Metagenomic Characterization of Soil Microbial Communities in the Luquillo Experimental Forest (Puerto Rico) and Implications for Nitrogen Cycling


School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA; School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA; Center for Environmental Biotechnology, Department of Microbiology, Department of Civil and Environmental Engineering, Department of Biosystems Engineering and Soil Science, University of Tennessee, Knoxville, Tennessee, USA; Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA; Department of Biology, University of Puerto Rico at Humacao, Humacao, Puerto Rico, USA; USDA Forest Service, International Institute of Tropical Forestry, Río Piedras, Puerto Rico, USA

ABSTRACT

The phylogenetic and functional diversities of microbial communities in tropical rainforests and how these differ from those of temperate communities remain poorly described but are directly related to the increased fluxes of greenhouse gases such as nitrous oxide (N₂O) from the tropics. Toward closing these knowledge gaps, we analyzed replicated shotgun metagenomes representing distinct life zones and an elevation gradient from four locations in the Luquillo Experimental Forest (LEF), Puerto Rico. These soils had a distinct microbial community composition and lower species diversity compared to those of temperate grasslands or agricultural soils. In contrast to the overall distinct community composition, the relative abundances and nucleotide sequences of N₂O reductases (nosZ) were highly similar between tropical forest and temperate soils. However, respiratory NO reductase (norB) was 2-fold more abundant in the tropical soils, which might be relatable to their greater N₂O emissions. Nitrogen fixation (nifH) also showed higher relative abundance in rainforest than in temperate soils, i.e., 20% versus 0.1 to 0.3% of bacterial genomes in each soil type harbored the gene, respectively. Finally, unlike temperate soils, LEF soils showed little stratification with depth in the first 0 to 30 cm, with ~45% of community composition differences explained solely by location. Collectively, these results advance our understanding of spatial diversity and metabolic repertoire of tropical forest soil communities and should facilitate future ecological studies of these ecosystems.

IMPORTANCE

Tropical rainforests are the largest terrestrial sinks of atmospheric CO₂ and the largest natural source of N₂O emissions, two greenhouse gases that are critical for the climate. The microbial communities of rainforest soils that directly or indirectly, through affecting plant growth, contribute to these fluxes remain poorly described by cultured-independent methods. To close this knowledge gap, the present study applied shotgun metagenomics to samples selected from three distinct life zones within the Puerto Rico rainforest. The results advance our understanding of microbial community diversity in rainforest soils and should facilitate future studies of natural or manipulated perturbations of these critical ecosystems.

KEYWORDS

Puerto Rico, diversity, nitrous oxide, nosZ, soil microbial communities

Soil microbiomes are some of the most complex ecosystems owing to microenvironments and steep physicochemical gradients, which can change on a micrometer or millimeter scale (1–3). Tropical rainforests (“forests” hereafter) are characterized by...
humid and wet climate patterns and account for a large portion of the world’s total forest cover (4). These forests have high levels of primary productivity (~30% of the total global production) due to large amounts of precipitation coupled with year-long warm temperatures and high levels of light (5). Consequently, high levels of biodiversity are observed in these forests, above- and belowground. The soil microbial communities of these forests, in particular, harbor taxa and genes that are exclusive to these habitats/locations, along with only a few cosmopolitan taxa that are shared with other (non-tropical forest) habitats (6, 7). Although tropical forest soils are critical ecosystems that host a plethora of distinct ecological niches, little is known about the metabolic potential of tropical soils, especially across elevation and depth gradients. Describing this metabolic diversity is important for studying and monitoring the microbial activities related to greenhouse gas fluxes, namely, nitrous oxide (N₂O) and carbon dioxide (CO₂), from the tropical soils (8).

Notably, tropical forests represent the largest terrestrial sinks of atmospheric CO₂ and the largest natural source of N₂O emissions (9–12). Natural soils have been reported to contribute over 43% of the total global N₂O emissions, with tropical ecosystems being the highest contributors, having 2 to 4 times higher contributions than natural temperate ecosystems (13–16). These soils are also responsible for about 70% of terrestrial nitrogen fixation, which underlies, at least in part, their high rates of net primary productivity (8, 17).

Microbially mediated nitrification and denitrification are the biotic processes contributing the most to global soil N₂O emissions (60 to 70%) (16, 18, 19), although chemodenitrification, i.e., ferrous iron generated by ferric iron-reducing bacteria reacting with nitrite to produce N₂O abiotically, is also likely high in iron-rich tropical soils (20). In soils, N₂O is biologically produced as a result of DNRA (dissimilatory nitrite reduction to ammonium) or denitrification respiratory pathways (19, 21, 22). Respiratory nitric oxide reductase (nor) is a key contributor to the microbial production of N₂O and is commonly found in the genomes of denitrifying bacteria as well as in that of some ammonia-oxidizing organisms (19, 23–27).

While both biotic and abiotic processes contribute to N₂O production, consumption of N₂O is exclusively mediated by microbial N₂O reductase (NosZ) activity (22, 28, 29). Yet, whether the denitrifying microorganisms in the tropical forest soils differ from their counterparts in temperate soils and whether their functional genes present in the community reflect the high nitrogen fluxes remain unanswered questions despite their apparent importance for better management and modeling of tropical soil ecosystems. It has also been demonstrated that tropical forests have significantly higher rates of nitrogen fixation (~70% of total terrestrial nitrogen fixation) than other ecosystems, significantly affecting the nitrogen budgets in these ecosystems (3, 30–32). How these ecosystem rates translate to the nitrogen-fixing microbial (sub)community diversity and gene relative abundance remains unclear.

The Luquillo Experimental Forest (LEF), also known as the El Yunque National Forest in Puerto Rico (PR), has been a long-term ecological research (LTER) site since 1988. The site is dedicated to the assessment of the effects of climate drivers on the biota and biogeochemistry. The forest has been subjected to several disturbance regimes over the last few decades, mostly natural and, to a smaller extent, anthropogenic, such as tourism and experimental manipulations (33, 34). This site encompasses distinct “life zones” characterized by sharp environmental gradients even across small spatial scales (33, 35, 36). The broad life zones based on the Holdridge classification system include the rain forest, wet forest, lower montane wet forest, and lower montane rain forest. These life zones are distinguished by elevation and temperature and rainfall patterns, in addition to other edaphic factors (37–40). The elevation and rainfall patterns also tend to influence oxygen availability, redox potential, nutrient uptake, and organic decomposition rates (37, 40, 41). The dynamic interplay of existing physicochemical gradients and climatic factors gives rise to a complex mosaic of biodiversity patterns observed in this forest (38). Hence, LEF represents an ideal environment to study
tropical microbial community diversity patterns and their effect on carbon and nitrogen cycling. The four sampling sites of this study (see below) were chosen to represent the distinct vegetation and life zones within the LEF.

Previous studies in the LEF, and in similar forest regions, have mostly focused on the effects of redox dynamics, litter decomposition, nitrogen (N), and other nutrient fertilization on microbial community activity through enzyme assays and biochemical process rate measurements. Few studies have examined microbial diversity patterns across an elevation gradient, and those were based only on low-resolution techniques such as terminal restriction fragment length polymorphism analysis (4, 10, 11, 42–44). Furthermore, studies linking marker-gene abundances (related to nitrogen cycling) with in situ flux measurements have been met with mixed success in the forest soils, despite the high N$_2$O fluxes often measured in these ecosystems (45). A possible underlying reason for this mixed success could be that the commonly used nosZ primers target only the clade I (typical) nosZ sequences and miss the often numerically more dominant clade II (atypical) sequences (46), a bias that can be circumvented by employing metagenomic analyses (47).

With recent developments in next-generation DNA sequencing and associated bioinformatics genome binning algorithms, nearly complete metagenome-assembled genomes (MAGs) can be recovered without cultivation (48, 49), opening new windows into studying soil microbial communities. Here, shotgun metagenomes originating from soils from the four different locations/life zones and three different depths in the LEF were analyzed to describe the microbial community diversity, biogeographical patterns, and metabolic potential differences across samples. Furthermore, the metagenomic data obtained from these soils were compared to similar data from temperate grasslands in Oklahoma (OK) (50) and agricultural soils from Illinois (IL), USA (49) previously obtained by our team. By analyzing nearly complete MAGs, we show that the most abundant microbial populations (based on number of reads recruited) at each of the sampling locations represent sequence-discrete populations, similar to those observed in other habitats (45). Using such sequence-discrete populations as the fundamental unit of microbial communities, we subsequently assess the population distribution at high resolution across the sampling sites (biogeography) and the gene content they carry, with a focus on nitrogen metabolism.

RESULTS

Sampling sites. Soil samples were collected on February 2016 from four locations/sites across the LEF (18.3° N, 65.80° W). The four sites, namely, Sabana, El Verde field station, Palm Nido, and Pico del Este, each located at different elevations from the mean sea level, i.e., 265, 434, 634, and 953 m, respectively, were chosen due to their unique landscapes and rainfall patterns, thereby creating distinct ecological niches (Fig. 1A).

The El Yunque forest is categorized into four distinct vegetation zones, namely, the Tabonuco, Palo Colorado, Sierra Palm, and Dwarf/Elfín forests. Sites Sabana and El Verde, which are located at the lowest elevation among the four sites within the LEF, fall under the Tabonuco forest category in terms of vegetation, dominated by the tree species Dacryodes excelsa (native to Puerto Rico). They are characterized by canopy cover and low light intensities at the ground level which account for the sparsely vegetated forest floor. However, these sites still harbor the richest flora of all sites (51). Palm Nido is characterized by unstable, wetter soils, steeper slopes, and vegetation that is dominated by the Sierra Palm (Prestoea montana). The site at the highest elevation, Pico del Este (dwarf forest ecosystem or “elfín woodlands”) is characterized by higher winds, lower temperatures, and vegetation that is enveloped by clouds (34, 52), and its main vegetation is comprised of moss and epiphytes. Furthermore, highly acidic soil and continuously water-saturated soils deficient in oxygen are some major characteristics of this ecosystem, with most mineral inputs for plants being dissolved in the rain and fog.

Three adjacent soil profiles were taken from each of the four LEF sites (4 sites encompassing 3 life zones; Palo Colorado was not sampled). For each profile, individual
soil cores were taken at each depth (0 to 5 cm, 5 to 20 cm, 20 to 30 cm) using a 3-cm diameter by 15-cm length soil corer (AMS Inc, ID) that was decontaminated between samplings by washing with 70% ethanol. The three cores at each sampling depth were pooled for community DNA extraction, producing a total of 12 samples across the four sites.

FIG 1 Sampling location map and microbial community diversity among the study sites. (A) Map of the four sampling sites within the Luquillo Experimental Forest (LEF). (B) Principal coordinate analysis (PCoA) plots based on Mash distances, colored by sampling site. (C) Nonmetric multidimensional scaling (NMDS) plot with the soil physicochemical parameters incorporated. The arrow lengths are proportional to the strength of the correlations obtained between measured soil physicochemical parameters and each ordination axis. The coordinates for the Luquillo Experimental Forest (LEF) were obtained from DEIMS-SDR (Dynamic Ecological Information Management System, site and dataset registry: https://deims.org/bd0b5bcf-4f2e-4038-8275-629ffa5bf2aa) and visualized using the terrain feature in Google Maps. The map was refined and annotated in Adobe Illustrator.
Diversity of forest microbial communities. The LEF soil communities were compared to those of intensively studied ecosystems, namely, the Oklahoma temperate grassland (OK) (1, 50) and Illinois agricultural soils (IL) (49), which were previously characterized with similar shotgun metagenomics approaches. Shotgun metagenomic sequencing recovered a total of 370 million reads across the 4 sites (see Table S2 in the supplemental material). Nonpareil 2.0 (53) was used to estimate sequence coverage, i.e., what fraction of the total extracted community DNA was sequenced. Nonpareil analysis of community diversity (Fig. S1) showed that the agricultural Urbana (IL) site had the highest diversity of all the soils compared ($N_d$ diversity 24.02; note that $N_d$ values are given in log scale) and consequently the lowest sequence coverage at (only) 37.23%. El Verde and Pico del Este (20 to 30 cm) were the least diverse or most completely sequenced with 87.1% and 73.4% coverage, respectively (Nonpareil diversity of 19.6 and 20.6, respectively, or about 2 to 3 orders of magnitude less diverse). Overall, OK and IL soils appear to be more diverse than the PR soils by about 2 orders of magnitude, on average, with an average Nonpareil value of 22.75 ± 0.37. Nearly complete coverage for El Verde and Pico del Este (20 to 30 cm samples) would require $2.402 \times 10^9$ bp and $8.735 \times 10^9$ bp, respectively, and for the same level of coverage, the more complex communities in Urbana (IL) would require a substantially higher sequencing effort of $1.282 \times 10^{12}$ bp. The OK soils had an estimated sequencing depth of $2.063 \times 10^{11}$ bp.

Community composition variation across the forest sites based on 16S rRNA gene sequences. The number of total 16S rRNA gene-based OTUs (operational taxonomic units) observed in each metagenome and the Chao1 estimate of total OTUs reflected the degree of undersampling at each site (Fig. S1 and S2) and were also consistent with the Nonpareil coverage estimates (Fig. 1). When Puerto Rico (PR) tropical soils were compared with the agricultural and grassland soils from the United States at the phylum level, Proteobacteria, Acidobacteria, and Actinobacteria were the most abundant taxa across all ecosystems. However, in the forest soils, a few highly abundant OTUs dominated the entire soil community, whereas in the OK and IL soils, OTUs were more evenly distributed (Fig. S2), consistent with the Nonpareil diversity results. Only 1.28% of the total detected OTUs (out of a total 8,019, non-singleton OTUs) were shared among all PR samples, while 49.95% of OTUs were exclusive to a particular sampling site in PR, partly reflecting the undersampling of the extant diversity by sequencing. Only 0.37% of the OTUs (out of a total 13,760, non-singleton OTUs) were shared among all the sites across all 3 ecosystems, all of which were assignable to Alphaproteobacteria, Acidobacteria, Verrucomicrobia, and Actinobacteria.

It is important to note that the above-mentioned data sets were obtained based on different DNA extraction methods and at different time points and laboratories. To test for the possible effect of the DNA extraction method on the derived results, we applied four additional DNA extraction methods on a selected subset of our PR samples, including two manual phenol-chloroform-based methods that are often advantageous for iron-rich soils like those in tropical forests. Our results revealed comparable levels of diversity among the different methods tested, showing that the method used here for extraction performed comparably to if not better than non-kit-based (manual) extraction methods. For example, $N_d$ values for soils from the four sampling sites ranged between 19 and 22, and the MP Bio FastDNA SPIN kit method used for the samples reported herein provided among the highest diversity values, especially for mid- and high-altitude samples (Fig. S3). Hence, the diversity patterns reported here are presumably independent of the DNA extraction method used. We would also like to note that given the differences in the soil compositions across the different sampling regions in PR soils alone, a single/uniform extraction protocol would have some inherent biases (i.e., some soils had higher clay/iron content than others in the same forest).

Factors driving community diversity in the forest soils: multidimensional scaling analysis of beta diversity. The principal coordinate analysis (PCoA) plots, constructed based on the Mash distances among whole metagenomes, showed a clustering pattern that was primarily governed by site/location. Accordingly, site explained
45.22% of the total diversity (Fig. 1B). The nonmetric multidimensional scaling (NMDS) analysis of the data revealed only site, pH, and soil moisture to be statistically significant physicochemical parameters in explaining the observed community diversity (Fig. 1C; Table S3). Analysis of similarity (ANOSIM) values also indicated site to be a more important factor than depth, with $P$ values of 0.001 and 0.94, respectively. Based on the distance-based redundancy analysis (dbRDA), site was the most significant factor, even when the interplay between site and sampling depth was accounted for (Table S4). Table 1 shows the partitioning of the variance between the proportion that is explained by constrained axes (i.e., environmental variables measured) and the proportion explained by unconstrained axes (i.e., variance not explained by environmental variables measured). The total variance explained by all (measured) environmental variables was 80.2% (Table 1), which is remarkably high for a soil ecosystem (54).

**Major N-cycling pathways.** Genes encoding proteins involved in denitrification and nitrogen fixation were the most abundant nitrogen (N) cycling pathway genes detected at different sites. Overall, the forest soils harbored an $\sim$2- to 3-fold higher abundance of denitrification genes, i.e., *narG, nirK, and norB* (catalyzing the reduction of nitrate, nitrite, and nitric oxide, respectively), than the grassland and agricultural soils (Fig. 2A). For instance, the *norB* gene abundance was found to be at the highest abundance among the denitrification genes, with $\sim$37% (standard deviation [SD] 9.5%) of the genomes in the PR soils predicted to carry a *norB* gene, compared to $\sim$17% (SD 4%) and $\sim$14% (SD 1.3%) at IL and OK, respectively. Similarly, *narG* showed a 3-fold higher abundance in the PR soils than in IL and OK soils (Fig. 2B). While denitrification gene abundances appeared higher in the tropical soils, the relative abundance of *nosZ* gene (i.e., 11.6% [SD 3%] of the total genomes across the four locations in the LEF were predicted to carry *nosZ*) was similar to *nosZ* relative abundance in IL and OK soils, i.e., 11.75% (SD 5%) and 11.08% (SD 3%), respectively (not statistically significant at $P = 0.05$). Similar to those of *nosZ*, DNRA gene abundances (namely, *nrfA*) were comparable across all sites studied herein (9%, SD 1.9%).

**Predominant nosZ clades are shared among soil ecosystems.** Placing *nosZ*-carrying reads to a reference *nosZ* phylogenetic tree revealed that clade II *nosZ*, affiliated predominantly with *Opitutus, Anaeromyxobacter*, and other closely related genera, dominated the *nosZ* gene pool in the tropical forests (Fig. 3; Figs. S4 to S7). In contrast, a very small fraction of reads (<10% of total *nosZ* reads) were recruited to clade I *nosZ*. Members belonging to the clade II *nosZ* dominated the *nosZ* gene pool in OK and IL soils as well, with IL agricultural soils showing the greatest *nosZ* sequence diversity among the three regions. Notably, *Opitutus terrae*-affiliated sequences represented the most abundant subclade (*nosZ* OTUs/subclades were defined at the 95% nucleotide sequence identity level) in all regions. Furthermore, most of the *O. terrae*-affiliated reads in the forest soil data set appeared to be assigned to a single subclade, while their counterparts in the OK and IL soils appeared to be more evenly distributed among several closely related *nosZ* subclades, i.e., showing higher sequence diversity (Fig. 3; Fig. S4 to S7). *O. terrae* (strain DSM 11246/PB90-1) *nosZ* reads at $\geq$95% identity made up between 20% and 60% of the total *nosZ* reads recovered from the El Verde site and, together with the second most abundant subclade from *Anaeromyxobacter* sp., contributed over 30% of the total *nosZ* reads across all four PR locations (Fig. 3). Despite the significant taxonomic diversity observed in these soils (Fig. S2), the soils from PR shared several abundant *nosZ* gene sequences/subclades at $\geq$95% nucleotide identity.

<table>
<thead>
<tr>
<th>Axis type</th>
<th>Inertia</th>
<th>Proportion</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constrained</td>
<td>0.0876</td>
<td>0.8021</td>
<td>6</td>
</tr>
<tr>
<td>Unconstrained</td>
<td>0.02161</td>
<td>0.1978</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>0.1092</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

*Site, sampling depth, pH, total nitrogen, total carbon, and moisture data were considered in the analysis.
Abundance of N-cycling genes and their distribution across soil ecosystems. (A) Abundance of hallmark genes for denitrification, DNRA, and nitrogen fixation pathways, represented as genome equivalents (% of total bacterial genomes sampled that carry the gene) in the metagenomes studied (see figure key). (B) Frequency of genomes carrying the respective denitrifying gene across the three ecosystems studied. Genes denoted by the same letter are not statistically significantly different between ecosystems (analysis of variance [ANOVA] Tukey test). Statistical significance reported at $P < 0.05$. Note that nitrification genes were not detected in any of the Puerto Rico sites.

FIG 2 Abundance of N-cycling genes and their distribution across soil ecosystems. (A) Abundance of hallmark genes for denitrification, DNRA, and nitrogen fixation pathways, represented as genome equivalents (% of total bacterial genomes sampled that carry the gene) in the metagenomes studied (see figure key). (B) Frequency of genomes carrying the respective denitrifying gene across the three ecosystems studied. Genes denoted by the same letter are not statistically significantly different between ecosystems (analysis of variance [ANOVA] Tukey test). Statistical significance reported at $P < 0.05$. Note that nitrification genes were not detected in any of the Puerto Rico sites.
Phylogenetic diversity of nosZ-carrying metagenomic reads recovered in each soil ecosystem. nosZ sequences were identified by the ROCker pipeline and placed in a reference nosZ phylogeny as described in Materials and Methods. The radii (Continued on next page)
identity with soils in OK and IL (Fig. 3). Furthermore, in order to compare the predominant nosZ sequence variants across the samples shown here, a new phylogenetic reference tree was constructed based on almost-full-length sequences obtained from the assemblies/MAGs obtained from the metagenomes studied here (namely, PR, OK, IL). The short reads identified as nosZ from the PR soils were placed on this tree and show that the majority of these reads are recruited by the nosZ sequences obtained from these assemblies/MAGs, indicating that the nosZ sequences across the ecosystems studied here are similar (Fig. S8.)

**Nitrogen fixation potential.** The nitrogen fixation genes (mainly nifH) were present at a much lower abundance in the lower altitude forest samples. For instance, only 1 to 3% of all genomes in the lower altitude samples were predicted to carry nifH, compared to ~20% of the genomes in the higher-elevation samples (Pico del Este) (Fig. 2A), and almost none of the reads from IL and OK metagenomes appeared to carry nifH (~0.1%). Therefore, nitrogen fixation gene abundance patterns indicated a much stronger selection for nitrogen fixation in the tropical forest relative to that in temperate agricultural or natural prairie soils, especially at higher elevations. Furthermore, no ammonia-oxidizing genes (amoA) were detected in any of the soils except for Urbana soils (IL), which had a history of fertilizer (N) input.

**Recovery of metagenome-assembled genomes (MAGs) representative of each site.** In order to test the prevalence of taxa and genes across our sampling sites, the distribution of abundant MAGs recovered from each PR sampling site (assembly and MAG statistics provided in Table S6) was assessed across the sites using read-recruitment plots (55). Taxonomic assignment using the Microbial Genomes Atlas, or MiGA (56), revealed that the most abundant MAG was at site El Verde (lowest elevation), representing 4.39% of the total metagenome, and was affiliated with an unclassified Verrucomicrobia. The second most abundant (1.8% of total) was likely a member of the genus Candidatus Koribacter (Acidobacteria) followed by an unclassified member of Acidobacteria (1.45% of total). The Verrucomicrobiun MAG was found at an abundance of 1.03% of the total population at Sabana and at 0.07% and 0.03% in Palm Nido and Pico del Este (highest elevation), respectively. Uneven coverage across the length of the reference sequence and nucleotide sequence identities were observed in the recruitment of short reads from Palm Nido and Pico del Este as well as with all OK data sets, indicating that the related populations in the latter samples were divergent from the reference MAG (Fig. S10). Therefore, at least this abundant low-elevation verrucomicrobial population did not appear to be widespread in the other samples analyzed here (Fig. S10). Similarly, the other abundant MAGs from other sites in the forest soils were unique to the corresponding sites (elevation) from which they were recovered. Almost all MAGs used in the analyses were assignable to a novel family, if not a higher taxonomic rank, according to MiGA analysis (compared to 11,566 classified isolate genomes available in the NCBI prokaryotic genome database), underscoring the large unexplored diversity harbored by the PR tropical rainforest soils. The sequence diversity/complexity as well as sequencing depth limited large-scale recovery of high-quality MAGS.

**Functional gene content of the MAGs.** The genome sequences of the most abundant MAGs from each location (n = 6) were analyzed in more detail to assess the functions they encoded, especially with respect to N-cycling pathways (Fig. 4). MAGs from Pico del Este (highest elevation) showed a high abundance of N metabolism related genes compared to MAGs from other sites (Fig. 4). Most notably, genes related to nitrogen fixation were found only in the Pico del Este MAG, which was consistent
with the short read analysis data sets showing greater relative abundance of *nifH* at this site. Nitrification (ammonia oxidation related genes) gene clusters were not detected in any of the recovered MAGs. *norB* and *nosZ* genes were found in three out of the six abundant MAGs analyzed. The most abundant El Verde MAG, most closely related to *O. terrae* (average amino acid identity [AAI] = 40%), possessed a *nosZ* gene, which was congruent with the *nosZ* phylogeny described above (i.e., ~60% of the *nosZ*-encoded reads from El Verde had a closest match to *O. terrae* *nosZ* sequences).
DISCUSSION

The present study reported the taxonomic and gene content diversities of poorly characterized tropical rainforest soils by using whole-community, shotgun metagenomic sequencing of samples from the Luquillo Experimental Forest (LEF), Puerto Rico. The recovered nearly complete MAGs represented several abundant and widespread organisms within this ecosystem that could serve as model organisms for future studies. Furthermore, since the LEF is subjected to various natural and experimental (e.g., warming, phosphorus fertilization) perturbations, our study could also provide a baseline for these perturbations and future soil microbial studies at LEF. Our results revealed that the LEF soils harbor distinct microbial communities at sites with elevation distinct from sea level. In contrast, and unlike several other soil ecosystems, sampling depth did not have a substantial impact on structuring community diversity (Fig. 1B and C), revealing no depth stratification in the LEF soils, at least for the depths sampled here (5 to 30 cm). This is most likely due to the lack of distinct soil horizons within the first 30 cm of the sampling sites and indicates that the soil formation and/or physicochemical properties in these ecosystems could differ markedly from those in their temperate counterparts (37).

A recent study examining the dominant bacterial (16S rRNA gene-based) phylotypes across the globe found that the predominant phylotypes were widespread across ecosystems. The only exception to this pattern was the forest tropical soils which harbor distinct phylotypes (7). Consistent with these conclusions, the majority of MAGs recovered from each LEF site appeared to be site-specific (e.g., see Fig. S10 in the supplemental material) and represented at least novel species and genera, further underlining the undertapped microbial diversity harbored by tropical forest soils. Currently, the environmental factors driving these diversity patterns remain poorly understood for tropical forest soils (7), but our study provided several new insights into this issue.

In particular, sites El Verde and Sabana (lowest elevation sites) had community structure and diversity similar to those of the two higher-elevation sampling sites, with certain MAGs being present at both sites but not in any of the other (higher-elevation) sites examined. This is presumably attributable to both sites having similar climate and vegetation patterns (i.e., Tabonuco forest). On the other hand, Pico del Este was the highest-elevation site and experiences almost continuous cloud cover as well as horizontal precipitation. The unique topology of Pico del Este was reflected in distinct and deeply novel MAGs and gene content, which differed markedly from those of the other three sampling sites within the LEF (PCoA plots, Fig. 2B). The high water content of the Pico del Este soils gives rise to a unique ecosystem dominated by epiphytes (e.g., moss) among other plant species (57). The epiphytic community has presumably significant effects on nutrient (e.g., nitrogen) cycling (58) and influences the water input to the soil, thereby shaping a unique habitat/niche for the soil microbes. Free-living microbes have been shown to be one of the highest contributors to biological N fixation in these forests, with high rates of nitrogenase activity associated with the presence of moss/epiphytes (44, 59). Furthermore, lower N-cycling rates and high water content in soils of cloud-immersed forests have been shown to be responsible for the overall low nitrogen levels in these (nitrogen-limited) ecosystems (60).

Consistent with these previous results and interpretations, the Pico del Este showed a high potential for nitrogen fixation, i.e., it was estimated that 1/5 of the total bacterial genomes sampled possessed genes for N fixation, which is at least 10 times greater nitrogenase gene/DNA abundance than that of any other site evaluated herein. Accordingly, we found that site (location) alone explained about half (45%) of the beta diversity differences observed among the four sampling sites, which reached ~80% when a few physicochemical parameters, namely, pH and moisture, were also included in the analyses (Fig. 2B; Table 1). This is a notably high fraction of beta diversity explained by measured parameters for a soil ecosystem (54) and likely reflected that location and the physical properties that characterized different locations within LEF...
structured diversity are much stronger than those in other soil ecosystems. Tropical forests have also been shown to have significantly higher rates of nitrogen fixation than other ecosystems, which can exceed the N retention capacity of the soil resulting in large N loss as N₂O (61). The findings reported here on denitrification gene abundances were generally consistent with these previous observations as well.

Links between soil community structure and nitrogen cycling can help close the knowledge gaps on how the forest ecosystems affect the release and mitigation of certain highly potent greenhouse gases such as N₂O. The gene abundances observed here, e.g., more than 2-fold higher abundance of norB (associated with NO reduction to N₂O) and similar nosZ (N₂O consumption) abundances in tropical soils relative to those in temperate soils, were consistent with higher N₂O emissions observed previously from the tropics. Further, in acidic soils such as the tropical forest soils evaluated in this study, with the possible exception of the cloud-immersed site, lack of N limitation can suppress complete denitrification, thereby leading to higher N₂O release compared to that of other soil ecosystems (30). These interpretations were consistent with our observation that the PR soils harbored a relatively high abundance of respiratory (related to denitrification) norB genes as well. Previous studies have also suggested that many denitrifying bacterial genomes possess the genes required to reduce nitrate to nitrous oxide but do not possess the gene responsible for the last step, i.e., N₂O reduction to N₂, leading to the release of N₂O gas (62), consistent with the findings of our study. Alternatively, the higher relative abundance of norB relative to that of nosZ or other denitrification pathway genes may be associated, at least partly, with detoxification of NO and/or dismutation of NO to N₂ and O₂ and not energy-yielding NO reduction to N₂O (reviewed in reference 62). However, this explanation is less likely, because we specifically focused on the respiratory reductase (cnorB), not its detoxifying homolog (qnorB), in our analysis (see Materials and Methods for further details).

It has been established that tropical forest soils are the single highest contributor of natural N₂O emissions. While several abiotic and microbial processes can contribute to soil N₂O, N₂O consumption is an exclusively microbial process, catalyzed by the enzyme product of the nosZ genes (29). Based on the assessment of the nosZ gene phylogeny, it appears that almost all of the nosZ genes from the tropical forest soils studied here belong to a previously overlooked clade II nosZ (29, 63). This clade consists mainly of nondenitrifying N₂O reducers (i.e., organisms possessing N₂O reductases but lacking all or some of the other reductases in the denitrification pathway). Despite the unique phylogenetic species- and gene-level diversity harbored by tropical soils in general, the nosZ gene sequence diversity appears to be shared between temperate and agricultural soils (Fig. 4). These findings imply strong selection pressure for conservation of nitrous oxide reductase sequences across tropical and temperate soil ecosystems that may not necessarily be applicable to other N-cycling genes and pathways, which warrants further attention in the future.

It would also be interesting to assess how the findings reported here for the LEF apply (or not) to other tropical forests especially because our study is based on a relatively small sample size. Despite the relatively small sample size, however, our results showed statistically significant differences along the elevation gradient sampled at the LEF that are independent of DNA extraction (Fig. S3) or sequencing methods and consistent with our metadata (Fig. 2) and previous process rate measurements. While the diversity in the Puerto Rico soils appears to be lower than that in temperate grassland and agricultural soils and different DNA extraction methods, including phenol-chloroform- and kit-based, provided similar results with our LEF samples (Fig. S3), it is important to note that the DNA of the temperate soil data sets used in our comparisons was extracted using different methods by the original studies (e.g., OK soils were extracted using the PowerSoil kit). Therefore, it would be important to confirm the preliminary findings in terms of α-diversity reported by our study by using the exact same DNA extraction and sequencing procedures in all soils.

Integration of functional (e.g., gene expression) data with in situ rate measurements will provide a more complete picture of the composition and functioning in tropical
forest soils. The identification of certain biomarker genes such as nosZ sequences in our study could facilitate future investigations on biogeochemical N-cycling and greenhouse gas emissions. For instance, the assembled MAGs and gene sequences provided here could be useful for the design of specific PCR assays for assessing transcript levels (activity), allowing potential linking of carbon dioxide, methane, nitrogen, and soil organic matter (SOM) turnover to the activity of individual populations and/or their genes. As the gradients at the LEF also provide a natural setting to interpret the potential ramifications of climate change scenarios such as altered precipitation patterns, the DNA sequences provided here could facilitate future manipulation experiments with an emphasis on understanding and predicting the effects of climate change on microbial community dynamics along the elevation gradient.

MATERIALS AND METHODS

Physiochemical analysis of soil samples. Soil samples were stored in sterile Whirl-Pak bags and kept on ice during transport and until storage at −80°C. Soil pH was determined using an automated LabFit AS-3000 pH analyzer, and soil extractable P, K, Ca, Mg, Mn, and Zn were extracted using the Mehlich-1 method and measured using an inductively coupled plasma spectrophotograph at the University of Georgia Agricultural and Environmental Services Laboratories (Athens, GA, USA). Soil extractable P using this method is interpreted as the bioavailable fraction of P. NH₄⁺-N and NO₃⁻-N were measured by first extracting them from soil samples with 0.1 N KCl, followed by the colorimetric phenate method for NH₄⁺ and the cadmium reduction method for NO₃⁻. The physicochemical conditions at the sites during the time of sampling are provided in Table S1.

Community DNA extraction and sequencing. Total DNA from soil was extracted using the FastDNA SPIN kit (MP Biomedicals, Solon, OH) following the manufacturer’s procedure with the following modifications (64). Soils were air dried under aseptic conditions followed by grinding with a mortar and pestle. Cells were lysed by bead beating and DNA was eluted in 50 μl of sterile H₂O. DNA sequencing libraries were prepared using the Illumina Nextera XT DNA library prep kit according to the manufacturer’s instructions, except the protocol was terminated after isolation of cleaned double-stranded libraries. Library concentrations were determined by fluorescent quantification using a Qubit HS DNA kit and Qubit 2.0 fluorometer (ThermoFisher Scientific), and samples were run on a high sensitivity DNA chip using the Bioanalyzer 2100 instrument (Agilent) to determine library insert sizes. An equimolar pool of the sequencing libraries was sequenced on an Illumina HiSeq 2500 instrument (located in the School of Biological Sciences, Georgia Institute of Technology) using the HiSeq Rapid PE cluster kit v2 and HiSeq Rapid SBS kit v2 (Illumina) for 300 cycles (2 by 150 bp paired end). Adapter trimming and demultiplexing of sequenced samples was carried out by the HiSeq instrument. In total, 12 metagenomic data sets were generated (3 per site for the three depths), and statistic details on each data set are provided in Table S2.

In order to test for any DNA extraction biases of the kit used above, especially for the high iron/clay content that characterizes tropical forest soils and is known to affect the extraction step, four additional DNA extraction methods were performed in parallel on a small subset of samples collected in 2018 from the same sites (6 samples per extraction method for 5 extraction methods covering the 4 sites). The methods included two manual (as opposed to kit-based) phenol-chloroform-based methods (65, 66) as well as two other kit-based methods, namely, DNeasy PowerSoil and DNeasy PowerSoil Pro (Qiagen Inc.). For this evaluation, the soils were first homogenized and subsequently divided into five subsamples to use with each method (including the FastDNA SPIN kit-based method mentioned above). The libraries were constructed and sequenced the same way as described above for the FastDNA SPIN kit method.

Bioinformatics analysis of metagenomic reads and MAGs. The paired end reads were trimmed and quality checked using the SolexaQA (67) package with a cutoff of Phred Q value of ≥20 (>99% accuracy per base-position) and a minimum trimmed length of 50 bp.

Assembly and population genome binning. Coassembly of the short reads from the same location was performed using IDBA-UD (68), and only resulting contigs longer than 500 bp in length were used for downstream analysis (e.g., functional annotation and MyTaxa classification). Genes were predicted on the coassembled contigs using MetaGeneMark (69) and the predicted protein-coding regions were searched against the NCBI All Genome database using BLASTP (70). Since the assembly of individual data sets resulted mostly in short contigs (data not shown), the contigs from the coassembly (combining metagenomes from the three sampling depths, for each site) were used for population genome binning. Contigs longer than 1 kb were binned using MaxBin (71) to recover individual MAGs (default settings). The resulting bins were quality checked for contamination and completeness using CheckM (72) and were further evaluated for their intrapopulation diversity and sequence discreteness using fragment recruitment analysis scripts as part of the Enveomics collection (55) essentially as previously described (73).

Functional annotation of MAGs. Genes were predicted for each MAG using MetaGeneMark and the predicted protein-coding regions were searched against the curated Swiss-Prot (74) protein database using BLASTP (70). Matches with a bitscore higher than 60 or amino acid identities higher than 40% were used in subsequent analysis. The Swiss-Prot database identifiers were mapped to their corresponding metabolic function based on the hierarchical classification subsystems of the SEED subsystem category (level 1) (75). The relative abundance of genes mapping to each function was calculated based on the number of predicted genes from each MAG assigned to the function (for read-based assessment,
see below). Relative abundance data were plotted in R using the “superheat” package (https://arxiv.org/abs/1512.01524). Individual biomarker genes for each step of the nitrogen-cycling pathway were manually verified by visually checking the alignment of the identified sequences by the pipeline outlined above against verified reference sequences.

Functional annotation of short reads. Protein-coding sequences present in short reads were predicted using FragGeneScan (76) using the 1% Illumina error model. The predicted genes were then searched against the Swiss-Prot database using BLASTP (best match). Low-quality matches (bitscore < 60) were excluded, and relative abundance of genes mapping to each function was determined as described in the previous section.

Community diversity estimation. (i) Nonpareil. Nonpareil (53) was used to estimate sequence coverage, i.e., what fraction of the total extracted community DNA was sequenced, and predict the sequencing effort required to achieve “nearly complete coverage” (≥95%). The default parameters in Nonpareil were used for all data sets. Only one of the two paired reads (forward) for each data set was used to avoid duplication of the paired reads, which can bias Nonpareil estimates (53).

(ii) Mash and multidimensional scaling. Mash, a tool employing the MinHash dimensionality reduction technique to compare sample-to-sample sequence composition based on k-mers (77), was used to compute pairwise distances between whole metagenomic data sets and construct the distance matrix to be used in multidimensional scaling. Pairwise Mash distances between the metagenomic data sets were computed from the size-reduced sketches (default parameters). PCoA and NMDS were employed to visualize the distance matrix and evaluate the physicochemical parameters driving community diversity, respectively. Furthermore, dbRDA (distance-based redundancy analysis) was used to obtain a finer resolution on the observed compositional variation. All of the above statistical analyses were performed using the vegan package in R (78), with default settings.

16S rRNA gene fragments recovered from shotgun metagenomes. 16S ribosomal rRNA (16S) gene fragments were extracted from the metagenomic data sets using Parallel-META (79). 16S-carrying reads were classified taxonomically using the GreenGenes database.

Recovered 16S fragments were clustered (“closed-reference OTU picking” strategy using UCLUST (80)) and taxonomically classified based on their best match in the GreenGenes database (81) at an ID of ≥97% in QIME (82, 83). The relative abundances of the OTUs were calculated based on the number of reads assigned to each OTU. Community composition was assessed based on OTU taxonomic assignments at the genus and the phylum ranks and was compared between the sites based on the relative abundance of OTUs at each site.

Identification of N-cycling genes using ROCker. ROCker (84) was employed for a precise identification and quantification of nosZ (encoding nitrous oxide reductase), norB (encoding respiratory nitric oxide reductase), nirK (encoding nitrite reductase), nrfA (encoding nitrite reductase), nrfB (encoding respiratory nitric oxide reductase), and nirS (encoding nitrous oxide reductase) genes. The NosZ reference protein sequences were aligned using CLUSTAL Omega (85) and a maximum likelihood reference tree was created using RAxML v8.0.19 (86) with a general time reversible model option, gamma parameter optimization, and “-f a” for the algorithm. The ROCker identified NosZ-encoding reads were extracted from all data sets, translated into protein sequences using FragGeneScan, and then added to the reference alignment using Mafft (87). The reads were placed in the phylogenetic tree using RAxML EPA algorithm and visualized using iTOL (88).

Intrapopulation diversity assessment based on recovered MAGs. The taxonomic affiliation of individual contig sequences of the MAG was evaluated based on MyTaxa, a homology-based classification tool (89). The MiGA (Microbial Genomes Atlas, www.microbial-genomes.org) webserver was used for the taxonomic classification of the whole MAG using the average nucleotide identity/stand amino acid identity (ANI/AI) concept. To assess intrapopulation diversity and sequence discreteness, each target population MAG was searched against all the reads from each location by BLASTN (only contigs longer than 2 kbp were used). Fragment recruitment plots were constructed based on the BLASTN matches (threshold values: nucleotide identity of ≥75% and alignment length of ≥80 bp) using the Enveomics collection of scripts (55). The evenness of coverage and sequence diversity of the reads across the length of the reference genome sequence were used to evaluate the presence and discreteness of the population in the chosen data set.

Data availability. All metagenomic data sets were deposited in the European Nucleotide Archive (ENA) under project PRJEB26500. Additional data are available at http://enve-omics.ce.gatech.edu/data/prsolls.

SUPPLEMENTAL MATERIAL
Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 1.8 MB.
ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomic Science Program (award DE-SC0006662) and US National Science Foundation (award 1831582). GG was supported by the Luquillo Critical Zone Observatory (National Science Foundation grant EAR-1331841) and the Luquillo Long-Term Ecological Research Site (National Science Foundation grant DEB-1239764).

All research at the USDA Forest Service International Institute of Tropical Forestry (IITF) is done in collaboration with the University of Puerto Rico. We thank Maria Rivera and Humberto Robles from IITF for their help in soil sampling.

We declare no conflict of interest.

REFERENCES


