Global biogeography of microbial nitrogen-cycling traits in soil

Michaeline B. Nelson[a], Adam C. Martiny[a,b], and Jennifer B. H. Martiny[a,1]

[a]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697; and [b]Department of Earth System Science, University of California, Irvine, CA 92697

Edited by Francisco J. Ayala, University of California, Irvine, CA, and approved April 21, 2016 (received for review February 12, 2016)

Microorganisms drive much of the Earth's nitrogen (N) cycle, but we still lack a global overview of the abundance and composition of the microorganisms carrying out soil N processes. To address this gap, we characterized the biogeography of microbial N traits, defined as eight N-cycling pathways, using publically available soil metagenomes. The relative frequency of N pathways varied consistently across soils, such that the frequencies of the individual N pathways were positively correlated across the soil samples. Habitat type, soil carbon, and soil N largely explained the total N pathway frequency in a sample. In contrast, we could not identify major drivers of the taxonomic composition of the N functional groups. Further, the dominant genera encoding a pathway were generally similar among habitat types. The soil samples also revealed an unexpectedly high frequency of bacteria carrying the pathways required for dissimilatory nitrate reduction to ammonium, a little-studied N process in soil. Finally, phylogenetic analysis showed that some microbial groups seem to be N-cycling specialists or generalists. For instance, taxa within the Deltaproteobacteria encoded all eight N pathways, whereas those within the Cyanobacteria primarily encoded three pathways. Overall, this trait-based approach provides a baseline for investigating the relationship between microbial diversity and N cycling across global soils.

nitrification | nitrogen fixation | ammonia assimilation | metagenomics | dissimilatory nitrite reduction

A grand challenge for this century is to predict how environmental change will alter global biogeochemical cycles. The field of biogeography has an important role to play in this effort (1). Environmental change is altering the distribution of biodiversity, which in turn is a key driver of biogeochemical processes (2, 3). Historically, biogeography has viewed biodiversity through a taxonomic lens, primarily resolving species distributions. However, a focus on traits—particularly those involved in ecosystem processes—may offer a clearer link between biodiversity patterns and biogeochemistry (4–6).

These ideas are particularly relevant for microorganisms. Microbes catalyze most of the biological transformations of the major elements of life (7), and because of their sheer abundance they account for a large pool of elements in living matter (8). Furthermore, like plants and animals, microbial taxonomic composition varies over space (9, 10), and this variation can influence ecosystem processes (11–14). Thus, a consideration of microbial traits should improve efforts to connect biogeographic patterns and ecosystem processes (15).

Here, we provide a first characterization of the global biogeographic patterns of microbial nitrogen (N) cycling traits in soil. Microbially driven transformations regulate biologically available N through exchange with the atmosphere (via N fixation and denitrification) and loss by nitrate leaching. They also influence the forms of N available for plant uptake. At the same time, human activities have altered, and continue to alter, the N cycle by increasing the amount of reactive N in the biosphere (16, 17). At local scales, N addition consistently shifts microbial composition in soils and other ecosystems (18, 19). The distribution of microbial traits might therefore be relevant for understanding current and future N cycling.

The taxonomic composition of soil microorganisms is correlated with spatial variation in climate, plant diversity, pH, disturbance, and many other factors (20–23). These biogeographic patterns help to identify factors that select on the entire suite of microbial traits. In this study, we reverse this direction of inquiry. We first characterize the patterns and drivers of just handful of traits associated with N cycling and then ask which taxa comprise these functional groups.

To quantify the abundance and composition of N-cycling traits, we analyzed ∼2.4 billion short-read sequences from 365 soil metagenomes sampled from around the globe. From this dataset, we identified sequences that indicate the potential for a microorganism to perform one of eight N pathways that convert inorganic N to other inorganic forms or microbial biomass. We then quantified the frequency and taxonomic association of microorganisms carrying these pathways in each sample. If a gene from a pathway was detected, we assumed the presence of the entire pathway in the organism. To compare the frequencies among the N pathways, we standardized for the number of genes (2–20) in each pathway. Although metagenomic sequences provide a measure of a community's trait diversity (24), the presence of a trait does not indicate how it is being used in the community. Thus, we cannot determine whether genes in the N pathways are expressed or the rate at which N is being transformed. However, assaying traits based on metagenomic sequences are parallel to other trait metrics used to describe an organism's functional potential, such as nutrient uptake affinity or temperature optimum for growth.

The global N trait dataset allowed us to address four main questions. First, what are the overall frequencies of the different N pathways in soil? We expected the frequencies to vary greatly by pathway. Indeed, the ability to perform nitrification is restricted to few microbial taxa, whereas ammonia assimilation is probably present in almost all taxa. Second, what drives variation in the frequencies of N pathways among soil samples? We hypothesized that N pathway frequencies would vary primarily by habitat type, which reflects major differences in plant communities and therefore N inputs into soils. Third, what are the main taxa encoding each N pathway? Surprisingly little is known about the dominant lineages encoding N-cycling traits across global soils. We therefore expected to find previously unrecognized, prominent players, particularly for the less-studied pathways such as dissimilatory nitrate to ammonium (DNRA). Finally, what underlies compositional variation among soil samples in microorganisms encoding N pathways? We hypothesized that the taxa responsible for each pathway would vary greatly by habitat type, because the habitat would select for specialized taxa. We

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "In the Light of Evolution X: Comparative Phylogeography," held January 8–9, 2016, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/ILE_X_Comparative_Phylogeography.

Author contributions: M.B.N., A.C.M., and J.B.H.M. designed research; M.B.N. performed research; A.C.M. contributed new reagents/analytic tools; M.B.N. and J.B.H.M. analyzed data; and M.B.N., A.C.M., and J.B.H.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: jmartiny@uci.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1601070113/-/DCSupplemental.

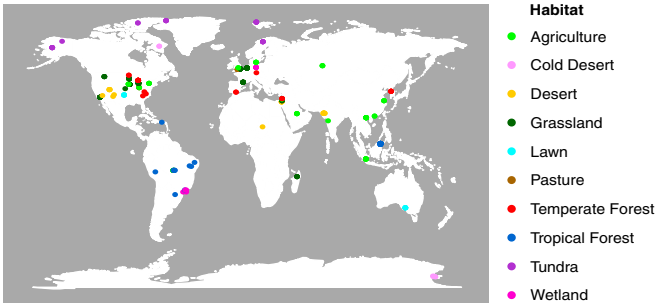

Fig. 1. The locations (n = 118) sampled to create the soil metagenomic libraries (n = 365) used in this analysis. The samples represent 10 distinct habitats including agriculture (n = 19), cold desert (n = 6), desert (n = 15), grassland (n = 14), lawn (n = 4), pasture (n = 2), temperate forest (n = 12), tropical forest (n = 34), tundra (n = 7), and wetland (n = 5).

further predicted that soil pH—previously identified as an important driver of soil composition (25, 26)—would also influence compositional variation within microorganisms encoding N-cycling traits.

Results

Metagenomic data from surface soil samples were retrieved from the metagenomics analysis server (MG-RAST) (27). After curating the samples for sequence and metadata quality, the final 365 samples represented 118 unique locations from 10 distinct habitat types covering natural and human-dominated systems (Fig. 1 and Dataset S1). Sequencing depth varied greatly among the samples but was not overtly biased toward any particular habitat type (Fig. S1). To standardize for sequencing depth, we report the abundance of each N pathway as its frequency in a sample. The trends observed were similar whether pathway frequency was normalized as the number detected per annotated sequence or per marker gene (based on 30 conserved, single-copy genes) (Fig. S1).

Bacteria dominated the metagenomic libraries, comprising 95% of all sequences, followed by 3% for Fungi and only 2% for Archaea. The fraction of fungal sequences in metagenomic libraries is known to be lower than their contribution to soil microbial biomass (10). We therefore concentrate our analyses on Bacteria and Archaea and report only general trends for Fungi. For instance, the proportion of total sequences of Bacteria, Archaea, and Fungi varied across habitat type (G-test of independence; $P << 0.001$) (Fig. S2). Archaea ranged from 0.9 to 11% of all sequences by habitat, with the highest percentage detected in deserts. The ratio of fungal to bacterial sequences was particularly high in temperate forest soil, as previously observed (28).

Frequency of Soil N Pathways. On average, 0.5% of all annotated sequences in a soil sample were associated with one of the eight N pathways (Fig. 2*A*), or an average of 3.3 and 4.7 N pathways per marker gene for Bacteria and Archaea, respectively. The frequency of the individual pathways varied by several orders of magnitude (one-way ANOVA $P < 0.001$; $F = 74.21$, df = 7) (Fig. 2*B*). Bacteria and Archaea displayed similar trends in their relative frequency of N pathways except for the absence of the dissimilatory nitrite reduction to ammonium pathway in Archaea. Fungal sequences were only associated with assimilatory pathways, including ammonia assimilation, assimilatory nitrate to nitrite, and assimilatory nitrite to ammonium.

Across all domains, the most common pathway was ammonia assimilation (Fig. 2*B*). For instance, among the Bacteria, an average of 280 ammonia assimilation pathways were detected for every million annotated bacterial sequences. In comparison, nitrification and N fixation were the least common pathways and detected only 6.1 and 4.6 times per million sequences, respectively. Notably, the relatively unstudied dissimilatory nitrite reduction to ammonium pathway was slightly more common that these two pathways, detected on average 9.3 times per million sequences.

Across all soil samples, N pathway frequencies were overwhelmingly positively correlated for both the Bacteria and Archaea (Fig. 3 *A* and *B*). To examine differences in pathways

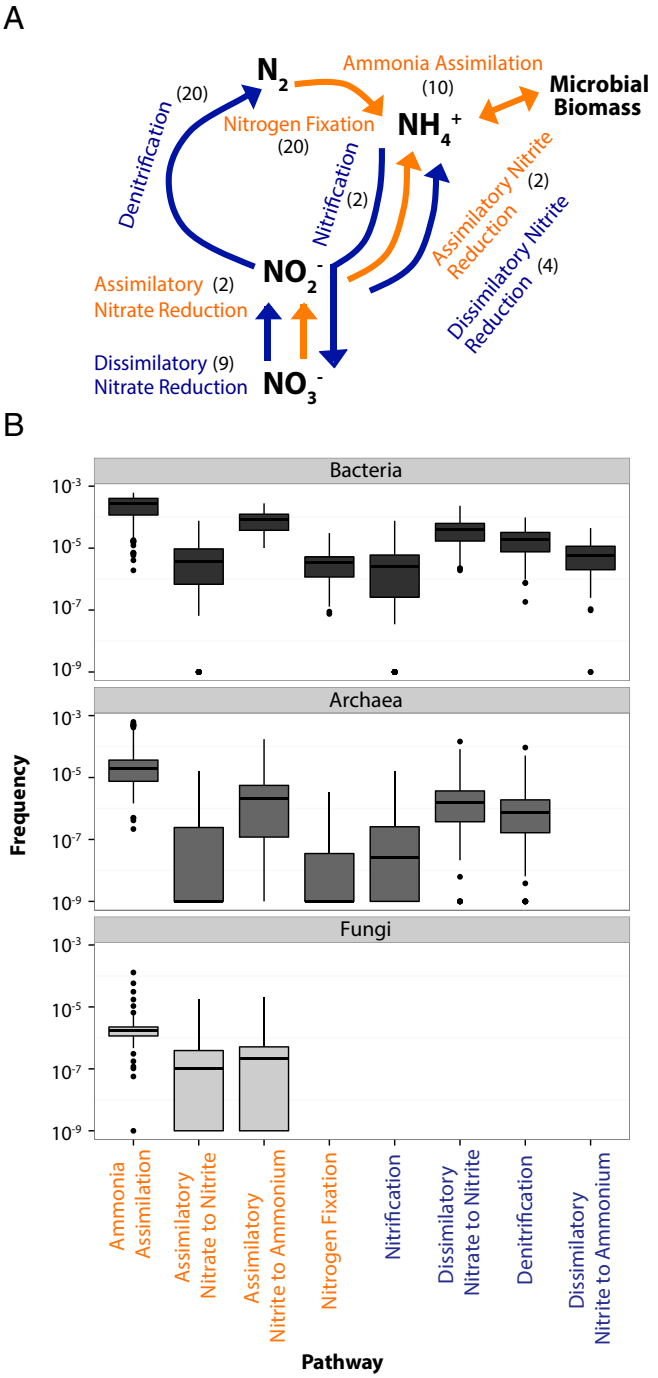

Fig. 2. N pathways and their frequencies. (*A*) N pathways considered in this study. The numbers in parentheses are the number of genes targeted for each pathway. Assimilatory pathways are in orange and dissimilatory pathways in blue. (*B*) Box plot of the frequency of each N pathway in a metagenomic library for Bacteria, Archaea, and Fungi. To compare across domains, frequencies are calculated as per annotated sequence in each domain. The upper and lower bounds of boxes correspond to the 25th and 75th percentiles, with a median line shown. Whiskers represent 1.5*IQR (interquartile range). Dots represent outliers.

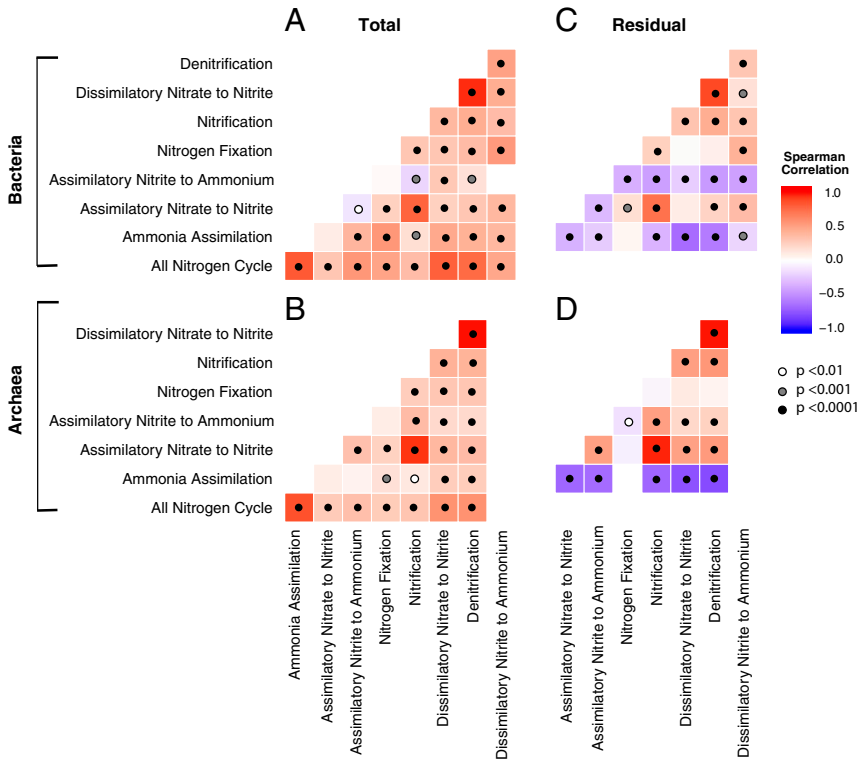

Fig. 3. The relationships between N pathway frequencies. Correlations between N pathways encoded by Bacteria (*A*) and Archaea (*B*) across the samples. (*C* and *D*) Correlations between the residuals of each pathway regressed against the total frequency of all N pathways.

beyond the trends shared by all, we calculated the residuals of the frequency of each pathway regressed against the frequency of all N pathways in a sample. This residual variation was also significantly correlated among many of the N pathways (Fig. 3 *C* and *D*). For instance, denitrification was highly positively correlated with dissimilatory nitrate reduction to nitrite within both Bacteria and Archaea (R^2 = 0.86 and 0.97, respectively, $P \leq 0.001$). This relationship is expected, because dissimilatory nitrate reduction to nitrite is the first step of the complete denitrification process; however, we separated the two steps here, because nitrate reduction to nitrate is also the first step in DNRA (29). Similarly, we separated DNRA into its two pathways: dissimilatory nitrate reduction to nitrite and dissimilatory nitrite reduction to ammonium (Fig. 2*A*). Among Bacteria, the assimilatory nitrite to ammonium pathway residual was negatively correlated with all other pathways. Likewise, the residual frequency of the ammonia assimilation pathway was negatively correlated with all other N pathways in both Bacteria and Archaea. N fixation generally showed weak or no correlation with other pathways.

Drivers of N Pathway Frequencies. The frequency of all N-cycling traits (summing across all pathways) varied greatly among soil samples, and initial analyses revealed broad biogeographic patterns. On average, the highest frequencies of total N pathways were detected in tropical forest and human-dominated (pasture, lawn, and agriculture) soils, whereas the lowest frequency was observed in cold deserts (Fig. S3). Total N pathway frequency also tended to decrease with increasing latitude (R^2 = 0.22, $P < 0.05$; Fig. S4).

To disentangle the drivers behind these patterns, we performed a multivariate regression analysis including habitat type and environmental parameters known to influence microbial abundance and composition (30, 31). Local measurements were not available for most samples; instead, we estimated these variables from secondary sources. For Bacteria, the regression model explained a large and significant proportion of the variability in the frequency of total N pathways (R^2 = 0.58, $P << 0.001$; Table 1). Habitat type contributed most to this model, both directly (positively related to total N pathways) and through interactions with soil carbon and N. The regression model for Archaea explained less variability in total N pathway frequency than for Bacteria (R^2 = 0.43, $P < 0.001$; Table 1). An interactive effect between carbon and N contributed the most to the model, and habitat was only important through an interactive effect with temperature.

We next examined the drivers of individual N pathway frequencies. Due to high covariance between pathways (Fig. 3 *A* and *B*), we fitted regression models to the total-frequency-corrected residuals for each pathway. These models varied greatly in their ability to explain this additional variation (Table 1). For example, the models for the N fixation pathway explained 80% and 63% of the variation among samples in Bacteria and Archaea, respectively ($P << 0.001$). In contrast, the same parameters did not explain any variation in the frequency of the dissimilatory nitrite reduction to ammonium pathway in Bacteria.

Among the significant models, habitat type was an important predictor of the individual pathway frequencies (Table 1). Habitat also interacted with other factors including precipitation, temperature, and soil N to influence the frequency of some pathways. For instance, denitrification frequency increased with temperature in deserts but decreased with temperature in tropical forests. Similarly, ammonia assimilation frequency increased with soil N in temperate forests but decreased with soil N in tropical forests. Soil carbon, which seemed to be a primary driver of total N pathway frequency, did not explain differences in the frequency of individual pathways in Bacteria. Including estimates of N deposition in these models only improved the denitrification model (R^2 increased from 0.41 to 0.48); denitrification frequency increased with increasing N deposition.

The models for individual pathway frequencies in Archaea generally explained less variation than those for Bacteria, perhaps due to the lower number of sequences per sample (Dataset S1). However, for the significant models, the individual N pathways were often best explained by the same parameters as the Bacteria. For instance, habitat type and habitat by temperature were the most

Table 1. Variation explained by the environmental variables in the regression models of the frequency of all (total) and individual N pathways

		Individual pathways (residuals)							
Environmental variables	Total	Ammonia assimilation	Assimilatory nitrate to nitrite	Assimilatory nitrite to ammonia	N fixation	Nitrification	Dissimilatory nitrate to nitrite	Denitrification	Dissimilatory nitrite to ammonia
Bacteria									
Habitat (H)	0.14	0.02	0.23	0.07	0.29	0.11	0.06	0.09	
Precipitation (P)		<0.01							
Temperature (T)		<0.01			0.02			0.02	
pH									
Organic carbon (C)	0.12					<0.01			
Total N	0.05				0.13				
H × P	<0.01		0.07			0.08	0.32		
H × T			0.09	0.23	0.31	0.21	0.03	0.31	
H × pH	<0.01		0.09		0.05				
H × C	0.1								
H × N	0.17	0.49		0.06					
P × T			0.02			0.05		<0.01	
C × N							<0.01		
Adjusted R^2	0.58	0.51	0.5	0.36	0.8	0.45	0.41	0.41	NS
Archaea									
Habitat			0.08		0.09	0.03		0.12	
Precipitation					<0.01				
Temperature						0.02		0.03	
pH								<0.01	
Organic carbon									
Total N			0.05			0.04			
H × P			0.21		0.09				
H × T	0.09		0.18		0.33	0.13			
H × pH					0.12			0.06	
H × C									
H × N									
P × T									
C × N	0.34								
Adjusted R^2	0.43	NS	0.52	NS	0.63	0.22	NS	0.21	NA

The models for the individual pathways are based on the residual frequencies of the pathway after correcting for the Total N pathway frequency (see text). Estimates of the fraction of explained variation are only reported for significant variables ($P < 0.05$). Samples were only included when all environmental variables could be obtained for that location ($n = 99$). NA, not assessed; NS, not statistically significant.

important predictors of N fixation frequency within both domains. Likewise, habitat, habitat by precipitation, and habitat by temperature contributed to the variation in assimilatory nitrate to nitrite frequency in both Archaea and Bacteria.

Taxonomic and Phylogenetic Distribution of N Pathways. A diverse range of microorganisms, encompassing 402 bacterial and 53 archaeal genera, encoded the N pathways. We first investigated the association of pathways within the same genera (Fig. 4 and Fig. S5). All genera for which we detected over 10 sequences carried the ammonia assimilation pathway. Genera carrying the pathway to complete the second half of denitrification also generally carried the first half of the pathway, dissimilatory nitrate to nitrite reduction. The same genera carrying these denitrification pathways sometimes, but not always, carried the dissimilatory nitrite reduction to ammonium pathway, or the second part of the complete DNRA process (Fig. 4 and Fig. S5). Some genera within the Gamma-, Delta-, and Epsilonproteobacteria (e.g., *Edwardsiella*, *Wolinella*, and *Anaeromyxobacter*) contained all three pathways. Indeed, denitrification and DNRA has recently been shown to be present and functional in the same bacteria (29, 32). We also detected genera that only carried the dissimilatory nitrite to ammonium pathway (in addition to ammonia assimilation), as was the case for five genera within the phylum Bacteriodetes.

More broadly, soil genera, and the phyla they fall into, varied in their degree of pathway specialization. Genera within the Cyanobacteria seemed to be specialists, carrying primarily the assimilatory nitrite to ammonium and N fixation pathways. In contrast, genera within the Deltaproteobacteria seemed to be N-cycling generalists, harboring up to six pathways (in addition to ammonia assimilation). Note, however, that these patterns do not distinguish between whether these genera are made up of generalists that encode many pathways or multiple specialists that encode specific pathways.

Focusing on each pathway individually revealed the most prominent taxa carrying the pathway across all soil samples. Here we consider two contrasting pathways, both in terms of their taxonomic distribution and the degree to which they have been studied. First, the abundance of the N fixation pathway in the soil samples was distributed broadly among both Archaea and Bacteria (Fig. 4 and Fig. S5). The most abundant N fixers detected were concentrated within the phylum Proteobacteria, with notable exceptions among the Chlorobi, Firmicutes, and Cyanobacteria (Fig. 5*A*). Most sequences were closely related to N-fixing genera that might be predicted to be common in soil, such as *Bradyrhizobium* and *Burkholderia*. Other abundant genera were less expected. For example, *Azoarcus* is an organism studied for its abilities to degrade soil contaminants (33), and *Pectobacterium* (Gammaproteobacteria) is known primarily as a plant pathogen (34). Indeed, although

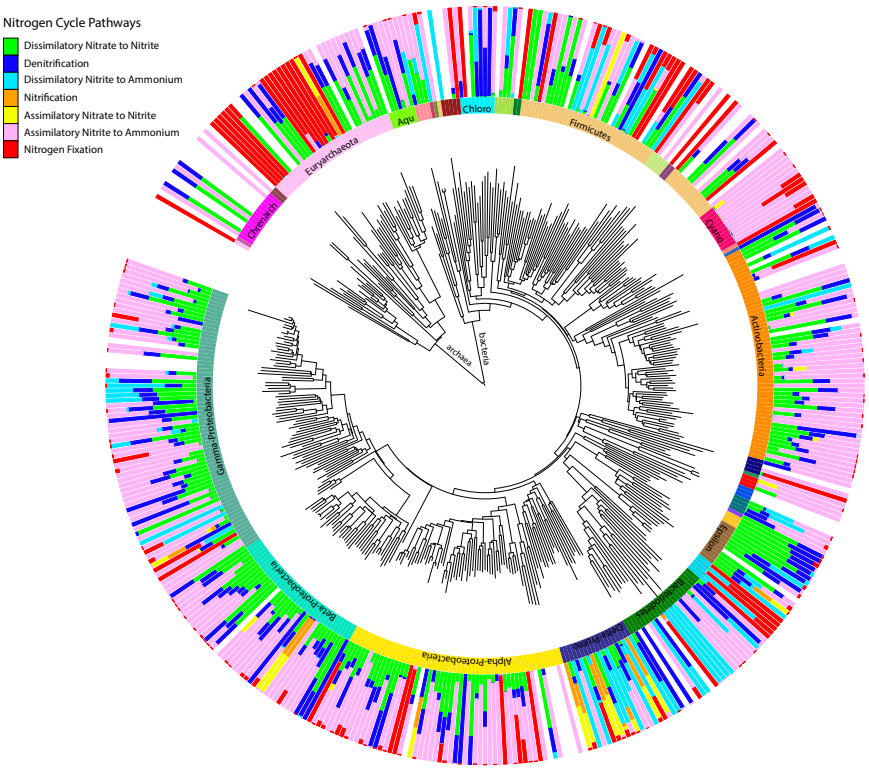

Fig. 4. Phylogenetic distribution of N pathways in the soil metagenomes. A neighbor-joining tree was constructed using 16S rRNA sequences (*Materials and Methods*) and includes all archaea and bacteria genera associated with N cycling sequences in the dataset. The outer circle plots the proportion of N cycle reads assigned to each pathway within the genus. The ammonia assimilation pathway is excluded, because it was found in all genera represented by at least 10 sequences. The inner circle indicates major classes and phyla. See Fig. S5 for a high-resolution figure with genus labels.

it is known that *Pectobacterium* encodes the suite of N fixation genes, it remains unclear whether they are functional (35).

Second, the pathway encoding dissimilatory nitrite reduction to ammonium was also broadly distributed across soil bacteria (Fig. 4), as noted before (36). However, the dominant soil taxa were restricted to two phyla, the Deltaproteobacteria and Verrucomicrobia (Fig. 5*B*). Verrucomicrobia are known to be abundant in soils, but their ecological role remains unclear (37, 38). The pathway's most abundant genus, *Anaeromyxobacter* (phylum Deltaproteobacteria), is common in agricultural soil and has recently been shown to carry out a previously unrecognized process of nondenitrifying N_2O reduction to N_2 (39). The relative abundances of genera encoding the other six N pathways in the soil samples are reported in Fig. S6.

Drivers of Taxonomic Composition by N Pathway. The same environmental variables that explained the overall frequency of the N pathways well explained much less of the variation in the taxonomic composition of the organisms encoding the pathways. For the eight pathways, the models only explained 7–19% of the composition variation of the individual N pathways (Table S1). However, as for pathway frequency, habitat type was the best predictor of composition, explaining up to 14% of the compositional variation in the assimilatory nitrite to ammonium pathway. Temperature also explained 11% of the compositional variation for the nitrification pathway. All other predictors, including pH, explained at most 3% of the variation for any pathway.

A closer examination of two pathways confirms weak compositional differences between the habitats. The 15 most abundant genera carrying the N fixation pathway were similarly abundant across all habitats except in cold deserts (Fig. 5*A*). The most abundant genera encoding the dissimilatory nitrite reduction to ammonium pathway displayed greater variability among habitats (confirming the model results in Table S1), but of these only one genus (*Chlorobium*) seemed specialized on a habitat (wetland) (Fig. 5*B*).

Discussion

Here, we used metagenomic data to characterize the biogeographic patterns of microbial N cycling traits in soil. The advantage of this approach is that it allows us to identify the traits—and the organisms harboring them—involved in many key functions at once. Specifically, the analysis provides a comprehensive map of the dominant lineages involved in eight N processes. The approach also allowed us to search all known genes in a pathway, while avoiding primer biases toward particular lineages (40).

The overall structure of microbial N traits—the relative frequency of the eight pathways—seems to be quite consistent across soils. This is not unexpected but had not been previously tested. For instance, the ammonia assimilation pathway was relatively common, and the pathways for N fixation and nitrification were relatively rare, as observed previously in soil and other environments (41–44). Less expected, however, was that N pathway frequencies within a soil sample were overwhelmingly positively correlated (Fig. 3). This result suggests that soil communities with high numbers of cells able to use one N pathway also generally support higher numbers of cells that can use other N pathways. Greater numbers of metagenomic sequences associated with nutrient cycles have previously been interpreted to be indicative of faster nutrient cycling rates (10). The positive correlations between pathways within the N cycle would seem to support this hypothesis. We also found a high frequency of Bacteria encoding the dissimilatory nitrite reduction to ammonium pathway, which leads to recycling of N in soils. The balance between DNRA and denitrification, which leads to the loss of N to the atmosphere, is thought to be key to soil N budgets. Our results confirm previous studies suggesting that this pathway may be more common than previously thought (45, 46), but the taxa encoding the process in soil environments remain to be carefully characterized (47).

The frequency of N traits further displayed clear biogeographic patterns. At the broadest scale, N trait frequency in Bacteria tended to decrease at higher latitudes, perhaps reflecting a general trend in N limitation in high-latitude ecosystems (48). Beyond latitude, the frequency of N cycling traits in soil communities depended largely

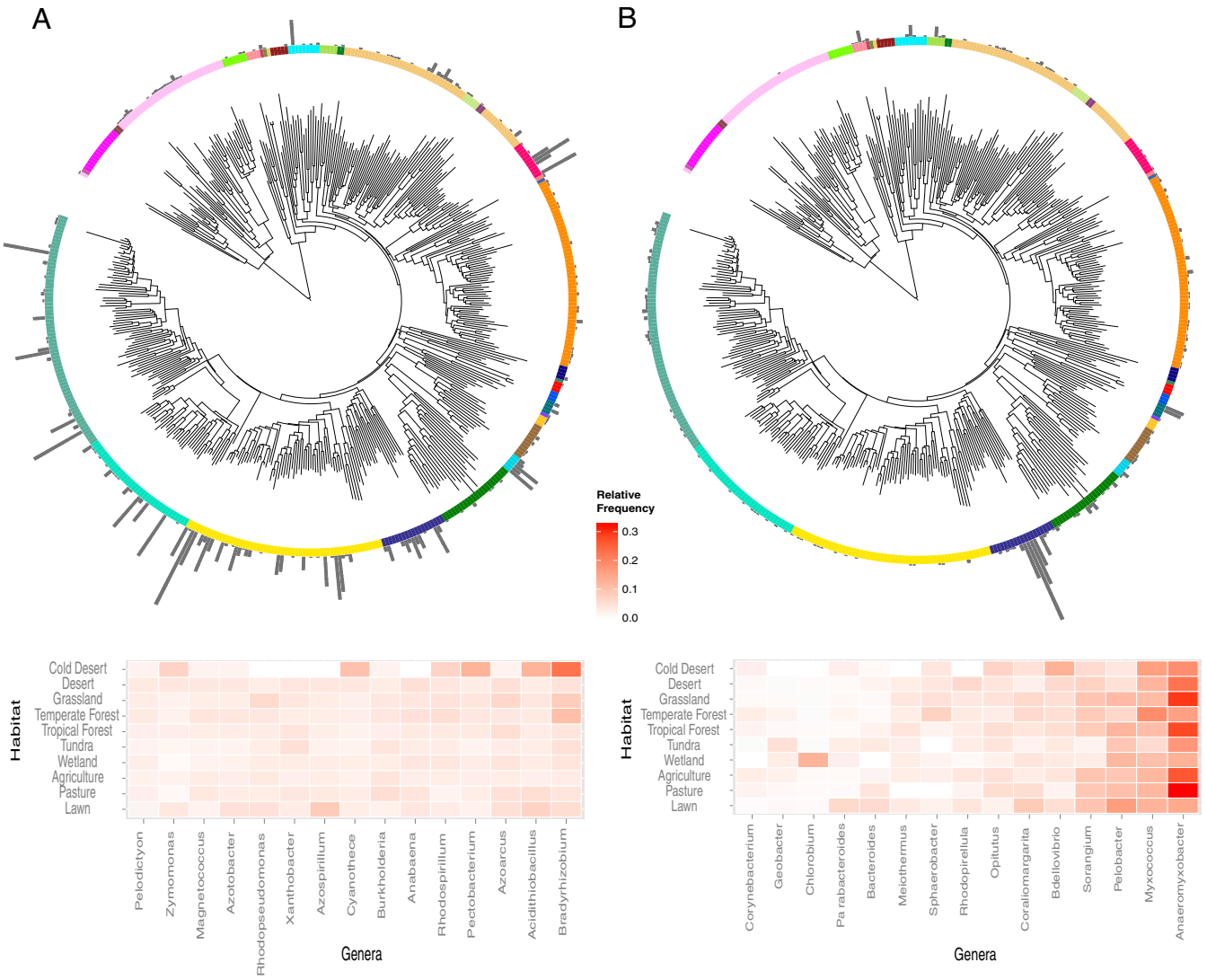

Fig. 5. Phylogenetic distribution of genera encoding specific pathways. The neighbor-joining tree was constructed using 16S rRNA data (*Materials and Methods*). The relative abundance of genera associated with (*A*) N fixation and (*B*) dissimilatory nitrite reduction to ammonium. Within a pathway, the proportion of sequences associated with each genera was calculated for each sample. The proportions were averaged across libraries within each location and then averaged across the 10 habitat types (to provide equal weighting to the habitats). The heat maps give the relative frequency by habitat for the 15 most abundant genera associated with each pathway. See Fig. S6 for plots of all pathways with genus labels.

on habitat type as well as soil carbon and N concentrations. N traits were highest in human-dominated habitats, where N inputs tend to be high, and tropical forests, which are generally thought to be less limited by N than temperate ecosystems (49). In contrast, N traits were lowest in cold deserts (Antarctic and Arctic), which are highly nutrient-limited (48, 50). However, given the low sample numbers for some habitat types, it will be important to retest these patterns as more data accumulate.

Contrary to our hypothesis, the taxa responsible for each N pathway did not vary greatly by habitat type. Within a pathway, genera that were dominant in one habitat tended to be dominant in all habitats. More generally, the environmental variables in our analyses were poor predictors of the compositional variation of the N functional groups. One possible reason for this result is that environmental preferences are conserved below the genus level and therefore would not be detected by our analysis. However, this reasoning does not explain why soil pH seems to have little influence on composition, because pH preference seems to be conserved at a broader taxonomic level (22, 51). Perhaps N functional groups are less specialized for a particular pH environment than microorganisms with other functional roles, but distinct pH-associated lineages in ammonia-oxidizing Archaea indicate that this is not always the case (52). Alternatively, the estimates of soil pH might have been too spatially coarse to detect a pattern.

A well-recognized issue in calculating the frequencies of genes or pathways from metagenomic data is how to normalize for overall genome abundance in the library (53). This normalization step is prone to uncertainties related to variation in mean genome size among communities. To address this issue, we estimated the frequencies of N pathways in two ways: using a set of conserved marker genes as well as the total number of annotated sequences within a domain. The first approach should be sensitive to differences in genome size, whereas the second approach includes more sequence reads and is thus more statistically robust. Because the two approaches led to similar findings, we conclude that the overall patterns in N pathway frequencies are likely not an artifact of normalization.

In sum, this study provides a foundation for future trait-based investigations of soil N cycling but also highlights two major challenges. First, we still know very little about how variability in the frequency and composition of microbial N traits will affect process

rates in soil environments (54). Indeed, a recent review found little correlation between an individual gene's abundance and the process rates that such genes encode. However, assessment of these links using metagenomic datasets is still needed (55). Second, assigning function and taxonomy from short-read sequences is limited by genomic databases where annotations in some cases may be sparse and/or erroneous (56, 57). The N cycle is an archetype of this problem, because new N processes and lineages continue to be identified (39, 58–61). Despite these challenges, the application of metagenomic data to a trait-based framework offers a powerful avenue for elucidating the role that microbial communities play in regulating biogeochemical processes (24, 62).

Materials and Methods

Dataset and Curation. Metagenomic samples (sequencing type "whole genome sequencing" and environmental package "soil", $n = 809$) in the MG-RAST database (27) were classified into one of 10 habitat types (desert, cold desert, grassland, temperate forest, tropical forest, tundra, wetland, agriculture, pasture, and lawn). Samples that could not be classified into these habitats (e.g., oil spill, mines, and microbial mats) were not considered further.

Global Positioning System coordinates and sample date associated with each metagenome identification were downloaded from MG-RAST via the R package matR (63, 64). To minimize the problem of pseudoreplication, we only considered samples from one date per location (the date with the most samples). Based on the statistics provided by MG-RAST, we further removed samples if (*i*) the number of uploaded sequences was equal to the number of post-QC sequences, which seemed to indicate a preprocessing step; (*ii*) the number of identified protein features was <10,000; or (*iii*) the total bacterial reads was <10,000. The remaining metagenomic libraries ($n = 365$) encompassed 118 unique locations. These were downloaded using the MG-RAST API version 3.2 with KEGG database annotations. Each sequence was assigned to the closest related genus in the database using an e-value of $\leq 10^{-5}$.

Data Standardization Across Metagenomic Libraries. Because sequencing effort varied greatly among samples, we standardized the bacterial and archaeal sequences by a suite of conserved, single-copy (i.e., marker) genes to control for possible variation in average genome size among samples (65) (Fig. S1). The Kegg orthology numbers for 30 Bacteria and Archaea marker genes (65) were matched to MD5 IDs using the nonredundant M5nr database. We then searched for these MD5 IDs in the samples annotated by the MG-RAST server.

The number of marker genes was also highly correlated with the total number of annotated sequences in a sample ($R^2 = 0.86$; Fig. S1). Thus, when comparing across Archaea, Bacteria, and Fungi, we standardized the samples by total annotated sequences. Sequencing effort varied greatly among the samples but was not overtly biased toward any particular habitat type (Fig. S1).

Identification of N Cycle Pathways. In each metagenomic library, we searched for sequences from eight N pathways, defined previously in ref. 46. These pathways included nitrification (number of genes targeted: $n = 2$), N fixation ($n = 20$), denitrification ($n = 20$), dissimilatory nitrate to nitrite reduction ($n = 9$), dissimilatory nitrite to ammonia reduction ($n = 4$), assimilatory nitrate to nitrite reduction ($n = 2$), assimilatory nitrite to ammonia reduction ($n = 2$), and ammonia assimilation ($n = 10$) (Fig. 2*A*). If a gene from a pathway was detected, we assumed the presence of the entire pathway.

Environmental Metadata. Environmental data were retrieved from a variety of publically available sources. In all cases, gridded spatial data files were downloaded, and data were extracted using the R packages raster, rdgal, and sp (66, 67). The data included average precipitation (millimeters) and temperature (degrees Celsius) from the month of sampling (68), soil pH (69), total organic carbon (kilograms per square meter) (69), total organic N (grams per square meter) (70), and N deposition (milligrams of N per square meter per year) (71). Approximate data grid resolution for precipitation and temperature was 0.01°, for soil pH and organic carbon was 0.5°, for total organic N was 0.1°, and for N deposition was 4.0°. Environmental metadata were assigned to each sample using the associated latitude and longitude coordinates. Where data were categorized into ranges (soil pH and total organic carbon), the average value from the range was used.

Statistical Analyses. To compare the relative abundance of N pathways across samples, we calculated the frequency of each pathway in a sample for both the Bacteria and Archaea. This frequency is the estimated number of times the pathway was detected per marker gene detected, or [number of pathway reads/number of pathway genes searched]/[number of marker gene reads/30]. Thus, a pathway's frequency of detection was also standardized for the number of genes in the pathway.

To test for differences in the frequency across pathways, we used a one-way analysis of variance, using the aov function in R. To test for correlations between the frequencies of the individual pathways within a sample, we used Spearman's correlation coefficient. To calculate the total N pathway frequency of each sample, we summed the frequency of all eight pathways. We used lm in R to calculate the residuals of each N pathway against a sample's total N pathway frequency.

To tease apart the relative importance of environmental variables on the frequency of N pathways, we used a multiple regression model (lm function in R) including the following variables: habitat type, temperature, precipitation, soil pH, organic carbon, and total N. For this analysis, we averaged data across multiple samples from the same location at just one sampling time, yielding 118 datasets. Based on a priori expectations (72), we also included the following interaction terms: habitat by temperature, habitat by precipitation, habitat by soil pH, habitat by organic carbon, habitat by total N, precipitation by temperature, and organic carbon by total N. To determine the relative importance of the various significant environmental factors from our model in contributing to variation in the frequency of N pathways across samples, we used a backward selection procedure (72, 73). Starting with the significant terms ($P < 0.01$) from our original model, we removed variables one at a time; the differences in R^2 values between each step were used to calculate the relative importance of the independent variable removed from the model. If there was no change or only a marginal change in R^2 when the term was removed, the term was assigned a relative importance of <0.01. After the initial analysis, N deposition was added to test whether this parameter improved the model.

To analyze the composition within each pathway, we calculated the proportional abundance of the genera in a sample and averaged these proportions across multiple samples from the same location. We then calculated a Bray–Curtis distance matrix for all sample locations. We used a distance-based linear model [DISTLM; PRIMER v6; PERMANOVA ++ (74, 75)] to test the significance and importance (an estimate of the proportion of R^2 explained) of the predictor variables for each pathway's composition, using a forward selection procedure.

Phylogenetic Visualization. We constructed a phylogenetic tree including a representative species from all genera encoding N sequences using 16S rRNA amplicon data (chosen for their sequence quality and length of ~1,400 bp) from the SILVA database (76). We aligned the sequences using SINA (77) and created a neighbor-joining tree with the default parameters in Geneious v9.0.5. We used the Interactive Tree of Life (iTOL) (78) to plot (*i*) the proportion of N pathways (excluding ammonia assimilation) detected within each genus and (*ii*) the relative abundance of genera encoding each individual pathway across the unique sampling locations ($n = 118$). For the N fixation and dissimilatory nitrate reduction pathways, we used the ggplot2 package (79) in R to plot heat maps of the relative frequencies of the 15 most abundant genera by habitat.

ACKNOWLEDGMENTS. We thank John Avise, Francisco Ayala, and Brian Bowen for the invitation to participate in this colloquium and Alex Chase for helpful feedback on earlier drafts of the manuscript. This work was supported by a US Department of Education Graduate Assistance in Areas of National Need Fellowship (to M.B.N.) and US Department of Energy, Office of Science, Office of Biological and Environmental Research Grant DE-PS02-09ER09-25.

1. Violle C, Reich PB, Pacala SW, Enquist BJ, Kattge J (2014) The emergence and promise of functional biogeography. *Proc Natl Acad Sci USA* 111(38):13690–13696.
2. Naeem S, Wright JP (2003) Disentangling biodiversity effects on ecosystem functioning: Deriving solutions to a seemingly insurmountable problem. *Ecol Lett* 6(6):567–579.
3. Cardinale BJ, et al. (2012) Biodiversity loss and its impact on humanity. *Nature* 486(7401):59–67.
4. Diaz S, Cabido M (2001) Vive la difference: Plant functional diversity matters to ecosystem processes. *Trends Ecol Evol* 16(11):646–655.
5. Reichstein M, Bahn M, Mahecha MD, Kattge J, Baldocchi DD (2014) Linking plant and ecosystem functional biogeography. *Proc Natl Acad Sci USA* 111(38):13697–13702.
6. McGill BJ, Enquist BJ, Weiher E, Westoby M (2006) Rebuilding community ecology from functional traits. *Trends Ecol Evol* 21(4):178–185.
7. Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* 320(5879):1034–1039.
8. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA* 95(12):6578–6583.

9. Martiny JBH, et al. (2006) Microbial biogeography: Putting microorganisms on the map. *Nat Rev Microbiol* 4(2):102–112.
10. Fierer N, et al. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci USA* 109(52):21390–21395.
11. Reed HE, Martiny JBH (2013) Microbial composition affects the functioning of estuarine sediments. *ISME J* 7(4):868–879.
12. van der Heijden MGA, Bardgett RD, van Straalen NM (2008) The unseen majority: Soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol Lett* 11(3):296–310.
13. Strickland MS, Lauber C, Fierer N, Bradford MA (2009) Testing the functional significance of microbial community composition. *Ecology* 90(2):441–451.
14. Schimel JP, Schaeffer SM (2012) Microbial control over carbon cycling in soil. *Front Microbiol* 3:348.
15. Green JL, Bohannan BJM, Whitaker RJ (2008) Microbial biogeography: From taxonomy to traits. *Science* 320(5879):1039–1043.
16. Vitousek PM, et al. (1997) Human alteration of the global nitrogen cycle: Sources and consequences. *Ecol Appl* 7(3):737–750.
17. Fowler D, et al. (2013) The global nitrogen cycle in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* 368(1621):20130164.
18. Allison SD, Martiny JBH (2008) Colloquium paper: Resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad Sci USA* 105(Suppl 1):11512–11519.
19. Ramirez KS, Craine JM, Fierer N (2012) Consistent effects of nitrogen amendments on soil microbial communities and processes across biomes. *Glob Change Biol* 18(6): 1918–1927.
20. Tedersoo L, et al. (2014) Fungal biogeography. Global diversity and geography of soil fungi. *Science* 346(6213):1256688.
21. Prober SM, et al. (2015) Plant diversity predicts beta but not alpha diversity of soil microbes across grasslands worldwide. *Ecol Lett* 18(1):85–95.
22. Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 75(15):5111–5120.
23. Philippot L, et al. (2009) Spatial patterns of bacterial taxa in nature reflect ecological traits of deep branches of the 16S rRNA bacterial tree. *Environ Microbiol* 11(12): 3096–3104.
24. Barberán A, Fernández-Guerra A, Bohannan BJM, Casamayor EO (2012) Exploration of community traits as ecological markers in microbial metagenomes. *Mol Ecol* 21(8): 1909–1917.
25. Rousk J, et al. (2010) Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J* 4(10):1340–1351.
26. Tsiknia M, Paranychianakis NV, Varouchakis EA, Nikolaidis NP (2015) Environmental drivers of the distribution of nitrogen functional genes at a watershed scale. *FEMS Microbiol Ecol* 91(6):fiv052.
27. Meyer F, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386.
28. Fierer N, Strickland MS, Liptzin D, Bradford MA, Cleveland CC (2009) Global patterns in belowground communities. *Ecol Lett* 12(11):1238–1249.
29. Yoon S, Cruz-García C, Sanford R, Ritalahti KM, Löffler FE (2015) Denitrification versus respiratory ammonification: Environmental controls of two competing dissimilatory NO3(-)/NO2(-) reduction pathways in Shewanella loihica strain PV-4. *ISME J* 9(5):1093–1104.
30. Fierer N, Jackson RB (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* 103(3):626–631.
31. Bru D, et al. (2011) Determinants of the distribution of nitrogen-cycling microbial communities at the landscape scale. *ISME J* 5(3):532–542.
32. Mania D, Heylen K, van Spanning RJM, Frostegård A (2014) The nitrate-ammonifying and nosZ-carrying bacterium Bacillus vireti is a potent source and sink for nitric and nitrous oxide under high nitrate conditions. *Environ Microbiol* 16(10):3196–3210.
33. Sun W, Cupples AM (2012) Diversity of five anaerobic toluene-degrading microbial communities investigated using stable isotope probing. *Appl Environ Microbiol* 78(4): 972–980.
34. Ma B, et al. (2007) Host range and molecular phylogenies of the soft rot enterobacterial genera *pectobacterium* and *dickeya*. *Phytopathology* 97(9):1150–1163.
35. Toth I, Humphris S, Campbell E, Pritchard L (2015) Why genomics research on *Pectobacterium* and *Dickeya* makes a difference. *Am J Potato Res* 92(2):218–222.
36. Welsh A, Chee-Sanford JC, Connor LM, Löffler FE, Sanford RA (2014) Refined NrfA phylogeny improves PCR-based nrfA gene detection. *Appl Environ Microbiol* 80(7):2110–2119.
37. Bergmann GT, et al. (2011) The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol Biochem* 43(7):1450–1455.
38. Fierer N, et al. (2013) Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* 342(6158):621–624.
39. Sanford RA, et al. (2012) Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. *Proc Natl Acad Sci USA* 109(48):19709–19714.
40. Myrold DD, Zeglin LH, Jansson JK (2014) The potential of metagenomic approaches for understanding soil microbial processes. *Soil Sci Soc Am J* 78(1):3–10.
41. Varin T, Lovejoy C, Jungblut AD, Vincent WF, Corbeil J (2010) Metagenomic profiling of Arctic microbial mat communities as nutrient scavenging and recycling systems. *Limnol Oceanogr* 55(5):1901–1911.
42. Souza RC, et al. (2015) Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes. *Appl Soil Ecol* 86:106–112.
43. Quinn RA, et al. (2014) Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung. *MBio* 5(2):e00956–e13.
44. Martiny AC, Treseder K, Pusch G (2013) Phylogenetic conservatism of functional traits in microorganisms. *ISME J* 7(4):830–838.
45. Rutting T, Boeckx P, Muller C, Klemedtsson L (2011) Assessment of the importance of dissimilatory nitrate reduction to ammonium for the terrestrial nitrogen cycle. *Biogeosciences* 8(7):1779–1791.
46. Nelson MB, Berlemont R, Martiny AC, Martiny JBH (2015) Nitrogen cycling potential of a grassland litter microbial community. *Appl Environ Microbiol* 81(20):7012–7022.
47. Kraft B, Strous M, Tegetmeyer HE (2011) Microbial nitrate respiration–genes, enzymes and environmental distribution. *J Biotechnol* 155(1):104–117.
48. Yergeau E, Kang S, He Z, Zhou J, Kowalchuk GA (2007) Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal transect. *ISME J* 1(2):163–179.
49. Vitousek PM (1984) Litterfall, nutrient cycling, and nutrient limitation in tropical forests. *Ecology* 65(1):285–298.
50. Jonasson S, Michelsen A, Schmidt IK (1999) Coupling of nutrient cycling and carbon dynamics in the Arctic, integration of soil microbial and plant processes. *Appl Soil Ecol* 11(2-3):135–146.
51. Martiny JBH, Jones SE, Lennon JT, Martiny AC (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* 350(6261):aac9323.
52. Gubry-Rangin C, et al. (2011) Niche specialization of terrestrial archaeal ammonia oxidizers. *Proc Natl Acad Sci USA* 108(52):21206–21211.
53. Manor O, Borenstein E (2015) MUSiCC: A marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol* 16:53.
54. Prosser JI (2015) Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology. *Nat Rev Microbiol* 13(7):439–446.
55. Rocca JD, et al. (2015) Relationships between protein-encoding gene abundance and corresponding process are commonly assumed yet rarely observed. *ISME J* 9(8): 1693–1699.
56. Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* 2(1):3.
57. Wu D, et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462(7276):1056–1060.
58. Strous M, et al. (1999) Missing lithotroph identified as new planctomycete. *Nature* 400(6743):446–449.
59. Könneke M, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437(7058):543–546.
60. Farnelid H, et al. (2011) Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One* 6(4):e19223.
61. van Kessel MAHJ, et al. (2015) Complete nitrification by a single microorganism. *Nature* 528(7583):555–559.
62. Fierer N, Barberán A, Laughlin DC (2014) Seeing the forest for the genes: Using metagenomics to infer the aggregated traits of microbial communities. *Front Microbiol* 5:614.
63. Team RDC (2011) *R: A Language and Environment for Statistical Computing* (The R Foundation for Statistical Computing, Vienna).
64. Braithwaite DT, Keegan KP (2013) matR: Metagenomics analysis tools for R. R package version 0.9.9.
65. Nayfach S, Pollard KS (2015) Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 16:51.
66. Hijmans RJ, van Etten J (2012) raster: Geographic analysis and modeling with raster data.
67. Bivand RS, Pebesma E, Gomez-Rubio V (2013) *Applied Spatial Data Analysis with R* (Springer, New York), 2nd Ed.
68. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25(15):1965–1978.
69. Batjes NH (2000) Global data set of derived soil properties, 0.5-degree grid (ISRIC-WISE). Available at daac.ornl.gov/.
70. Group GSDT (2000) Global gridded surfaces of selected soil characteristics (IGBP-DIS) (Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, TN).
71. Dentener F, et al. (2006) Nitrogen and sulfur deposition on regional and global scales: A multimodel evaluation. *Global Biogeochem Cy* 20(4):GB4003.
72. Ramette A (2007) Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol* 62(2):142–160.
73. Mac Nally R (2002) Multiple regression and inference in ecology and conservation biology: Further comments on identifying important predictor variables. *Biodivers Conserv* 11(8):1397–1401.
74. Clarke KR, Warwick RM (2001) *Change in Marine Communities: An Approach to Statistical Analysis and Interpretation* (PRIMER-E, Plymouth, UK), 2nd Ed.
75. Anderson MJ, Gorley RN, Clarke KR (2008) *PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods* (PRIMER-E, Plymouth, UK).
76. Quast C, et al. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 41(Database issue, D1): D590–D596.
77. Pruesse E, Peplies J, Glöckner FO (2012) SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28(14):1823–1829.
78. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1):127–128.
79. Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York).