Ecological Patterns of *nifH* Genes in Four Terrestrial Climatic Zones Explored with Targeted Metagenomics Using FrameBot, a New Informatics Tool

Qiong Wang,[a] John F. Quensen III,[a] Jordan A. Fish,[a] Tae Kwon Lee,[b] Yanni Sun,[c] James M. Tiedje,[a] James R. Cole[a]

Center for Microbial Ecology, Michigan State University, East Lansing, Michigan, USA[a]; School of Civil and Environmental Engineering, Yonsei University, Seoul, Republic of Korea[b]; Computer Science and Engineering Department, Michigan State University, East Lansing, Michigan, USA[c]

ABSTRACT **Biological nitrogen fixation is an important component of sustainable soil fertility and a key component of the nitrogen cycle. We used targeted metagenomics to study the nitrogen fixation-capable terrestrial bacterial community by targeting the gene for nitrogenase reductase (*nifH*). We obtained 1.1 million *nifH* 454 amplicon sequences from 222 soil samples collected from 4 National Ecological Observatory Network (NEON) sites in Alaska, Hawaii, Utah, and Florida. To accurately detect and correct frameshifts caused by indel sequencing errors, we developed FrameBot, a tool for frameshift correction and nearest-neighbor classification, and compared its accuracy to that of two other rapid frameshift correction tools. We found FrameBot was, in general, more accurate as long as a reference protein sequence with 80% or greater identity to a query was available, as was the case for virtually all *nifH* reads for the 4 NEON sites. Frameshifts were present in 12.7% of the reads. Those *nifH* sequences related to the *Proteobacteria* phylum were most abundant, followed by those for *Cyanobacteria* in the Alaska and Utah sites. Predominant genera with *nifH* sequences similar to reads included *Azospirillum*, *Bradyrhizobium*, and *Rhizobium*, the latter two without obvious plant hosts at the sites. Surprisingly, 80% of the sequences had greater than 95% amino acid identity to known *nifH* gene sequences. These samples were grouped by site and correlated with soil environmental factors, especially drainage, light intensity, mean annual temperature, and mean annual precipitation. FrameBot was tested successfully on three ecofunctional genes but should be applicable to any.**

IMPORTANCE **High-throughput phylogenetic analysis of microbial communities using rRNA-targeted sequencing is now commonplace; however, such data often allow little inference with respect to either the presence or the diversity of genes involved in most important ecological processes. To study the gene pool for these processes, it is more straightforward to assess the genes directly responsible for the ecological function (ecofunctional genes). However, analyzing these genes involves technical challenges beyond those seen for rRNA. In particular, frameshift errors cause garbled downstream protein translations. Our FrameBot tool described here both corrects frameshift errors in query reads and determines their closest matching protein sequences in a set of reference sequences. We validated this new tool with sequences from defined communities and demonstrated the tool's utility on *nifH* gene fragments sequenced from soils in well-characterized and major terrestrial ecosystem types.**

Received 5 August 2013 **Accepted** 16 August 2013 **Published** 17 September 2013

Citation Wang Q, Quensen JF, III, Fish JA, Lee TK, Sun Y, Tiedje JM, Cole JR. 2013. Ecological patterns of *nifH* genes in four terrestrial climatic zones explored with targeted metagenomics using FrameBot, a new informatics tool. mBio 4(5):e00592-13. doi:10.1128/mBio.00592-13.

Editor Jizhong Zhou, University of Oklahoma

Address correspondence to James M. Tiedje, tiedjej@msu.edu.

High-throughput sequencing analysis of 16S rRNA genes is now an established method of interrogating microbial diversity in environmental samples. But the resolution of data from the rRNA gene is limited (1, 2), so finer-grained taxonomic information is lost. Also, the phylogeny derived from genes involved in many important ecological functions and 16S phylogeny often do not match due to horizontal gene transfer (see reference 3). Therefore, direct tracking of ecofunctional genes may provide better insight into important aspects of bacterial diversity and function. Analysis of ecofunctional gene amplicon data presents some challenges distinct from those posed by 16S rRNA amplicon data. Because protein-coding genes often evolve at a higher rate than rRNA, while the encoded protein sequence evolves at a lower rate, it can be advantageous to compare protein sequences. However, indels, which are common sequencing artifacts, cause frameshifts and lead to a corrupt protein translation downstream from the artifact.

Several tools are available to detect and correct frameshift sequencing errors in next-generation-sequencing (NGS) short reads (see discussion in reference 4). Two recent programs were designed to use a dynamic programming approach to detect frameshifts in high-volume NGS data. FragGeneScan was developed to find complete and partial open reading frames in short metagenomic sequences. It requires no training on a particular gene of interest (5). Instead, FragGeneScan uses a Hidden Markov Model (HMM) trained on general codon usage bias. Because the HMM models an open reading frame at the DNA level, allowing nucleotide insertions and deletions, FragGeneScan allows transitions

TABLE 1 Frameshifts detected with FrameBot for regions from three genes amplified from defined communities

Gene	Strain	No. of sequences passing FrameBot[a]	No. of sequences with frameshifts[b]	No. of frameshifts[c]
nifH[d]	*Desulfitobacterium hafniense* DCB-2	5,726	772	1,056
nifH	*Nostoc* sp. PCC	4,053	309	399
nifH	*Burkholderia xenovorans* LB400	9,140	1,947	2,224
but[e]	*Roseburia intestinalis* L1-82	407	206	325
but	*Roseburia inulinivorans* DSM 16841	291	99	154
bphA[f]	*Polaromonas naphthalenivorans* CJ2	724	146	242
bphA	*Rhodococcus jostii* RHA1	1,028	719	1,026

[a] Number of reads with a FrameBot best match of greater than 30% identity to a (known) defined community sequence.

[b] Number of reads with one or more frameshifts detected by FrameBot.

[c] Total number of frameshifts detected by FrameBot.

[d] *nifH*, nitrogenase reductase.

[e] *but*, butyryl-CoA: acetate CoA-transferase.

[f] *bphA*, biphenyl dioxygenase alpha subunit.

between reading frames. HMMFrame was developed as a protein domain classification tool for metagenomic data (4). As with HMMER (6) and other protein profile HMM annotation tools, HMMFrame uses a set of protein family models from Pfam (7) or other sources to scan metagenomic data. Unlike HMMER, HMMFrame incorporates error models for specific sequencing technologies and is able to match across frameshift errors in the input DNA fragments.

Unlike shotgun metagenomic data, a specific gene and gene region are targeted with amplicon sequencing, but the raw sequences are still subject to the same frameshift artifacts. Also, instead of classifying the frameshift-corrected reads into those coding for different protein families, for amplicons it is often useful to compare these reads to closely related well-studied reference gene sequences. For those genes with little horizontal transfer, identifying the nearest neighbors can be the first step in taxonomic assignment. Even for 16S rRNA, where many other tools are available, such a pairwise nearest-neighbor approach can be the method of choice for highly variable regions, such as the Global Alignment for Sequence Taxonomy (GAST) tools for taxonomic assignment of amplicons of the V6 hypervariable region (8). Since standard pairwise alignment of each query against many rRNA gene sequences would be prohibitively slow, GAST uses a heuristic to identify a small subset of sequences for alignment.

We combined the two functions of frameshift correction and nearest-neighbor identification into a single routine, FrameBot. This tool uses a dynamic programming algorithm, as do HMMFrame and FragGeneScan, but instead of comparing reads to a general protein model as FragGeneScan does or to a protein family profile HMM as HMMFrame does, FrameBot detects and corrects frameshifts during pairwise comparisons to reference sequences using metric indexing to minimize the number of comparisons required to find the nearest reference sequence. Minor adjustments to the amino acid substitution matrix were necessary in order to satisfy a mathematical requirement of the index structure (9).

We used this tool to assess the taxonomic groups and ecological patterns inferred from a key gene in ecosystem function, *nifH*, in four contrasting ecosystems that are part of NEON, the NSF National Ecological Observatory Network.

RESULTS

FrameBot algorithm. FrameBot is based on the algorithm of Guan and Uberbacher (10) with modifications. Briefly, FrameBot uses a dynamic programming technique similar to that used for pairwise protein alignment but uses three sets of two-dimensional matrices, one for each potential reading frame. When determining the entry in a matrix, the scoring algorithm considers diagonal, horizontal, and vertical transitions corresponding to an amino acid match/substitution, an amino acid insertion, and an amino acid deletion, respectively, as do standard pairwise dynamic programming algorithms. In addition, the FrameBot algorithm considers transitions from positions in the two other matrices, corresponding to a one- or two-nucleotide insertion or deletion (frameshift errors). For nucleotide deletions, the published algorithm of Guan and Uberbacher reuses nucleotides from the adjacent codon, resulting in a codon with a variable and potentially very poor amino acid substitution score. Instead, we implemented a consistent amino acid penalty for nucleotide deletions. Also, the original algorithm described only a Smith-Waterman (local)-style algorithm; we extended it to global and overlap alignment algorithms. The global algorithm was used in this work. To speed up FrameBot for the NEON nearest-neighbor analysis (below), we applied the Approximating and Eliminating Search Algorithm metric indexing algorithm (11) with a modified Blosum 62 metric matrix (9) in order to reduce the number of comparisons (see Methods S1 in the supplemental material).

FrameBot performance. We compared the performance of FrameBot to that of FragGeneScan (5) and HMMFrame (4). To test the 3 tools for accuracy, we used sets of 454 pyrosequencing reads for the amplified parts of 3 different genes from defined communities of organisms with known genome sequences. For the *nifH* gene, the defined community carried seven genes from the *nifH* family, but not all were amplified under the conditions used, including those not demonstrated to have nitrogenase reductase function (see Table S1 in the supplemental material). Reads from three *nifH* samples, each amplified from a single organism that made up the defined community, were filtered through the amplicon read initial processing step. Of those passing, 0.18% shared less than 30% protein identity with any sequence in the nifH_tit reference set (see Methods S1) using FrameBot and were discarded as artifacts. Overall, 16% of the reads in this set had one or more frameshift errors (Table 1). Similar initial processing was performed for defined communities that contained butyrate-producing genes important to intestinal health, butyryl-coenzyme A (CoA)–acetate CoA-transferase (*but*) and the polychlorinated biphenyl (PCB)-degrading gene biphenyl

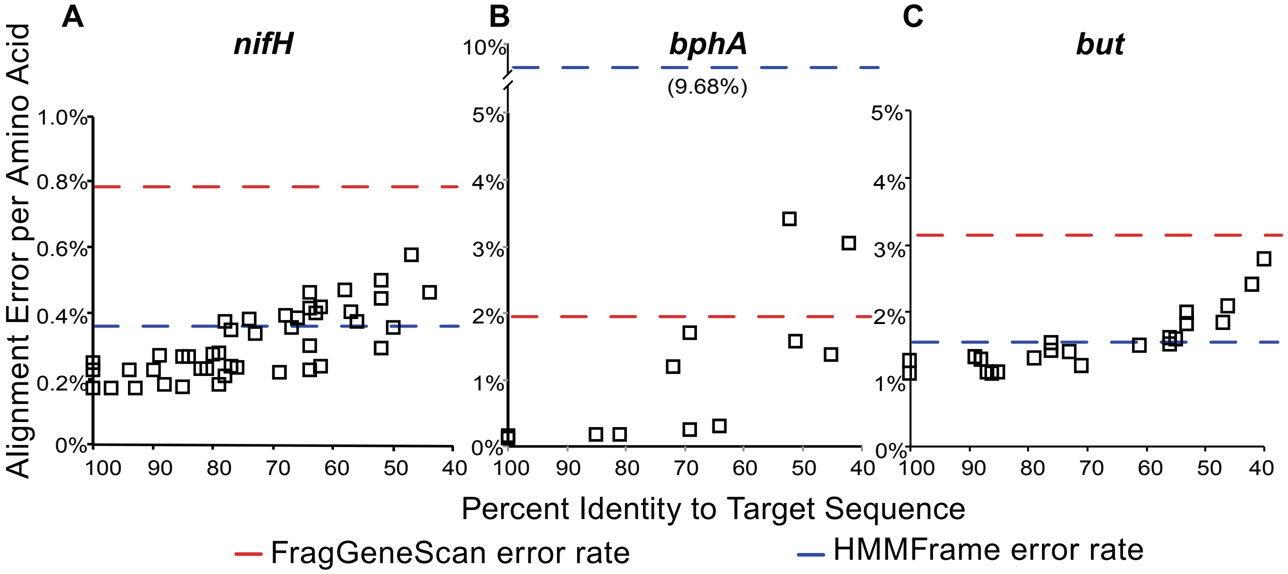

FIG 1 FrameBot performance using reference sequences at various percentages of identity to query sequences. Target protein sequences were chosen from the FunGene site (http://fungene.cme.msu.edu) at various distances from the known defined community sequences. The error rates at 100% identity represent baseline sequencing errors. The test genes are *nifH* (nitrogenase reductase) (A); *bphA* (biphenyl dioxygenase alpha subunit) (B); and *but* (butyryl-CoA: acetate CoA-transferase) (C). Dotted lines represent the overall error rates for FragGeneScan and HMMFrame on the same amplicon data. The error rate from HMMFrame for *nifH* shown here (0.36%) is calculated from an HMM trained on the group I, II, and III sequences from the augmented Zehr reference set. When trained on the entire augmented Zehr reference set, the error rate rose to 0.67%, and when trained on the group I-only sequences, the error rate was 0.34%.

dioxygenase alpha subunit (*bphA*) (Table 1). These genes were chosen to represent a range of characteristics to test analysis.

These sequences were subjected to frameshift correction using FrameBot, FragGeneScan, and HMMFrame. The resulting protein sequences from all three tools were then compared to the known parent protein sequences using a custom tool implementing the overlap Needleman-Wunsch algorithm (12). Amino acid mismatches, insertions, and deletions were counted as errors. For FrameBot, we chose reference sequences at different degrees of identity to the community members and plotted the average number of alignment errors per amino acid for each reference sequence-community member combination. With FragGeneScan and HMMFrame, no reference sequence is required (after training for HMMFrame), so we calculated a single average error rate for all reads in a defined community (see Methods S1 in the supplemental material). For the *nifH*-defined community, the error rate with FrameBot increased as the distance between the query sequence and the reference sequence increased but always remained low (Fig. 1). FrameBot outperformed FragGeneScan at all reference-read distances tested and outperformed HMMFrame at 80% or higher identity, while HMMFrame outperformed FragGeneScan. For the *but*-defined community, FrameBot and HMMFrame performed similarly and outperformed FragGeneScan with reference sequences of at least 50% identity. For the *bphA* gene, FrameBot outperformed the other 2 programs in most cases when a reference sequence with 50% or higher identity was available, while HMMFrame performed poorly.

We also compared FrameBot with AmpliconNoise (13), a general pyrosequencing error correction tool. Using the same *nifH*-defined community data, after processing the reads with AmpliconNoise, most of the frameshifts had been removed for two of the three defined community members, but for *Desulfitobacterium hafniense* the fraction of sequences with frameshifts actually increased. In addition, many reads had been truncated by AmpliconNoise and, hence, would need to be discarded in order to compare equivalent regions of the *nifH* gene (see Table S2 in the supplemental material).

NEON sample analysis. The NEON *nifH* data set contained 278,561 unique DNA reads out of a total of 1.1 million sequences obtained using the Poly primers (14) from 222 soil samples taken from 4 NEON sites (Table 2). It required 11.5 h to frameshift correct and calculate the nearest neighbors for these unique reads using a single thread on a Mac Pro with a 2.5-GHz Intel Core i5 processor. On average, each query required 39 comparisons out of the 675 reference sequences in the augmented Zehr reference set (see Materials and Methods) to find the nearest match. Slightly less than 0.03% of the reads shared less than 30% protein identity with any sequence in the reference set. These likely represented aberrant sequencing artifacts and were excluded from further analysis. For the remaining reads, 90.6% had a nearest match in the reference set with 90% identity or above, while 99.97% matched with 80% identity or above. About one-third (254) of the reference sequences were a closest match to at least one read. Overall, 12.7% of the reads contained at least one frameshift. On average, reads with frameshifts contained 1.3 frameshifts, with a maximum of 11. Before further analysis, we removed 6 samples that had fewer than 1,000 sequences after frameshift correction, 2 from Alaska (AK) and 4 from Hawaii (HI).

The well-characterized NifH proteins form three coherent clusters (groups I to III); a fourth group of *nifH*-like genes (group IV) consists of deep-branching genes with mostly unknown function, while another related cluster (group V) contains bacteriochlorophyll (*bchL*) and related genes (15). We found that, overall, 92.8%, 7.0%, 0.16%, and 0.004% of the reads belonged to *nifH* groups I, II, and III, and to groups IV and V, respectively, with some variation between sites (see Table S3 in the supplemental

TABLE 2 Characteristics of NEON sites and samples used for *nifH* amplicon analysis[a]

	Result for indicated site			
Site parameter	Alaska (AK)	Florida (FL)	Hawaii (HI)	Utah (UT)
Ecological region	Boreal forest/taiga	Subtropical/dry forest	Subtropical/lower montane wet forest	Grassland/shrubland
Soil type	Goldstream silt loam	Candler fine sand	Akaka silty clay loam	Taylorsflat loam
Soil taxonomy	Coarse-silty mixed, superactive, subgelic typic Hitoturbels	Hyperthermic, uncoated Lamellic Quartzipsamments	Hydrous, ferrihydritic, isothermic Acrudoxic Hydrudands	Fine-loamy, mixed, superactive, mesic, xeric Haplocalcids
Drainage	Very poor	Excessive	Moderately well	Well
MAP (mm)	260	750	4,000	274
MAT (°C)	−3	20	16	8.9
Latitute, longitude	65.15N 147.49W	29.69N 81.99W	19.93N 155.28W	40.17N 112.45W
Light intensity[b]	3	4	2	5
% OM	18.3 ± 11.6	1.2 ± 0.3	51.4 ± 12.3	1.5 ± 0.3
pH	4.6 ± 0.8	5.0 ± 0.6	4.9 ± 0.7	8.0 ± 0.3
Water content (g)	5.5 ± 1.6	0.25 ± 0.18	0.74 ± 0.07	0.52 ± 0.56
Ca (ppm)	724 ± 337	121 ± 60	564 ± 401	5110 ± 651
Na (ppm)	15.7 ± 2.9	10.7 ± 1.6	33.1 ± 8.5	30.1 ± 4.6
Microbial biomass (mg C/kg)	4.7 ± 3	1.0 ± 0.5	31.1 ± 18.7	8.9 ± 3.5
No. of samples	26	17	171	8
No. of sequences	125,294	79,619	896,824	19,105

[a] MAP and MAT were measured for each site; % OM was calculated as the average of the % OM of samples from the same site. The % OM, Water content, Ca, Na and microbial biomass data show mean and standard deviations.

[b] Ordered 1 to 5 by perceived relative sunlight exposure.

material; see Dataset S2 for detailed results for each sample). Reads with best matches to proteobacterial reference sequences predominated in all four sites (Fig. 2). Alphaproteobacterial matches made up, on average, from 43.3% in AK to 72.4% in Florida (FL). Considerable numbers of reads similar to *Azospirillum* reference sequences were found in all samples, along with *Bradyrhizobium* and *Rhizobium* matches. Gammaproteobacterial matches were highest in AK, while deltaproteobacterial matches were lower in FL than in AK and Utah (UT). There are 12 reference genera, all *Proteobacteria*, with greater than 5% best-matching reads for at least one site (see Fig. S1 in the supplemental material). Of these, matches to *nifH* from *Sideroxydans*, *Methylobacterium*, and *Bradyrhizobium* were more common in AK, while *Anaeromyxobacter* matches were more common in UT. *Cyanobacteria* constituted the next most frequently matched phylum in the AK and UT sites but were present in much lower proportions in FL and HI. Matches to the cyanobacterial genus *Nostoc* were significantly ($P < 0.001$) more common in UT.

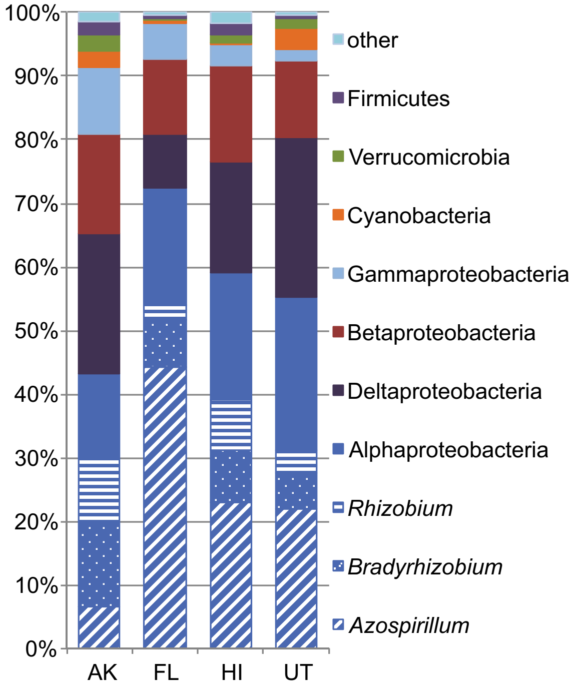

FIG 2 Relative abundances of NEON reads grouped by nearest matches at the phylum and class levels, averaged for each site (observatory) as indicated by state. The three most dominant genera in alphaproteobacteria are also shown. Other, all phyla with less than 0.5% nearest matches from any site.

Ordination. We visualized the differences in samples using principal component analysis (PCA) to view the community grouping among the *nifH* samples with the corresponding environmental factors (Fig. 3). The first principal component appears to separate the sites with respect to in-site variations, especially within the HI samples, which are present in the highest numbers. The sites were well separated by the second and third principal components, except for overlap between the FL and UT samples. The AK sites are poorly drained and have the lowest mean annual temperature (MAT), while HI has the highest mean annual temperature and the highest mean annual precipitation (MAP).

Indicator species. Using the nearest-neighbor sequence assignments as "species" assignments, we calculated the four-way Dufrene-Legendre "indicator value" (16), treating each site's samples as a class (see Dataset S3 in the supplemental material). For this test, the sample sizes were normalized by randomly selecting a subset of reads from each sample equal to the number in the least abundant sample. There were 15, 4, 7, and 6 indicator values above 0.5 for the AK, FL, HI, and UT sites, respectively. Most of these corresponded to *Proteobacteria*, but they included diverse bacte-

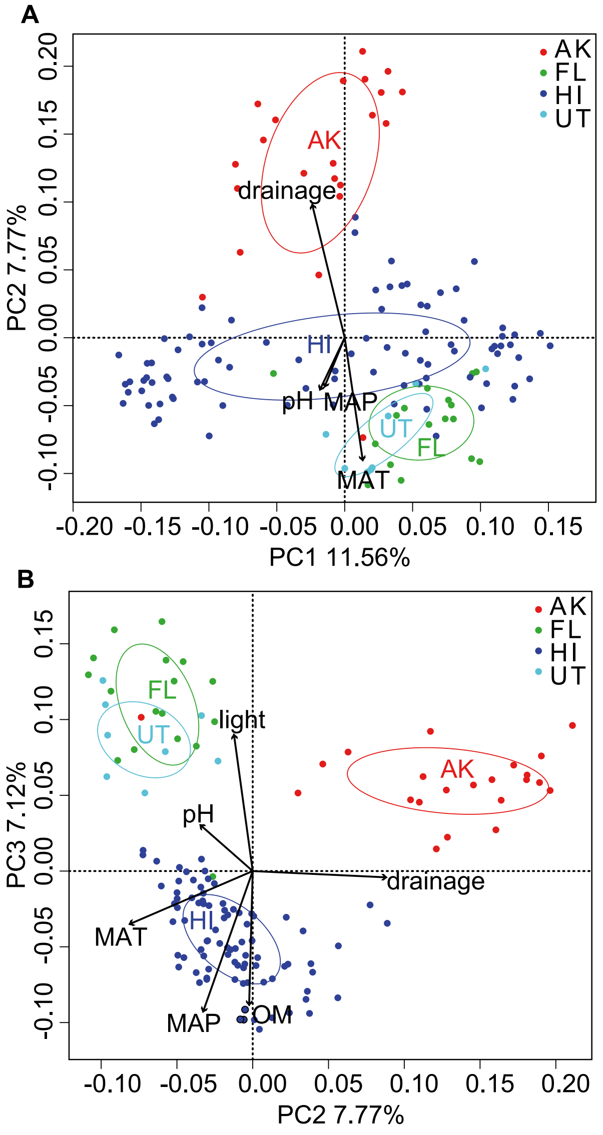

FIG 3 Principal component analysis of NEON samples. (A) PC1 and PC2. (B) PC2 and PC3. The input data were standardized using the Wisconsin square root normalization as implemented in R. Ellipses represent 1 standard deviation of the points from the centroid. The soil environmental variables were fitted to the ordination using the envfit method from the labdsv R package. Arrows were plotted for variables with significance of fit ≤ 0.01.

rial phyla and *Archaea*, most notably several *Cyanobacteria* indicators for the AK and UT sites.

DISCUSSION

Ecologically important functions, such as nitrogen fixation, are not necessarily preserved in related organisms. Directly tracking key genes in the pathway, such as the *nifH* gene, should offer a better insight into nitrogen fixation capacity in an environment than would be obtained by extrapolating from 16S rRNA data. To track such genes requires correcting coding sequence frameshifts caused by sequence artifacts. General homology search tools, such as BLASTx (17), cannot correct frameshift artifacts. The FrameBot program presented here does so and is fast enough to use on the large datasets produced by NGS technology.

FrameBot performed better than two other frameshift correction tools developed for high-throughput sequencing in tests using amplicon sequence data from defined communities for three different functional genes. Sequences of one of these genes (*bphA*) are very diverse, and a large fraction of amplicons retrieved from environmental samples share less than 40% identity with sequences of known genes (18). FragGeneScan, which requires no information about the specific gene of interest, outperformed HMMFrame, which relies on gene-specific information, and performed nearly as well as FrameBot on this gene. Our speculation is that the HMMFrame protein profile HMM does not strongly differentiate between amino acids when the sequences are too diverse. In such a case, a model based on codon usage, as incorporated in FragGeneScan, may outperform tools using gene-specific information. We also found that decreasing the HMMFrame training set diversity for *nifH* from a model that included both *nifH* and *nifH*-like genes to one that included only *nifH* genes decreased the error rate. However, further decreasing the diversity to group I genes had little additional effect on HMMFrame error rates (Fig. 1).

Biological nitrogen fixation is an important driver of primary productivity and is carried out exclusively by nitrogen-fixing *Bacteria* and *Archaea*. The *nifH* gene, coding for nitrogenase reductase, has been used as a marker for the nitrogen fixation pathway and is one of the best-studied ecofunctional genes. The pathway for nitrogen fixation has been subject to horizontal transfer, potentially in several early events, making some deep branches in gene phylogeny different from the underlying rRNA-based phylogeny (15). For such processes, it may make sense to map the genes back to the underlying taxonomy thorough nearest-neighbor comparisons, as was done here using FrameBot. FrameBot uses a reference set of sequences for the gene of interest and is well suited for genes such as *nifH*, where relatively extensive reference sets are available. For broad gene families, such as the family of *nifH* and *nifH*-like genes, the inclusion of *nifH*-like genes in the reference set simplified unambiguous identification of these non-target-gene amplicons. If the reference database is well curated, as is the case for *nifH* here, the accuracy of assignment is much better than that from a general-purpose database such as GenBank.

The four NEON sites were chosen to represent very different soil and climatic conditions: (i) the taiga of AK is very cold, wet, and rich in organic matter due to its overlying permafrost, with moderate light penetration due to a short, sparse black spruce vegetation; (ii) the rain forest of HI is also wet and rich in organic matter but at a constant warm temperature and under a dense, tall forest cover, allowing minimal light at the soil surface; (iii) the Great Basin of UT is desert-like, with shrub vegetation, low levels of organic matter, higher levels of salts, a wide annual temperature cycle, and extensive light exposure; and (iv) the subtropical forest of FL is on sandy soil low in organic matter and water-holding capacity and is warm, with an annual temperature cycle, and its xeric character results in sparse tree vegetation, allowing substantial light at the soil surface. These very different soil conditions test the *nifH* gene-targeted method and its analysis and would be expected to select different nitrogen-fixing microbial communities.

For all 4 sites, over 90% of reads were most similar to proteobacterial reference sequences. There was high variation from sample to sample in the percentages of best matches to the individual reference sequences; still, the sites were well separated by principal component analysis, except for FL and UT (Fig. 3), both well-drained xeric sites with high levels of light exposure and low levels of organic matter (Table 2).

Reads with nearest matches to strains of *Azospirillum* spp. were at high levels in all 4 sites and averaged 44%, 23%, and 22% of the reads in the FL, HI, and UT samples, respectively, and were the most common nearest matches for these 3 sites (Fig. 2). *Azospirillum* are well-known plant growth-promoting rhizobacteria and are produced commercially as biofertilizer for both plant hormonal effects and nitrogen fixation on nonnodulating crops (INTX Microbials product literature). In FL and HI samples, the majority of these reads were within 95% amino acid identity (AAI) to the *nifH* gene of *A. amazonense*, which has been shown to benefit rice growth mainly through increased nitrogen fixation (19).

Most surprising were the large number of reads in the four sites with nearest matches to strains of *Rhizobium*, ranging from 2% to 10%, and to strains of *Bradyrhizobium*. The *Bradyrhizobium* matches averaged between 6% and 13%, and most were within 5% amino acid identity to *Bradyrhizobium* sp. ORS278, a member of the non-Nod factor, photosynthetic stem-nodulating symbionts of the *Aeschynomene* (20). These *Bradyrhizobium* have a narrow host range and form a separate 16S rRNA group from *B. japonicum*, which nodulates soybean, and other root-nodulating *Bradyrhizobium* species (21). Their *Aeschynomene* hosts are normally associated with warm climates, but interestingly, the ORS278-related reads were significantly more common in the AK samples (12%). Hawaii is the only site to have a legume plant known to host rhizobia, namely, the root- and canopy-nodulating *Bradyrhizobium* of the *Acacia koa* tree, although it was a minor member of the sampled forest. Hawaii is known, however, to have a number of native nodulating legumes, and many more were introduced by Europeans, such that both *Rhizobium* and *Bradyrhizobium* would have had hosts on the island though not at the sampled site. Best matches to known symbiont *Rhizobium* spp. were also common at all four sites. For all of these *Rhizobiales* matches, we have no evidence that the reads actually come from plant symbionts. In fact, *Methylobacterium nodulans* ORS 2060, another nodulating symbiont member of the order *Rhizobiales* and the only known nitrogen-fixing *Methylobacterium*, was the closest reference for 8.6% of the Alaska reads. However, the *M. nodulans nifH* gene sequence is very similar to that of another reference, *Gluconacetobacter diazotrophicus* PA1 5, a sugarcane endophyte in the order *Rhodospirillales*, with only 1 amino acid difference between the 2 *nifH* gene regions, while the AK reads had a median of 6 differences. The prominence of *nifH* sequences very similar to those of the legume-nodulating rhizobia in the four sites without plant hosts suggests that these rhizobia have a broad-ranging, free-living lifestyle or that they are not rhizobia but that the genes are from dominant unrelated and uncultured soil bacteria whose *nifH* genes may have been horizontally transferred.

Beyond the alphaproteobacteria, matches to the deltaproteobacterial genus *Geobacter* were common, making up 14% of the total in AK, while *Anaeromyxobacter* matches made up 14% of the total in UT (see Fig. S1 in the supplemental material). It may not be surprising that an anaerobe such as *Geobacter* would be common at the AK site where the soil is classified as very poorly drained. Although the one known *Anaeromyxobacter* sp. has not been shown to form spores, in general, members of the order *Myxococcales* are known for the formation of desiccation-resistant myxospores, potentially an important survival trait in the normally arid Utah environment.

Beyond the *Proteobacteria*, the phylum *Cyanobacteria* was the second most frequently represented in AK and UT, while cyanobacterial nearest matches were much less common in HI and FL. Also, the indicator "species" with the second-highest indicator value for AK and UT sites were nearest matches to cyanobacterial strains. These are two of the three sites with light exposure. The taiga (boreal forest) site (Caribou Poker Flat Watershed) is populated by short black spruce, allowing substantial light penetration to the soil surface. It also has a thick and dense ground cover of *sphagnum* moss and reindeer lichen and is moist most of the time due to the underlying permafrost and high levels of soil organic matter, conditions expected for algal growth during the long summer days. Cyanobacteria could be both free-living and associated with lichens and moss. Moss-*Cyanobacteria* associations have been found to contribute significantly to fixed nitrogen pools in boreal forests (22). The majority of the reads showed the most similarity to *Anabaena* sequences and the corresponding organisms are likely free-living, but a small percentage (0.05% of total AK reads) were associated with *Nostoc*, a genus containing free-living as well as lichen- and moss-associated members. In the UT site, arid conditions would at first seem to preclude the presence of a high concentration of *Cyanobacteria*; however, the UT "indicator" *Nostoc punctiforme* produces a desiccation-resistant form and its filaments are known to be important in holding together desert soil crusts (23). In addition, many *Nostoc* strains are reported to be important in nitrogen-fixing desert soil crusts (24, 25). The FL site also has substantial light exposure but it is a xeric site without algal crusts and is not conducive to cyanobacterial growth. The HI site has suitable moisture, but the dense forest severely limits the amount of light reaching the soil surface.

Like all supervised methods, FrameBot is unable to correctly classify novel gene sequences, but by examining sequence similarity, FrameBot is able to flag potentially novel sequences. In addition, FrameBot avoids the sequencing-error-induced diversity explosion that plagues *de novo* clustering (26). For the NEON sites, virtually all environmental *nifH* gene sequences shared 80% or greater amino acid identity to a known reference sequence, and over 90% shared 90% or greater identity. This is striking and was not the case for other genes we have examined, as they were more diverse. It suggests that the diversity of the amino acid sequence for this region is more constrained and that most of the diversity in this gene is already known, at least for group I. This leads to the inference that we also know most of the terrestrial N_2-fixing genera or, if not, that such a result must be due to horizontal gene transfer; e.g., if the often dominant soil phylum *Acidobacteria* has N_2-fixing types, it must be due to horizontal gene transfer or the presence of an as-yet-unrecognized outlier sequence type.

Conversely, for the group I reference sequences, those most expected to be amplified by the Poly primers, more than 80% matched reads at an amino acid identity of 95% or greater (Table 3; see also Table S5 in the supplemental material), a level of genome-wide average amino acid identity expected between strains of the same species (27). Gibbons et al. recently demonstrated that most of the rRNA gene-based diversity seen in a broad collection of marine sites could be found at a single deeply sequenced site (28). They hypothesized that in marine environments, these majority rare taxa act as a "seed bank," allowing rapid shifts in populations in response to changing environmental con-

TABLE 3 Number of Zehr (augmented) reference sequences matching the Poly primers and NEON reads

Group(s)[a]	Total	No. of matching primers[b]	No. of matching reads[c]
I	183	168	145
II	88	44	21
III	48	17	8
IV and V	356	13	0

[a] Group I consists of primarily aerobes, including cyanobacteria and proteobacteria, with typical Mo–dependent nitrogenases; group II consists of anaerobes and *Archaea* and group III those with alternative metal nitrogenases (15).

[b] At most two mismatches to forward and reverse primers. A total of 118 group I reference sequences had 0 or 1 mismatch.

[c] Reference sequences with at least 95% amino acid identity to one or more reads from the NEON samples.

ditions. Such analysis tracks diversity at the core genome level (97% rRNA gene fragment identity) but can group populations with widely different gene content (29). Finding that most *nifH* genes in this gene reference collection, derived from organisms found over a period of many years and from diverse environments, have close matches at these four sites is consistent with the presence of a soil seed bank storing most of the extant *nifH* protein-level diversity, irrespective of the core genome (rRNA) background.

MATERIALS AND METHODS

FrameBot. FrameBot is coded in the Java programming language and has been tested using the Java 1.6 run time from Apple under the OSX 10.6 and 10.7 operating systems and using the Java version 1.6 OpenJDK Run time Environment (IcedTea6 1.10.6) under the CentOs release 5.8 operating system. FrameBot requires an amino acid substitution scoring matrix. Two different matrices were used in this work. The Blosum 62 matrix (30) was used except for work with the metric index, where a metric modification of the Blosum 62 matrix was used (9). The online version of FrameBot is available on RDP's FunGene site (http://fungene.cme.msu.edu). The command line FrameBot program, source code, and manual are distributed under the terms of the GNU GPLv3 and are freely available on SourceForge (https://sourceforge.net/projects/rdpframebot).

Zehr *nifH* reference set. For analysis of NEON ecological observatory samples, the tree_Genomes_AA_Dec2011_MASK_GENOME data set was downloaded from the 17 February 2012 release *nifH* database of Zehr et al. (31; http://www.es.ucsc.edu/~wwwzehr/research/database). An additional 218 *nifH* and *nifH*-like genes (2 belonging to group I, the remainder from groups IV and V) not matching Zehr reference sequences were extracted from IMG (http://img.jgi.doe.gov) based on annotation and added to the reference set to create the augmented Zehr reference set. From these, we extracted the protein region (excluding primers) corresponding to that amplified by the Poly primers, and for each protein sequence we obtained the corresponding DNA sequence region from GenBank. After removing duplicates, 675 protein sequences remained (see Table S4 in the supplemental material). The DNA sequences were compared to the forward and reverse Poly primers using the RDP Initial Process Tool (2), allowing a maximum of 2 mismatches to either primer. This curated reference set is more accurate than a set extracted from GenBank, especially for groups IV and V, members of which are often incorrectly annotated as *nifH* in GenBank.

Preparation of *nifH* amplicon libraries for pyrosequencing. DNA extracted from the soil samples served as the template in triplicate PCRs performed using a Roche High Fidelity PCR system (Roche Diagnostics GmbH, Mannheim, Germany). Each 20-μl reaction volume contained 1.8 mM $MgCl_2$, a 400 nM concentration of each primer, a 200 μM concentration of each deoxynucleoside triphosphate (dNTP) (Invitrogen, Eugene, OR), 4 μg of bovine serum albumin (BSA) (New England Biolabs, Ipswich, MA), 1 U of *Taq* polymerase, and 50 ng of template DNA. The forward primer consisted of the 25-bp 454-A adapter and a 10-bp bar code followed by the 20-bp primer PolF (14) (5′-CGT ATC GCC TCC CTC GCG CCA TCA G-bar code-TGC GAY CCS AAR GCB GAC TC-3′). The reverse primer consisted of the 25 bp 454-B adapter and the 20-bp primer PolR (5′-CTA TGC GCC TTG CCA GCC CGC TCA GAT SGC CAT CAT YTC RCC GGA-3′). These primers target an approximately 320-bp region of the *nifH* gene. Amplifications were performed on an Eppendorf Mastercycler thermocycler (Eppendorf North America, Hauppauge, NY) using the following temperature program: 3 min at 95°C and 30 cycles of 45 s at 95°C, 45 s at 62°C, and 1 min at 72°C, followed by a final extension for 7 min at 72°C.

PCR amplicon libraries were purified on a 1.2% agarose gel. Amplicons were visualized using SYBR Safe Gel stain (Invitrogen) and extracted from the gel using a QIAquick gel extraction kit (Qiagen) per the manufacturer's directions. DNA was eluted from the filter with 30 μl of the elution buffer and purified a second time using a Qiagen PCR purification kit. Each purified reaction sample was quantified using a PicoGreen double-stranded DNA (dsDNA) assay kit (Invitrogen) and a Qubit fluorometer (Invitrogen) according to the manufacturer's instructions. Equal DNA masses of each sample were combined, the total concentration was adjusted to 20 ng/μl, and the pooled sample was sent to the Research Technology Support Facility (RTSF) at Michigan State University (East Lansing) for emulsion PCR (emPCR), GS amplicon library preparation, and pyrosequencing on a 454 Life Sciences GS-FLX machine (Roche).

Defined communities. Three different ecofunctional genes were chosen to represent different types of test characteristics. For *nifH*, genomic DNAs from three different organisms (Table 1) were amplified separately with separate sample bar codes using the Poly primers as described above. These primers have broad coverage for most phyla we expect to be common in soils if up to two mismatches are allowed, although missing the nitrogen-fixing *Euryarchaeota* and many of the *Firmicutes* (see Table S5 in the supplemental material). This is consistent with previous simulations allowing no mismatches that showed twice the percentage of coverage for soil-related sequences versus for all *nifH* sequences tested (51% versus 25%; 32). For the *but*-defined community, genomic DNA was amplified as previously described using three separate primer sets (33). Genomic DNAs from the following organisms were obtained from DSMZ and mixed prior to amplification: *Roseburia intestinalis* L1-82, *R. inulinivorans* DSM 16841, *Faecalibacterium prausnitzii* A2-165, *Eubacterium hallii* DSM 3353, and *E. rectale* DSM 17629. For the *bphA*-defined community, genomic DNAs from *Polaromonas naphthalenivorans* CJ2 and *Rhodococcus jostii* RHA1 were amplified separately using different bar codes and the BPHD-f3 and BPHD-r1 primers (18).

NEON soil samples. The National Ecological Observatory Network (NEON) consists of 20 ecosystem types (termed Domains) from the tundra to the tropics and provides a continent-scale system for observing biological change (http://www.neoninc.org). We obtained 222 soil samples collected from 4 of the NEON observatories representing very different soil and climatic conditions (Table 2; see also Text S1 in the supplemental material) along with their soil metadata. Soil DNA was extracted and *nifH* gene fragments were amplified as described above. An array of soil chemical and physical attributes were measured for these samples and were provided by Jacob Parnell at NEON Inc., Boulder, CO (see Dataset S1 in the supplemental material).

Amplicon read initial processing. Amplicon sequences from *nifH* samples were filtered through initial bar code matching and quality control steps using the RDP Pyro Initial Process tool (2) with the following parameters: forward primer maximum mismatches two, reverse primer maximum mismatches zero, minimum length 300, number of N's zero, minimum average quality score 20. Those sequences passing these initial quality steps with 30% amino acid identity or greater to the closest reference sequence using FrameBot and with a minimum length of 100 amino acids were accepted as valid *nifH* family sequences. Reads that shared less than 30% identity with any sequence in the reference set likely represented aberrant sequencing artifacts and were excluded from further analysis. Amplicon reads from *but* and *bphA* samples were processed in a manner similar to that described in Text S1 in the supplemental material.

Defined community error analysis at the DNA level. Reads from defined communities that passed the initial quality and FrameBot filters were compared to the expected nucleotide sequences using the RDP Defined Community Analysis Tool (http://fungene.cme.msu.edu/) implementing the Needleman-Wunsch algorithm (12). This algorithm performs a global alignment between the query sequence and the reference sequence. It requires that the sequences cover the same region of the gene. Before error measurement, the *bphA* amplicon reads that passed initial processing were truncated to correspond to nucleotide positions 163475 to 163885 of *Rhodococcus jostii* RHA1 NC_008269.1 (corresponding to amino acid positions 218 to 354 of *Rhodococcus jostii* RHA1 BphA protein sequence YP_707265.1). The overall error rates (nucleotide insertions, deletions, and mismatches) per base were 0.24%, 0.75%, and 0.22% for the *nifH*-, *but*-, and *bphA*-defined communities, respectively.

Ecological analysis. We used the FrameBot nearest-neighbor assignments as "species" and samples as "sites" for PCA. Most HI soil samples were split after collection, with one set of replicates used to test soil storage conditions. These replicates were not included in the ordination analysis. We included 6 soil environmental factors (drainage, light intensity, MAP, MAT, percent organic matter [% OM], and pH) that were measured for the majority of samples and which we felt were likely important for driving bacterial community composition and were not correlated to other typically measured environmental attributes (see Dataset S1 in the supplemental material). Vegan package version 2.0-3 in R 2.15 was used to perform PCA. Indicator species were identified using labdsv package version 1.5-0. Samples from the same site were treated as a "group" for indicator species analysis. The one-way analysis of variance (ANOVA) test, the Tukey's honestly significant difference (HSD) test, and box and whisker plots were produced in R. We first combined replicates from the same sampling location (latitude and longitude) by taking the average FrameBot nearest-neighbor match counts. The reference sequences and their nearest-match relative abundances were used as variables for ANOVA. When genera were compared, the relative abundances of matches to reference sequences within the same genus were summed before ANOVA testing.

Availability of supporting data. The sequence data from this study have been submitted to the ENA Short Read Archive (http://www.ebi.ac.uk/ena/) under accession no. ERP002231, ERP002028, ERP002042, and ERP002032.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00592-13/-/DCSupplemental.

Dataset S1, XLS file, 0.1 MB.
Dataset S2, XLS file, 0.4 MB.
Dataset S3, XLS file, 0.1 MB.
Text S1, PDF file, 0.1 MB.
Figure S1, EPS file, 0.5 MB.
Table S1, DOCX file, 0.1 MB.
Table S2, DOCX file, 0.1 MB.
Table S3, DOCX file, 0.1 MB.
Table S4, DOCX file, 0.1 MB.
Table S5, DOCX file, 0.1 MB.

ACKNOWLEDGMENTS

This research was supported by the Office of Science (BER), U.S. Department of Energy, grant No. DE-FG02-99ER62848, DE-FG02-07ER64398, and DE-FC02-07ER64494 (DOE Great Lakes Bioenergy Research Center), NIH Human Microbiome Project Demonstration Project (UH3 DK083993), and NEON Inc.

We thank Marius Vital for the laboratory work for *but*-defined community samples and Jacob Parnell, James Leary, Christian Giardina, Jeremy B. Jones, Rachel Gallery, Eric Triplett, Stephen Coates, and Chien-Lu Ping for providing the site and environmental metadata for the NEON samples.

REFERENCES

1. **Stackebrandt E, Ebers J.** 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today **33**:153–155.
2. **Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM.** 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. **37**:D141–D145.
3. **Dagan T, Artzy-Randrup Y, Martin W.** 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc. Natl. Acad. Sci. U. S. A. **105**:10039–10044.
4. **Zhang Y, Sun Y.** 2011. HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. BMC Bioinformatics **12**:198. doi:10.1186/1471-2105-12-198.
5. **Rho M, Tang H, Ye Y.** 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. **38**:e191. doi:10.1093/nar/gkq747.
6. **Eddy SR.** 2011. Accelerated profile HMM Searches. PLOS Comput. Biol. **7**:e1002195. doi:10.1371/journal.pcbi.1002195.
7. **Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD.** 2012. The Pfam protein families database. Nucleic Acids Res. **40**:D290–D301.
8. **Huse SM, Welch DM, Morrison HG, Sogin ML.** 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. **12**:1889–1898.
9. **Xu W.** 2006. On integrating biological sequence analysis with metric distance based database management systems. Ph.D. thesis. University of Texas at Austin, Austin, TX.
10. **Guan X, Uberbacher EC.** 1996. Alignments of DNA and protein sequences containing frameshift errors. Comput. Appl. Biosci. **12**:31–40.
11. **Vidal E.** 1986. An algorithm for finding nearest neighbours in (approximately) constant average time. Pattern Recognition Lett. **4**:145–157.
12. **Needleman SB, Wunsch CD.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.
13. **Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ.** 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics **12**:38. doi:10.1186/1471-2105-12-38.
14. **Poly F, Monrozier LJ, Bally R.** 2001. Improvement in the RFLP procedure for studying the diversity of nifHnifH genes in communities of nitrogen fixers in soil. Res. Microbiol. **152**:95–103.
15. **Raymond J, Siefert JL, Staples CR, Blankenship RE.** 2004. The natural history of nitrogen fixation. Mol. Biol. Evol. **21**:541–554.
16. **Dufrene M, Legendre P.** 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. **67**:345–366.
17. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.
18. **Iwai S, Chai B, Sul WJ, Cole JR, Hashsham SA, Tiedje JM.** 2010. Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. ISME J. **4**:279–285.
19. **Rodrigues EP, Rodrigues LS, Martinez de Oliveria AL, Baldani VLD, dos Santos Teixeira KR, Urquiaga S, Reis VM.** 2008. Azospirillum amazonense inoculation: effects on growth yield and N_2 fixation of rice (*Oryza sativa* L.). Plant Soil **302**:249–261.
20. **Giraud E, Moulin L, Vallenet D, Barbe V, Cytryn E, Avarre JC, Jaubert M, Simon D, Cartieaux F, Prin Y, Bena G, Hannibal L, Fardoux J, Kojadinovic M, Vuillet L, Lajus A, Cruveiller S, Rouy Z, Mangenot S, Segurens B, Dossat C, Franck WL, Chang WS, Saunders E, Bruce D, Richardson P, Normand P, Dreyfus B, Pignol D, Stacey G, Emerich D, Verméglio A, Médigue C, Sadowsky M.** 2007. Legumes symbiosis: Absence of nod genes in photosynthetic bradyrhizobia. Science **316**:1307–1312.
21. **Molouba F, Lorquin J, Willems A, Hoste B, Giraud E, Dreyfus B, Gillis M, de Lajudie P, Masson-Boivin C.** 1999. Photosynthetic bradyrhizobia from *Aeschynomene* spp. are specific to stem-nodulated species and form a separate 16S ribosomal DNA restriction fragment length polymorphism group. Appl. Environ. Microbiol. **65**:3084–3094.
22. **Rousk K, Jones DL, DeLuca TH.** 2013. Moss-cyanobacteria associations

as biogenic sources of nitrogen in boreal forest ecosystems. Front. Microbiol. **4**:150. doi:10.3389/fmicb.2013.00150.

23. **Fleming ED, Castenholz RW.** 2007. Effects of periodic desiccation on the synthesis of the UV-screening compound, scytonemin, in cyanobacteria. Environ. Microbiol. **9**:1448–1455.
24. **Belnap J.** 2002. Nitrogen fixation in biological soil crusts from southeast Utah, USA. Biol. Fertil. Soils **35**:128–135.
25. **Yeager CM, Kornosky JL, Morgan RE, Cain EC, Garcia-Pichel F, Housman DC, Belnap J, Kuske CR.** 2007. Three distinct clades of cultured heterocystous cyanobacteria constitute the dominant N2-fixing members of biological soil crusts of the Colorado Plateau, USA. FEMS Microbiol. Ecol. **60**:85–97.
26. **Reeder J, Knight R.** 2009. The "rare biosphere": a reality check. Nat. Methods **6**:636–637.
27. **Konstantinidis KT, Tiedje JM.** 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. **187**:6258–6264.
28. **Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA.** 2013. Evidence for a persistent microbial seed bank throughout the global ocean. Proc. Natl. Acad. Sci. U. S. A. **110**:4651–4655.
29. **Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R.** 2005. The microbial pan-genome. Curr. Opin. Genet. Dev. **15**:589–594.
30. **Henikoff S, Henikoff JG.** 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. **89**:10915–10919.
31. **Zehr JP, Jenkins BD, Short SM, Steward GF.** 2003. Nitrogenase gene diversity and microbial community structure: a cross-system comparison. Environ. Microbiol. **5**:539–554.
32. **Gaby JC, Buckley DH.** 2012. A comprehensive evaluation of PCR primers to amplify the *nifH* gene of nitrogenase. PLOS One **7**:e42149. doi:10.1371/journal.pone.0042149.
33. **Vital M, Penton CR, Wang Q, Young VB, Antonopoulos DA, Sogin ML, Morrison HG, Raffals L, Chang EB, Huffnagle GB, Schmidt TM, Cole JR, Tiedje JM.** 2013. A gene-targeted approach to investigate the intestinal butyrate-producing bacterial community. Microbiome **1**:8. doi:10.1186/2049-2618-1-8.