environmental microbiology

Environmental Microbiology (2011) **13**(7), 1790–1799 doi:10.1111/j.1462-2920.2011.02488.x

A global census of nitrogenase diversity

John Christian Gaby and Daniel H. Buckley*
Department of Crop and Soil Sciences, Cornell University, Ithaca, NY 14853, USA.

Summary

The global diversity of nitrogen-fixing microorganisms was assessed through construction and analysis of an aligned database of 16 989 *nifH* sequences. We conclude that the diversity of diazotrophs is still poorly described and that many organisms remain to be discovered. Our analyses indicate that diversity is not distributed evenly across phylogenetic groups or across environments and that some of the most diverse assemblages and environments remain the most poorly characterized. The majority of OTUs were rare, falling in the long tail of the frequency distribution. The most dominant OTUs fell into either the *Cyanobacteria* or the α, β, and γ *Proteobacteria*, and five of these dominant OTUs do not have any representatives cultivated in isolation. Soils contained the greatest diversity of *nifH* sequences of all of the environments surveyed. Cluster III, which is dominated by *nifH* sequences from obligate anaerobes, was found to contain the greatest diversity of all *nifH* lineages and is also the group for which diversity is the least sampled. Our findings provide context for ongoing efforts to explore diazotroph diversity, indicating specific groups and environments that remain poorly characterized.

Introduction

Nitrogen limits primary production in both terrestrial and marine ecosystems (Vitousek and Howarth, 1991), and biological nitrogen fixation is the dominant natural process by which ecosystems obtain nitrogen. Nitrogen fixation is mediated solely by *Bacteria* and *Archaea* that possess the enzyme nitrogenase. The nitrogenase enzyme complex is encoded by the genes *nifH, nifD* and *nifK*, with *nifH* encoding the dinitrogenase reductase subunit (as reviewed in Rubio and Ludden, 2002). Zehr and McReynolds were the first to develop PCR primers for amplification of *nifH* genes, which they used to examine *Trichodesmium thiebautii* distribution and diversity in the Caribbean Sea (Zehr and McReynolds, 1989). This approach was later applied using diverse *nifH* primer sets to target a wide range of environments (as reviewed in Zehr *et al*., 2003a). The *nifH* gene has remained the signature gene for studying the diversity of nitrogen-fixing organisms. Zehr and colleagues (2003a) performed the last major evaluation of *nifH* sequences in public databases, using the approximately 1500 *nifH* sequences then available. In that effort the authors examined the phylogeny of *nifH* sequences and performed a cross-system comparison of the distribution of *nifH* lineages across different environments (Zehr *et al*., 2003a). We have used the large and ever-growing body of *nifH* sequences available in public databases to determine the degree to which current sampling efforts have captured the global diversity of nitrogen-fixing organisms, and to assess our current understanding of diazotroph diversity in different environments and in different *nifH* lineages.

Five major clusters with homology to *nifH* have been described (Young, 1992; Chien and Zinder, 1994; Zehr *et al*., 2003a; Raymond *et al*., 2004). The majority of known *nifH* sequences fall into cluster I. Cluster I is composed entirely of *nifH* genes from the conventional FeMo nitrogenase of bacteria. This cluster contains genes from most *Proteobacteria*, all *Cyanobacteria* and certain *Firmicutes* (*Paenibacillus*) and *Actinobacteria* (*Frankia*) (Zehr *et al*., 2003a). Cluster II contains a relatively small number of sequences that belong to the alternative FeV and FeFe nitrogenases as well as sequences belonging to certain methanogenic *Archaea* (Zehr *et al*., 2003a). Cluster III is dominated by *nifH* sequences from anaerobic members of the *Bacteria* and *Archaea* including: spirochetes, methanogens, acetogens, sulfate-reducing bacteria, green sulfur bacteria and clostridia (Zehr *et al*., 2003a). Clusters IV and V are composed of *nifH* paralogues that are not involved in nitrogen fixation and include genes of various functions including some involved in photopigment biosynthesis and certain electron transport reactions (Young, 2005). The phylogenies of *nifD*, *K*, *E* and *N* all generally agree with the nitrogenase clusters that have been determined using *nifH* (Young, 2005).

Any effort to make cross-system comparisons of diversity are confronted by a variety of potential biases, and it is important to consider these constraints before interpreting analyses made at this scale. First, PCR primer bias can impact the diversity and relative abundance of *nifH* genes detected. Such biases can either exclude discovery of certain lineages (Bürgmann *et al*., 2004) or can

Received 23 December, 2010; accepted 17 March, 2011. *For correspondence. E-mail dhb28@cornell.edu; Tel. (+1) 607 255 1716; Fax (+1) 607 255 8615.

alter the ratios of sequence abundance in the products relative to the templates of PCR (Suzuki and Giovannoni, 1996; Sipos *et al*., 2007). Primer sets for *nifH* are designed to be either universal (Ueda *et al*., 1995; Marusina *et al*., 2001; Poly *et al*., 2001) or group-specific (Bürgmann *et al*., 2004), but in practice each primer set exhibits a unique range of specificities. Because each primer set has its own specificity or bias, and different research groups may favour different sets of primers and different PCR reaction conditions, there exists considerable potential for variation in results between laboratories. A second concern is that sampling efforts have been uneven with many studies performed on soils and the photic zone of the ocean and fewer performed on extreme or anoxic environments (sediments, microbial mats, gut contents, etc.). We have chosen to focus our analyses on taxonomic richness and have chosen to use the Chao1 richness estimator for making diversity estimates. The value of the Chao1 estimate is calculated using the frequency of sequences that occur exactly one or two times. This estimator has the disadvantage of providing richness estimates that are strongly dependent on the number of sequences analysed, but offers the advantage of providing confidence intervals that allow robust inter-sample comparisons provided that the same number of sequences is sub-sampled from each collection of sequences (Hughes *et al*., 2002).

The existence of *nifH* paralogues that can be mis-annotated as *nifH* is also a cause for concern when conducting analyses of putative *nifH* sequences obtained from environmental sources. These paralogues can be split into two groups: those encoding subunits of alternative nitrogenases, and those not involved in nitrogen fixation. While the conventional nitrogenase contains a FeMo metal cluster, the alternative nitrogenases contain either FeV or FeFe metal clusters and the dinitrogenase reductase subunits of these enzymes are encoded, respectively, by *vnfH* and *anfH*. While *anfH* sequences fall into cluster II, *vnfH* sequences do not form a distinct phylogenetic cluster apart from *nifH* (Zehr *et al*., 2003a), and in the current study no effort was made to discriminate the *vnfH* sequences from the *nifH* sequences. Additionally, a range of *nifH* paralogues are commonly mis-annotated as *nifH* in sequence databases and assembled genomes although they have no role in nitrogen fixation (Souillard *et al*., 1988; Staples *et al*., 2007). When subject to phylogenetic analysis these paralogues fall out into cluster IV and V and can clearly be resolved from true nitrogenases (Young, 2005); these sequences have been excluded from our *nifH* diversity analyses.

In this report we analyse the current status of the *nifH* gene census through creation of a phylogenetically organized database of 16 989 *nifH* sequences. We used the database to make estimates of nitrogenase richness and estimates of coverage in different lineages and in different environments. This synthesis of sequence information provides context on our current understanding of nitrogenase diversity that can be used to direct future studies or formulate new hypotheses. Given the limitations and biases associated with examining data present in public databases it is important to recognize that this report cannot make conclusive statements about absolute differences in diversity. This report provides a glimpse at the current status of the census for nitrogen-fixing organisms and characterizes our current understanding of diazotroph diversity.

Results

Richness estimates for environmental categories

The majority of available *nifH* sequences provide only partial coverage of the gene and many of these partial sequences do not overlap completely. To address this issue we calculated diversity estimates using the range of columns in the sequence alignment that provided the greatest degree of overlap, consisting of 10 833 sequences that spanned a region of 322 nucleotide positions (as described in *Experimental procedures*). These 10 833 sequences comprised 8193 unique sequences, 3358 $OTU_{0.05}$, 2341 $OTU_{0.10}$, 809 $OTU_{0.20}$ and 23 $OTU_{0.40}$ (Fig. 1). The accumulation curves at $OTU_{0.20}$, and $OTU_{0.40}$ appeared to level off while those at other OTU cut-offs did not (Fig. 1), indicating that we still have an incomplete census of diazotroph species but that sampling of the major lineages of diazotrophs is fairly complete. We

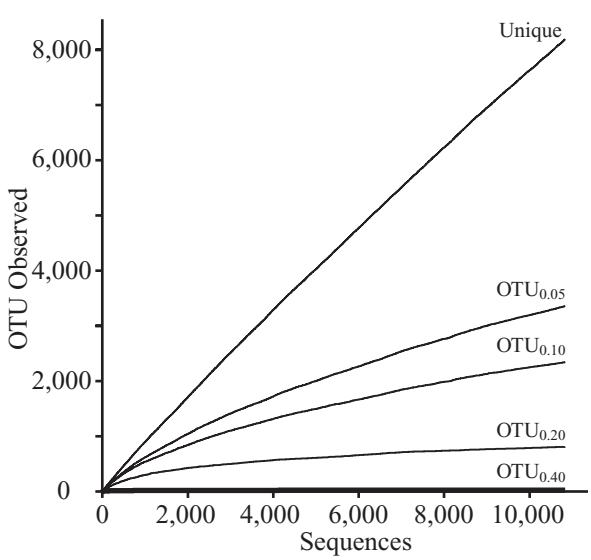

Fig. 1. Collector's curves for 10 833 *nifH* sequences that shared a common frame of 322 nucleotides (*A. vinelandii* positions 133 to 454). Lines indicate the number of OTUs detected using different similarity cut-offs as indicated in the figure.

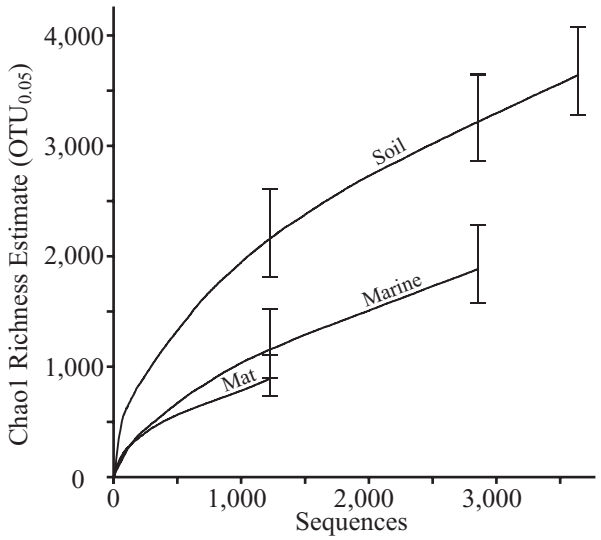

Fig. 2. Chao1 richness estimates for *nifH* sequences belonging to different environmental categories. Bars indicate 95% confidence intervals plotted at the sampling endpoints of each curve. The environmental categories are indicated by labels in the figure.

observed that an $OTU_{0.40}$ roughly delineates the major *nifH* sequence clusters. The greatest number of these 10 833 sequences came from either soil (3644) or marine sources (2855). Soil contained a greater diversity of diazotrophs than any other environment represented in the database (Fig. 2). When the analysis is limited to 2855 sequences to equalize sampling intensity for each sequence set, the Chao1 richness estimate for soil was 3216 $OTU_{0.05}$ [2864 lower 95% confidence interval (LCI), and 3646 higher 95% confidence interval (HCI)] and that for marine systems was 1884 $OTU_{0.05}$ (1581 LCI, 2286 HCI) and this difference is statistically significant. Likewise, the diversity of soil is still the highest when microbial mat communities are included in the analysis (which requires assessing richness at a sampling intensity of 1222 sequences) (Fig. 2). At this sampling intensity the richness of microbial mats (890 $OTU_{0.05}$; 737 LCI, 1110 HCI) was not significantly different from that of marine systems (1154 $OTU_{0.05}$; 903 LCI, 1522 HCI), but both of these estimates were lower than the richness estimate for soil (2156 $OTU_{0.05}$; 1809 LCI, 2612 HCI) and these differences are significant. The Chao1 richness estimator systematically underestimates richness at low levels of sampling but permits robust comparisons of relative richness between environments (Hughes *et al.*, 2002), thus Chao1 richness estimates should be treated as a lower bound.

Richness estimates for phylogenetic clusters

Chao1 richness estimates were calculated for phylogenetic groups and the greatest sequence richness was observed in cluster III and the lowest for the *Cyanobacteria* (Fig. 3). When assessed at common sampling intensity (2089 sequences), the Chao1 richness estimate for the *Cyanobacteria* (480 $OTU_{0.05}$; 410 LCI, 590 HCI) was significantly lower than that for the α, β, and γ *Proteobacteria* (1818 $OTU_{0.05}$; 1587 LCI, 2117 HCI), which in turn was significantly lower than that for cluster III (2829 $OTU_{0.05}$; 2483 LCI, 3259 HCI). Subcluster IA, which contains δ *Proteobacteria* of the *Desulfuromonadales,* is a distinct and divergent group that represents approximately 8% of the *nifH* sequences in the database (Table 1). Thus, we treated subcluster IA as a distinct group in our analysis. To make comparisons of Chao1 richness estimates that include subcluster IA requires limiting the sampling intensity to 873 sequences. At this sampling intensity the Chao1 richness estimate for subcluster IA (1382 $OTU_{0.05}$; 1081 LCI, 1819 HCI) was significantly higher than that for the *Cyanobacteria* (358 $OTU_{0.05}$; 283 LCI, 485 HCI) but was not significantly different from estimates obtained for the α, β, and γ *Proteobacteria* or cluster III.

Richness was also evaluated at a deeper level of phylogenetic resolution ($OTU_{0.20}$) (Fig. S1). This cut-off provides an objective estimate for the number of major subgroups present within each of the larger clusters and subclusters examined. When a uniform sampling intensity of 2089 sequences was applied, cluster III was estimated to contain by far the greatest level of diversity (639 $OTU_{0.20}$; 596 LCI, 702 HCI) followed by the α, β, and γ *Proteobacteria* (154 $OTU_{0.20}$; 138 LCI, 195 HCI) and *Cyanobacteria* (47 $OTU_{0.20}$; 44 LCI, 66 HCI) and these

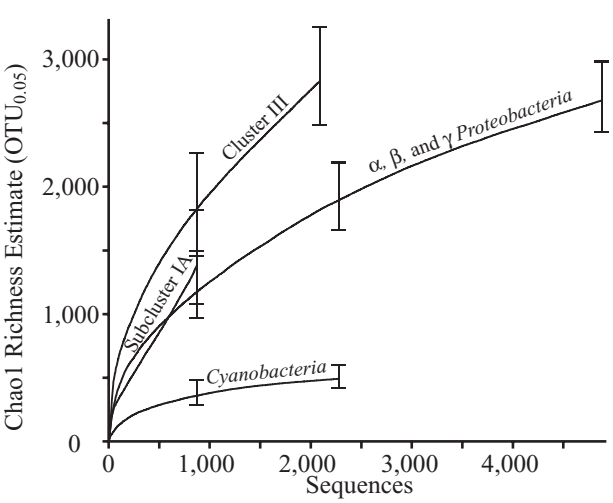

Fig. 3. Chao1 richness estimates at $OTU_{0.05}$ for *nifH* sequences belonging to different phylogenetic clusters as indicated by labels in the figure. Selection of clusters was informed by the OTU groupings established by DOTUR analysis at $OTU_{0.40}$. Bars indicate 95% confidence intervals plotted at the sampling endpoints of each curve.

Table 1. Count of *nifH* sequences as a function of phylogenetic cluster and source.

	Termite		Mat		Marine		Soil		Other[c]		Total[d]
	Seq.[a]	%[b]	Seq.[a]	%[b]	Seq.[a]	%[b]	Seq.[a]	%[b]	Seq.[a]	%[b]	Seq.[a]
Subcluster IA	0	0	17	1	143	4	822	14	364	6	1 346
Cluster III	300	57	571	39	395	13	644	11	798	13	2 708
Cyanobacteria	0	0	719	49	957	30	562	10	635	10	2 873
α, β, and γ *Proteobacteria*	6	1	148	10	1437	46	3219	56	3203	53	8 013
Other[e]	223	42	12	1	209	7	501	9	1104	18	2 049
Total[f]	529	100	1467	100	3141	100	5748	100	6104	100	16 989

a. The number of *nifH* sequences from each *nifH* cluster in the environment specified.
b. The percentage of *nifH* sequences from each *nifH* cluster in the environment specified.
c. Indicates *nifH* data from environments not listed in preceding columns as well as sequences that could not be attributed to an environmental source.
d. The total number of sequences in the database belonging to each *nifH* cluster.
e. Indicates *nifH* data from sequence clusters not represented in preceding rows.
f. Indicates the total number of sequences in the database associated with each environment.

differences are significant. The richness of subcluster IA could only be assessed at a sampling intensity of 873 sequences (56 $OTU_{0.20}$; 40 LCI, 122 HCI). At this level of sampling the richness of subcluster IA did not differ significantly from that of the *Cyanobacteria*, but it was lower than both that of cluster III (566 $OTU_{0.20}$; 492 LCI, 676 HCI) and the α, β, and γ *Proteobacteria* (129 $OTU_{0.20}$; 114 LCI, 166 HCI) and these differences were significant. In addition, collector's curves generated using the $OTU_{0.20}$ criterion demonstrate that the discovery of new groups within the α, β, and γ *Proteobacteria,* the *Cyanobacteria,* and within subcluster IA is increasingly unlikely, while diversity within cluster III remains poorly sampled (Fig. S1).

Rank-abundance distribution of nifH sequences

A rank-abundance plot made at $OTU_{0.05}$ for the 10 833 *nifH* sequences shows a long tail of rare sequences (Fig. 4). When clustered at $OTU_{0.05}$, there are 2097 sequences that were observed only once, comprising 19% of the sequences analysed (Fig. 4). The observation of a long tail of rare species was likewise observed for most clusters when frequency distributions were expressed for individual phylogenetic clusters (Fig. 5).

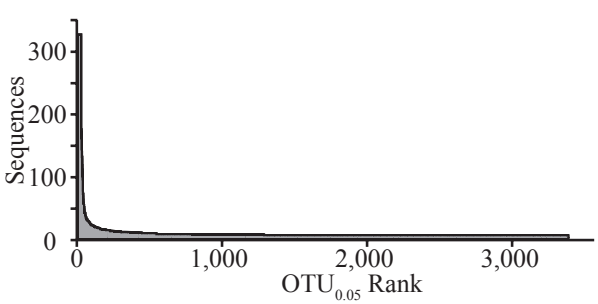

Fig. 4. Rank-abundance distribution calculated for all $OTU_{0.05}$ in the *nifH* database.

The OTU frequency distribution for the *Cyanobacteria*, however, displayed a striking pattern of dominance, with a small number of dominant taxa, which distinguished it from the other phylogenetic clusters (Fig. 5). This pattern of dominance can be expressed as a reduction in evenness and quantified using Pielou's J. As expected, the value of Pielou's J for the *Cyanobacteria* (0.727) was lower than that of any other cluster [subcluster IA (0.922), α, β, and γ *Proteobacteria* (0.857), and cluster III (0.932)] consistent with the strong pattern of dominance.

We also identified the nitrogen-fixing taxa (defined at $OTU_{0.05}$) that are most abundant in the database (Tables 2 and S1). These data are strongly influenced by site selection bias in the surveys conducted to date, and may also be influenced by PCR-related biases, and they should not

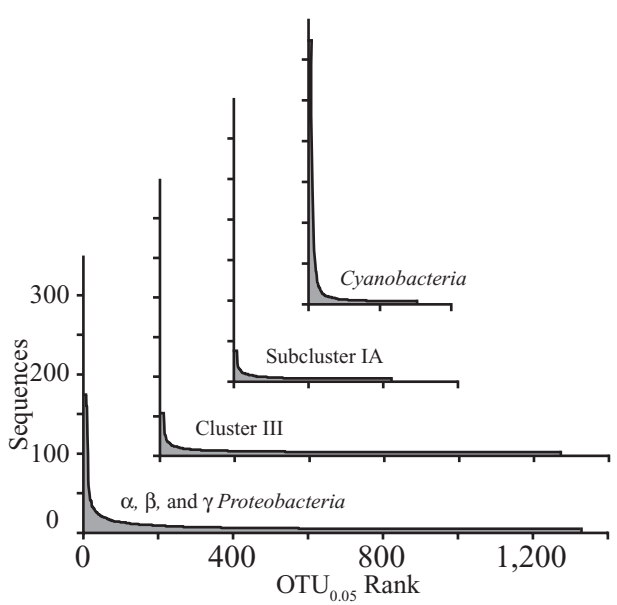

Fig. 5. Rank-abundance distribution for $OTU_{0.05}$ within each of the major *nifH* clusters and subclusters.

Table 2. The 10 most frequent $OTU_{0.05}$ observed in the *nifH* database.

Rank	Seq.[a]	Source	*nifH* cluster	Cultivated member
1	320	Marine	*Cyanobacteria*	*Katagnymene spiralis*
2	172	Marine	*Cyanobacteria*	None
3	170	Marine	α, β, and γ *Proteobacteria*	None
4	156	Soil	α, β, and γ *Proteobacteria*	None
5	156	Marine	α, β, and γ *Proteobacteria*	*Rhodobacter sphaeroides*
6	139	Freshwater	*Cyanobacteria*	*Lyngbya wollei*
7	138	Marine	α, β, and γ *Proteobacteria*	None
8	126	Hypersaline	*Cyanobacteria*	None
9	117	PCR reagent contaminant, soil, marine	α, β, and γ *Proteobacteria*	*Stenotrophomonas maltophilia*
10	107	Marine	*Cyanobacteria*	*Trichodesmium thiebautii*

a. The number of sequences in each OTU.

be taken as a measure of global abundance. Regardless, they provide information on the nitrogen-fixing organisms most commonly observed in sequence databases. Five of the 10 most observed taxa (at $OTU_{0.05}$) are from the *Cyanobacteria*, and the other five are from the α, β, and γ *Proteobacteria*. These 10 taxa represent a total of 1601 sequences, or 15% of the 10 833 sequences assessed with 864 sequences (8%) belonging to *Cyanobacteria* and 737 sequences (7%) belonging to *Proteobacteria*. Five of the 10 OTUs do not contain cultivated representatives, and six of the 10 were observed primarily in marine systems. The most frequently observed OTU (consisting of 320 sequences documented in 14 submissions, Table S1) contained the species *Katagnymene spiralis* (Lundgren *et al*., 2001), which has been recommended for incorporation into the genus *Trichodesmium* (Orcutt *et al*., 2002). Another of the most frequently observed OTUs contained the species *T. thiebautii* (107 sequences documented in 11 submissions, Table S1). The second most frequent OTU corresponds to the unicellular cyanobacterial group A (UCYN-A) (172 sequences documented in 20 studies, Table S1). Another of the most frequently observed OTUs contains the species *Stenotrophomonas maltophila* (117 sequences, documented in 20 studies, Table S1). Sequences from this OTU have been documented as contaminants in PCR reagents (Zehr *et al*., 2003b), and *S. maltophila* has been indicated as a common contaminant of ultrapure water systems (Kulakov *et al*., 2002).

The interaction of environment and phylogeny

The environmental affiliation of the sequences was examined with respect to phylogeny (Table 1 and Fig. S2). Termite guts were dominated by sequences that belonged to *nifH* cluster III (Table 1). Microbial mats contained primarily *nifH* sequences from cluster III and the *Cyanobacteria* (Table 1, Fig. S2). Marine systems were dominated by *nifH* sequences from the α, β, and γ proteobacterial cluster and the *Cyanobacteria*, while soils were dominated primarily by sequences from the α, β, and γ proteobacterial cluster (Table 1).

Cluster III *nifH* sequences have been recovered from a wide range of environments, with 15% of all cluster III sequences observed in marine systems, 24% observed in soils, and 21% observed in microbial mats and 11% in termite guts. Cyanobacterial *nifH* sequences were common in marine, soil and mat systems with 33% of all cyanobacterial sequences observed in marine systems, 20% observed in soils, 25% observed in microbial mats and none observed in termite guts. The α, β, and γ proteobacterial cluster were most common in soil and marine environments, with 18% of all sequences from the α, β, and γ *Proteobacteria* observed in marine systems, 40% observed in soils, 2% observed in microbial mats and $< 1\%$ in termite guts. The majority of *nifH* sequences belonging to subcluster IA were observed in soils, with 11% of all subcluster IA sequences observed in marine systems, 61% observed in soils, 1% observed in microbial mats and none observed in termite guts. Subcluster IA sequences also represented 14% of all *nifH* sequences observed in soils (Table 1).

Discussion

The database we assembled consists of 16 989 *nifH* sequences from numerous independent studies conducted in a range of environments and sites across the globe. Accumulation curves calculated from the database suggest that while we have a near complete census of the major lineages of diazotrophs (as defined at $OTU_{0.40}$ or $OTU_{0.20}$) the discovery of new taxa (as defined at $OTU_{0.05}$) continues at a brisk pace (Fig. 1). Nearly one out of every five taxa observed (at $OTU_{0.05}$) is represented by only a single sequence in the database generating a long tail of rare OTUs (Fig. 4). This pattern of distribution has been documented previously in microbial communities giving rise to the idea of the 'rare biosphere' (Sogin *et al*., 2006). Soils account for both the greatest number and the greatest diversity of sequences present in the database

(Fig. 2). In contrast, the marine environment is characterized by lower diversity and clear dominance by a relatively small number of nitrogen-fixing *Cyanobacteria*, most notably *Katagnymene spralis*, *T. thiebautii* and UCYN-A (Tables 2 and S1), which represent some of the most frequently observed sequence types in the database.

The major phylogenetic clusters are not distributed equally across environments (Table 1). Most striking are the results for cluster III. This cluster is composed entirely of anaerobic organisms including the genera *Treponema*, *Spirochaeta*, *Clostridium* and *Desulfovibrio*, and various lineages of methanogens as has been described previously (Zehr *et al*., 2003a). Termite gut communities were observed to contain many lineages from cluster III, likely relating to the anoxic nature of the termite gut (Brune and Friedrich, 2000). Likewise, cluster III *nifH* genes were second only to *Cyanobacteria* in microbial mat communities (Table 1). Microbial mats are frequently home to sulfate-reducing bacteria like *Desulfovibrio* as well as fermentative spirochetes. Members of cluster III could also be observed in marine sediments and in certain soils. Most remarkable, cluster III was observed to contain 639 groups at $OTU_{0.20}$ (in contrast to only 154 for the α, β, and γ proteobacterial cluster and 47 for the *Cyanobacteria*, when 2089 sequences were sampled from each cluster). Thus, it seems clear that the highest potential for discovery of novel nitrogen-fixing organisms resides among anoxic environments. Most studies performed on nitrogen-fixing organisms to date have focused either on soils or on the photic zone of the ocean, and anoxic environments have received somewhat less attention. In addition, most universal *nifH* primer sets have been designed using sequences from *Proteobacteria* and *Cyanobacteria*. Efforts to characterize the diversity of diazotrophs in anoxic systems would likely benefit from the creation of new primer sets designed to encompass the diversity of cluster III sequences.

While cluster III *nifH* sequences were found primarily in anoxic environments, the other clusters each demonstrated different patterns of environmental affinity. Most sequences from the *Cyanobacteria* were found in marine systems and microbial mats (Table 1) as would be expected (Zehr *et al*., 2003a). Cyanobacterial *nifH* sequences were also observed in association with soils, mostly from soil crusts, photosynthetic communities that live at the soil surface in certain arid ecosystems. In contrast, most sequences from the α, β, and γ *Proteobacteria* were recovered from soils, with a large number also recovered from marine systems and a minority observed in microbial mats (Table 1, Fig. S2). Sequences from subcluster IA were primarily observed in soils (Table 1). This cluster contains members of the δ *Proteobacteria* from the orders *Desulfuromonadales* and *Myxococcales*. Cultivated representatives include iron and sulfur-reducing anaerobic bacteria from the genera *Geobacter, Pelobacter, Desulfuromonas* and *Anaeromyxobacter*. While these cultivated representatives are found in select subgroups within subcluster IA, the majority of the subgroups within IA do not contain any cultivated representatives. Members of the IA group have been shown by $^{15}N_2$ stable isotope probing to be actively engaged in nitrogen fixation in soil (Buckley *et al*., 2007), and this group can account for as much as half of the *nifH* genes observed in certain soils (Hsu and Buckley, 2009).

Six of the 10 $OTU_{0.05}$ that were observed most frequently in the database have been observed primarily in marine systems, and three of these are *Cyanobacteria*. Nitrogen-fixing *Cyanobacteria* are widespread and provide an important source of N in many marine systems (as reviewed in Paerl and Zehr, 2000). The frequency with which these taxa are observed in the database may owe to their global distribution in marine environments, but their frequency of observation may also be a function of the number of studies conducted on nitrogen fixation in marine systems. The most abundant $OTU_{0.05}$ containing 320 sequences and observed in 14 studies includes the cultivated representative *K. spiralis*, which is closely related to *Trichodesmium* species (Lundgren *et al*., 2001; Orcutt *et al*., 2002). Sequences related to *T. thiebautii* are also common in the database (Tables 2 and S1). Both of these organisms are prevalent in tropical and subtropical ocean waters. In particular, *Trichodesmium* has been shown to account for a substantial portion of primary production in these oceanic zones (Karl *et al*., 2002) and has been estimated to contribute 38% of nitrogen fixation in Atlantic, Pacific and Indian Ocean waters above 25°C (Mahaffey *et al*., 2005). The second most abundant OTU (Tables 1 and S1) corresponds to the unicellular cyanobacterial group A (UCYN-A). UCYN-A was initially discovered in oligotrophic ocean waters (Zehr *et al*., 1998), and was later shown to be both abundant and to express its nitrogenase in the subtropical North Pacific Ocean (Zehr *et al*., 2001). Subsequently, it has been determined that UCYN-A exhibits a low level of sequence divergence (Tripp *et al*., 2010) and is widespread in the oceans (Langlois *et al*., 2005; 2008; Needoba *et al*., 2007; Moisander *et al*., 2010). UCYN-A also follows a diel pattern of nitrogenase expression where highest expression occurs during the day (Church *et al*., 2005; Needoba *et al*., 2007; Zehr *et al*., 2007). UCYN-A, although yet-to-be cultivated, has been shown through metagenomic approaches to lack oxygenic photosystem II (Zehr *et al*., 2008) and instead has a photofermentative metabolism (Tripp *et al*., 2010) possibly explaining the ability of the organism to fix nitrogen during the daytime.

The frequency with which sequences are observed in the database is unlikely to reflect their actual abundance in the environment, although we would expect abundant

and widespread organisms to be observed in multiple studies and for these organisms to be common in the sequence database as a result. Clearly, inter-study differences in PCR protocols and sequencing intensity provide opportunities for OTUs from certain environments to be highly overrepresented in sequence databases. Another interesting observation is that sequences affiliated with *Stenotrophomonas maltophilia* comprise one of the OTUs seen most frequently in the database [observed in 20 studies (Table S1)]. Sequences from this OTU have been identified as a common contaminant in PCR reagents (Zehr *et al*., 2003b), and *S. maltophilia* isolates have been documented as contaminants in ultrapure water systems (Kulakov *et al*., 2002). This finding suggests that the frequency with which this OTU has been observed may have more to do with frequency of finding DNA from this organism in the laboratory environment than the frequency with which this organism occurs in natural environments.

We have undertaken the first comprehensive assessment of *nifH* richness as a function of environment and phylogenetic affiliation. Our findings reveal that much diversity still awaits discovery, particularly among anaerobic nitrogen fixers and in the soil environment. The database we have created is a resource that may be used for further exploration of the sequence data as well as to develop and evaluate PCR primers to target undersampled phylogenetic groups and environments.

Experimental procedures

Constructing the nifH *database*

Construction of the *nifH* database began by downloading all sequence records containing a *nifH* related entry from the GenBank Nucleotide Database (Benson *et al*., 2008). A simple search for the term *nifH* was insufficient to recover all *nifH* sequences, and the search query also included the terms dinitrogenase reductase and nitrogenase iron protein in various arrangements. These records were manually vetted to eliminate non-*nifH* sequences caught by the search. Additionally, *nifH* sequences were extracted from genomes and multi-CDS records. The *nifH* sequences were imported into ARB (Ludwig *et al*., 2004) along with a seed alignment used to guide the sequence alignment process. The seed alignment was based on the Fer4_NifH (PF00142) protein family alignment obtained from Pfam (Finn *et al*., 2010). The Fer4_NifH seed alignment was reverse-translated into nucleotide sequences using BioEdit (Hall, 1999). The Fer4_NifH seed alignment was used to construct a PT_server in ARB, and this PT_server was used to align the *nifH* records imported from GenBank. The reverse-translated sequences of the Pfam starter alignment were then deleted from the database. The alignments were visually inspected and manually corrected. Poorly aligned or difficult-to-align sequences were detected through phylogenetic analysis in later stages of database construction and were subsequently removed. The database contains all *nifH* sequences submitted to GenBank until 2/4/2009 and is available for download in ARB database format at http://www.css.cornell.edu/faculty/buckley/nifH_database_2_4_09.arb.

Richness estimates

The diversity of *nifH* sequences in the database was evaluated using DOTUR (Schloss and Handelsman, 2005) to cluster sequences based on nucleic acid sequence similarity into OTUs, and the cut-offs used herein are represented with the convention $OTU_{0.05}$, in which the OTU cut-off criterion is provided in subscript. The $OTU_{0.05}$ nucleotide sequence cut-off for conserved protein encoding sequences is expected to correspond very roughly to the level of microbial species (Konstantinidis *et al*., 2006). The *nifH* sequence fragments in the database vary in length and position in the gene alignment complicating efforts to make a single distance matrix that includes all sequences. To solve this problem, we identified a portion of the sequence alignment that provided the greatest number of overlapping sequences of sufficient length for substantive analysis. The region identified for DOTUR analyses spanned the nucleotide positions 133 to 454 (numbering based on the *nifH* gene sequence of *Azotobacter vinelandii,* ACCN# M20568). As a result, a total of 10 833 sequences were ultimately used in DOTUR analyses (paralogous sequences that belong to clusters IV and V were excluded from DOTUR analyses and are not included in this total). The richness estimates obtained using positions 133 to 454 matches closely with estimates made with other regions of the *nifH* gene as determined by analyses of the Chao1 richness estimate for different regions of the gene sequence (Fig. S3). The distance matrix was calculated in ARB and exported for use in DOTUR analyses. DOTUR's default, furthest neighbour clustering method was used as well as randomized input order and rarefaction. Pielou's J, or evenness, was calculated from the DOTUR output by dividing the final sampling value of the Shannon diversity index for each of the phylogenetic clusters by the natural logarithm of the number of species sampled.

The association of sequences with different environmental categories was achieved by searching for environment-associated keywords in all the fields of the GenBank entries (Table 1, Fig. S2). Sequences of each type were marked and then manually verified. Sequences that could not be associated with an environment were excluded from the environment analysis. Search terms were combined in logical expressions for each query to recover relevant sequence entries while excluding confounding entries. For example, sequence entries for the marine category included sequences from coastal and open ocean water column samples but excluded marine sediments, mats, estuaries and wetlands. The soils category includes rhizosphere and bulk soil samples from terrestrial habitats but excludes soils in wetlands and estuaries. Further details about the environmental sources of sequences in these categories can be found by examining the saved configurations of these categories in the ARB *nifH* database. Whereas the different environmental and phylogenetic categories that we evaluated each contain different numbers of sequences, and the Chao1 richness estimator is influenced strongly by the number of sequences sampled, we controlled for sampling intensity in

categorical comparisons by using the lowest number of sequences common to each category when calculating Chao1 estimates. As described above, sequences used to calculate Chao1 were drawn randomly without replacement from the set of sequences in each category.

Phylogenetic analyses

Phylogenetic analyses were complicated by the presence of non-overlapping sequence fragments in the database, as described in the above section. Two approaches were used to construct phylogenetic trees and to determine the phylogenetic affiliation of individual sequences. Initially, a comprehensive guide tree was created in ARB to facilitate database management and selection of sequences. The guide tree was created by first using neighbour joining to construct a tree from a non-redundant set of 6878 sequences that had sequence information between nucleotide positions 133 to 454. Additional sequences were added to the tree in batches using the quick add by parsimony function in ARB. The final tree contained 16 989 sequences. The guide tree was used primarily for database navigational purposes.

We based selection of phylogenetic clusters upon several considerations. First, wherever possible we identified clusters in a manner consistent with affiliations identified in Zehr and colleagues (2003a). Second, in order to identify objective criteria for defining phylogenetic clusters, we examined OTUs formed at different levels of sequence similarity and found that an $OTU_{0.40}$ cut-off roughly corresponded to the major distinct monophyletic lineages we observed in phylogenetic analyses. This $OTU_{0.40}$ criterion generated subclusters within cluster I that corresponded to the α, β, and γ *Proteobacteria*, the *Cyanobacteria* and subcluster IA. Subcluster IA was first described by Zehr and colleagues (2003a) at which time it was composed of 65 sequences and contained no cultivated isolates. Our analyses show that this cluster contains δ *Proteobacteria* of the *Desulfuromonadales* and *Myxococcales*. The $OTU_{0.40}$ criterion, however, yielded a large number of subclusters when applied to cluster III, the membership of these subclusters in cluster III was too small to allow robust independent analysis. Thus, we chose to maintain cluster III as a single entity in our analyses as it has been described by previous groups (Chien and Zinder, 1994; Raymond *et al*. 2004; Young, 2005). Finally, we did not perform analyses of diversity within cluster II, or within several $OTU_{0.40}$ groups that are present at the base of the *Cyanobacteria* (containing *Frankia*, *Paenibacillus* and ε *Proteobacteria*) because these groups contained too few sequences to permit robust analysis.

Acknowledgements

This project was supported by National Research Initiative Competitive Grant no. 2005-35107-15266 from the USDA Cooperative State Research, Education, and Extension Service, and by the National Science Foundation under Grant No. MCB-0447586.

References

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D. (2008) Genbank. *Nucleic Acids Res* **36** (database issue)**:** D25–D30.

Brune, A., and Friedrich, M. (2000) Microecology of the termite gut: structure and function on a microscale. *Curr Opin Microbiol* **3:** 263–269.

Buckley, D.H., Huangyutitham, V., Hsu, S.-F., and Nelson, T.A. (2007) Stable isotope probing with $^{15}N_2$ reveals novel noncultivated diazotrophs in soil. *Appl Environ Microbiol* **73:** 3196–3204.

Bürgmann, H., Widmer, F., Von Sigler, W., and Zeyer, J. (2004) New molecular screening tools for analysis of free-living diazotrophs in soil. *Appl Environ Microbiol* **70:** 240–247.

Chien, Y.-T., and Zinder, S.H. (1994) Cloning, DNA-sequencing, and characterization of a *nifD*-homologous gene from the archaeon *Methanosarcina barkeri* 227 which resembles *nifD1* from the eubacterium *Clostridium pasteurianum*. *J Bacteriol* **176:** 6590–6598.

Church, M.J., Short, C.M., Jenkins, B.D., Karl, D.M., and Zehr, J.P. (2005) Temporal patterns of nitrogenase gene (*nifH*) expression in the oligotrophic North Pacific Ocean. *Appl Environ Microbiol* **71:** 5362–5370.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res* **38:** D211–D222.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symp Ser* **41:** 95–98.

Hsu, S.-F., and Buckley, D.H. (2009) Evidence for the functional significance of diazotroph community structure in soil. *ISME J* **3:** 124–136.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J.M. (2002) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67:** 4399–4406.

Karl, D., Michaels, A., Bergman, B., Capone, D., Carpenter, E., Letelier, R., *et al*. (2002) Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58:** 47–98.

Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* **361:** 1929–1940.

Kulakov, L.A., McAlister, M.B., Ogden, K.L., Larkin, M.L., and O'Hanlon, J.F. (2002) Analysis of bacteria contaminating ultrapure water in industrial systems. *Appl Environ Microbiol* **68:** 1548–1555.

Langlois, R.J., LaRoche, J., and Raab, P.A. (2005) Diazotrophic diversity and distribution in the Tropical and Subtropical Atlantic Ocean. *Appl Environ Microbiol* **71:** 7910–7919.

Langlois, R.J., Hümmer, D., and LaRoche, J. (2008) Abundances and distributions of the dominant *nifH* phylotypes in the Northern Atlantic Ocean. *Appl Environ Microbiol* **74:** 1922–1931.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, *et al*. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32:** 1363–1371.

Lundgren, P., Söderbäck, E., Singer, A., Carpenter, E., and Bergman, B. (2001) *Katagnymene*: characterization of a novel marine diazotroph. *J Phycol* **37:** 1052–1062.

Mahaffey, C., Michaels, A.F., and Capone, D.G. (2005) The conundrum of marine N_2 fixation. *Am J Sci* **305:** 546–595.

Marusina, A.I., Boulygina, E.S., Kuznetsov, B.B., Tourova, T.P., Kravchenko, I.K., and Gal'chenko, V.F. (2001) A system of oligonucleotide primers for the amplification of *nifH* genes of different taxonomic groups of prokaryotes. *Mikrobiologiya* **70:** 86–91.

Moisander, P.H., Beinart, R.A., Hewson, I., White, A.E., Johnson, K.S., Carlson, C.A., *et al*. (2010) Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science* **327:** 1512–1514.

Needoba, J.A., Foster, R.A., Sakamoto, C., Zehr, J.P., and Johnson, K.S. (2007) Nitrogen fixation by unicellular diazotrophic cyanobacteria in the temperate oligotrophic North Pacific Ocean. *Limnol Oceanogr* **52:** 1317–1327.

Orcutt, K.M., Rasmussen, U., Webb, E.A., Waterbury, J.B., Gundersen, K., and Bergman, B. (2002) Characterization of *Trichodesmium* spp. by genetic techniques. *Appl Environ Microbiol* **68:** 2236–2245.

Paerl, H.W., and Zehr, J.P. (2000) Marine nitrogen fixation. In *Microbial Ecology of the Oceans*. Kirchman, D.L. (ed.). New York, USA: Wiley-Liss, pp. 387–426.

Poly, F., Monrozier, L.J., and Bally, R. (2001) Improvement in the RFLP procedure for studying the diversity of *nifH* genes in communities of nitrogen fixers in soil. *Res Microbiol* **152:** 95–103.

Raymond, J., Siefert, J.L., Staples, C.R., and Blankenship, R.E. (2004) The natural history of nitrogen fixation. *Mol Biol Evol* **21:** 541–554.

Rubio, L.M., and Ludden, P.W. (2002) The gene products of the *nif* regulon. In *Nitrogen Fixation at the Millennium*. Leigh, G.J. (ed.). Amsterdam, The Netherlands: Elsevier Science B.V., pp. 101–136.

Schloss, P.D., and Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71:** 1501–1506.

Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60:** 341–350.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., *et al*. (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103:** 12115–12120.

Souillard, N., Magot, M., Possot, O., and Sibold, L. (1988) Nucleotide sequence of regions homologous to *nifH* (nitrogenase Fe protein) from the nitrogen-fixing archaebacteria *Methanococcus thermolithotrophicus* and *Methanobacterium ivanovii*: evolutionary implications. *J Mol Evol* **27:** 65–76.

Staples, C.R., Lahiri, S., Raymond, J., Von Herbulis, L., Mukhophadhyay, B., and Blankenship, R.E. (2007) Expression and association of group IV nitrogenase NifD and NifH homologs in the non-nitrogen-fixing archaeon *Methanocaldococcus jannaschii*. *J Bacteriol* **189:** 7392–7398.

Suzuki, M.T., and Giovannoni, S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62:** 625–630.

Tripp, H.J., Bench, S.R., Turk, K.A., Foster, R.A., Desany, B.A., Niazi, F., *et al*. (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464:** 90–94.

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. *J Bacteriol* **177:** 1414–1417.

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

Young, J.P.W. (1992) Phylogenetic classification of nitrogen-fixing organisms. In *Biological Nitrogen Fixation*. Stacey, G., Burris, R.H., and Evans, H.J. (eds). New York, USA: Chapman and Hall, pp. 43–86.

Young, J.P.W. (2005) The phylogeny and evolution of nitrogenases. In *Genomes and Genomics of Nitrogen-Fixing Organisms*. Palacios, R., and Newton, W.E. (eds). Dordrecht, The Netherlands: Springer, pp. 221–241.

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

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

Zehr, J.P., Waterbury, J.B., Turner, P.J., Montoya, J.P., Omoregie, E., Steward, G.F., *et al*. (2001) Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. *Nature* **412:** 635–638.

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

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

Zehr, J.P., Montoya, J.P., Jenkins, B.D., Hewson, I., Mondragon, E., Short, C.M., *et al*. (2007) Experiments linking nitrogenase gene expression to nitrogen fixation in the North Pacific subtropical gyre. *Limnol Oceanogr* **52:** 169–183.

Zehr, J.P., Bench, S.R., Carter, B.J., Hewson, I., Niazi, F., Shi, T., *et al*. (2008) Globally distributed uncultivated oceanic N_2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322:** 1110–1112.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Chao1 richness estimates at $OTU_{0.20}$ for *nifH* sequences belonging to different phylogenetic clusters as indicated by labels in the figure. The 95% confidence intervals are plotted at the sampling endpoints for each curve.

Fig. S2. Phylogenetic distribution of *nifH* sequences associated with different source environments. The figure contains three copies of the same radial tree representing all 16 989 sequences in the database (creation of the tree is described

in methods). Red shading indicates branches with sequences that match the environmental origin indicated below each tree (the numbers indicate the number of sequences labelled in each tree). For convenience, groups are labelled on the left-most copy of the tree: (A) Cluster II; (B) Cluster IV; (C) Cluster III; (D) Subcluster IA; (E) *Cyanobacteria*; (F) α, β, and γ *Proteobacteria*. The branches that fall between Subcluster IA and the *Cyanobacteria* in the figure are unlabelled and include *Frankia*, *Paenibacillus* and members of the ε *Proteobacteria*.

Fig. S3. The figure evaluates the degree to which richness estimates are sensitive to the region of the *nifH* alignment being analysed. The analysis was performed by extracting different regions of the alignment from a common set of full-length sequences ($n = 144$). Each alignment region was then used to generate Chao1 richness estimates. The numbers in the legend correspond to positions in the alignment. The alignment positions 496 to 1189 correspond to *A. vinelandii* nucleotide positions 133 to 454.

Table S1. Identifying information for the most frequent $OTU_{0.05}$ observed in the *nifH* database.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.