consuming and computationally demanding "placing on the tree" approach (Matsen et al.,
64 2010, Price et al., 2010).

65 In addition to the above phylogeny-based sequence characterization, sequence similarity
66 and sequence composition are also utilized to classify sequences into established taxonomic
67 groups (Bazinet and Cummings, 2012). Basic Local Alignment Search Tool (BLAST) matches
68 sequences against an annotated database (Altschul et al., 1990); however, in this sequence
69 similarity-based approach, sequences without close relatives are likely to be misidentified. Naïve
70 Bayesian Classifier is a fast sequence composition-based technique that calculates
71 oligonucleotide (8-mer) frequencies. It is implemented and widely used for rRNA sequence
72 classification in the Ribosomal Database Project (Wang et al., 2007), but was found to be inferior
73 to BLAST, for example, in classifying sequences of *pmoA*, a functional marker gene of
74 methanotrophs (Dumont et al., 2014).

75 Since the introduction of *nifH* as a molecular marker for nitrogen fixation, there has been
76 exponential growth in the number of *nifH* genes deposited into the NCBI GenBank.
77 Furthermore, this gene is increasingly being used in next generation sequencing studies (Farnelid
78 et al, 2011, Collavino et al., 2014, Bentzon-Tilia et al., 2014). A rapid and easily-implemented
79 approach is needed to facilitate and standardize the essential step of classifying environmental
80 *nifH* sequences.

81 We hypothesized that a handful of single positions in the *nifH* amino acid sequence
82 contain sufficient information to classify *nifH* amplicons into phylogenetic groups. Our graphical
83 exploration with WebLogo (Crooks et al., 2004) confirmed previously reported conserved
84 residues in the *nifH* sequence (Schlessman et al., 1998); strings of single letters clearly set apart
85 four formerly named regions: P loop (*Azotobacter vinelandii* residues A9 - A19), Switch I (A38-

A48), Metal Cluster Coordination (A86-A102), and Switch II (A125-A142). Between these extended constant regions, the sequence contains variable positions, including the previously identified 60's loop (Schlessman et al., 1998), a potential for amino acid signatures of phylogenetic groups.

Decision tree-based statistical modeling is capable of identifying signature residues and utilizing them for sequence annotation. Classification And Regression Trees (CART), the most popular tree-based methodology in data mining and machine learning, was specifically designed to handle large complex data sets (Breiman et al., 1983). The CART model consists of a hierarchy of simple decision rules, each based on a single predictor – in our case a position in the amino acid sequence – which are organized and graphically presented as a binary decision tree. Among its many applications, CART has been used in various ecological studies to model abundance data and correlate environmental and biological parameters (De'ath and Fabricius, 2000, Pesch et al., 2011, Clarke et al., 2008, Usio et al., 2006), but has not been tested for environmental amplicon classification.

This study presents a comparative diazotroph community analysis utilizing a novel cluster assignment of *nifH* amino acid sequences based on CART decision trees. The statistical models selected signature residues that contain sufficient information to distinguish among the established phylogenetic clusters as well as to screen for two groups of nitrogen-fixing marine cyanobacteria, *Trichodesmium* and *Candidatus* Atelocyanobacterium thalassa (UCYN-A). Our rapid phylogenetic cluster annotation was tested on a wide range of environmental sequence sets and was utilized in cross-ecosystem analyses that identified widely-distributed and habitat-specific nitrogen-fixers and revealed key differences between diazotroph communities in marine and terrestrial ecosystems.

Results and Discussion

Main cluster annotation

CART statistical modeling was used to develop a decision tree that successfully assigns *nifH* amino acid sequences into the well-established four main clusters. A large annotated training set (details in Appendix S1) was obtained from the above discussed *nifH* sequence database with residues labeled according to the *Azotobacter vinelandii* sequence from A1 to A290 (Schlessman et al., 1998). Instead of sequence similarity and phylogeny calculations, the CART model labels sequences based on just few residues selected by an iterative algorithm (details in Appendix S2). The decision tree for main cluster annotation contains only three sequence positions (A109, A49, and A53) and four terminal nodes, each corresponding to a main cluster (Figure 1). For example, cluster I is primarily identified by the signature phenylalanine (F) at position A109, which is replaced by a similarly hydrophobic leucine or methionine in clusters II and III. The predicted cluster labels matched the database annotation with high (98%) accuracy (Table 1). The model was evaluated by ten-fold cross-validation, as well as by predicting main clusters in a model-independent data set (details in Appendix S1) derived from soil samples (Collavino et al., 2014).

Subcluster annotation

Decision trees based on a handful of signature residues were also found effective in assigning sequences into subclusters. Cluster I, mainly Proteobacteria and Cyanobacteria, is split into several groups that exhibit approximate correspondence with the 16S rRNA phylogeny (Zehr et al., 2003). The classification tree contains twelve terminal nodes, one for each subcluster label in the database (Figure S1). The decision nodes involve only eight residues, most in the so called 60's loop located at the interface between the two nitrogenase components in the 3D

protein structure. Classification accuracy is above 96% in most subclusters (Table 1). The highest error rate is due to confusion between subclusters 1K and 1J, which are neighboring branches of the phylogenetic tree, both containing α- and β-Proteobacteria.

Subcluster 1B, composed exclusively of cyanobacterial *nifH* sequences, holds special interest in ecological studies in aquatic as well as terrestrial environments. Cyanobacteria sequences can be distinguished from other cluster I sequences with 97% accuracy based on a single decision node, position A103. It is located in the same alpha helix as the signature residue A109 of cluster I. These two residues contain sufficient information for a simple screen: if A109=F (phenylalanine) and A103=I (isoleucine), then with high probability the sequence belongs to a Cyanobacteria. This algorithm resulted in 7% false negative (3,087/3,304 Cyanobacteria identified) and 1% false positive (138 / 13,263) in the training set. Most (128) of the sequences erroneously marked as Cyanobacteria are from cluster 1E, a cluster made up primarily of Firmicutes from the *Paenibacillus* genera. These results are consistent with reports that *nifH* genes and homologues from some *Paenibacillus* species appear to cluster with cyanobacterial *nifH* (Choo et al., 2003), which underscores the need for additional screening in environments where cluster 1E organisms are expected to be present.

A CART model also assigns cluster II sequences, mainly from organisms with alternative nitrogenase, into subclusters with high accuracy (Table S1)**.** The tree contains four decision nodes (A54, A67, A115, and A117) and five terminal nodes corresponding to each subcluster (Figure S2).

Subcluster annotation is problematic within cluster III, which is composed of sequences mostly from anaerobic organisms belonging to diverse Archaea and Bacteria taxa. Cluster III is characterized by long branch lengths and deep bifurcations, and is less congruent with 16S rRNA

phylogeny than cluster I. Most primary positions in the decision nodes are between A76 and A87 (Figure S3), a region that straddles two beta sheets towards the edge of the 3D structure and is distinct from the 60's loop utilized in annotating cluster I and II sequences. Building upon original cluster III annotations (Zehr et al., 2003), the database currently defines eighteen subclusters. The low accuracy of our model (Table S1) indicates that the current subcluster designations are an overfit of sequence variation and suggests that the phylogeny-based subcluster definition in this group needs to be revisited.

Cluster IV contains the most divergent sequences that belong to non-nitrogen-fixing Archaea and Bacteria, but does not include the protochlorophyllide reductases, which are filtered out during creation of the ARB database (Heller et al., 2014). The primary positions in the decision nodes are located considerably further from the N-terminus than those selected in models for the other main clusters (Figure S4). Despite the high amino acid variability at most positions, our subcluster labels match the database annotation with 97% accuracy (Table S1).

Annotation of targeted cyanobacterial groups

With an appropriately annotated training set, similar decision trees can be developed in order to identify sequences at genus, species, strain, or ecotype levels. To demonstrate this, we targeted *Candidatus* Atelocyanobacterium thalassa (UCYN-A) and *Trichodesmium* spp., two cyanobacterial groups that are important nitrogen-fixers in the oligotrophic marine environment (Zehr, 2011). Because sequences are labeled only by main and subclusters in the *nifH* database, training set with genus-level annotation was obtained via calculating operational taxonomic unit (OTU) groups from the 3,304 cyanobacterial sequences pulled from the *nifH* database utilizing the rapid screen discussed above. We used our own grouping algorithm (details in Appendix S3) coded in the R environment (R Development Core Team, 2013), because it provided us full

understanding and control of the results and seamless interface with network graphs for visualization and CART modeling, both performed in R. Protein BLAST search against the reference protein database identified the representative sequence in each OTU group.

Binning at 95% amino acid sequence similarity resulted in 179 cyanobacterial OTUs. The largest group was composed of 222 sequences, and their representative sequence was identified by BLAST as a *Trichodesmium* sequence; hence, we obtained a training set with two classes of sequences, *Trichodesmium* and non-*Trichodesmium*. In addition to correctly labeling the 222 sequences, our CART model with two decision nodes marked eight others as *Trichodesmium*; indeed, their closest match found by BLAST was *Trichodesmium* but only at 92-94% identity. From any mix of *nifH* amino acid sequences, a rapid screen based on just 4 residues identifies a sequence as *Trichodesmium* if: A109 = F and A103 = I and A78 = K and A52 = A.

The second largest group contained 176 sequences with the representative sequence identified by BLAST as UCYN-A. In addition to correctly labeling all sequences in this UCYN-A annotated OTU, a CART model identified two more sequences as UCYN-A; their closest relative in the protein database was confirmed as UCYN-A, but only at 93% identity. A rapid screen for UCYN-A is also simple: if A109 = F and A103 = I and A78 = I and A85 = L, then the sequence is likely from UCYN-A. Thus the decision tree approach is a powerful way to pull out specific sequence types from a mix of *nifH* sequences, and shows promise for quickly screening results from large datasets, such as those from next generation sequencing runs.

Evaluation of CART annotation

Performance of the CART model-based cluster assignment was further evaluated by analyzing *nifH* sequence sets deposited in NCBI. We selected twenty-five studies producing a total of 6,170 sequences and covering a wide range of environments: open ocean and sea surface,

hydrothermal vent, soil, rhizosphere, phyllosphere, sediment, and termite symbionts (Table 2). In the original papers, sequences were assigned to main clusters using non-standardized "placing on the tree" based approaches. As in the training set, the dominant portion (82%) of these environmental sequences belongs to cluster I, and only few sets cover all four main clusters (Table 2). Main cluster labels predicted by CART match the large groups originally identified in these publications. Almost all marine sequence sets contained *Trichodesmium* and/or UCYN-A sequences that were successfully identified by our CART model-based rapid screening (Table 2). Direct evaluation of our subcluster assignment is not possible because each study uses different annotations, which are not deposited as metadata in GenBank. Thus, comparing CART derived subcluster labels with group labels defined in these studies is challenging and can only be done qualitatively by comparing group proportions.

We compared cluster annotations resulting from our CART models with those obtained through building neighbor joining trees (TREE) and by protein BLAST search (details in Appendix S4) using sequence data from the twenty-five studies. In both cases, we used *nifH* sequences from 600 annotated genomes as a reference set, as assigned in the curated *nifH* database (Heller et al., 2014). Comparison of the three different cluster annotation methods is summarized on Venn-like diagrams (Figure 2.) For 99% of the sequences, all three methods assigned the same main cluster label. This almost perfect agreement drops to 87% when considering subcluster labels. Note, that 3% of the sequences were assigned three different subcluster labels by the three annotation methods. If we join groups 1K and 1J, as suggested earlier, then the three-way agreement increases to 90%. Furthermore, if we disregard the subcluster assignments within cluster III and annotate all cluster III sequences with a single label, based on the previously discussed problem with this main cluster, the three-way match

reaches 94%. Although these three commonly used methods have the potential to yield quite similar results, the time required for each analysis in terms of computational resource usage (details in Table S2 and Appendix S4) and manual analysis varies substantially. Although the CART model and modern phylogenetic programs, e.g. FastTree (Price et al., 2009) can analyze very large sets of sequences in a short amount of time, the CART model output provides *nifH* cluster annotations, while the task of inferring cluster identities from a phylogenetic tree with >100,000 nodes requires further bioinformatic analysis. No pipelines currently exist to streamline this process for functional gene trees.

We also evaluated CART annotations by binning each sequence set into OTUs and visualizing each set on a network graph. A network of dissimilar sequences, e.g. the *grass* set, contains many singletons and small OTUs of 2-5 sequences (Figure 3A), whereas a network of more similar sequences, e.g. the *M2* set, is dominated by few OTUs that comprise the majority of sequences (Figure 3B). OTUs were generated at 98, 95, 90, 85, and 80 percent amino acid sequence similarity levels (Table 2). In each set, the CART annotation error rate, i.e. proportion of sequences mislabeled by CART, was calculated. A sequence was flagged as mislabeled if its CART-derived cluster assignment did not match the cluster label of the majority of its associated OTU (see example in Figure 3A).

In order to quantify the correspondence between cluster / subcluster annotation and OTUs delimited at each similarity cutoff, we calculated the Gini impurity index (Breiman et al., 1983), i.e. a weighted average of cluster or subcluster label impurity (ranging 0 – 1) in each OTU (Figure 4). At 98% similarity, the error rate and the label impurity were very low in each set and, as expected, both statistics increased with decreasing similarity cutoff. As observed in the annotation method comparison, high error rate and impurity were often due to binning together

1K and 1J sequences or mislabeling cluster III sequences. In general, main cluster labels match OTUs generated up to 80% similarity cutoff, while subcluster labels become incongruent with OTUs above 95% similarity cutoff (Figure 4).

Widely-distributed and habitat-specific diazotrophs

Our novel annotation technique enabled us to identify widely-distributed and habitat-specific diazotrophs by examining phylogenetic cluster structures across twenty-five ecosystems (Table 2). We annotated all 6,170 sequences by main cluster and by cluster I subcluster labels, resulting in fifteen phylogenetic groups. Due to the small proportion of cluster II, III, and IV sequences, they were not annotated at a finer level. The largest group, labeled 1K, contains 1,464 sequences, followed by 1B with 992 and 1G with 882 sequences. Sequences within each cluster group were binned at 98% amino acid sequence similarity, a cut-off typically used for species level identification. Only positions A45 – A153, which corresponds to the most commonly used *nifH* primer sets (Gaby and Buckley, 2012), were considered in this analysis. Each of the fifteen clusters was visually explored on network graphs and representative sequences of the largest twenty-five OTUs were identified by protein BLAST (Table 3).

Intra-OTU sequence origin and representative sequence identity revealed diazotrophs present in a wide range of ecosystems, as well as organisms unique to a particular marine or terrestrial environment. The representative of the largest OTU, labeled 1K, matched three organisms at 100% identity: *Burkholderia xenovorans*, *Sphaerotilus natans*, and *Methyloversatilis discipulorum*. Four additional large 1K OTUs contain sequences of mixed origin and were identified at 98 – 100% identity (Table 3). The recovery of these sequences in multiple habitats suggests that they are either sourced from PCR reagent contaminants or from truly ubiquitous diazotrophs. Like PCRs targeting the 16S rRNA gene, *nifH* PCRs are highly

subject to contamination from genomic DNA present in reagents used in the extraction of nucleic acids, the laboratory environment, and/or the PCR reagents (Zehr et al., 2003). Common *nifH* contaminants include *Burkholderia* spp. Some studies take care to remove these potential sequences by analyzing negative controls (e.g. Farnelid et al., 2011), but it is widely assumed that contaminant-sourced sequences are submitted to GenBank. Further work is needed to determine whether these mixed origin OTUs are indeed contaminant sequences; however, it is striking that most 1K OTUs do not come from mixed origins, which strengthens the argument that they are habitat specific.

The two largest 1B OTUs originating from various marine environments were identified at 100% identity as *Trichodesmium erythraeum* and UCYN-A, respectively (Table 3). The largest OTUs of cluster 1G also come from diverse marine environments, but we cannot determine the source organism because they have only 94-97% amino acid identity to cultivated γ-Proteobacteria (Table 3). While several of the large OTUs were of strictly marine origin, there were only two terrestrial-only large OTUs: the second largest OTU in cluster 1J (99% similar to *Acidithiobacillus ferrivorans*) and the largest OTU in cluster II (99% similar to *Dickeya paradisiacal*) (Table 3). Diazotrophs unique to a specific environment were identified in *Sponge, leaf, Baltic, rhiz,* and *soilD* sets.

Difference between marine and terrestrial ecosystems

The true power of uniformly applied cluster labels becomes evident when comparing diazotroph communities across various ecosystems. We hypothesized that diazotroph taxa distribute unevenly across ecosystems with the main contrast being between marine and terrestrial habitats. We explored similarities and differences among diazotroph assemblages by comparing cluster proportions calculated from sequence counts of twenty-five ecosystems and

fifteen cluster labels as annotated by CART. Chi-squared test supports that ecosystems and clusters are not independent, i.e. there is a significant ($p < 2.2e-16$) difference in cluster proportions across habitats. Similar test on a 2*15 table (marine vs. terrestrial aggregates as rows) also shows significant row – column dependence ($p < 2.2e-16$). Correspondence analysis projection on the first two components indicates a clear separation of marine and terrestrial environments (Figure 5). Sequence sets from open ocean surfaces (dark blue circles) group at top left surrounded by sets derived from sea environments (light blue circles). These marine ecosystems are mainly characterized by a high proportion of 1B and 1G labeled sequences (red triangles). In contrast, the non-marine sets spread diagonally with the termite symbiont set (yellow circle) at the bottom containing largely sequences from clusters 4 and 3 (red triangles), followed by three closely grouped sediment sets (brown circles) dominated by anaerobic cluster 3 sequences. The terrestrial sets (green circles) at top right are distinguished by a high proportion of clusters 1K, 1J, 1F, and 1E (red triangles). Cluster 2, projected close to the center (0, 0) of the two component plot, dominates the third component and sets apart the *Deep* ocean environment along a third axis (not displayed). There are three outlier marine sets that project together with terrestrial sets: *Baltic, M1,* and *P3*. In all three cases, the high proportion of cluster 1K sequences gives these sets a terrestrial profile, which may be due in part to the lack of recovery of the two most dominant marine diazotroph OTUs, *Trichodesmium* and UCYN-A, in these studies. The *phylo* ecosystem, a rainforest phyllosphere dominated by 1B sequences, is an outlier showing a marine rather than a terrestrial characteristic. With some explainable exceptions, marine and terrestrial diazotroph communities are distinct from each other and dominated by different phylogenetic clusters. ANOSIM analysis, which compares intra- and inter-group variances, confirmed the above qualitative assessment. The test indicates significant

difference between marine and terrestrial habitats (R = 0.12, p = 0.041) as well as significant difference among the five types of habitats (R = 0.14, p = 0.048). Differential analysis of gene count data based on negative binomial distribution (DESeq2) provided further support for our hypothesis that certain clusters are more prevalent in marine while others are more typical in terrestrial ecosystems. Significant increase in clusters 1B (p = 0.081) and 1G (p = 0.0003) were found in marine habitats, while significant increase in clusters 1A (p = 0.034) and 1E (p = 0.006) were found in terrestrial habitats. In sedimentary habitats significant increase was found in cluster 3 (p = 0.035) and significant decrease in clusters 1B (p = 0.0004) and 1K (p = 0.001).

Conclusion

With statistical modeling, we supported our hypothesis that the *nifH* amino acid sequence contains signature residues with sufficient information for phylogenetic cluster membership prediction. Similar classification models could be developed for other functional genes, making use of available annotated training sets. Although subcluster divisions have been applied to characterize sequences from a wide range of ecosystems (Bonnet et al., 2013, Duc et al., 2009, Mohamed et al., 2008, Hamersley et al., 2011, Moisander et al., 2008, Collavino et al., 2014), these phylogenetically defined groups are not prevalent in the diazotroph studies. Instead, diversity and sample similarity analyses are often based on operational taxonomic units defined at various similarity levels (Hamilton et al., 2011, Hsu and Buckley, 2009, Turk et al., 2011, Gaby and Buckley, 2011), or on study specific sequence groups called clades (Deslippe and Egger, 2006), operational protein units (Lema et al., 2012), or simply groups (Man-Aharonovich et al., 2007). As demonstrated in our cross-ecosystem analysis, uniformly applied sequence characterization reveals information not present in individual studies. Furthermore, our novel annotation method, which is available for general use in the form of Python scripts at

339 www.jzehrlab.com under the *nifH* tab, does not require the computationally demanding
340 calculation of phylogenies and it can be accomplished with less resources and expertise; hence, it
341 would greatly facilitate the exploration and comparison of diazotroph communities.

342 ## Acknowledgements

343 We thank Julie Robidart for helpful discussions, Zak Peters for a first draft of a Python script,
344 and Ed Boring (UCSC) for IT support. This research was funded by a Gordon and Betty Moore
345 Foundation Marine Investigator award (J.P.Z.) and the National Science Foundation Science
346 Center for Microbial Oceanography Research and Education (C-MORE, grant no. EF-0424599).

Figure Legends

Figure 1.

Graphical representation of the CART classification model that successfully assigns sequences into main clusters based on three residues. The four terminal nodes correspond to clusters I, II, III, and IV. Each decision node lists the sequence position and the amino acids in the left group of sequences. For example, if a sequence has phenylalanine (F), tryptophan (W), or tyrosine (Y) at position A109, then it belongs to cluster I.

Figure 2.

Venn-like graphical summary of match among cluster labels obtained from three annotation methods: CART, protein BLAST, and phylogenetic analysis (TREE). Three- and two-way matches are reported in terms of percentage of labeled sequences from 25 data sets.

Figure 3.

Network graphs of two *nifH* sequence sets, each binned at 98% amino acid similarity. Sequences, represented by circles, are color coded according to their predicted main cluster (I=black, II=red, III=green, IV=blue) and labeled by their predicted subcluster. Sequences binned together into an OTU are shown as connected circles. Identical subcluster labels within OTUs support correct annotation by CART.
A: *grass* set of 67 sequences grouped into 41 OTUs; note a mislabeled 4G sequence.
B: *M2* set of 65 sequences grouped into 11 OTUs.

Figure 4.

Gini impurity indices plotted in function of OTU sequence similarity cutoffs (80, 85, 90, 95, and 98 percent). Each point is calculated as weighted average of cluster (top row) or subcluster (bottom row) label impurities of individual OTUs at given similarity cutoff. For better visualization, points are labeled according to ecosystems and the 25 data sets are grouped into three origins: ocean (6), sea (8), and terrestrial (11).

Figure 5.

Ecosystem similarities in terms of diazotroph community composition are visualized on a correspondence analysis projection (first two components cover 24% and 20% variance). The twenty-five environmental sequence sets are represented by circles, color coded by ecosystems (dark blue = open ocean, light blue = sea, green = terrestrial, brown = sediment, and yellow = symbiont), while the fifteen *nifH* clusters are symbolized by red triangles.

Table 1.

CART classification accuracy calculated on the training set (Fit%), on a model-independent test set (Test%), and estimated by ten-fold cross-validation (Pred%). The column labeled Total includes percent of accurate annotation of all sequences, i.e. all clusters or all subclusters aggregated.

A: Accuracy of main cluster annotation.

	Cluster				Total
	I	**II**	**III**	**IV**	
Size	17,321	542	3,876	758	22,497
Fit%	99	98	95	96	98
Pred%	99	96	94	95	98
Test%	99	100	92	100	98

B: Accuracy of subcluster annotation in cluster I.

	Subcluster												Total
	1	**1A**	**1B**	**1C**	**1D**	**1E**	**1F**	**1G**	**1J**	**1K**	**1O**	**1P**	
Size	4	1,436	3,304	260	566	131	60	1,461	1,451	3,239	304	421	12,637
Fit%	100	99	96	98	99	96	98	98	94	93	91	96	96
Pred%	75	99	95	97	99	97	95	98	94	93	89	95	96
Test%	0	99	100	100	17	94	--	100	90	82	100	100	91

Table 2.

Origin and structure of the twenty-five environmental *nifH* sequence sets. Column UCYN shows number of sequences annotated as UCYN-A and column Tricho has number of sequences annotated as *Trichodesmium* in each set by our CART models.

Set	Reference Environment	Seq Unique	N. of OTUs					Main Clusters					
			98	95	90	85	80	I	II	III	IV	UCYN	Tricho
A1	Turk et al., 2011 ocean	603 281	66	41	26	12	5	564	10	29	0	44	75
A2	Langlois et al., 2005 ocean	175 128	40	20	10	5	3	172	1	2	0	34	106
Arab	Jayakumar et al., 2012 sea	132 37	16	11	7	7	5	125	4	3	0	0	5
arctic	Deslippe and Egger, 2006 terrestrial	42 30	20	15	9	5	5	33	0	0	9	0	0
Baltic	Farnelid et al., 2009 sea	433 215	58	32	20	13	8	341	11	81	0	0	0
bay	Burns et al., 2002 ESM	17 17	17	16	11	5	3	6	0	11	0	0	0
Deep	Mehta et al., 2003 ocean	120 85	39	21	17	13	9	6	68	32	14	0	0
geoth	Hamilton et al., 2011 terrestrial	66 57	27	20	11	7	3	60	1	5	0	0	0
glacier	Duc et al., 2009 terrestrial	318 139	56	39	21	14	7	254	4	60	0	0	0
grass	Bagwell et al., 2002 ESM	67 63	41	27	12	4	3	32	0	32	3	0	0
leaf	Reed et al., 2010 terrestrial	296 127	25	6	5	4	4	188	76	32	0	0	0
M1	Man-Aharonovich et al., 2007 sea	191 103	38	31	15	9	4	122	33	35	1	11	0
M2	Yogev et al., 2011 sea	65 34	11	8	5	5	3	63	2	0	0	6	5
P1	Zehr et al., 2007 ocean	86 60	23	10	7	5	4	84	2	0	0	9	20
P2	Halm et al., 2012 ocean	106 100	79	53	21	8	6	90	2	13	1	8	0
P3	Fernandez et al., 2011 ocean	693 408	53	17	10	6	5	674	5	11	3	0	0
phylo	Furnkranz et al., 2008 terrestrial	137 103	28	11	5	4	4	108	25	2	2	0	0
rhiz	Lovell et al., 2008 ESM	455 266	172	112	53	20	11	164	0	278	12	0	0
S1	Bombar et al., 2011 sea	57 40	24	14	7	4	3	49	0	8	0	0	12
S2	Moisander et al., 2008 sea	382 203	37	16	10	7	3	375	0	7	0	1	150
S3	Kong et al., 2011 sea	287 167	69	33	17	4	2	253	3	29	2	11	25
soilA	Hsu and Buckley, 2009 terrestrial	415 162	55	30	8	6	3	414	0	1	0	0	0
soilD	Pereira e Silva et al., 2011 terrestrial	646 290	136	69	29	14	7	620	17	8	0	0	0
Sponge	Mohamed et al., 2008 sea	347 123	26	12	7	7	2	243	13	91	0	0	16
termite	Du et al., 2012 symbiont	34 33	27	24	18	12	9	0	7	16	11	0	0
Total		6170						5039	284	786	58	123	414

Table 3.

Largest OTUs resulting from joint binning of twenty-five environmental sequence sets at 98% amino acid sequence similarity. Cluster labels were predicted by CART and OTUs were calculated in each subcluster separately. OTUs are identified by their cluster label, size, and origin of their sequences. An OTU is labeled "unique" if all its sequences belong to a single data set, and "mixed" if sequences originate from terrestrial and marine habitats. Closest relatives of representative sequences were identified by protein BLAST.

OTU	Size	Origin	Type	Representative	Ident
1K(1)	480	A1, A2, Baltic, geoth, glacier, grass, M1, M2, P1, P3, S2, soilA, soilD, Sponge	mixed	*Burkholderia xenovorans* *Sphaerotilus natans* *Methyloversatilis discipulorum*	100%
1B(1)	342	A1, A2, Arab, M2, P1, S1, S2, S3, Sponge	marine	*Trichodesmium erythraeum*	100%
1G(1)	250	A1, A2, M1, P1, S1, S2, S3	marine	*Marinobacterium* spp. *Azotobacter* spp.; *Vibrio* spp. *Pseudomonas* spp.	94%
1K(2)	244	A1, Arab, arctic, Baltic, P1, P3, soilA, soilD	mixed	*Bradyrhizobium* spp.	100%
1K(3)	233	Arab, P3, S2, S3, soilA	mixed	*Novosphingobium malaysiense*	100%
1J(1)	182	Arab, P3, S2	marine	*Rhodovulum* spp. *Sinorhizobium meliloti* *Confluentimicrobium* spp.	97%
1G(2)	122	A1, P2, P3, S2, S3	marine	*Teredinibacter* spp. *Marinobacterium* spp. *Pseudomonas* spp.; *Vibrio* spp.	96%
1G(3)	110	A1, Arab	marine	*Marinobacterium* spp. *Azotobacter* spp. *Gynuella* spp.	97%
1B(2)	104	A1, A2, M1, M2, P1, P2, S2, S3	marine	UCYN-A	100%
1G(4)	80	A1, Arab	marine	*Marinobacterium* spp. *Pseudomonas* spp.	97%
1J(2)	80	geoth, glacier, soilD	terrestrial	*Acidithiobacillus ferrivorans*	99%
II(1)	77	leaf, phylo	terrestrial	*Dickeya paradisiacal*	99%
III(1)	75	Sponge	unique	*Desulfobulbus mediterraneus* *Desulfovibrio oxyclinae*	94%
1B(3)	73	Sponge	unique	Endosymb. of *Epithemia turgida*	94%
1J(3)	69	leaf	unique	*Gluconacetobacter diazotrophicus* *Rubrivivax gelatinosus*	97%
1A(1)	63	A1, M1, S1, S2, S3	marine	*Desulfuromonas acetoxidans*	96%
1G(5)	57	A1, A2, P1, P2, S3	marine	*Marinobacterium* spp. *Sedimenticola* spp. *Pseudomonas* spp.	95%
1K(4)	53	A1, P3, soilD	mixed	*Xanthobacter* spp. *Bradyrhizobium* spp. *Hyphomicrobium* spp.	99%
III(2)	51	Baltic	unique	*Verrucomicrobiae bacterium*	96%
1B(4)	51	geoth, glacier, Sponge	mixed	*Leptolynobya* spp. *Oscillatoriophycideae* spp. *Nodosilinea* spp.	98%
1A(2)	46	rhiz	unique	*Pelobacter carbinolicus*	97%
1B(5)	45	A1, A2, M2, Sponge	marine	*Trichodesmium erythraeum*	97%
1J(4)	45	soilD	unique	*Rhizobium acidisoli* *Rhizobium etli*	100%
1P(1)	45	A1, A2, M1, P2, S3, soilD	mixed	*Methylomonas koyamae*	97%
1K(5)	44	A1, Baltic, M2, P1, P3, soilD, Sponge	mixed	*Derxia gummosa* *Aquabacterium* spp. *Azohydromonas australica*	98%

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology,* 215(3), pp. 403-410.

Bagwell, C. E., La Rocque, J. R., Smith, G. W., Polson, S. W., Friez, M. J., Longshore, J. W. and Lovell, C. R. (2002) 'Molecular diversity of diazotrophs in oligotrophic tropical seagrass bed communities', *Fems Microbiology Ecology,* 39(2), pp. 113-119.

Bazinet, A. L. and Cummings, M. P. (2012) 'A comparative evaluation of sequence classification programs', *Bmc Bioinformatics,* 13.

Bentzon-Tilia, M., Farnelid, H., Jurgens, K. and Riemann, L. (2014) ' Cultivation and isolation of N2-fixing bacteria from suboxic waters in the Baltic Sea', *Fems Microbiology Ecology,* 88(2), pp. 358-371.

Betancourt, D. A., Loveless, T. M., Brown, J. W. and Bishop, P. E. (2008) 'Characterization of diazotrophs containing Mo-independent nitrogenases, isolated from diverse natural environments', *Applied and Environmental Microbiology,* 74(11), pp. 3471-3480.

Bombar, D., Moisander, P. H., Dippner, J. W., Foster, R. A., Voss, M., Karfeld, B. and Zehr, J. P. (2011) 'Distribution of diazotrophic microorganisms and nifH gene expression in the Mekong River plume during intermonsoon', *Marine Ecology Progress Series,* 424, pp. 39-U55.

Bonnet, S., Dekaezemacker, J., Turk-Kubo, K. A., Moutin, T., Hamersley, R. M., Grosso, O., Zehr, J. P. and Capone, D. G. (2013) 'Aphotic N-2 Fixation in the Eastern Tropical South Pacific Ocean', *Plos One,* 8(12).

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1983) *Classification and Regression Trees.* Wadsworth.

Burns, J. A., Zehr, J. P. and Capone, D. G. (2002) 'Nitrogen-fixing phylotypes of Chesapeake Bay and Neuse River estuary sediments', *Microbial Ecology,* 44(4), pp. 336-343.

Chien, Y. T., Auerbuch, V., Brabban, A. D. and Zinder, S. H. (2000) 'Analysis of genes encoding an alternative nitrogenase in the archaeon Methanosarcina barkeri 227', *Journal of Bacteriology,* 182(11), pp. 3247-3253.

Chien, Y. T. and Zinder, S. H. (1994) 'Cloning, DNA sequencing, and characterization of *nifD*-homologous gene from the archaeon *Methanosarcina barkeri* 227 which resembles *nifD*1 from the eubacterium *Clostridium pasteurianum*', *Journal of Bacteriology,* 176(21), pp. 6590-6598.

Choo, Quok-Cheong, Mohd-Razip Samian, and Nazalan Najimudin. "Phylogeny and characterization of three nifH-homologous genes from Paenibacillus azotofixans." *Applied and environmental microbiology* 69.6 (2003): 3658-3662.

Clarke, K. R., Somerfield, P. J. and Gorley, R. N. (2008) 'Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage', *Journal of Experimental Marine Biology and Ecology,* 366(1-2), pp. 56-69.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M. and Tiedje, J. M. (2009) 'The Ribosomal Database Project: improved alignments and new tools for rRNA analysis', *Nucleic Acids Research,* 37, pp. D141-D145.

Collavino, M., Tripp, J., Frank, I., Vidoz, M., Calderoli, P., Donato, M., Zehr, J. and Aguilar, M. (2014) 'nifH pyrosequencing reveals the potential for location-specific soil chemistry to influence N2-fixing community dynamics. Environmental Microbiology', *Environmental Microbiology*, 16(10), pp 3211-3223.

Crooks, G. E., Hon, G., Chandonia, J. M. and Brenner, S. E. (2004) 'WebLogo: A sequence logo generator', *Genome Research,* 14(6), pp. 1188-1190.

De'ath, G. and Fabricius, K. E. (2000) 'Classification and regression trees: A powerful yet simple technique for ecological data analysis', *Ecology,* 81(11), pp. 3178-3192.

DeLong, E. F. (2009) 'The microbial ocean from genomes to biomes', *Nature,* 459(7244), pp. 200-206.

Desai, M. S., Assig, K. and Dattagupta, S. (2013) 'Nitrogen fixation in distinct microbial niches within a chemoautotrophy-driven cave ecosystem', *Isme Journal,* 7(12), pp. 2411-2423.

Desai, M. S. and Brune, A. (2012) 'Bacteroidales ectosymbionts of gut flagellates shape the nitrogen-fixing community in dry-wood termites', *Isme Journal,* 6(7), pp. 1302-1313.

Deslippe, J. R. and Egger, K. N. (2006) 'Molecular diversity of nifH genes from bacteria associated with high arctic dwarf shrubs', *Microbial Ecology,* 51(4), pp. 516-525.

Du, X., Li, X. J., Wang, Y., Peng, J. X., Hong, H. Z. and Yang, H. (2012) 'Phylogenetic Diversity of Nitrogen Fixation Genes in the Intestinal Tract of Reticulitermes chinensis Snyder', *Current Microbiology,* 65(5), pp. 547-551.

Duc, L., Noll, M., Meier, B. E., Burgmann, H. and Zeyer, J. (2009) 'High Diversity of Diazotrophs in the Forefield of a Receding Alpine Glacier', *Microbial Ecology,* 57(1), pp. 179-190.

Dumont, M. G., Luke, C., Deng, Y. C. and Frenzel, P. (2014) 'Classification of pmoA amplicon pyrosequences using BLAST and the lowest common ancestor method in MEGAN', *Frontiers in Microbiology,* 5.

Edgar, R. C. (2010) 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics,* 26(19), pp. 2460-2461.

Falkowski, P. G. (1997) 'Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean', *Nature,* 387(6630), pp. 272-275.

Farnelid, H., Andersson, A. F., Bertilsson, S., Abu Al-Soud, W., Hansen, L. H., Sorensen, S., Steward, G. F., Hagstrom, A. and Riemann, L. (2011) 'Nitrogenase Gene Amplicons from Global Marine Surface Waters Are Dominated by Genes of Non-Cyanobacteria', *Plos One,* 6(4).

Farnelid, H., Oberg, T. and Riemann, L. (2009) 'Identity and dynamics of putative N-2-fixing picoplankton in the Baltic Sea proper suggest complex patterns of regulation', *Environmental Microbiology Reports,* 1(2), pp. 145-154.

Fernandez, C., Farias, L. and Ulloa, O. (2011) 'Nitrogen Fixation in Denitrified Marine Waters', *Plos One,* 6(6).

Fong, A. A., Karl, D. M., Lukas, R., Letelier, R. M., Zehr, J. P. and Church, M. J. (2008) 'Nitrogen fixation in an anticyclonic eddy in the oligotrophic North Pacific Ocean', *Isme Journal,* 2(6), pp. 663-676.

Furnkranz, M., Wanek, W., Richter, A., Abell, G., Rasche, F. and Sessitsch, A. (2008) 'Nitrogen fixation by phyllosphere bacteria associated with higher plants and their colonizing epiphytes of a tropical lowland rainforest of Costa Rica', *Isme Journal,* 2(5), pp. 561-570.

Gaby, J. C. and Buckley, D. H. (2011) 'A global census of nitrogenase diversity', *Environmental Microbiology,* 13(7), pp. 1790-1799.

Gaby, J. C. and Buckley, D. H. (2012) 'A Comprehensive Evaluation of PCR Primers to Amplify the nifH Gene of Nitrogenase', *Plos One,* 7(7).

Gaby, J. C. and Buckley, D. H. (2014) 'A comprehensive aligned nifH gene database: a multipurpose tool for studies of nitrogen- fixing bacteria', *Database-the Journal of Biological Databases and Curation*.

Halm, H., Lam, P., Ferdelman, T. G., Lavik, G., Dittmar, T., LaRoche, J., D'Hondt, S. and Kuypers, M. M. M. (2012) 'Heterotrophic organisms dominate nitrogen fixation in the South Pacific Gyre', *Isme Journal,* 6(6), pp. 1238-1249.

Hamersley, M. R., Turk, K. A., Leinweber, A., Gruber, N., Zehr, J. P., Gunderson, T. and Capone, D. G. (2011) 'Nitrogen fixation within the water column associated with two hypoxic basins in the Southern California Bight', *Aquatic Microbial Ecology,* 63(2), pp. 193-+.

Hamilton, T. L., Boyd, E. S. and Peters, J. W. (2011) 'Environmental Constraints Underpin the Distribution and Phylogenetic Diversity of nifH in the Yellowstone Geothermal Complex', *Microbial Ecology,* 61(4), pp. 860-870.

Heller, P., Tripp, H. J., Turk-Kubo, K. and Zehr, J. P. (2014) 'ARBitrator: a software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank', *Bioinformatics,* 30(20), pp. 2883-2890.

Hsu, S. F. and Buckley, D. H. (2009) 'Evidence for the functional significance of diazotroph community structure in soil', *Isme Journal,* 3(1), pp. 124-136.

Jayakumar, A., Al-Rshaidat, M. M. D., Ward, B. B. and Mulholland, M. R. (2012) 'Diversity, distribution, and expression of diazotroph nifH genes in oxygen-deficient waters of the Arabian Sea', *Fems Microbiology Ecology,* 82(3), pp. 597-606.

Kong, L. L., Jing, H. M., Kataoka, T., Sun, J. and Liu, H. B. (2011) 'Phylogenetic diversity and spatio-temporal distribution of nitrogenase genes (nifH) in the northern South China Sea', *Aquatic Microbial Ecology,* 65(1), pp. 15-27.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L. and Pace, N. R. (1985) 'Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses', *Proceedings of the National Academy of Sciences of the United States of America,* 82(20), pp. 6955-6959.

Langlois, R. J., LaRoche, J. and Raab, P. A. (2005) 'Diazotrophic diversity and distribution in the tropical and subtropical Atlantic ocean', *Applied and Environmental Microbiology,* 71(12), pp. 7910-7919.

Lema, K. A., Willis, B. L. and Bourne, D. G. (2012) 'Corals Form Characteristic Associations with Symbiotic Nitrogen-Fixing Bacteria', *Applied and Environmental Microbiology,* 78(9), pp. 3136-3144.

Li, W. Z. and Godzik, A. (2006) 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics,* 22(13), pp. 1658-1659.

Lovell, C. R., Decker, P. V., Bagwell, C. E., Thompson, S. and Matsui, G. Y. (2008) 'Analysis of a diverse assemblage of diazotrophic bacteria from Spartina alterniflora using DGGE and clone library screening', *Journal of Microbiological Methods,* 73(2), pp. 160-171.

Man-Aharonovich, D., Kress, N., Bar Zeev, E., Berman-Frank, I. and Beja, O. (2007) 'Molecular ecology of nifH genes and transcripts in the eastern Mediterranean Sea', *Environmental Microbiology,* 9(9), pp. 2354-2363.

Matsen, F. A., Kodner, R. B. and Armbrust, E. V. (2010) 'pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree', *BMC Bioinformatics*, 11, pp 538-554.

Mehta, M. P., Butterfield, D. A. and Baross, J. A. (2003) 'Phylogenetic diversity of nitrogenase (nifH) genes in deep-sea and hydrothermal vent environments of the Juan de Fuca ridge', *Applied and Environmental Microbiology,* 69(2), pp. 960-970.

Mohamed, N. M., Colman, A. S., Tal, Y. and Hill, R. T. (2008) 'Diversity and expression of nitrogen fixation genes in bacterial symbionts of marine sponges', *Environmental Microbiology,* 10(11), pp. 2910-2921.

Moisander, P. H., Beinart, R. A., Voss, M. and Zehr, J. P. (2008) 'Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon', *Isme Journal,* 2(9), pp. 954-967.

Moisander, P. H., Morrison, A. E., Ward, B. B., Jenkins, B. D. and Zehr, J. P. (2007) 'Spatial-temporal variability in diazotroph assemblages in Chesapeake Bay using an oligonucleotide nifH microarray', *Environmental Microbiology,* 9(7), pp. 1823-1835.

Pereira e Silva, M. C., Semenov, A. V., van Elsas, J. D. and Salles, J. F. (2011) 'Seasonal variations in the diversity and abundance of diazotrophic communities across soils', *Fems Microbiology Ecology,* 77(1), pp. 57-68.

Pesch, R., Schmidt, G., Schroeder, W. and Weustermann, I. (2011) 'Application of CART in ecological landscape mapping: Two case studies', *Ecological Indicators,* 11(1), pp. 115-122.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2010) 'FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments', *Plos One,* 5(3).

Price, M.N., Dehal, P.S. and Arkin, A.P. (2009) 'FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix', *Molecular Biology and Evolution*, 26, pp.1641-1650.

R Development Core Team (2013) *R: A language and environment for statistical computing*. Available at: http://www.R-project.org/.

Rahav, E., Bar-Zeev, E., Ohayon, S., Elifantz, H., Belkin, N., Herut, B., Mulholland, M. R. and Berman-Frank, I. (2013) 'Dinitrogen fixation in aphotic oxygenated marine environments', *Frontiers in Microbiology,* 4.

Rappe, M. S. and Giovannoni, S. J. (2003) 'The uncultured microbial majority', *Annual Review of Microbiology,* 57, pp. 369-394.

Raymond, J., Siefert, J. L., Staples, C. R. and Blankenship, R. E. (2004) 'The natural history of nitrogen fixation', *Molecular Biology and Evolution,* 21(3), pp. 541-554.

Reed, S. C., Townsend, A. R., Cleveland, C. C. and Nemergut, D. R. (2010) 'Microbial community shifts influence patterns in tropical forest nitrogen fixation', *Oecologia,* 164(2), pp. 521-531.

Schlessman, J. L., Woo, D., Joshua-Tor, L., Howard, J. B. and Rees, D. C. (1998) 'Conformational variability in structures of the nitrogenase iron proteins from Azotobacter vinelandii and Clostridium pasteurianum', *Journal of Molecular Biology,* 280(4), pp. 669-685.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. and Weber, C. F. (2009) 'Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities', *Applied and Environmental Microbiology,* 75(23), pp. 7537-7541.

Steward, G. F., Zehr, J. P., Jellison, R., Montoya, J. P. and Hollibaugh, J. T. (2004) 'Vertical distribution of nitrogen-fixing phylotypes in a meromictic, hypersaline lake', *Microbial Ecology,* 47(1), pp. 30-40.

Turk, K. A., Rees, A. P., Zehr, J. P., Pereira, N., Swift, P., Shelley, R., Lohan, M., Woodward, E. M. S. and Gilbert, J. (2011) 'Nitrogen fixation and nitrogenase (nifH) expression in tropical waters of the eastern North Atlantic', *Isme Journal,* 5(7), pp. 1201-1212.

Ueda, T., Suga, Y., Yahiro, N. and Matsuguchi, T. (1995) 'Remarkable N_2 fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of *nifH* gene sequences', *Journal of Bacteriology,* 177(5), pp. 1414-1417.

Usio, N., Nakajima, H., Kamiyama, R., Wakana, I., Hiruta, S. and Takamura, N. (2006) 'Predicting the distribution of invasive crayfish (Pacifastacus leniusculus) in a Kusiro Moor marsh (Japan) using classification and regression trees', *Ecological Research,* 21(2), pp. 271-277.

Vitousek, P. M. and Howarth, R. W. (1991) 'Nitrogen limitation on land and in the sea: how can it occur?', *Biogeochemistry,* 13(2), pp. 87-115.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (2007) 'Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy', *Applied and Environmental Microbiology,* 73(16), pp. 5261-5267.

Yamada, A., Inoue, T., Noda, S., Hongoh, Y. and Ohkuma, M. (2007) 'Evolutionary trend of phylogenetic diversity of nitrogen fixation genes in the gut community of wood-feeding termites', *Molecular Ecology,* 16(18), pp. 3768-3777.

Yogev, T., Rahav, E., Bar-Zeev, E., Man-Aharonovich, D., Stambler, N., Kress, N., Beja, O., Mulholland, M. R., Herut, B. and Berman-Frank, I. (2011) 'Is dinitrogen fixation significant in the Levantine Basin, East Mediterranean Sea?', *Environmental Microbiology,* 13(4), pp. 854-871.

Zehr, J. P. (2011) 'Nitrogen fixation by marine cyanobacteria', *Trends in Microbiology,* 19(4), pp. 162-173.

Zehr, J. P., Hewson, I. and Moisander, P. (2009) 'Molecular biology techniques and applications for ocean sensing', *Ocean Science,* 5(2), pp. 101-113.

Zehr, J. P., Jenkins, B. D., Short, S. M. and Steward, G. F. (2003) 'Nitrogenase gene diversity and microbial community structure: a cross-system comparison', *Environmental Microbiology,* 5(7), pp. 539-554.

Zehr, J. P. and McReynolds, L. A. (1989) 'Use of Degenerate Oligonucleotides for Amplification of the *nifH* Gene from the Marine Cyanobacterium *Trichodesmium thiebautii*', *Applied and Environmental Microbiology,* 55(10), pp. 2522-2526.

Zehr, J. P., Mellon, M., Braun, S., Litaker, W., Steppe, T. and Paerl, H. W. (1995) 'Diversity of Heterotrophic Nitrogen-fixation Genes in a Marine Cyanobacterial Mat', *Applied and Environmental Microbiology,* 61(7), pp. 2527-2532.

Zehr, J. P., Montoya, J. P., Jenkins, B. D., Hewson, I., Mondragon, E., Short, C. M., Church, M. J., Hansen, A. and Karl, D. M. (2007) 'Experiments linking nitrogenase gene expression to nitrogen fixation in the North Pacific subtropical gyre', *Limnology and Oceanography,* 52(1), pp. 169-183.

Rapid annotation of *nifH* gene sequences using Classification and Regression Trees (CART) facilitates environmental functional gene analysis

Supporting Information

Ildiko E. Frank, Kendra A. Turk-Kubo, Jonathan P. Zehr

Figure S1.

Graphical representation of the CART classification model that assigns sequences into subclusters within cluster I. Terminal nodes correspond to the twelve subclusters defined in the database.

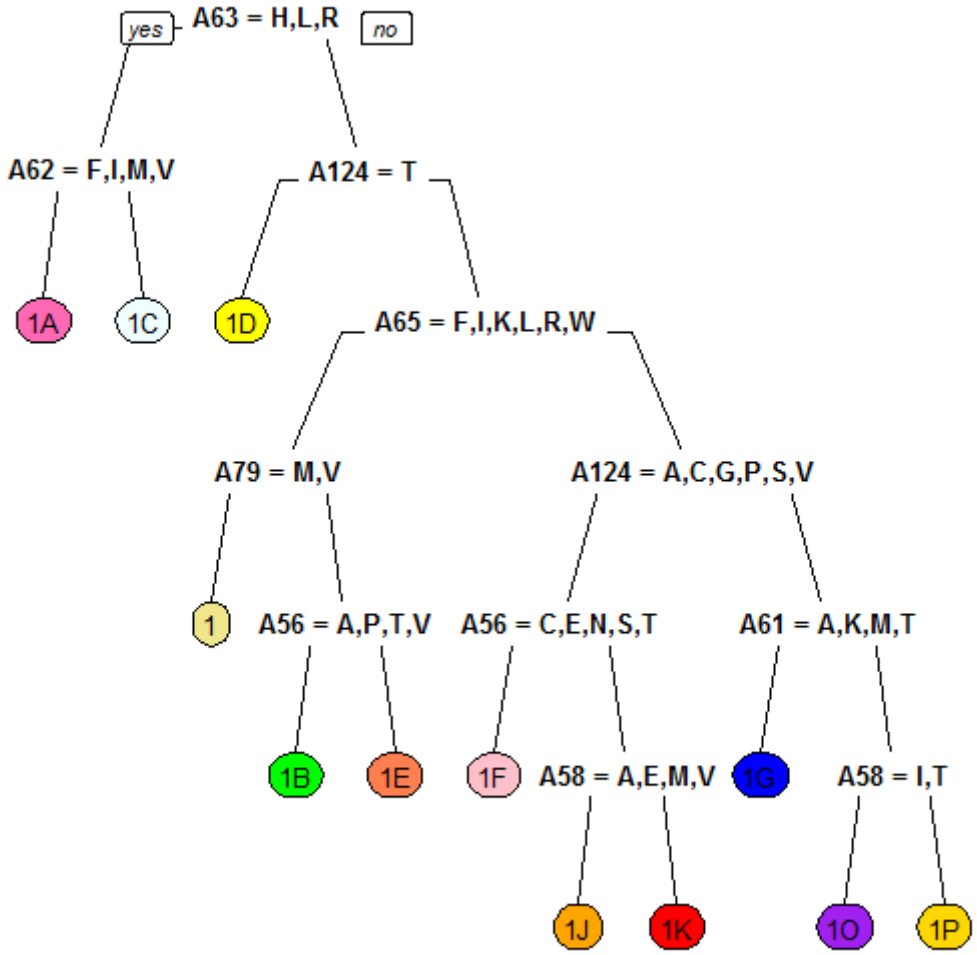
yes
A63 = H,L,R
no
A62 = F,I,M,V
A124 = T
1A
1C
1D
A65 = F,I,K,L,R,W
A79 = M,V
A124 = A,C,G,P,S,V
1
A56 = A,P,T,V
A56 = C,E,N,S,T
A61 = A,K,M,T
1B
1E
1F
A58 = A,E,M,V
1G
A58 = I,T
1J
1K
1O
1P

Figure S2.

Graphical representation of the CART classification model that assigns sequences into subclusters within cluster II. Terminal nodes correspond to the five subclusters defined in the database.

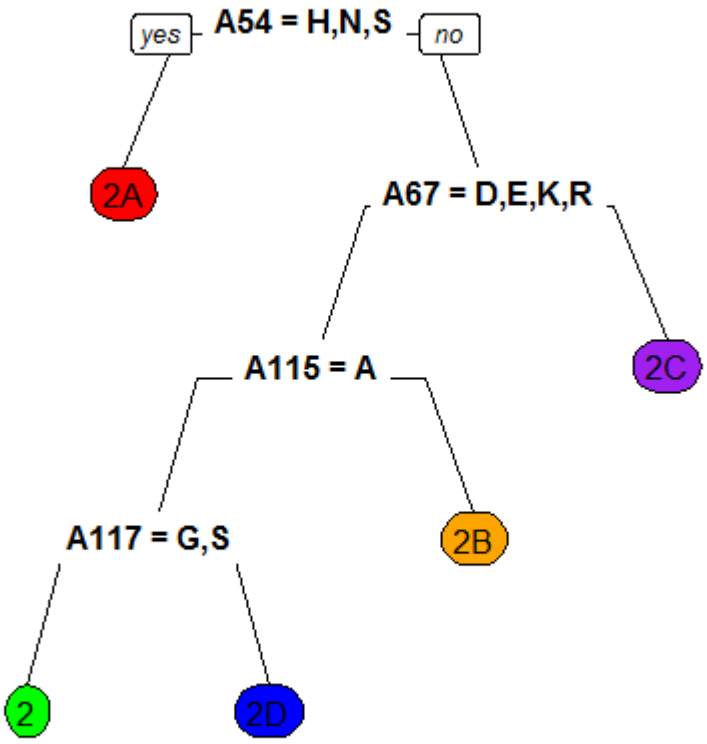

Figure S3.

Graphical representation of the CART classification model that assigns sequences into subclusters within cluster III. One of the eighteen subclusters defined in the database, 3Q, is represented by two terminal nodes.

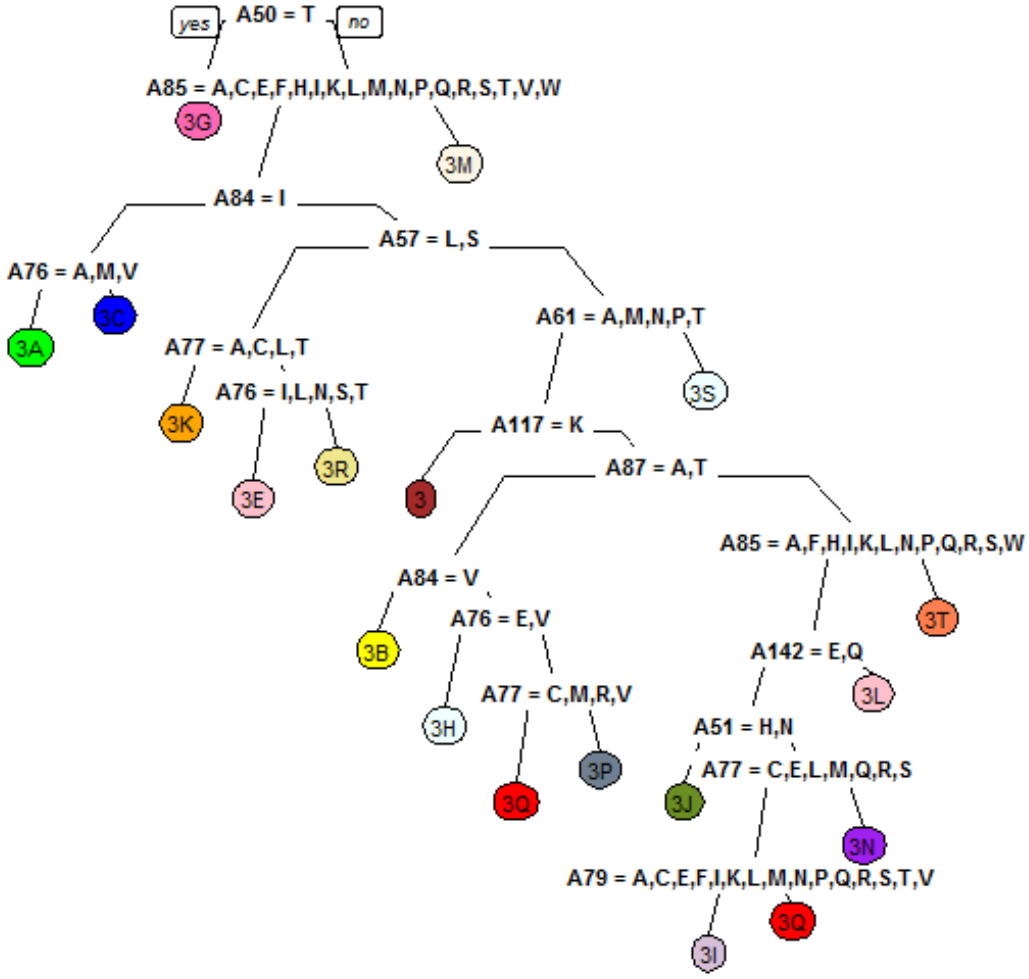

Figure S4.

Graphical representation of the CART classification model that assigns sequences into subclusters within cluster IV. Terminal nodes correspond to the eight subclusters defined in the database.

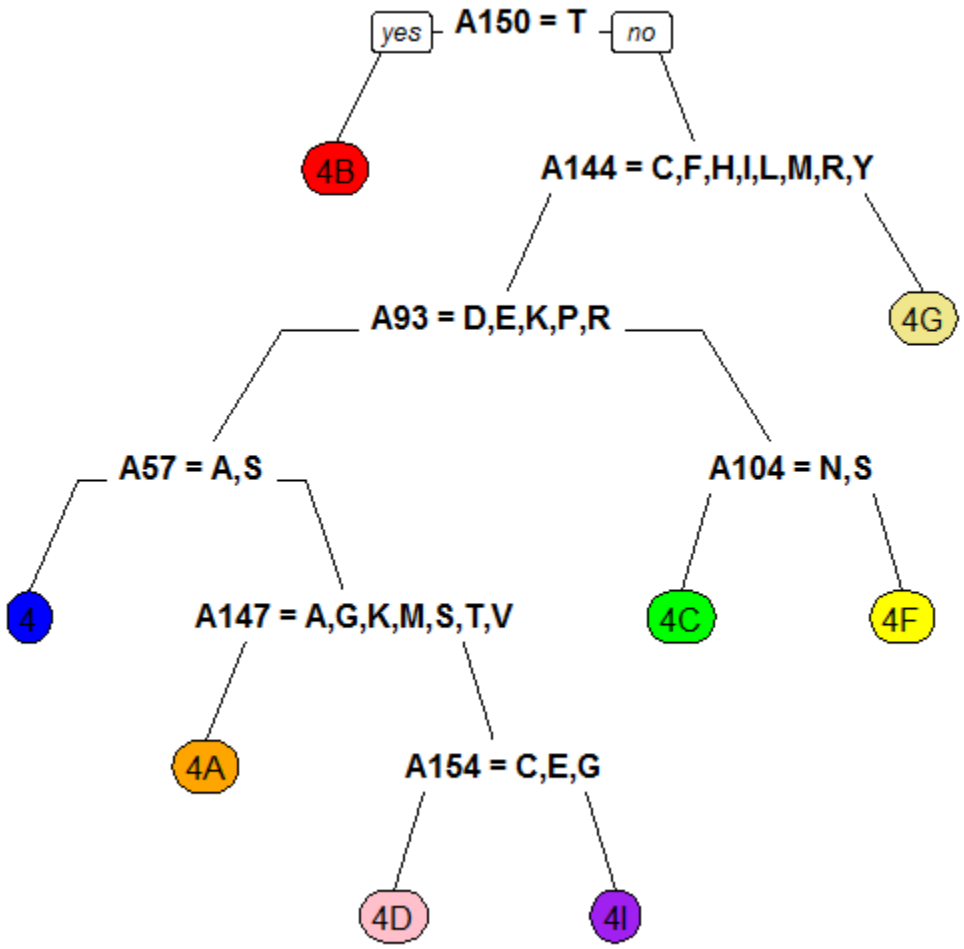

Table S1.

CART classification accuracy calculated on the training set (Fit%) and estimated by ten-fold cross-validation (Pred%).

A: Accuracy of subcluster annotation in cluster II.

Cluster	2	2A	2B	2C	2D	TOTAL
Size	6	80	78	171	11	346
Fit%	83	99	97	99	100	99
Pred%	0	99	95	99	36	95

B: Accuracy of subcluster annotation in cluster III.

Cluster	3	3A	3B	3C	3E	3G	3H	3I	3J	3K
Size	10	167	5	228	771	255	192	341	196	50
Fit%	80	92	100	77	84	90	93	54	74	90
Pred%	70	91	80	77	82	91	92	66	76	86

Cluster	3L	3M	3N	3P	3Q	3R	3S	3T	TOTAL
Size	263	28	114	331	50	34	14	72	3,121
Fit%	73	100	74	44	90	97	86	90	76
Pred%	75	98	75	35	66	97	79	81	76

C: Accuracy of subcluster annotation in cluster IV.

Cluster	4	4A	4B	4C	4D	4F	4G	4I	TOTAL
Size	9	62	88	24	94	150	12	24	463
Fit%	100	87	100	100	95	100	100	96	97
Pred%	89	77	100	100	95	97	100	83	94

Table S2.

Computation resource usage analysis for three methods used to annotate amino acid sequences with *nifH* cluster labels. Details are in Appendix S3.

	CART				BLAST				TREE			
	1000 sequences	**10,000 sequences**	**100,000 sequences**	**1,000,000 sequences**	**1000 sequences**	**10,000 sequences**	**100,000 sequences**	**1,000,000 sequences**	**1000 sequences**	**10,000 sequences**	**100,000 sequences**	**1,000,000 sequences**
User time (seconds)	0.38±0.01	2.18±0.00	20.03±0.09	199.95±1.12	69.35±1.01	699.60±10.13	6937.90±107.9	68733.00±65.04	6.47±0.04	8.86±0.01	9.85±0.02	233.101±0.06
System time (seconds)	0.01±0.01	0.08±0.01	0.19±0.02	1.19±0.08	0.02±0.01	0.13±0.02	1.46±0.09	16.55±0.08	0.0±0.0	0.0±0.0	0.02±0.01	0.21±0.02
Elapsed (wall clock) time (h:mm:ss)	00:00.6±0:00.0	00:02.4±0:00.0	00:20.5±0:00.0	03:21.4±0:01.1	01:09.4±0:01.1	11:39.5±00:10.1	1:55:38±0:01.51	19:05:17±0:01.14	00:06.5±0:00.0	00:08.9±0:00.0	00:09.9±0:00.0	03:53.5±0:00.0
Maximum resident set size (Kbytes)	13457±2.3	13457±2.3	13456±4.0	13456±0.0	28882±712.9	31539±336.8	34238±92.3	43536±0	4609±2.3	7964±2.3	44860±4	442129±2.3

Appendix S1: Data Sets and Bioinformatics

Training Set for CART Modeling

A publicly available and manually curated *nifH* sequence database (Heller et al., 2014) was utilized to train the CART models. In January 2013, it contained 22,497 sequences annotated by main clusters I (17,321), II (542), III (3,876), and IV (758). Representative sequences (16,567), identified by grouping at 98% amino acid identity using the CD-HIT suite (Li & Godzik, 2006), were manually assigned to 43 subclusters of uneven sizes based on where they were found after creating a neighbor joining tree that also contained genome sequences with cluster designations. The largest groups were 1B (3,304) and 1K (3,239), whereas the smallest subclusters 1, 2, 2D, 3, 3B, 3S, 4, and 4G contained less than 15 sequences.

CART models were trained with amino acid sequences, where positions were labeled according to the *Azotobacter vinelandii* residues from A1 to A290. Sequence coverage in the training set that varies among residues may affect which positions are included in a model. Dominance of environmental sequence fragments – only 663 sequences were obtained from fully sequenced genomes – explains the observed high coverage between positions A45 and A153, which corresponds to the targeted region of the most commonly used PCR primer sets (Gaby & Buckley, 2012). The number of sequences including positions before A39 (start of the nifH3 primer) and after position A153 (end of the PolR primer) is extremely low. There are notable dips in the number of sequences at positions A67 and A68, and especially at position A119. The first two anomalies are primarily due to deletions in cluster III sequences, whereas the gap at position A119 occurs predominantly in cluster I.

Test Set for CART Evaluation

A set of 1,558 unique *nifH* sequences derived from soil samples were imported into the above discussed *nifH* database and manually assigned to four main and seventeen subclusters (Collavino et al., 2014). As in the training set, most sequences (90%) belong to cluster I. This training set-independent data, composed of sequence fragments covering positions between A45 and A153, were used to evaluate the CART models' cluster prediction accuracy for main clusters and cluster I subclusters.

Bioinformatics

Aligned and cluster-annotated sequences were exported in fasta format from a *nifH* gene sequence database (Heller et al., 2014) stored in ARB (Ludwig et al., 2004). Statistical analysis was performed in R, an open source data analysis environment (R Development Core Team, 2013). Sequences were imported into R using package "seqinr" (Charif et al., 2012). Correspondence analysis to visualize ecosystem similarity was performed with R package "ca" (Nenadic & Greenacre, 2007). ANOSIM test was calculated using the "vegan" package (Oksanen et al. 2015) and differential gene analysis was performed with the "DESeq2" package (Anders & Huber 2010).

Appendix S2: CART Modeling

Classification And Regression Trees (CART) models (Breiman, Friedman, Olshen, & Stone, 1983) were used to predict cluster assignments of *nifH* amino acid sequences. One model assigns sequences into main clusters, and four separate models further annotate sequences by subclusters. Positions in the amino acid sequence, which may have twenty different amino acids as levels, were used as categorical predictor variables. Cluster assignment defined in the database was the categorical response to be predicted. Each decision node of the tree was defined in terms of a primary sequence position and a list of amino acids that determined how sequences traversed down the tree all the way to the terminal nodes corresponding to *nifH* clusters. When a primary position was missing from an amino acid sequence, cluster prediction was based on the corresponding surrogate position. Such "backup" positions, identified for each decision node in the model, are highly correlated with the primary positions and their use does not diminish the classification performance. Due to the uneven sizes of the categories (main clusters or subclusters) categories were weighted in inverse proportion to their size. Ten-fold cross-validation was applied to quantify the predictive power of each model. R packages "rpart" (Therneau, Atkinson, & Ripley, 2014) and "rpart.plot" (Milborrow, 2014) were used to train, evaluate, and display the CART models.

Appendix S3: OTU Calculation

Similarity between sequence pairs was quantified as normalized Hamming distance (number of sequence positions with different amino acids divided by the sequence length) on the A45 - A153 position range. This measure that ranges from 0 to 1, where 0 indicates identical sequences, was calculated by the command "daisy" in R package "cluster" (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2014). OTUs were defined by the following algorithm:

1. Calculate distances between all sequence pairs: D_{ij} for i and j = 1, nseq.
2. Define sequence connectivity d_{ij} at specified similarity level set by dmax
   (e.g. 98% similarity corresponds to dmax = 0.02): if $D_{ij} < dmax$ then $d_{ij} = 1$ (sequence pair connected), else $d_{ij} = 0$.
3. Loop through the following steps until all sequences are assigned to an OTU:
   - count number of connections for each sequence: $\Sigma_j d_{ij}$ for i = 1, nseq;
   - select the sequence with the largest number of connections as representative of the next OTU;
   - representative and all its connections form the next OTU; exclude them from further grouping.
4. Update sequence connectivity by removing inter-OTU connections.

This algorithm assures that within an OTU, similarity between a member and a representative sequence is equal to or higher than the specified level. Furthermore, each sequence is connected to one and only one OTU representative, and similarity between representatives from different OTUs is always less than the specified level.

The resulting sequence connectivity matrices were transformed into networks where vertices represent sequences and edges indicate intra-OTU sequence connections. Networks were calculated and plotted with package "network" (Butts, Handcock, & Hunter, 2014).

Appendix S4: Evaluating CART-derived annotations against Blastp and tree placement approaches

We selected the same twenty-five studies used in our cross-ecosystem analysis to compare CART-derived annotations, to those derived using two other common approaches – Blastp (BLAST) and tree placement (TREE). For Blastp analyses, we created a custom blastable database of 600 genome-derived *nifH* sequences with trusted *nifH* cluster annotation (available in the curated *nifH* database described in Heller et al., 2014). Sequences were blasted against the custom database using command line BLAST+ (Camacho et al., 2008) using arguments to recover the top Blastp hit for each query (-max_target_seqs 1) and an output format that allows easy manipulation (-outfmt 10). Due to the small size of the datasets, the cluster annotation could be manually assigned in excel using the cluster annotation of the top *nifH* genome hit for each sequence. For the tree placement approach (TREE), neighbor-joining amino acid trees that contained both the 600 genome-derived *nifH* sequences and the environmental sequences were constructed in ARB (Ludwig et al., 2004) using custom masks to select for the *nifH* amplicon region generated in each study. No bootstrapping was used. The time consuming portion of this analysis is manually determining which *nifH* cluster an unknown sequence falls into, based on placement on the resulting tree, and this is also a technique vulnerable to human error. This has been a common approach for many studies, but is only tractable when you have relatively few sequences (e.g. clone-library based studies).

In order to test computation resource usage needed to perform these three different analyses, we generated sequence files containing 1000, 10,000, 100,000 and 1,000,000 random aligned *nifH* fragments, and analyzed each sequence file in triplicate using each of the three methods (CART, BLAST, and TREE) on a computer with a Supermicro X9DR3-F motherboard,

2 Intel Xeon E5-2609 @ 2.40GHz processors with 4 cores each, and 32 GB of RAM. To generate neighbor-joining phylogenetic trees using the most rapid approach available, we selected FastTree (Price et. al., 2009). The results are presented in Table S2.

Computational resources for both CART and TREE approaches are comparable, while BLAST is much more resource intensive. However, it is important to note that this test only measures the resources needed to run the core analysis, and in the case of the TREE and BLAST approaches, does not include the time needed downstream to manually assign *nifH* cluster annotations based on the output of each analysis. As the number of sequences in a dataset grows, the downstream analyses become more time intensive for BLAST or TREE approaches, and in the case of the TREE approach, no pipelines currently exist to streamline this process for functional gene trees. The CART model requires no additional analysis, as the output includes *nifH* cluster annotation for each sequence.

References

Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* 11(10), R106.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1983). *Classification and Regression Trees*: Wadsworth.

Butts, C., Handcock, M., & Hunter, D. (2014). network: Classes for Relational Data (Version R package version 1.9.0.).

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T.L. (2008). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Charif, D., Lobry, J., Necsulea, A., Palmeira, L., Penel, S., & Perriere, G. (2012). seqinr: Biological Sequences Retrieval and Analysis (Version R package version 3.0-6).

Collavino, M., Tripp, J., Frank, I., Vidoz, M., Calderoli, P., Donato, M., .Aguilar, M. (2014). nifH pyrosequencing reveals the potential for location-specific soil chemistry to influence N2-fixing community dynamics. *Environmental Microbiology* 16(10), pp 3211-3223.

Gaby, J. C., & Buckley, D. H. (2012). A Comprehensive Evaluation of PCR Primers to Amplify the nifH Gene of Nitrogenase. *Plos One, 7*(7).

Heller, P., Tripp, H. J., Turk-Kubo, K., & Zehr, J. P. (2014). ARBitrator: a software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank. *Bioinformatics, 30*(20), 2883-2890.

Li, W. Z., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics, 22*(13), 1658-1659.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, . . . Schleifer, K. H. (2004). ARB: a software environment for sequence data. *Nucleic Acids Research, 32*(4), 1363-1371.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2014). cluster: Cluster Analysis Basics and Extensions (Version R package version 1.15.2.).

Milborrow, S. (2014). Plot rpart models. An enhanced version of plot.rpart (Version R package version 1.4-4.).

Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software, 20*(3).

Oksanen, J, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Helene Wagner (2015). vegan: Community Ecology Package. R package version 2.2-1.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2009) FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26, 1641-1650.

R Development Core Team. (2013). R: A language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Therneau, T., Atkinson, B., & Ripley, B. (2014). rpart: Recursive Partitioning and Regression Trees (Version R package version 4.1-8.).