AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

ENVIRONMENTAL MICROBIOLOGY

Diazotroph Community Characterization via a High-Throughput *nifH* Amplicon Sequencing and Analysis Pipeline

John Christian Gaby,[a] Lavanya Rishishwar,[a,b,c] Lina C. Valderrama-Aguirre,[d,e] Stefan J. Green,[f] Augusto Valderrama-Aguirre,[c,g,h] I. King Jordan,[a,b,c] Joel E. Kostka[a,c]

[a]School of Biology, The Georgia Institute of Technology, Atlanta, Georgia, USA
[b]Applied Bioinformatics Laboratory, Atlanta, Georgia, USA
[c]PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia
[d]Laboratory of Microorganismal Production (Bioinoculums), Department of Field Research in Sugarcane, Incauca S.A.S, Cali, Valle del Cauca, Colombia
[e]School of Natural Resources and Environmental Engineering, PhD Program in Sanitary and Environmental Engineering, Universidad del Valle, Cali, Valle del Cauca, Colombia
[f]DNA Services Facility, Research Resources Center, University of Illinois at Chicago, Chicago, Illinois, USA
[g]Biomedical Research Institute, Universidad Libre, Cali, Valle del Cauca, Colombia
[h]Regenerar, Center of Excellence for Regenerative and Personalized Medicine, Valle del Cauca, Colombia

ABSTRACT The dinitrogenase reductase gene (*nifH*) is the most widely established molecular marker for the study of nitrogen-fixing prokaryotes in nature. A large number of PCR primer sets have been developed for *nifH* amplification, and the effective deployment of these approaches should be guided by a rapid, easy-to-use analysis protocol. Bioinformatic analysis of marker gene sequences also requires considerable expertise. In this study, we advance the state of the art for *nifH* analysis by evaluating *nifH* primer set performance, developing an improved amplicon sequencing workflow, and implementing a user-friendly bioinformatics pipeline. The developed amplicon sequencing workflow is a three-stage PCR-based approach that uses established technologies for incorporating sample-specific barcode sequences and sequencing adapters. Based on our primer evaluation, we recommend the Ando primer set be used with a modified annealing temperature of 58°C, as this approach captured the largest diversity of *nifH* templates, including paralog cluster IV/V sequences. To improve *nifH* sequence analysis, we developed a computational pipeline which infers taxonomy and optionally filters out paralog sequences. In addition, we employed an empirical model to derive optimal operational taxonomic unit (OTU) cutoffs for the *nifH* gene at the species, genus, and family levels. A comprehensive workflow script named TaxADivA (TAXonomy Assignment and DIVersity Assessment) is provided to ease processing and analysis of *nifH* amplicons. Our approach is then validated through characterization of diazotroph communities across environmental gradients in beach sands impacted by the Deepwater Horizon oil spill in the Gulf of Mexico, in a peat moss-dominated wetland, and in various plant compartments of a sugarcane field.

IMPORTANCE Nitrogen availability often limits ecosystem productivity, and nitrogen fixation, exclusive to prokaryotes, comprises a major source of nitrogen input that sustains food webs. The *nifH* gene, which codes for the iron protein of the nitrogenase enzyme, is the most widely established molecular marker for the study of nitrogen-fixing microorganisms (diazotrophs) in nature. In this study, a flexible sequencing/analysis pipeline, named TaxADivA, was developed for *nifH* amplicons produced by Illumina paired-end sequencing, and it enables an inference of taxonomy, performs clustering, and produces output in formats that may be used by programs that facilitate data exploration and analysis. Diazotroph diversity and community composition are linked to ecosystem functioning, and our results advance the phylo-

Received 26 July 2017 **Accepted** 21 November 2017

Accepted manuscript posted online 27 November 2017

Citation Gaby JC, Rishishwar L, Valderrama-Aguirre LC, Green SJ, Valderrama-Aguirre A, Jordan IK, Kostka JE. 2018. Diazotroph community characterization via a high-throughput *nifH* amplicon sequencing and analysis pipeline. Appl Environ Microbiol 84:e01512-17. https://doi.org/10.1128/AEM.01512-17.

Editor M. Julia Pettinari, University of Buenos Aires

Address correspondence to Joel E. Kostka, joel.kostka@biology.gatech.edu.

genetic characterization of diazotroph communities by providing empirically derived *nifH* similarity cutoffs for species, genus, and family levels. The utility of our pipeline is validated for diazotroph communities in a variety of ecosystems, including contaminated beach sands, peatland ecosystems, living plant tissues, and rhizosphere soil.

KEYWORDS bioinformatics, metagenomics, next-generation sequencing, nitrogen fixation, sequence analysis, taxonomy

Nitrogen is a biologically essential element, and a lack of fixed nitrogen can limit ecosystem productivity. Only microorganisms in the domains *Bacteria* and *Archaea* are known to be capable of fixing nitrogen, termed diazotrophy, and this reaction is performed by the nitrogenase enzyme. Nitrogenase is a complex of proteins encoded by the *nifH*, *nifD*, and *nifK* genes (1). Of the three genes, *nifH* has been the most widely used for investigating the diversity and composition of diazotroph communities, in part due to the fact that *nifH* is the more highly conserved (2). The phylogeny of the *nifH* gene and its paralogs may be divided into four clusters (I, II, III, and IV/V), as designated by Chien and Zinder (3). Clusters I and III encompass *nifH* from the conventional [MoFe]nitrogenase, whereas cluster II constitutes an alternative form of nitrogenase which has a different metal at its active site. Members of cluster IV are paralogs of the *nifH* gene and did not appear to function in nitrogen fixation (3) until a recent study demonstrated nitrogen fixation by *Endomicrobium proavitum*, which possesses only cluster IV nitrogenase (4). An additional cluster, V, has been defined, and this includes the *bchX*, *bchL*, and *chlL* genes, which are associated with photosynthetic pigment biosynthesis (5). Functional genes are more likely to be horizontally transferred than taxonomic genes (such as the rRNA gene) (6), and there is evidence to support the occurrence of horizontal gene transfer (HGT) of *nif* genes in certain diazotrophs (5, 7).

Diazotroph diversity has been described in a wide range of marine (8), terrestrial (9), and managed (10) ecosystems, including extreme environments, such as hydrothermal vents (11), and in host-associated environments, like the termite gut (12, 13). In the marine environment, the dominance of an uncharacterized heterotrophic nitrogen fixer, originally termed UCYN-A, was discovered in tropical and subtropical marine waters through *nifH* sequencing (14). UCYN-A was later found to be a symbiont of a marine eukaryote (15), thereby showing that nitrogen inputs in these tropical waters are in part provided via the symbiotic interaction of an alga and a heterotrophic cyanobacterium. In terrestrial ecosystems, the response of diazotrophs to land use change was studied in the Amazon rainforest, showing that while richness did not change due to forest conversion to pasture, there was a 10-fold increase in the abundance of diazotrophs in pasture and a change in community composition (16). In agricultural systems, diazotroph community composition has been shown to be influenced by management practices, such as nitrogen fertilization (17) and organic residue retention (18). Furthermore, *nifH* sequencing has identified the diazotroph species that occur in the microbiomes of crop plants and has shown them to be a subset of the soil diazotroph community (10). As the aforementioned examples show, the use of the *nifH* gene as a molecular marker has yielded significant advances in our understanding of the ecology and function of diazotrophs.

Most studies of the *nifH* gene have been carried out with the Sanger sequencing platform. In 2012, GenBank contained approximately 33,000 *nifH* sequences contributed by 1,211 studies (2) over nearly 20 years of work. At that time, the study with the single highest contribution of sequences generated 1,299 sequences (10). Recently, individual studies using next-generation sequencing technologies, like pyrosequencing, have generated from 80,000 to just over 300,000 sequences (19, 20), which is up to 10 times the number of *nifH* sequences generated by Sanger sequencing over the previous 2 decades (see Fig. S1 in the supplemental material). In fact, a single Illumina MiSeq 2 × 250 run can generate ~9 million merged amplicon sequences (21). Compared to the highest number of *nifH* reads ever generated in a Sanger sequencing

study, it is now possible to generate 1,000 times more sequences in a single study with next-generation sequencing technologies. Thus, advances in sequencing technology now permit unprecedented sequencing depth and sample coverage, thereby enabling rapid and robust marker gene surveys.

To make use of the vast amount of sequence data generated by new technologies, there is a need for effective, rapid, and easy-to-use analysis pipelines that are tailored to the requirements of each amplicon type. Quality control and filtering of error sequences are specific to the sequencing platform (22), and the platform used can vary depending on amplicon length and other factors. In the case of *nifH*, some high-coverage primer sets can amplify paralogous sequences (12) whose function may not be to fix nitrogen, in which case the paralogs would need to be identified and removed from the data set before conducting diversity analysis. Decisions must be made as to which operational taxonomic unit (OTU) cutoff is most appropriate for the sequence variability that exists in the gene of interest (23), although there exists the possibility of using clustering-independent approaches (24). There are a number of clustering algorithms available (e.g., see references 25–27), and each one has unique advantages and limitations. Identification and taxonomic description of organisms require a reference database, and although these are available for 16S (28), the taxonomic marker gene for bacteria, a database with the same advanced level of functionality and coverage has not been developed for most functional genes, including the *nifH* gene. Finally, there are a multitude of ways to analyze and visualize microbial diversity data, and selecting the most informative representations can be challenging and time-consuming.

We present a flexible sequencing and analysis pipeline tailored to the requirements of the *nifH* gene, and we reveal important considerations in order to ensure an accurate analysis. We present data assessing technical considerations, like primer choice and bias, and the inclusion of appropriate internal controls during sequencing. We also discuss approaches to reduce the cost of sequencing and sample multiplexing through the use of existing technologies, and we empirically define OTU cutoffs for the *nifH* gene which correspond to delineations of species, genus, and family. We then apply the pipeline to the analysis of diazotroph communities sampled from a range of environments, including oil-contaminated beach sands, peatland ecosystems, living *Sphagnum* tissues, sugarcane tissues, and rhizosphere soil. A workflow script named TaxADivA (TAXonomy Assignment and DIVersity Assessment) is provided that processes *nifH* amplicons produced by Illumina paired-end sequencing (Fig. S2). The script enables an inference of taxonomy, performs clustering, and produces output in formats that may be used by programs that facilitate data exploration and analysis.

RESULTS

In this study, the 5′ barcoding approach consisted of 3 libraries, each containing about 2.5 million reads. Given that there were 15 samples per library, if divided evenly, this would amount to an average of about 160,000 sequences per sample. With the Fluidigm approach, 9,002,212 indexed non-*phiX* reads were obtained, resulting in a mean of 23,443 reads per sample (for 384 samples).

An examination of internal controls showed that the demultiplexing resulted in a correct association of the control samples with their identifying barcodes, though there was a certain frequency with which apparent error sequences resulted in low-abundance, usually singleton, OTUs, and these could be eliminated by applying an abundance cutoff of 0.5% (Table S1). In the case of the internal control amplified from the genomic DNA template, 4 high-abundance OTUs remained after the abundance cutoff was applied, and BLAST results showed that the 3 OTUs of lesser abundance appeared to be due to the presence of *nifH* sequences most closely related to *Klebsiella variicola* DX120E. However, when performing beta-diversity analysis on environmental samples, the application of a 0.5% cutoff eliminates most OTUs, including low-abundance and rare taxa. A less stringent abundance cutoff of 0.05% would still eliminate many spurious OTUs in the controls and retain more OTUs representing real taxa in environmental samples (Table S1).

TABLE 1 Samples used for testing and optimizing the pipeline presented in this study

Project name	Sample name	No. of samples	No. of reads
Control	1 standard	3	146,046
	2 standards	1	58,584
	4 standards	1	33,212
	Subtotal	5	237,842
Peatland	Peat	5	698,301
	Sphagnum	2	533,672
	Subtotal	7	1,231,973
Sand	Clean	7	1,439,528
	Oil	7	1,065,570
	Subtotal	14	2,505,098
Sugarcane	Rhizosphere	9	1,072,175
	Root	3	102,098
	Stem	7	284,804
	Subtotal	19	1,459,077
Total		45	5,433,990

An effective sequence analysis pipeline was assembled using a combination of preexisting programs, alignment and taxonomy files, and a custom workflow script. Erroneous reads were culled at merging and primer trimming; for instance, in one sequencing run which included root, stem, and leaf samples, our sequence pipeline allowed recovery of 65% of rhizosphere and 3% of tissue initial reads prior to filtering of cluster IV/V sequences (Fig. S3). The number of starting reads associated with each library was highly consistent in the 5′-barcoding approach, with 2.62, 2.50, and 2.41 million sequences in libraries A, B, and C, respectively. The number of reads per sample within a library varied over 2 orders of magnitude from 4,000 to 600,000 reads (Table 1 and Fig. S4). For the Fluidigm barcoding approach, the number of reads per sample ranged from several thousand to 150,000, with a median of about 20,000 reads, and many of the low-read samples corresponded to those samples which had poor amplification. In addition, certain primer sets have a propensity to amplify cluster IV/V sequences present in particular environments, and our implementation of a filtering step detected these sequences for optional removal (Table 2 and Fig. 1).

The quality scores associated with the sequencing were excellent. After merging, the overlapping paired-end reads (e.g., there were approximately 100 bases of overlap for the Ando amplicons using the Fluidigm approach) increased in mean quality score, because the read merging program PEAR reassigns the product of the quality scores for the overlapping bases (29). Thus, the median sequence quality for all merged sequences was 55 by the 5′ barcoding approach and 37 by the Fluidigm barcoding approach, indicating an error frequency of less than 1 in 10^5 and 1 in 10^3 sequenced nucleotides, respectively.

Sequences that fall within cluster IV/V were amplified in high proportion with the Ueda and Ando primer sets from sand and sugarcane tissue samples (Table 2 and Fig. 1). Among the dominant OTUs (those OTUs above the abundance cutoff of 0.5%) for the clean sand, the Ueda primer set yielded 84% cluster IV/V OTUs, and for the Ando primer set, the same sample resulted in 58% cluster IV/V OTUs (Table 2). For sugarcane stem samples, cluster IV/V OTUs comprised up to 13% of OTUs with the Ando primer set and 50% with the Ueda primer set (Table 2 and Fig. 1). For the Poly primer set, only 1 of 9 samples had a single cluster IV/V OTU (Table 2 and Fig. 1), and this corroborates with *in silico* analysis, which shows that the primer set does not match to sequences in cluster IV/V as do the Ueda and Ando primer sets (Table 3).

We compared 4 *nifH* primer sets using the 179 nucleotides that overlapped the 4 amplicons and which also spanned positions 151 to 330 of the *nifH* gene (*Azotobacter vinelandii nifH* [GenBank accession no. M20568]). Diversity indices varied according to

TABLE 2 Percentage of cluster IV paralogs of *nifH* recovered as a result of environment sampled and *nifH* primer set used[a]

Sample name	Primer set name	No. of dominant OTUs	Cluster IV OTUs	
			No.	%
Clean sand I600	Poly	41	0	0
	Ando	40	23	58
	Fedorov	35	0	0
	Ueda	31	26	84
	ZM	29	0	0
Oil sand A	Ando	14	4	29
	Ueda	13	6	46
	Fedorov	9	0	0
	ZM	8	0	0
	Poly	4	0	0
Sugarcane rhizosphere (5 mo)	ZM	37	0	0
	Ando	26	1	4
	Fedorov	25	0	0
	Poly	24	0	0
	Ueda	24	2	8
Sugarcane stem (5 mo)	Ando	32	4	13
	ZM	27	6	22
	Ueda	24	12	50
	Poly	23	1	4
	Fedorov	11	1	9

[a]The primer sets for each sample are sorted by the number of dominant OTUs the primer set yielded.

the primer set used, with the ZM and Ando primer sets yielding the highest values of observed species, Shannon index, and phylogenetic diversity (Table 4). Cluster IV/V filtering may be preferred before diversity analysis, as the values for the diversity indices are likely to be influenced by the presence of cluster IV/V sequences, and some primer set and sample combinations recover a high percentage of cluster IV/V sequences (Table 2 and Fig. 1). A beta-diversity analysis of the same sample DNAs amplified with different primer sets shows a strong effect of primer (Fig. 2). The Poly and ZM primers cluster together, whereas the Ueda and Ando primers separate (Fig. 2B). Similarly, taxon abundances correlate more between samples amplified with the ZM and Poly primers, while the Ueda and Ando primers also exhibit higher correlation between themselves

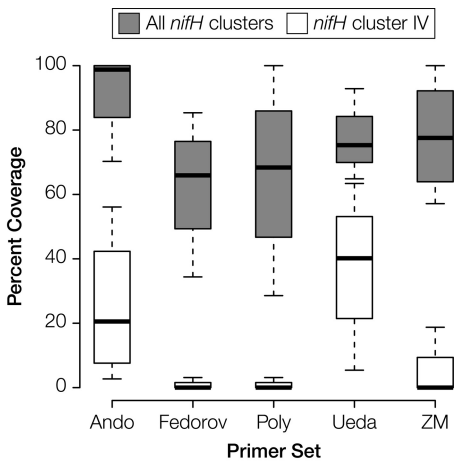

FIG 1 The average percentage of *nifH* and cluster IV paralogs of *nifH* in all the samples using the five primer sets evaluated here.

TABLE 3 *nifH* primer sets and their sequences used in this study, along with their experimental characteristics and *in silico* coverages

Primer set name	Reference	Primer set	Sequence (5′ to 3′)	PCR primer annealing temp (°C)	Concn of primer in PCR (nM)	Positions	T_m (°C)[a]	Deg[b]	Amplicon length produced (bases)	*nifH*[c]	IV[d]
Ando	52	IGK3	GCIWTHTAYGGIAARGGIGGIATHGGIAA	58	1,000	19–47	69.4–75.3	72	395	92	78
		DVV	ATIGCRAAICCICCRCAIACIACRTC	58	1,000	388–413	71.7–75.8	8	395	94	52
Fedorov	59	nifH-2F	GMRCCIGGIGTIGGYTGYGC	62	1,000	277–296	69.2–78.3	16	215	87	20
		nifH-3R	TTGTTGGCIGCRTASAKIGCCAT	62	1,000	469–491	68.5–72.1	8	215	48	3
Poly	58	polF	TGCGAYCCSAARGCBGACTC	62	200	115–134	63.8–70.1	24	362	39	3
		polR	ATSGCCATCATYTCRCCGGA	62	200	457–476	63.7–67.5	8	362	35	0
Ueda	57	Ueda19F	GCIWTYTAYGGIAARGGIGG	55	1,000	19–38	62.4–67.9	16	389	93	76
		Ueda407R	AAICCRCCRCAIACIACRTC	55	1,000	388–407	63.9–70.6	8	389	91	71
ZM	56	nifH2	TGYGAYCCNAARGCNGA	58	1,000	115–131	54.0–68.1	128	362	95	37
		nifH1	ADNGCCATCATYTCNCC	58	1,000	460–476	52.5–63.9	96	362	94	13

[a]The range of melting temperatures (T_m) for oligonucleotides in the degenerate primer mix.
[b]Primer degeneracy as the number of constituent oligonucleotides.
[c]Percent true *nifH* (clusters I and III) covered, as determined by *in silico* analysis.
[d]Percent true cluster IV (*nifH* paralogs) covered, as determined by *in silico* analysis.

(Fig. S5). The Poly and ZM primer sets overlap in their binding site on *nifH* and amplify nearly the same stretch of *nifH*.

A *nifH* alignment and a taxonomic reference set were assembled, one specifying clusters and the other taxonomic lineage based on cultivated representatives. These references were assembled from 32,954 *nifH* sequences that are part of a database in ARB format that is an update of previous database versions (2). An alignment was made from a dereplicated set of OTU representatives formed by clustering at 97% similarity that contains 8,058 sequences. This permitted the alignment of 83,480 sequence representatives of length 337 bases in 457 s. The taxonomic reference set split the *nifH* database into the principal taxonomic clusters to allow for identification and removal of cluster IV/V sequences.

The time required to complete each step of the pipeline was evaluated, and the most time-consuming step is phylogenetic tree construction for UNIFRAC analysis, which required 3,316 s using the QIIME wrapper. The second most time-consuming step is taxonomy assignment with QIIME to remove cluster IV/V sequences, which required 2,082 s to conduct a BLAST search for the 112,544 representative sequences against the taxonomic reference set containing 8,058 sequences. In particular, the

TABLE 4 Alpha-diversity measures by sample and primer set[a]

Alpha-diversity measure	Sample	Data by primer set (mean ± SE)			
		Ando	ZM	Ueda	Poly
Shannon	Oil sand A	1.9 ± 0.0	1.4 ± 0.0	1.4 ± 0.0	1.0 ± 0.0
	Clean sand I600	5.7 ± 0.0	5.9 ± 0.0	4.2 ± 0.0	5.5 ± 0.0
	5-mo rhizosphere	5.8 ± 0.0	6.1 ± 0.0	5.4 ± 0.0	5.3 ± 0.0
	5-mo stem	4.8 ± 0.0	3.4 ± 0.0	3.5 ± 0.0	3.1 ± 0.0
Observed species	Oil sand A	75.2 ± 1.4	77.6 ± 1.1	63.9 ± 0.6	52.6 ± 1.0
	Clean sand I600	289.3 ± 2.2	429.3 ± 4.1	254.5 ± 1.6	371.6 ± 3.9
	5-mo rhizosphere	490.2 ± 4.1	338.3 ± 2.1	466.8 ± 1.8	365.6 ± 4.3
	5-mo stem	205.5 ± 3.6	138.5 ± 2.4	158.0 ± 0.0	144.0 ± 2.8
Phylogenetic diversity	Oil sand A	17.9 ± 0.3	17.5 ± 0.3	15.1 ± 0.1	12.5 ± 0.3
	Clean sand I600	41.9 ± 0.3	60.2 ± 0.4	39.8 ± 0.2	48.7 ± 0.4
	5-mo rhizosphere	70.3 ± 0.5	49.5 ± 0.4	63.6 ± 0.2	48.7 ± 0.5
	5-mo stem	38.5 ± 0.3	18.1 ± 0.3	21.4 ± 0.0	18.6 ± 0.4

[a]The Ando and ZM primer sets were able to capture more *nifH* diversity in the four sample types tested here than the three other universal *nifH* primer sets. The alpha diversity of diazotrophs in each of the four samples was measured as the Shannon diversity index, observed species, and phylogenetic diversity.

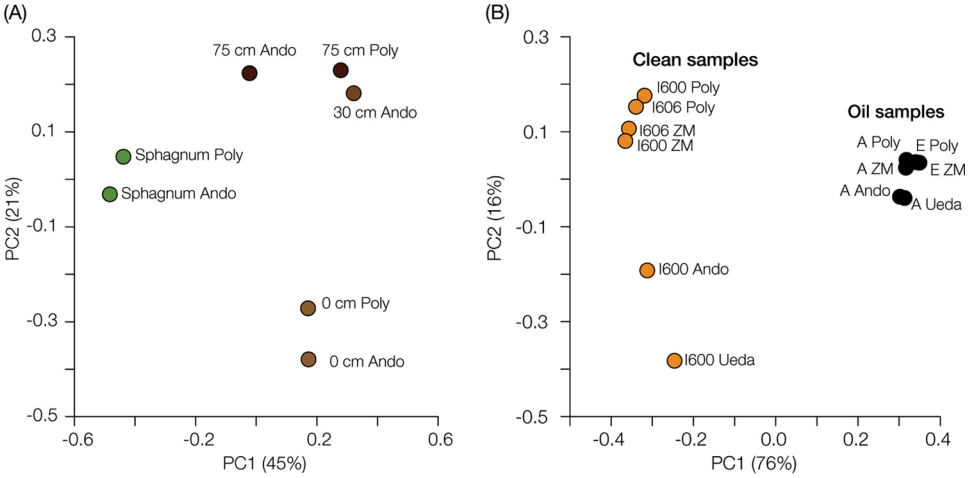

FIG 2 Clustering of sequencing samples based on their sampling source and amplification primer set used. Principal coordinates (PC1 and PC2) analysis of weighted UNIFRAC distances for the diazotroph community in *Sphagnum* (green) and peat from 0 (light brown), 30 (brown), and 75 cm (black) sampling depth (A) and in Pensacola Beach sands following oil contamination from the Deepwater Horizon oil spill (B). The sample labels denote the sample name and the *nifH* primer set used for amplification.

TaxADivA script was written to improve pipeline performance at the taxonomic assignment step.

The analysis pipeline was evaluated for samples from a contaminated beach, a peatland, and a sugarcane field. Diazotroph communities were distinct in each ecosystem and between sample types (Fig. 2 and 3). Beta-diversity analysis revealed a strong separation of diazotroph communities in oiled sands in comparison to clean sands (Fig. 2B), and alpha-diversity analysis showed that communities in heavily oiled sand were less diverse (Table 4). Peatland diazotroph communities exhibited a strong vertical stratification from the surface dominated by living *Sphagnum* plants down into deeper layers of degraded peat (Fig. 2A). A predominance of aerobic diazotrophic genera were associated with living peat moss (*Sphagnum*) tissues, including *Methylobacterium*, *Bradyrhizobium*, and *Nostoc*, whereas communities from degraded peat contained a substantial contribution of *nifH* sequences whose closest identified relatives were facultative or strict anaerobes, including *Methanomethylophilus*, *Thiocystis*, *Desulfobotu-*

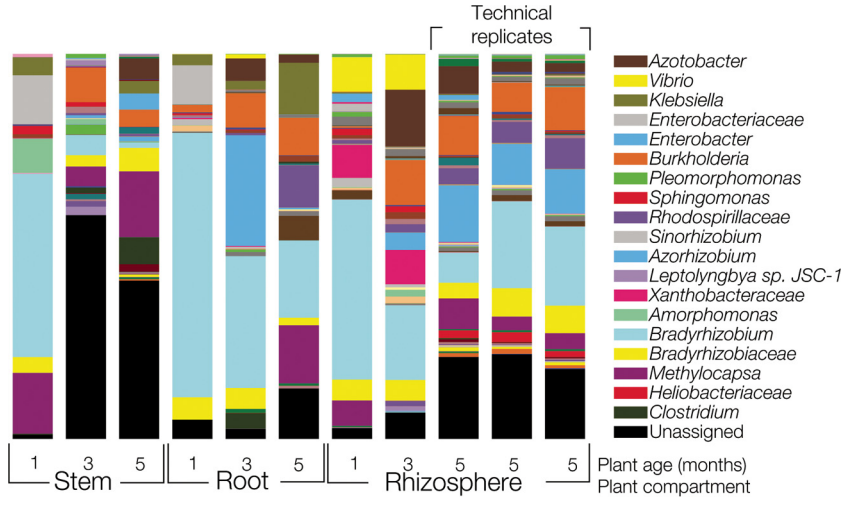

FIG 3 Microbial diversity captured in the stem, root, and rhizosphere compartments of sugarcane. Taxonomy bar plots are color-coded to indicate diazotroph genera according to the legend on the right. For taxa which could not be established at the genus level, the family is indicated. The Ando primer set (DVV/IGK3) was used to amplify the *nifH* gene from all samples shown.

lus, *Desulfarculus*, *Anaeromyxobacter*, *Geobacter*, *Rhodospirillum*, *Rhodopseudomonas*, *Desulfovibrio*, *Rhodoplanes*, *Rubrivivax*, and *Rhodoblastus*. The sugarcane rhizosphere soil diazotroph community was distinct from those of stem and root tissue samples (Fig. 3). The root and stem of sugarcane contained common plant-associated diazotroph genera, such as *Bradyrhizobium*, *Methylocapsa*, *Burkholderia*, and *Azotobacter*, while the rhizosphere soil contained these genera as well as *Azorhizobium* (Fig. 3). The diazotroph communities of the sugarcane stem were much less diverse than those of the rhizosphere soil (Table 4).

DISCUSSION

Nitrogen-fixing microorganisms (diazotrophs) are distributed throughout the majority of Earth's ecosystems. The diversity and community composition of diazotrophs are linked directly to ecosystem function. However, environmental complexity and past methodological constraints have often limited our ability to link specific diazotroph groups to ecosystem function. Since the vast majority of diazotrophs are not yet in cultivation, cultivation-independent molecular approaches are essential to the analysis of community dynamics, and the *nifH* gene is the most widely used molecular proxy for nitrogen fixation potential (30). The large volumes of sequence data generated by high-throughput technology present new research opportunities as well as data processing challenges. In this study, we present a custom-designed approach for the improved analysis of nitrogen-fixing microbial communities using next-generation sequencing on an Illumina platform. Our approach includes modifications to PCR conditions, cost savings through improved multiplexing, and a powerful, easy-to-use sequence analysis pipeline that we have termed "TaxADivA."

Although the expenses associated with high-throughput sequencing have been greatly reduced, cost remains as an important limitation. Thus, our protocol was optimized to capture the highest diversity of diazotrophs while minimizing the costs of next-generation sequencing through multiplexing. Two different barcoding approaches were investigated for sequencing on an Illumina MiSeq platform, one whereby the 5′ ends of the *nifH* primers are barcoded, and another whereby common sequences are synthesized onto the 5′ ends of the *nifH* primers, and then the common sequence is used to attach the amplicon to a set of barcodes and adapter sequences developed by Fluidigm (31). The 5′ approach was used to identify the best-performing primer sets. Based on our primer evaluation, we recommend the Ando primer set be used with a modified annealing temperature of 58°C, as this approach captured the largest diversity of *nifH* templates, including cluster IV sequences. Our approach then removes cluster IV sequences using TaxADivA. The *nifH* amplicons generated by the Ando primer set are just 395 bases in length, and thus there is an overlap of about 100 bases between the two paired-end reads, which allows for read merging to generate full-length *nifH* amplicons. The Ando set contains 5 inosines in each primer, which makes it expensive to synthesize, and given the need for multiple barcoded primers, we recommend the Fluidigm approach, whereby common sequences can be employed to greatly reduce costs. The Fluidigm approach also allows us to extend this method to other genes with the same barcode set simply by synthesizing a common sequence on the 5′ end of each target-specific primer. While we have recommended the Ando primer set, it should be noted that ecologically relevant trends, such as the difference in diazotroph community composition between oiled and clean beach sands, would still be apparent regardless of the *nifH* primer set used. The ZM primer set captures a diversity of *nifH* templates similar to that with the Ando primers. Thus, if investigators are concerned about contamination by cluster IV/V sequences, the ZM primer, which excludes most cluster IV OTUs, can be used, and TaxADivA could be modified for this primer set.

TaxADivA is a workflow script with a unique combination of programs for the processing and analysis of *nifH* amplicons generated by next-generation sequencing. At

least two other pipelines are available which can be used for the processing of *nifH* sequences, the Functional Gene Pipeline (FunGene [32]), and a more recent tool developed by Frank et al. (30). The benefits of TaxADivA over these other approaches include flexibility to define analysis criteria, improved sequence filtering and trimming, and consideration of full-length amplicon sequences. For example, the pipeline developed by Frank et al. achieves rapid classification of *nifH* sequences (30) by assigning sequences to phylogenetic subclusters. The approach uses classification and regression trees to identify amino acids in NifH that discriminate between the subclusters. While the principal advantage of the Frank et al. approach is that it is rapid because it employs a simple positional search based on one or a few amino acids, it may not always be as accurate as a consideration of the full-length sequence. In comparison to the FunGene tool, TaxADivA meets particular needs, such as filtering of cluster IV/V sequences (see below) and trimming of highly degenerate primer sequences. Trimming of primer sequences may fail during quality control (33), likely due to the extreme degeneracy of *nifH* primers, and thus we adapted the primer trimming function in the TaxADivA script to trim a specified number of bases off the flanking ends of the amplicon.

TaxADivA can be run as a standalone script on a desktop personal computer (PC) without the need to access Internet servers, and it has the flexibility of allowing the user to define criteria through command-line arguments, including the clustering cutoff for species and genus, minimum sequence depth per sample, and number of threads to use. The TaxADivA script generates output for several powerful but easy-to-use programs that permit an exploration of oligotypes, alpha and beta diversity, and differences in the relative abundances of diazotroph taxa between experimental groups, and these programs generate figures and reports in .html format for interactive exploration with a Web browser.

TaxADivA was employed along with our well-curated database to reevaluate the thresholds for the phylogenetic characterization of diazotrophs at the species, genus, and family levels. Our empirical model shows that the optimal sequence cutoffs for OTUs are 92% and 88% identity for delineating *nifH* diversity at the species and genus levels, respectively. In order to investigate the controls of nitrogen fixation in the environment, the proper identification of ecologically relevant taxa that mediate nitrogen fixation is critical (30, 34). Here, we provide improved target cutoffs supported by a curated database to accurately assess diazotroph diversity using *nifH* as a molecular proxy. Further, these cutoffs can be used to distinguish between functional and nonfunctional *nifH*.

The amplification of *nifH* paralog sequences (cluster IV/V) is problematic because their presence can skew the results of diversity analyses. Cluster IV sequences are paralogs of *nifH* which had yet to be demonstrated to fix nitrogen (35) until *Endomicrobium proavitum* was recently shown to possess only cluster IV and fix nitrogen (4). However, given that this is a single instance and that the exact phylogenetic extent of cluster IV diazotrophy has yet to be precisely delineated, we decided to provide users the option of filtering out cluster IV/V sequences in our pipeline. Once more information becomes available in the future, this feature can be further refined to identify particular subclusters within cluster IV that are known to fix nitrogen. Cluster IV sequences have been associated with methanogens (35, 36), which would be found primarily in anoxic environments. Other non-*nifH* paralogs that may be amplified by *nifH* primers include bacteriochlorophyll synthesis genes from photosynthetic bacteria (37) and their homologs in chloroplasts (38); together, these comprise cluster V. In particular, we show that the Ueda and Ando primer sets amplify a high proportion of cluster IV/V sequences. Previous studies using the Ueda primers also amplified a high proportion of cluster IV/V sequences from environments, like the termite gut (13). However, whether cluster IV/V sequences are amplified by a particular primer set requires first that the sequences be present in the sample. Environments with a high number of bacteriochlorophyll-producing bacteria could also presumably yield amplification of non-*nifH* paralogous sequences, as was seen with beach sands. Regardless of the primer set employed, there appears to be a lower prevalence of cluster IV/V

sequences in oxic soils. Moreover, our pipeline includes an optional filtering step to remove cluster IV/V sequences.

The inclusion of internal controls is recommended, as it allows for confirmation of correct demultiplexing and a means to evaluate sequencing errors. To completely eliminate the low-abundance OTUs which presumably occur as the result of sequencing errors, we applied a frequency cutoff of 0.005. Previous work with short subunit (SSU) rRNA gene amplicons led to the recommendation of a 0.00005 cutoff (39); however, this cutoff was to be applied to a composite sample set consisting of millions of sequences. In our case, since our read depth varied by 2 orders of magnitude, the cutoff would need to be applied to individual samples in order to prevent elimination of samples with low read depths from the combined sequence set. Thus, in cases where there is significant variation in read depth, we recommend application of a cutoff to each individual sample. Based upon our work, a frequency cutoff of 0.0005 would eliminate most spurious OTUs while leaving many OTUs representing the more abundant organisms, and this is generally consistent with the cutoff applied in other studies (40).

In order to evaluate our sequencing and analysis pipeline for application to studies of diazotroph ecology, we studied a range of ecosystems and environmental gradients. Our laboratory has investigated the impacts of oil contamination from the Deepwater Horizon (DWH) disaster in Florida beach sands in the Gulf of Mexico using amplicon-based and metagenomics approaches (41–43). Based largely on PCR amplification and sequencing of SSU rRNA genes, previous work has shown a pronounced decrease in the taxonomic diversity of microbial communities in marine environments exposed to oil contamination (41–43). Oil represents an effective carbon substrate for microbes that is poor in major nutrients (N and P), and an overall increase in microbial abundance is observed along with enrichment in known hydrocarbon-degrading bacterial groups in response to oiling. In this study, we utilized our new approach to reveal that a much lower taxonomic diversity of diazotrophs is observed in oil-contaminated sands than in clean sands. Our results indicate that the reduction in taxonomic diversity in response to oiling should be extended to diazotroph communities. One explanation for these results is that in response to oil input, nitrogen limitation drives the enrichment of hydrocarbon-degrading microbial groups which are also capable of nitrogen fixation. Further study is warranted on the identity and function of the enriched diazotroph microbial groups.

Peatlands are wetlands that tend to be extremely nutrient poor and nitrogen limited (44). In *Sphagnum*-dominated peat bogs of northern Minnesota, soil microbial communities were shown to stratify according to peat depth in parallel with the decomposition of organic matter and nutrient release from remineralization (45–47). In this study, we show that distinct diazotroph communities inhabit the living moss layer at the peat surface in comparison to degraded peat. Our observations are further corroborated by an analysis of *nifH* genes in metagenomes along vertical geochemical gradients in the same peatland (44, 48). Moreover, our results show vertical stratification to be the strongest ecological forcing of diazotrophic communities at the site. Since nitrogen fixation represents an important source of nitrogen to peatland ecosystems, these results begin to define which taxa are most important in supplying nitrogen.

In the sugarcane fields sampled, substantial shifts in diazotroph community composition were observed between plant compartments and in the rhizosphere. With our new approach, we show that the rhizosphere, well known to be a hot spot for microbes, contains the highest diversity of diazotrophs among the compartments of the plant microbiome that we studied. Moreover, the root compartment, in particular, appeared to select for certain diazotrophs, such as *Bradyrhizobium* spp., a microbial group which is often associated with sugarcane (34, 49–54). Sugarcane is a crop that does not contain nodulated diazotroph symbionts, and these results provide a starting point for future studies to identify free-living or associative diazotroph taxa that contribute to the fitness of this important crop plant. A substantial number of unassigned sequences were obtained from sugarcane, and this could owe to the presence of cluster IV/V or

other sequences not captured by the filtering step, or to the presence of uncharacterized diazotrophs in these environments. As diazotroph genomes are continually being added to public databases, a future update of our curated database to include the new diazotroph sequences may reduce the number of unassigned OTUs, and the inclusion of more cluster IV/V sequences in the taxonomy reference may help identify missed paralogs.

Here, we describe a new approach for investigating diazotroph ecology by targeting PCR amplification, next-generation sequencing, and bioinformatics analysis of *nifH* genes. Our approach is tailored for Illumina paired-end sequencing and includes modifications to PCR conditions, cost savings through improved multiplexing, and a powerful, easy-to-use sequence analysis pipeline that we have termed TaxADivA. The Ando primer set is recommended, and we emphasize best practices, such as cluster IV/V sequence filtering. Our results advance the phylogenetic characterization of diazotroph communities by providing empirically derived *nifH* similarity cutoffs for species, genus, and family levels. Our workflow script, TaxADivA, is flexible and may be extended for use with other functional genes and/or primer pairs.

MATERIALS AND METHODS

Sampling. Diazotroph communities were characterized from a range of ecosystems in order to properly evaluate the *nifH* analysis and sequencing pipeline. Samples were collected from 20 cm to 40 cm below the surface in the supratidal zone of the municipal beach (lat 30.32616, long −87.17450) at Pensacola Beach, FL, USA, in July 2010 and June 2011. Samples collected from 2010 were heavily impacted by oil that came ashore from the Deepwater Horizon (DWH) oil spill in the Gulf of Mexico in 2010, while the 2011 samples were collected after microbial communities had recovered and oil contamination was close to background levels (41, 42). This depth interval was chosen due the fact that oil was buried and persisted as a layer of contamination in the beach for 6 to 8 months in 2010 (42, 43). Samples representing a wetland or peatland ecosystem were collected from a *Sphagnum*-dominated ombrotrophic bog in July 2012 from the Marcel Experimental Forest operated by the USDA Forest Service near Grand Rapids, MN. Surface samples (living *Sphagnum* peat moss and partially degraded peat) and core samples (peat) were collected to 1 m below the surface, as previously described (44). Living peat moss and peat samples were immediately sectioned into depth intervals and frozen on dry ice. Sugarcane samples were collected from the Liberia farm operated by Incauca, S.A., in Valle del Cauca, Colombia, a tropical valley of the Andes Mountains, with a mean annual temperature of 23°C and precipitation ranging from <800 to 2,000 mm per year (55). Sprouting shoots were excavated from the soil, and shoots with attached roots and adherent soil were shipped intact to the laboratory. Once in the laboratory, the plants were subsampled to obtain the rhizosphere, root, and stem compartments. After processing, the samples were stored at −20°C or colder until DNA extraction. An overview of the experimental and computational pipeline is presented in Fig. 4.

DNA was extracted from sand, peat, and soil samples using a Mo Bio PowerSoil DNA isolation kit (Mo Bio Laboratories, Inc., Carlsbad, CA, USA). For all of these samples except sand, DNA was extracted from 0.25 g of material. DNA was extracted from 0.5 g of clean sand samples and from 0.15 g of material for the oiled sands, due to the fact that the extraction of greater amounts resulted in coextraction of inhibitors. For sugarcane microbiome (root and stem) samples, DNA was extracted from 0.050 g of material with a Mo Bio PowerPlant DNA isolation kit (Mo Bio Laboratories, Inc.). The bead-beating step of the DNA extractions was performed on a Talboys high-throughput homogenizer (Talboys, Thorofare, NJ, USA) for 2 min at maximum intensity.

PCR amplification, library preparation, and sequencing. Five *nifH* primer sets (Table 3) were selected for evaluation based on criteria of high coverage (i.e., Ando [52], ZM [56], and Ueda [57]), low template-specific primer bias (i.e., Poly [58]), or small amplicon size (i.e., Fedorov [59]). To perform sample-specific barcoding, degenerate forward primers were synthesized by the addition of a 12-base Golay barcode to the 5′ end of a *nifH* primer (i.e., "5′-barcoded primer approach"). Each unique forward primer was combined with its respective reverse primer in PCRs with each sample to generate barcoded amplicon sequences (Data Set 1). PCR amplification was carried out in 25-μl reaction mixtures with 0.625 units of DreamTaq DNA polymerase (Thermo Fisher Scientific, Inc., Pittsburgh, PA, USA), 1× buffer supplied by the polymerase manufacturer containing 2 mM Mg^{2+}, 200 μM dinucleoside triphosphate (dNTPs; Thermo Fisher Scientific, Inc.), 2 mg/ml nonacetylated bovine serum albumin (BSA; New England BioLabs, Ipswich, MA, USA), and primer concentrations given in Table 3. One microliter of the DNA extracts was added to each PCR except for sugarcane stem extracts, where up to 6 μl was added because of the apparently low concentration of *nifH* in the extract. A hot start PCR was performed with a 180-s hot start step at 95°C, followed by 40 cycles at 95°C for 30 s, the annealing temperature for each respective primer set (Table 3) for 30 s, and extension at 72°C for 60 s. The cycling program ended with a 600-s final extension at 72°C. Barcoded amplicons were pooled and sequencing adapters ligated as described below.

Subsequently, a revised amplification protocol was employed (i.e., Fluidigm approach; see the supplemental material for the detailed protocol) for the Ando primer only. Here, sample-specific barcoding and

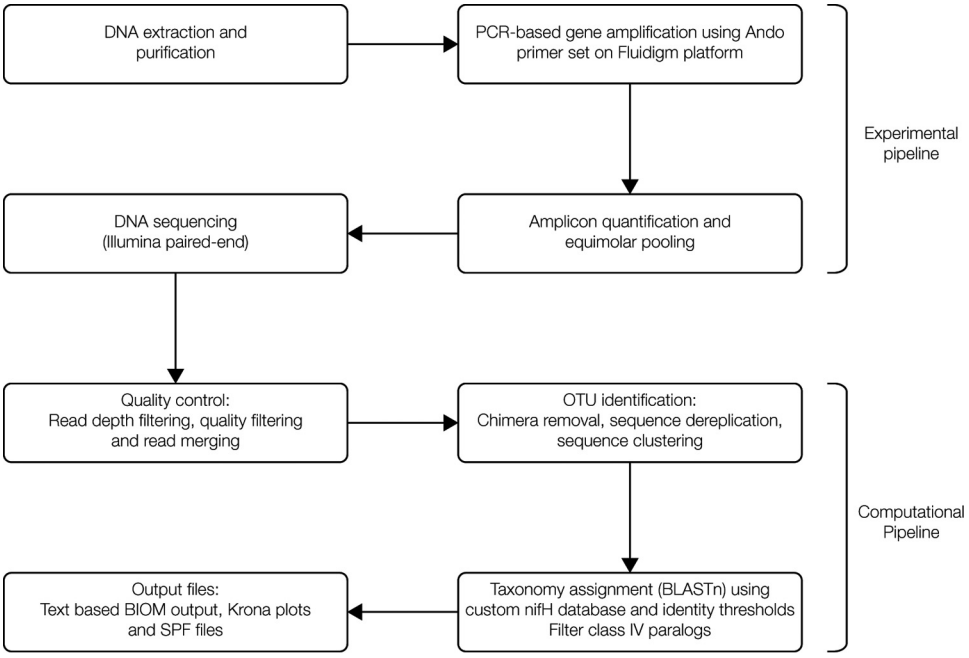

FIG 4 A schematic of the *nifH* amplicon experimental and computational (TaxADivA) workflow proposed here.

sequencing adapter incorporation were performed using only PCR, employing the targeted amplicon sequencing (TAS) approach described previously (e.g., see references 60–62). In our study, however, three stages of PCR were performed (see supplemental protocol 1 for reaction and cycling conditions), because the use of the primers with linker sequences (necessary for the Fluidigm approach) led to the generation of strong primer dimers. Thus, the first-stage amplification was performed with the Ando primers without linkers for 35 cycles. Subsequently, an additional 4 cycles of PCR were performed in a separate reaction with Ando primers having the 5′-linker sequences (common sequence 1 [CS1] and common sequence 2 [CS2] on the forward and reverse primers, respectively). In the final stage of PCR, amplicons from the second stage of PCR amplification were amplified using AccessArray primers (Fluidigm AccessArray barcode library; Fluidigm, San Francisco, CA, USA) to incorporate sample-specific barcodes and Illumina sequencing adapters. In stages 2 and 3, 1 μl of product from the previous stage was used as the template without purification.

For amplicons generated using the 5′-barcoded primer approach, library preparation was performed at the DNA Services Facility at the University of Illinois at Chicago, and sequencing was performed at the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign. For the amplicons prepared via the Fluidigm barcoding approach, sequencing was conducted at the Georgia Institute of Technology on a MiSeq system (Illumina, San Diego, CA, USA) with the MiSeq reagent kit version 2 (500 cycles) at 10 pM library concentration and with 15% PhiX DNA (Illumina) included in the sequencing run. Further information on quality control, purification, library preparation, and sequencing is provided in the supplemental material.

Sequence processing pipeline. Analyses were performed on a desktop computer with Intel Core i7-4770 CPU × 8 (3.40 GHz) with 16 Gb random access memory (RAM) running 64-bit Ubuntu 14.04 LTS BioLinux (63). Sequence data were received in fastq format. In the case of the approach using 5′-attached barcodes, the amplicons had been split into 3 libraries containing 15 samples per library. Samples from both approaches were evaluated for sequence quality with the program FastQC version 0.11.3 (64). The forward and reverse reads in each library were merged at the overlapping region using the program PEAR version 0.8.1 (29), with arguments specifying a minimum overlap of 35, a disabled statistical test, a maximum assembly length of 450, and a minimum assembly length of 215. The fastq format files were converted to FASTA and QUAL files, and because about half of the reads should by chance be in the reverse complement orientation, each FASTA file was also reverse complemented. Next, demultiplexing was performed on FASTA files for both orientations with the QIIME version 1.7.0 (65) script split_libraries.py, with arguments specifying disable primer usage, Golay of 12 barcodes, maximum barcode errors of 0, and maximum homopolymer of 10. A small percentage of sequences identified to have a barcode in the reverse-complement orientation were identical to sequence identifiers in the normal orientation, and these duplicates were removed from the reverse-complemented sequences. To trim the primer sequences, the trim.seqs command was used in the program MOTHUR version 1.31.2 (66). Cluster IV/V sequences were identified by using the QIIME script assign_taxonomy.py with a custom reference database created by identifying the cluster associated with each sequence in the *nifH* database of Gaby and Buckley (2), and all cluster IV/V paralogs were removed. Clustering was performed for each sample separately, with the implementation of USEARCH version 6.1 (67) in QIIME using the command pick_otus.py, with an $OTU_{0.95}$ similarity cutoff. OTU tables were made, and all singletons were eliminated from

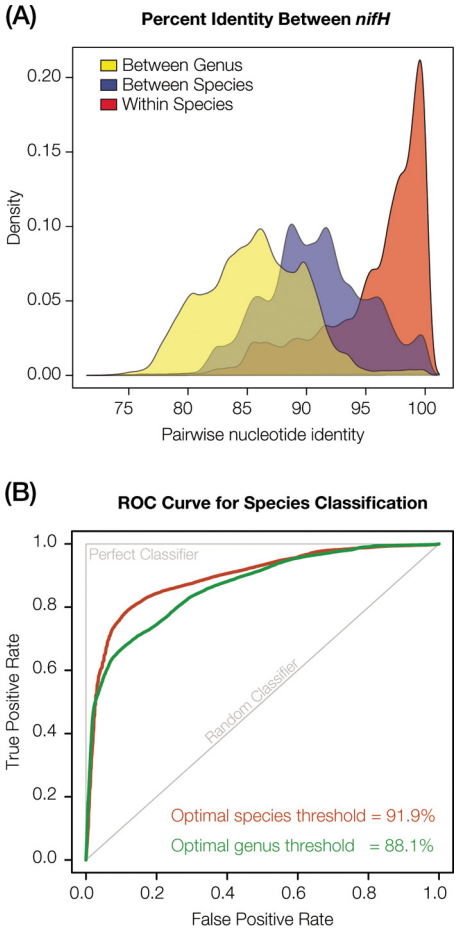

FIG 5 Empirical computation of optimal thresholds used for species and genus classifications. Percent identity of *nifH* genes from diazotrophic taxa as evaluated within species (red), between species (blue), and between genera (yellow) (A) and the receiver operating characteristic (ROC) curve for species classification (B) with *nifH* sequences. Optimal species and genus thresholds are given in panel B and are the default clustering cutoffs used in the TaxADivA workflow script.

the OTU tables. The time required to complete each command in the pipeline was evaluated with the BASH command 'time.'

A workflow script called TaxADivA (for TAXonomy Assignment and DIVersity Assessment) was written in Perl to facilitate the analysis of *nifH* amplicon sequences. The script uses multithreading to parallelize the processing of sequences and thereby reduce run time. Similar to the above-described pipeline, sequences are merged with PEAR (29), primers are trimmed, chimeras are removed and sequences are clustered with USEARCH (67), taxonomy is assigned with BLAST (68) by reference to a *nifH* taxonomy database, cluster IV/V sequences can be optionally removed, and a biom-compatible OTU table with taxonomy is produced, which may be easily converted for use with QIIME. The script also produces output for taxonomy exploration with the program Krona (69), using STAMP (70) in order to statistically test for differences in the relative abundance of taxa, using QIIME (65) to produce alpha- and beta-diversity metrics which may be visualized with Emperor (71), and using oligotyping analysis by Minimum Entropy Decomposition (72), which produces taxonomically labeled oligotype networks explorable with the network visualization tool Gephi (73). The source code, documentation, taxonomy files, and instructions for use of the script are available on GitHub (https://github.com/lavanyarishishwar/taxadiva).

Empirical determination of OTU cutoffs for species and genus. An empirical statistical model with a receiver operating characteristic (ROC) curve was employed to determine the optimal thresholds for species, genus, and family classifications. All validated *nifH* sequences were obtained from the curated *nifH* gene database (2) and compared (Fig. S2). Sequence identity values were computed for each pair of *nifH* sequences, and the resulting value was noted along with the corresponding relationship between the sequences, i.e., within species, between species, or between genera (Fig. 5A). A grid search approach was implemented to test all possible cutoff values between 70% and 100% sequence identity with a step-size of 0.1%. For each possible cutoff value, the number of sequences that were correctly (true positives [TP] and true negatives [TN]) and incorrectly (false positives [FP] and false negatives [FN]) placed were computed in the three taxonomy-level comparisons of within species, within genus, and within family. The overall true-positive rate [TPR = TP/(TP + FN)] and false-positive rate [FPR = FP/(FP + TN)]

were computed for each possible cutoff value, and an ROC curve was created for each of three taxonomy-level comparisons (Fig. 5B). The cutoff value closest to the top-left corner of the graph (TPR = 1 and FPR = 0) is the optimal cutoff value for the given data set. The optimal species cutoff of 91.9% (Fig. 5B) was set as the default clustering cutoff in the TaxADivA script. An OTU is assigned at either the species (BLAST hit, ≥91.9%), genus (≥88.1%), family (≥75%), or order (<75%) level; if no BLAST hit is returned, the OTU is labeled unclassified (Fig. S2).

Primer comparison. The sequences had to be aligned and then trimmed to the overlapping nucleotides common to the 4 evaluated primer sets, otherwise the clustering algorithm would generate spurious clusters because of length and overlap differences between amplicons generated with the different primer sets. To perform the alignment of all sequences, the QIIME command parallel_align_seqs_pynast.py was used with the arguments of minimum length set to 300 and minimum percent sequence identity to the closest BLAST hit in the template alignment set to 60. This script uses the PyNAST aligner (74) to align the sequences to a reference alignment based upon a set of sequences dereplicated at 97% similarity from the *nifH* database of Gaby and Buckley (2). Using ARB version 5.5 (75), we created a filter to indicate the overlapping alignment positions between the amplicon types that result from each primer set. The filter was used with the filter_alignment.py script of QIIME to trim the sequences in each sample. Then, all remaining gap characters were removed from the alignment, and the trimmed sequences thus consisted of the nucleotide positions which overlapped the products of the tested primer sets.

Diversity analysis. In order to conduct beta-diversity analysis using UNIFRAC, we reconstructed the phylogeny of representative sequences from each OTU. First, a representative set of sequences was picked using the QIIME command pick_rep_set.py, and these were aligned with parallel_align_seqs_pynast.py against the *nifH* reference alignment. Alignment columns which were all gaps were eliminated with the QIIME command filter_alignment.py, and a phylogenetic tree was made using the QIIME command make_phylogeny.py. Alpha- and beta-diversity metrics were calculated in QIIME. For creating a taxonomy bar plot, the TaxADivA script was used to assign taxonomy to OTUs and output an OTU table, which was then converted to hdf5 format and used as input with the QIIME script summarize_taxa_through_plots.py.

Accession number(s). All sequence data can be accessed from NCBI BioProject no. PRJNA418634.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/AEM.01512-17.

SUPPLEMENTAL FILE 1, PDF file, 0.7 MB.

SUPPLEMENTAL FILE 2, XLS file, 0.1 MB.

ACKNOWLEDGMENTS

This research was made possible in part by grants from Incauca S.A, The Gulf of Mexico Research Initiative, and the Terrestrial Ecosystem Science Program, under U.S. Department of Energy contracts DE-SC0007144 and DE-SC0012088.

We thank Neha Sarode for implementing the Fluidigm barcoding approach, the Georgia Institute of Technology's Department of Earth and Atmospheric Sciences and the laboratories of Frank Stewart and Konstantinos T. Konstantinidis for use of their Illumina MiSeq machine, and Jack Gilbert for helpful discussion on how to improve the pipeline.

REFERENCES

1. Rubio LM, Ludden PW. 2002. The gene products of the *nif* regulon, p 101–136. *In* Leigh GJ (ed), Nitrogen fixation at the millennium, 1st ed. Elsevier Science, Amsterdam, The Netherlands.
2. Gaby JC, Buckley DH. 2014. A comprehensive aligned *nifH* gene database: a multipurpose tool for studies of nitrogen-fixing bacteria. Database (Oxford) 2014:bau001. https://doi.org/10.1093/database/bau001.
3. Chien YT, Zinder SH. 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. https://doi.org/10.1128/jb.176.21.6590-6598.1994.
4. Zheng H, Dietrich C, Radek R, Brune A. 2016. *Endomicrobium proavitum*, the first isolate of *Endomicrobia* class. nov. (phylum *Elusimicrobia*)–an ultramicrobacterium with an unusual cell cycle that fixes nitrogen with a group IV nitrogenase. Environ Microbiol 18:191–204. https://doi.org/10.1111/1462-2920.12960.
5. Raymond J, Siefert JL, Staples CR, Blankenship RE. 2004. The natural history of nitrogen fixation. Mol Biol Evol 21:541–554. https://doi.org/10.1093/molbev/msh047.
6. Woese CR. 1987. Bacterial evolution. Microbiol Rev 51:221–271.
7. Bolhuis H, Severin I, Confurius-Guns V, Wollenzien UIA, Stal LJ. 2010. Horizontal transfer of the nitrogen fixation gene cluster in the cyanobacterium *Microcoleus chthonoplastes*. ISME J 4:121–130. https://doi.org/10.1038/ismej.2009.99.
8. Foster RA, Paytan A, Zehr JP. 2009. Seasonality of N_2 fixation and *nifH* gene diversity in the Gulf of Aqaba (Red Sea). Limnol Oceanogr 54:219–233. https://doi.org/10.4319/lo.2009.54.1.0219.
9. Yeager CM, Kornosky JL, Housman DC, Grote EE, Belnap J, Kuske CR. 2004. Diazotrophic community structure and function in two successional stages of biological soil crusts from the Colorado Plateau and Chihuahuan Desert. Appl Environ Microbiol 70:973–983. https://doi.org/10.1128/AEM.70.2.973-983.2004.
10. Roesch LFW, Camargo FAO, Bento FM, Triplett EW. 2008. Biodiversity of diazotrophic bacteria within the soil, root and stem of field-grown maize. Plant Soil 302:91–104. https://doi.org/10.1007/s11104-007-9458-3.
11. Mehta MP, Butterfield DA, Baross JA. 2003. Phylogenetic diversity of nitrogenase (*nifH*) genes in deep-sea and hydrothermal vent environments of the Juan de Fuca Ridge. Appl Environ Microbiol 69:960–970. https://doi.org/10.1128/AEM.69.2.960-970.2003.
12. Yamada A, Inoue T, Noda S, Hongoh Y, Ohkuma M. 2007. Evolutionary

trend of phylogenetic diversity of nitrogen fixation genes in the gut community of wood-feeding termites. Mol Ecol 16:3768–3777. https://doi.org/10.1111/j.1365-294X.2007.03326.x.
13. Ohkuma M, Noda S, Kudo T. 1999. Phylogenetic diversity of nitrogen fixation genes in the symbiotic microbial community in the gut of diverse termites. Appl Environ Microbiol 65:4926–4934.
14. Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T, Tripp JH, Affourtit JP. 2008. Globally distributed uncultivated oceanic N_2-fixing cyanobacteria lack oxygenic photosystem II. Science 322:1110–1112. https://doi.org/10.1126/science.1165340.
15. Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, Kuypers MMM, Zehr JP. 2012. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. Science 337:1546–1550. https://doi.org/10.1126/science.1222700.
16. Mirza BS, Potisap C, Nüsslein K, Bohannan BJM, Rodrigues JLM. 2014. Response of free-living nitrogen-fixing microorganisms to land use change in the amazon rainforest. Appl Environ Microbiol 80:281–288. https://doi.org/10.1128/AEM.02362-13.
17. Coelho MR, de Vos M, Carneiro NP, Marriel IE, Paiva E, Seldin L. 2008. Diversity of *nifH* gene pools in the rhizosphere of two cultivars of sorghum (*Sorghum bicolor*) treated with contrasting levels of nitrogen fertilizer. FEMS Microbiol Lett 279:15–22. https://doi.org/10.1111/j.1574-6968.2007.00975.x.
18. Hsu S-F, Buckley DH. 2009. Evidence for the functional significance of diazotroph community structure in soil. ISME J 3:124–136. https://doi.org/10.1038/ismej.2008.82.
19. Collavino MM, Tripp HJ, Frank IE, Vidoz ML, Calderoli PA, Donato M, Zehr JP, Aguilar OM. 2014. *nifH* pyrosequencing reveals the potential for location-specific soil chemistry to influence N_2-fixing community dynamics. Environ Microbiol 16:3211–3223. https://doi.org/10.1111/1462-2920.12423.
20. Farnelid H, Andersson AF, Bertilsson S, Al-Soud WA, Hansen LH, Sørensen S, Steward GF, Hagström Å, Riemann L. 2011. Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. PLoS One 6:e19223. https://doi.org/10.1371/journal.pone.0019223.
21. Illumina. 2015. MiSeq system. System specification sheet: sequencing. Publication no. 770-2011-001. Illumina, San Diego, CA. https://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf.
22. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341. https://doi.org/10.1186/1471-2164-13-341.
23. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91. https://doi.org/10.1099/ijs.0.64483-0.
24. Tikhonov M, Leach RW, Wingreen NS. 2015. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. ISME J 9:68–80. https://doi.org/10.1038/ismej.2014.117.
25. Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998. https://doi.org/10.1038/nmeth.2604.
26. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2:e593. https://doi.org/10.7717/peerj.593.
27. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584. https://doi.org/10.7717/peerj.2584.
28. Maidak BL, Cole JR, Lilburn TG, Parker CT, Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM. 2001. The RDP-II (Ribosomal Database Project). Nucleic Acids Res 29:173–174. https://doi.org/10.1093/nar/29.1.173.
29. Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620. https://doi.org/10.1093/bioinformatics/btt593.
30. Frank IE, Turk-Kubo KA, Zehr JP. 2016. Rapid annotation of *nifH* gene sequences using classification and regression trees facilitates environmental functional gene analysis. Environ Microbiol Rep 8:905–916. https://doi.org/10.1111/1758-2229.12455.
31. Bhatia S, Batra N, Pathak A, Green SJ, Joshi A, Chauhan A. 2015. Metagenomic evaluation of bacterial and archaeal diversity in the geothermal hot springs of Manikaran, India. Genome Announc 3:e01544–14. https://doi.org/10.1128/genomeA.01544-14.
32. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, Cole JR. 2013. FunGene: the functional gene pipeline and repository. Front Microbiol 4:291. https://doi.org/10.3389/fmicb.2013.00291.
33. Zhang B, Ryan Penton C, Xue C, Wang Q, Zheng T, Tiedje JM. 2015. Evaluation of the Ion Torrent Personal Genome Machine for gene-targeted studies using amplicons of the nitrogenase gene *nifH*. Appl Environ Microbiol 81:4536–4545. https://doi.org/10.1128/AEM.00111-15.
34. Yeoh YK, Paungfoo-Lonhienne C, Dennis PG, Robinson N, Ragan MA, Schmidt S, Hugenholtz P. 2016. The core root microbiome of sugarcanes cultivated under varying nitrogen fertilizer application. Environ Microbiol 18:1338–1351. https://doi.org/10.1111/1462-2920.12925.
35. Staples CR, Lahiri S, Raymond J, Von Herbulis L, Mukhophadhyay B, Blankenship RE. 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. https://doi.org/10.1128/JB.00876-07.
36. Souillard N, Magot M, Possot O, 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. https://doi.org/10.1007/BF02099731.
37. Nomata J, Mizoguchi T, Tamiaki H, Fujita Y. 2006. A second nitrogenase-like enzyme for bacteriochlorophyll biosynthesis: reconstitution of chlorophyllide a reductase with purified X-protein (BchX) and YZ-protein (BchY-BchZ) from *Rhodobacter capsulatus*. J Biol Chem 281:15021–15028. https://doi.org/10.1074/jbc.M601750200.
38. Suzuki JY, Bauer CE. 1992. Light-independent chlorophyll biosynthesis: involvement of the chloroplast gene *chlL* (*frxC*). Plant Cell 4:929–940.
39. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10:57–60. https://doi.org/10.1038/nmeth.2276.
40. Debenport SJ, Assigbetse K, Bayala R, Chapuis-Lardy L, Dick RP, McSpadden Gardener BB. 2015. Shifting populations in the root-zone microbiome of millet associated with enhanced crop productivity in the Sahel region (Africa). Appl Environ Microbiol 81:2841–2851. https://doi.org/10.1128/AEM.04122-14.
41. Rodriguez-R LM, Overholt WA, Hagan C, Huettel M, Kostka JE, Konstantinidis KT. 2015. Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. ISME J 9:1928–1940. https://doi.org/10.1038/ismej.2015.5.
42. Huettel M, Overholt WA, Kostka JE, Hagan C, Kaba J, Wells WB, Dudley S. 2018. Degradation of Deepwater Horizon oil buried in a Florida beach influenced by tidal pumping. Mar Pollut Bull 126:488–500. https://doi.org/10.1016/j.marpolbul.2017.10.061.
43. Kostka JE, Prakash O, Overholt WA, Green SJ, Freyer G, Canion A, Delgardio J, Norton N, Hazen TC, Huettel M. 2011. Hydrocarbon-degrading bacteria and the bacterial community response in Gulf of Mexico beach sands impacted by the Deepwater Horizon oil spill. Appl Environ Microbiol 77:7962–2974. https://doi.org/10.1128/AEM.05402-11.
44. Lin X, Tfaily MM, Green SJ, Steinweg JM, Chanton P, Imvittaya A, Chanton JP, Cooper W, Schadt C, Kostka JE. 2014. Microbial metabolic potential for carbon degradation and nutrient (nitrogen and phosphorus) acquisition in an ombrotrophic peatland. Appl Environ Microbiol 80:3531–3540. https://doi.org/10.1128/AEM.00206-14.
45. Tfaily MM, Cooper WT, Kostka JE, Chanton PR, Schadt CW, Hanson PJ, Iversen CM, Chanton JP. 2014. Organic matter transformation in the peat column at Marcell Experimental Forest: humification and vertical stratification. J Geophys Res Biogeosci 119:661–675. https://doi.org/10.1002/2013JG002492.
46. Wilson RM, Hopple AM, Tfaily MM, Sebestyen SD, Schadt CW, Pfeifer-Meister L, Medvedeff C, McFarlane KJ, Kostka JE, Kolton M, Kolka RK, Kluber LA, Keller JK, Guilderson TP, Griffiths NA, Chanton JP, Bridgham SD, Hanson PJ. 2016. Stability of peatland carbon to rising temperatures. Nat Commun 7:13723. https://doi.org/10.1038/ncomms13723.
47. Lin X, Tfaily MM, Steinweg JM, Chanton P, Esson K, Yang ZK, Chanton JP, Cooper W, Schadt CW, Kostka JE. 2014. Microbial community stratification linked to utilization of carbohydrates and phosphorus limitation in a boreal peatland at Marcell Experimental Forest, Minnesota, USA. Appl Environ Microbiol 80:3518–3530. https://doi.org/10.1128/AEM.00205-14.
48. Warren MJ, Lin X, Gaby JC, Kretz CB, Kolton M, Morton PL, Pett-Ridge J, Weston DJ, Schadt CW, Kostka JE, Glass JB. 2017. Molybdenum-based

diazotrophy in a *Sphagnum* peatland in northern Minnesota. Appl Environ Microbiol 83:e01174-17. https://doi.org/10.1128/AEM.01174-17.
49. Rouws LFM, Fischer D, Schmid M, Reis VM, Baldani JI, Hartmann A. 2015. Culture-independent assessment of diazotrophic bacteria in sugarcane and isolation of *Bradyrhizobium* spp. from field-grown sugarcane plants using legume trap plants, p 955–965. *In* de Bruijn FJ (ed), Biological Nitrogen Fixation. John Wiley & Sons, Inc., New York, NY.
50. de Souza RSC, Okura VK, Armanhi JSL, Jorrín B, Lozano N, da Silva MJ, González-Guerrero M, de Araújo LM, Verza NC, Bagheri HC, Imperial J, Arruda P. 2016. Unlocking the bacterial and fungal communities assemblages of sugarcane microbiome. Sci Rep 6:28774. https://doi.org/10.1038/srep28774.
51. Fischer D, Pfitzner B, Schmid M, Simões-Araújo JL, Reis VM, Pereira W, Ormeño-Orrillo E, Hai B, Hofmann A, Schloter M, Martinez-Romero E, Baldani JI, Hartmann A. 2011. Molecular characterisation of the diazotrophic bacterial community in uninoculated and inoculated field-grown sugarcane (Saccharum sp.). Plant Soil 356:83–99. https://doi.org/10.1007/s11104-011-0812-0.
52. Ando S, Goto M, Meunchang S, Thongra-ar P, Fujiwara T, Hayashi H, Yoneyama T. 2005. Detection of *nifH* sequences in sugarcane (*Saccharum officinarum* L.) and pineapple (*Ananas comosus* [L.] Merr.). Soil Sci Plant Nutr 51:303–308. https://doi.org/10.1111/j.1747-0765.2005.tb00034.x.
53. Thaweenut N, Hachisuka Y, Ando S, Yanagisawa S, Yoneyama T. 2011. Two seasons' study on *nifH* gene expression and nitrogen fixation by diazotrophic endophytes in sugarcane (*Saccharum* spp. hybrids): expression of *nifH* genes similar to those of rhizobia. Plant Soil 338:435–449. https://doi.org/10.1007/s11104-010-0557-1.
54. Burbano CS, Liu Y, Rösner KL, Reis VM, Caballero-Mellado J, Reinhold-Hurek B, Hurek T. 2011. Predominant *nifH* transcript phylotypes related to *Rhizobium rosettiformans* in field-grown sugarcane plants and in Norway spruce. Environ Microbiol Rep 3:383–389. https://doi.org/10.1111/j.1758-2229.2010.00238.x.
55. Carbonell González JA, Quintero Durán R, Torres Aguas JS, Osorio Murillo CA, Isaacs Echeverri CH, Victoria Kafure JI. 2011. Zonificación agroecológica para el cultivo de la caña de azúcar en el valle del río Cauca (cuarta aproximación). Principios metodológicos y aplicaciones. Serie técnica no. 38, 4th ed. Cenicaña, Cali, Valle del Cauca, Colombia.
56. Zehr JP, McReynolds LA. 1989. Use of degenerate oligonucleotides for amplification of the *nifH* gene from the marine cyanobacterium *Trichodesmium thiebautii*. Appl Environ Microbiol 55:2522–2526.
57. Ueda T, Suga Y, Yahiro N, 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. https://doi.org/10.1128/jb.177.5.1414-1417.1995.
58. Poly F, Monrozier LJ, 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. https://doi.org/10.1016/S0923-2508(00)01172-4.
59. Fedorov DN, Ivanova EG, Doronina NV, Trotsenko YA. 2008. A new system of degenerate oligonucleotide primers for detection and amplification of *nifHD* genes. Microbiology 77:247–249. https://doi.org/10.1134/S0026261708020215.
60. Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, Udall JA, Wilcox ER, Crandall KA. 2011. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. Genome Biol Evol 3:1312–1323. https://doi.org/10.1093/gbe/evr106.
61. Green SJ, Venkatramanan R, Naqib A. 2015. Deconstructing the polymerase chain reaction: understanding and correcting bias associated with primer degeneracies and primer-template mismatches. PLoS One 10:e0128122. https://doi.org/10.1371/journal.pone.0128122.
62. Herbold CW, Pelikan C, Kuzyk O, Hausmann B, Angel R, Berry D, Loy A. 2015. A flexible and economical barcoding approach for highly multiplexed amplicon sequencing of diverse target genes. Front Microbiol 6:731. https://doi.org/10.3389/fmicb.2015.00731.
63. Field D, Tiwari B, Booth T, Houten S, Swan D, Bertrand N, Thurston M. 2006. Open software for biologists: from famine to feast. Nat Biotechnol 24:801–803. https://doi.org/10.1038/nbt0706-801.
64. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
65. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koening JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336. https://doi.org/10.1038/nmeth.f.303.
66. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541. https://doi.org/10.1128/AEM.01541-09.
67. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.
68. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
69. Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12:385. https://doi.org/10.1186/1471-2105-12-385.
70. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. 2014. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 30:3123–3124. https://doi.org/10.1093/bioinformatics/btu494.
71. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2:16. https://doi.org/10.1186/2047-217X-2-16.
72. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. 2014. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J 9:968–979. https://doi.org/10.1038/ismej.2014.195.
73. Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. Third Int AAAI Conf Weblogs Soc Media, 17 to 20 May 2009, San Jose, CA.
74. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26:266–267. https://doi.org/10.1093/bioinformatics/btp636.
75. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüßmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer K-H. 2004. ARB: a software environment for sequence data. Nucleic Acids Res 32:1363–1371. https://doi.org/10.1093/nar/gkh293.