environmental microbiology

Environmental Microbiology (2010) doi:10.1111/j.1462-2920.2010.02315.x

# Global patterns in the biogeography of bacterial taxa

**Diana R. Nemergut,[1,2]* Elizabeth K. Costello,[3] Micah Hamady,[4] Catherine Lozupone,[3,5] Lin Jiang,[6] Steven K. Schmidt,[7] Noah Fierer,[7,8] Alan R. Townsend,[1,7] Cory C. Cleveland,[9] Lee Stanish[1,2] and Rob Knight[3]**

[1]*Institute of Arctic and Alpine Research,* [2]*Environmental Studies Program,* [3]*Department of Chemistry and Biochemistry,* [4]*Department of Computer Science,* [7]*Department of Ecology and Evolutionary Biology and* [8]*Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309, USA.*

[5]*Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108, USA.*

[6]*School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA.*

[9]*Department of Ecosystem and Conservation Sciences, University of Montana, Missoula, MT 59812.*

## Summary

**Bacteria control major nutrient cycles and directly influence plant, animal and human health. However, we know relatively little about the forces shaping their large-scale ecological ranges. Here, we reveal patterns in the distribution of individual bacterial taxa at multiple levels of phylogenetic resolution within and between Earth's major habitat types. Our analyses suggest that while macro-scale habitats structure bacterial distribution to some degree, abundant bacteria (i.e. detectable using 16S rRNA gene sequencing methods) are confined to single assemblages. Additionally, we show that the most cosmopolitan taxa are also the most abundant in individual assemblages. These results add to the growing body of data that support that the diversity of the overall bacterial metagenome is tremendous. The mechanisms governing microbial distribution remain poorly understood, but our analyses provide a framework with which to test the importance of macro-ecological environmental gradients, relative abundance, neutral processes and the ecological strategies of individual taxa in structuring microbial communities.**

Received 9 February, 2010; accepted 20 June, 2010. *For correspondence. E-mail nemergut@colorado.edu; Tel. (+1) 303 735 1239; Fax (+1) 303 492 6388.

## Introduction

Forces shaping the biogeography of macroorganisms – including dispersal limitations, habitat differentiation, competition and adaptive radiation – have been a central focus of ecology for more than a century (Brown and Lomolino, 1998). Yet, while microorganisms are the most abundant and diverse organisms on Earth (Whitman *et al*., 1998), relatively little is known about the patterns of, or controls over, microbial distribution within and between the planet's major habitat types. One common theory holds that the tremendous dispersal potential of microbes will lead to everything being everywhere (i.e. no dispersal limitations), with environmental selection determining which species are abundant (Martiny *et al*., 2006). However, until recently, methodological limitations have prevented large-scale tests of ideas about where certain microorganisms exist, and why (Hugenholtz *et al*., 1998; Prosser *et al*., 2007).

Over the last decades, however, molecular phylogenetic approaches have revolutionized microbiology, expanding our view of microbial diversity and our appreciation of the complexity of microbial communities (Hugenholtz *et al*., 1998). While these techniques do not provide an exhaustive sampling of any but the simplest microbial assemblages, they do provide information on the dominant members of the community, allowing ecologically meaningful questions to be addressed about the distribution of these lineages. These methods have been used to reveal that some microorganisms exhibit distinct biogeographical patterns (Horner-Devine *et al*., 2004; Green and Bohannan, 2006; Martiny *et al*., 2006), which appear to be controlled by differences in environmental variables in some cases (Horner-Devine *et al*., 2004), and geographical distance in others (Cho and Tiedje, 2000; Whitaker *et al*., 2003). Other work investigating overall community composition supports the role of environmental gradients in structuring both lake and soil bacterial assemblages (Fierer and Jackson, 2006; Van der Gucht *et al*., 2007). Biotic interactions may also be important in determining microbial community composition; a recent study showed that microbial communities exhibit more segregation of taxa than would be predicted by chance, suggesting that competitive interactions and/or niche specialization may be important in structuring bacterial biogeography (Horner-Devine *et al*., 2007).

To date, however, most studies of microbial biogeography have focused on a single habitat type or on a

© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd

phylogenetically restricted set of taxa; thus, broader patterns in the distribution of microorganisms among Earth's major ecosystems remain poorly understood. Recently, Lozupone and Knight (2007) demonstrated that salinity is the primary driver of community-level phylogenetic differentiation among bacterial assemblages sampled from different habitat types. Yet, we know that many bacterial phyla are widely distributed across multiple habitats (Madigan *et al*., 2000). Thus, to further investigate the role of macro-scale habitats in structuring the biogeographical patterns exhibited by individual bacterial taxa, we examined the distribution of taxa at multiple levels of phylogenetic resolution (98%, 95%, 92% and 89% 16S rRNA gene sequence identity), both within and across different habitat types. Here, we show that there is minimal taxon overlap between assemblages both within and between habitat types, and that the most abundant taxa are also the most widely distributed.

## Results

We examined the distribution of 16S rRNA genes in a data set of clone libraries assembled from a variety of habitat types (Lozupone and Knight, 2007) and expanded upon (Table S1). Operational taxonomic units (OTUs) were selected at four different levels of sequence identity: 98%, 95%, 92% and 89% and the collection of OTUs present in a given sample was considered an individual assemblage. Although there is controversy about the amount of sequence differentiation that constitutes a particular taxonomic ranking, there is some consensus that these levels of divergence are less than those required to differentiate phyla (Hugenholtz *et al*., 1998; Dojka *et al*., 2000) and have been used to correspond roughly to species, genus, family and order respectively (Stackebrandt and Goebel, 1994; Schloss and Handelsman, 2004).

We first examined the distribution of OTUs across all 238 assemblages examined. At the species level of sequence identity, more than 85% of all OTUs were not detected in more than one assemblage and no single OTU was observed in more than 12% of assemblages (Fig. 1A). At higher levels of sequence divergence, more OTUs were widespread; for example, at the order level of identity, 35% of OTUs were found in two or more assemblages. However, at all levels of phylogenetic resolution, distribution patterns featured a similar pattern with the majority of OTUs found in no more than one assemblage and small numbers of OTUs that were more highly distributed. All OTUs that were detected in greater than 20% of assemblages belonged to the *Proteobacteria*, specifically the $\alpha$-, $\beta$- and $\gamma$-proteobacteria. Although 'unclassifiable' OTUs comprised 10% of the original data set (Fig. 1B), no single unclassifiable OTU was observed in more than 6% of all assemblages for any OTU definition.

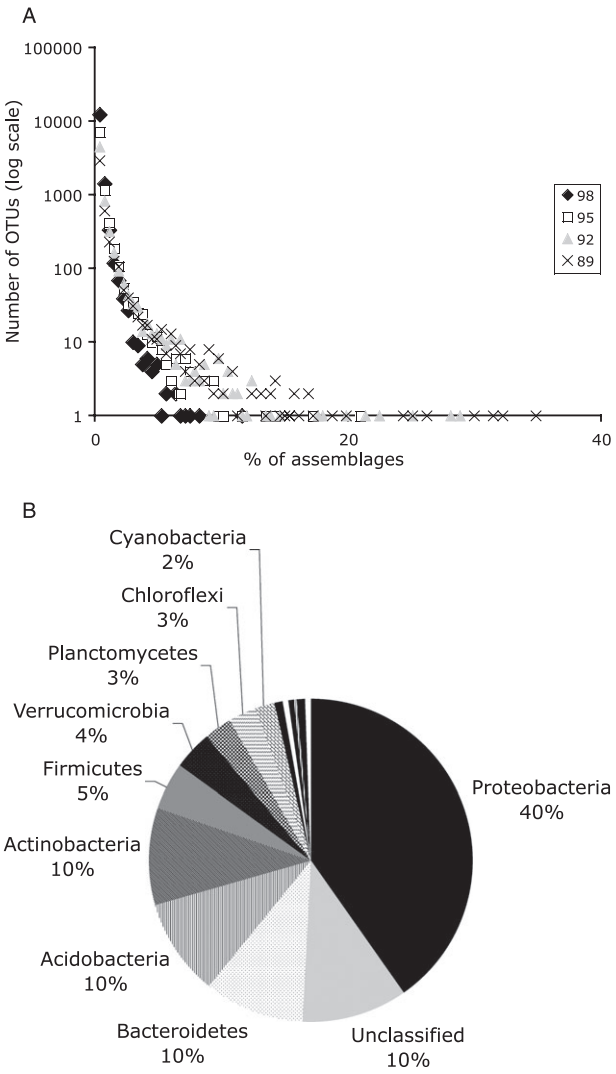


**Fig. 1.** A. The number of OTUs that were found in different proportions of assemblages within our clone library data set (Table S1), which contains 28 115 sequences and 238 assemblages. At all OTU definitions, the vast majority of lineages were observed in only a single assemblage.
B. The relative abundance of different phyla within the clone library data set. Phyla that represent at least 2% of all sequences are labelled.

To examine how much of this limited distribution was driven by environment type, we explored patterns of occurrence across the 14 different habitats. At the species level, 97% of OTUs were found in no more than one habitat type and no single OTU was detected in more than six habitats (Fig. 2). Although less pronounced, these patterns were also discernible at lower levels of phylogenetic resolution, with 92%, 88% and 84% of genus-, family- and order-level OTUs, respectively, detected in no more than a single habitat type. OTUs detected in more than five habitat types were related to the *Comamonadaceae*, *Pseudomonadaceae*, *Aeromonas*, *Staphylococcus* and *Propionibacterium* (Table 1). We performed PERMANOVAs

© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, 

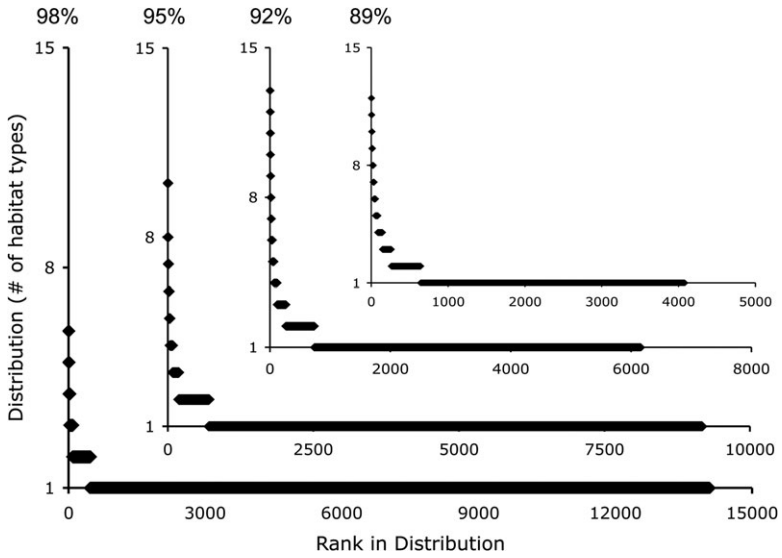


**Fig. 2.** Rank distribution plots displaying the presence of OTUs in different numbers of habitat types. At all OTU definitions, the vast majority of lineages were observed in only a single habitat type.

(Anderson, 2001) to test the hypothesis that habitat types structure the distribution of bacteria. This is an analysis of variance test that uses permutation to examine the significance of factors (in this case, habitat types) in partitioning variation within multivariate data sets (in this case, an assemblage by OTU presence–absence matrix). These analyses revealed that, while most variation in assemblage composition was accounted for within habitat types (83–95%), there was a significant amount (5–17%) of variation between different habitat types ($P < 0.001$).

The distribution of OTUs across assemblages was then examined within the six habitat types for which we had the most samples: soil, lakewater, freshwater sediment, seawater, marine sediment and insect-associated assemblages. Again, distribution was assayed at four different OTU cut-offs. At all levels of phylogenetic resolution, all habitat types revealed a distribution pattern featuring a few widely dispersed lineages and many more confined lineages (Fig. 3).

**Table 1.** Phylogenetic identities of ecologically 'widely distributed' OTUs at the 98% minimum sequence identity cut-off.

| OTU # | Number of habitat types | Classification |
|---|---|---|
| 467 | 6 | *Comamonadaceae* |
| 178 | 5 | *Pseudomonadaceae* |
| 243 | 5 | *Propionibacterium* |
| 469 | 5 | *Pseudomonas* |
| 63 | 5 | *Staphylococcus* |
| 107 | 5 | *Aeromonas* |
| 163 | 6 | *Pseudomonas* |
| 144 | 6 | *Comamonadaceae* |

Table shows the OTU identifier (OTU #), the number of habitat types it was detected in, and its classification. Note that our classification system did not allow all OTUs to be identified at the same level of phylogenetic resolution; some were resolved to the genus level while others were resolved to the family level.

We also examined OTU distribution across two pyrosequencing-based soils data sets including an inter-continental analysis of 88 samples (Lauber *et al*., 2009) and a study of 27 samples from within a single hectare of tropical rainforest. The advantage of examining both of these data sets separately is that we can look for similarities and differences in patterns of distribution that exist over both large (inter-continental) and small (intra-hectare) scales. Although we did not examine multiple OTU definitions for the short sequences generated via pyrosequencing because of known inconsistencies (Elshahed *et al*., 2008), both of these data sets contained more than 1000 16S rRNA gene sequences per soil, and thus are much better sampled than the clone library data. Another advantage of the pyrosequencing studies is that it is possible to consider the relationship between relative abundance and distribution patterns, which is impossible for the compiled data set because of irregularities in analysing and reporting abundance between studies (Lozupone and Knight, 2007).

The distribution of OTUs within the pyrosequencing data sets shows the same basic pattern as was seen in the clone library data (Fig. 4). For the large-scale data set, 75% of OTUs were not found in more than one soil (Lauber *et al*., 2009) while 68% from the smaller-scale tropical forest site were detected in a single soil sample. The top four most widely distributed OTUs from the large-scale data set were detected in between 70% and 88% of soils, and were all related to the *Bradyrhizobiales*. Three OTUs related to the $\alpha$-proteobacteria and one OTU related to the $\delta$-proteobacteria were detected in all of the tropical forest soils.

For each OTU, we plotted its total abundance across all assemblages against the number of assemblages in which it was detected, revealing a significant, positive relationship (Fig. 5A). We also plotted the average of the

© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, *Environmental Microbiology*

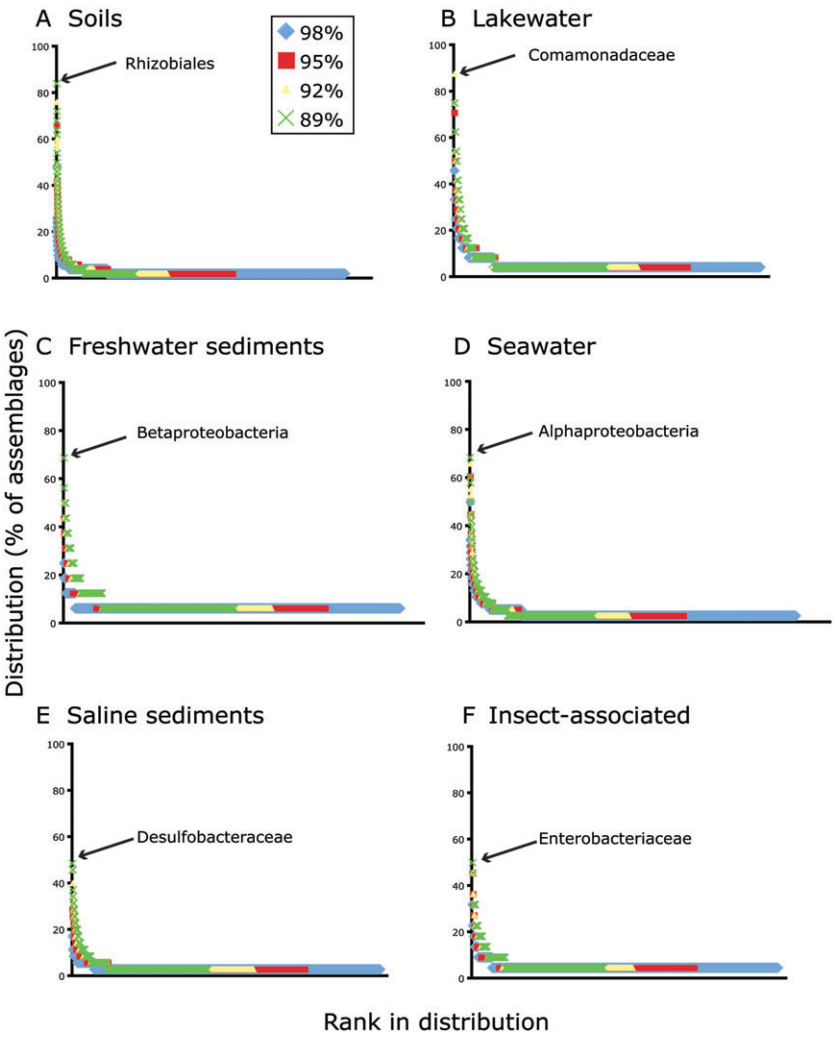


**Fig. 3.** The rank in distribution plotted against the per cent of assemblages each OTU was found in for (A) soils ($n = 49$), (B) lakewater ($n = 21$), (C) freshwater sediments ($n = 15$), (D) seawater ($n = 40$), (E) saline sediments ($n = 36$) and (F) insect-associated samples ($n = 15$) for the clone library data. Those OTUs that were most widely dispersed within habitat types are indicated. Within habitat types, some OTUs were widely distributed among assemblages while the majority were limited to only a few assemblages.

relative abundance of each OTU across all assemblages against its distribution, which did not change the shape of the function (data not shown). The top 10 most abundant OTUs from each sample were found in an average of 28% of assemblages (versus 2% for all OTUs) for the large-scale data set. For the tropical forest data set, the top 10 most abundant OTUs from each sample were found in an average of 69% of soils (versus 8% for all OTUs). Interestingly, the high abundance values for the overall top 10 OTUs in each data set were not driven by a few assemblages with high proportions of these sequences. Rather, they reflect moderate abundances (relative to the total abundance of that OTU within the data set) across many samples (Fig. 5B).

## Discussion

Our results support that, for the most abundant organisms from these assemblages, macro-scale habitats structure bacterial distribution (Fig. 2). Indeed, as has been shown in other work (Tanner *et al.*, 1998) close relatives of the eight OTUs that were detected in five or more habitat types (Table 1) are among the most abundant organisms found on human skin (e.g. *Staphylococcus*, *Propionibacterium*) or have been found in low-organic matter water supplies (e.g. *Comamonadaceae*, *Pseudomonadaceae*, *Aeromonas*) (Burtscher *et al.*, 2009; Costello *et al.*, 2009), suggesting that these 'widely distributed' bacteria actually may be contaminants introduced during sample processing or PCR amplification.

Other studies have shown that habitat types harbour different *overall communities of bacteria* (Lozupone and Knight, 2007) and archaea (Auguet *et al.*, 2010), but have not determined if habitat type also shapes the *distribution of individual microbial taxa*. For example, one possible explanation for the difference in community composition between habitat types may be that bacteria are widespread across multiple habitat types, but that different habitat types support different *combinations* of organisms. Our data support that, for the most part, abundant (i.e.

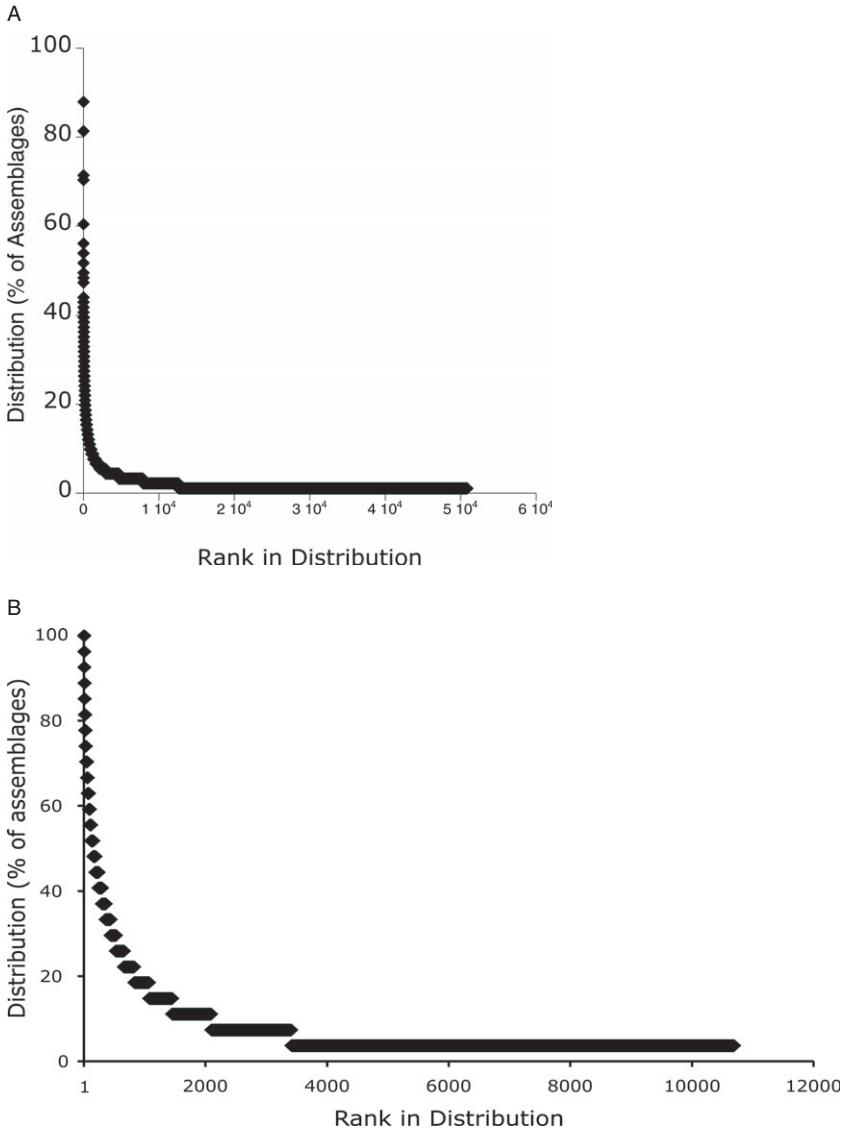


**Fig. 4.** The rank in distribution plotted against the per cent of assemblages each OTU was found in for (A) the inter-continental soils data set ($n = 88$) and (B) the tropical forest (intra-hectare) soils data set ($n = 27$).

detectable using 16S rRNA gene sequencing methods) bacteria are confined to specific environments and that a significant fraction of the variation in the distribution of bacteria is related to habitat type.

Although macro-scale habitats do structure bacterial distribution, our results also suggest that, within a habitat type, most bacterial taxa are still restricted to a relatively small number of assemblages (Figs 1, 3 and 4) and that there is a positive relationship between the relative abundance of an organism and its distribution across assemblages (Fig. 5). We discuss the implications of these observations below.

### *Most bacteria are confined to one assemblage*

We found that between 65% and 85% of OTUs at all levels of sequence identity examined were present in only a single assemblage (Fig. 1). This pattern of limited distribution has also been observed between assemblages within individual habitat types (e.g. Figs 4 and 5), including coastal waters (Pommier *et al*., 2007) and soils (Noguez *et al*., 2005; Fulthorpe *et al*., 2008), but has not been documented across habitats. For example, Fulthorpe and colleagues (2008) examined four soils from different sites in North and South America using pyrosequencing of SSU rRNA genes (Roesch *et al*., 2007). They generated between ~26 000 and 53 000 gene sequences per soil and showed that, at the 97% identity cut-off, 74% of OTUs were confined to a single assemblage. Likewise, in the Lauber and colleagues (2009) study 75% of sequences at the 97% OTU cut-off were detected in only a single sample (Fig. 5).

Other studies have used more sensitive methods to support endemism among particular groups of micro-

© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, *Environmental Microbiology*

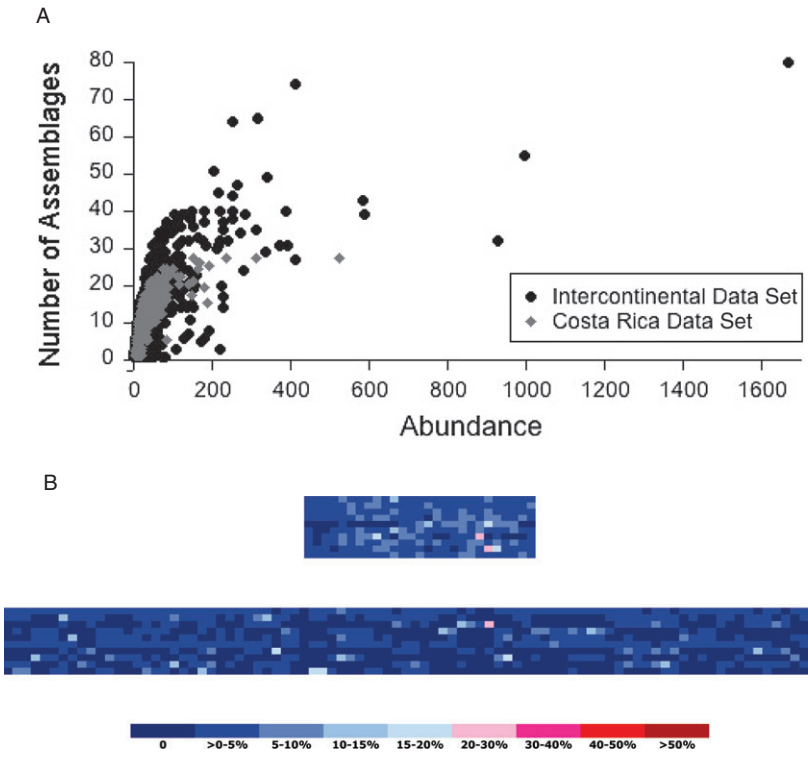


**Fig. 5.** The relationship between abundance and distribution of OTUs.
A. The grey diamonds represent OTUs from the tropical forest data set; the black circles represent OTUs from the *trans*-continental (Lauber *et al*., 2009) data set. Here, we plotted the total abundance (within the entire data set) of each OTU against its distribution. However, we also examined the relationship between the average of the relative abundance of each OTU within all assemblages against its distribution, which yielded similar results (data not shown).
B. Heatmaps of the top 10 most abundant OTUs for each study (tropical forest soils = top, intercontinental soils = bottom) showing the abundance of each OTU in each assemblage relative to its total abundance across the data set. Each column represents a different assemblage; each row represents a different OTU; the colour of the cells represents the relative abundance of that OTU within specific assemblages.

organisms. For example, Cho and Tiedje (2000) isolated fluorescent pseudomonads from 10 sites on four continents. Using a method for genomic fingerprinting, they revealed no overlap in genotypes between sites or between continents. Likewise, Wawrik and colleagues (2007) used tRFLPs to examine 16S rRNA and type II polyketide synthase genes of actinomycetes from soils collected in New Jersey and Asia and showed that fewer than 1% of phylotypes were found in more than 50% of soils that they examined. Geographical isolation has also been demonstrated for archaea in hot springs (Whitaker *et al*., 2003) and β-proteobacteria in sediments (Horner-Devine *et al*., 2004). Thus, several studies using a variety of methods support microbial endemism over a range of environments.

Although sampling limitations constrain our ability to conclude that the distribution patterns that we observed reflect microbial endemism, we can say that abundant bacteria exhibit a pattern of distribution both within (Figs 3 and 4) and between habitat types (Fig. 1A), with most organisms being found in no more than one assemblage. It is widely recognized that, *within* individual assemblages, few taxa are abundant and most taxa are rare (Curtis *et al*., 2002). Here we identify a similar pattern *across* assemblages: relatively few taxa are cosmopolitan and the vast majority are restricted to individual assemblages. That said, we caution that improved sequencing technology may alter these interpretations in the future. Indeed, Preston's 'Veil Line' concept suggests that many organisms exhibit a normal distribution pattern which can be obscured by undersampling (Preston, 1948). He suggests that many organisms with so-called rare distribution patterns will reveal a more intermediate distribution in exhaustively sampled assemblages.

### *Abundant bacteria are more widely distributed*

We observed that, across soil assemblages, abundant organisms were more likely to be widely distributed (Fig. 5). This pattern was observed at two very different spatial scales: within a single hectare of rainforest soil and within a variety of soils sampled across two continents. A positive relationship between abundance and distribution among soil bacteria is also apparent in an examination of the pyrosequencing data from the Fulthorpe and colleagues (2008) study: 50% of the top 10 most abundant organisms were found in more than one of the four soils that they analysed. Likewise, Pommier and colleagues (2007) discovered a positive relationship between OTU abundance and distribution across coastal seawater samples. Spain and colleagues (2009) reported a similar pattern in the analysis of their large 16S rRNA gene clone library data set from a grassland soil: they observed that the most abundant orders of *Proteobacteria* were more highly distributed among other environments. Sloan and colleagues (2006) also described a positive relationship between abundance in distribution in sewage treatment facilities, estuaries, lakewater and microbiome samples.

© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, *Environmental Microbiology*

As mentioned above, it is difficult to discern a particular organism's relative abundance from our compiled clone library data set because of inconsistent reporting and screening methods. However, it is noteworthy that many taxa that are cosmopolitan within habitat types (Fig. 3) have been identified as abundant members of their respective environments in other studies (Janssen, 2006; Newton *et al*., 2007; Rusch *et al*., 2007). This may be a general feature of all of life, as the positive correlation between the distribution and abundance of macroorganisms has been well documented (Brown, 1984). For macroorganisms, distribution varies in geographical scale by more than 12 orders of magnitude, and as techniques to sample microbial communities improve, it will permit us to assess if microbial distribution patterns exhibit a similar level of variation.

What could cause the positive relationship between abundance and distribution? We propose three, non-mutually exclusive possibilities. First, it could simply reflect the fact that these organisms are easier to detect within our current sampling limitations. Another possibility is that higher local population sizes enable wider dispersal potentials. A prevailing hypothesis for microbial biogeography states that population sizes are extremely large and thus dispersal is not limited (Fenchel and Finlay, 2004). However, not all organisms are abundant within a community; in fact, most organisms are rare (Sogin *et al*., 2006; Ashby *et al*., 2007). Thus, the larger population sizes of the most abundant organisms may facilitate their dispersal and help drive the positive relationship between abundance and distribution. Indeed, some calculations suggest that very rare soil organisms may be present at densities of one cell per 27 $km^2$ (Curtis *et al*., 2002), which would undoubtedly limit their distribution potential. Sloan and colleagues (2006) described a near-neutral model for microbial community assembly, in which the distribution of taxa is largely determined by immigration and chance. Random assembly processes would lead to a positive relationship between distribution and relative abundance (Sloan *et al*., 2006) and may partially explain the within-habitat distribution patterns observed here. Finally, the relationship between abundance and distribution may reflect a positive relationship between regional and global distributions, a pattern that has been shown for macroorganisms (reviewed in Prinzing *et al*., 2004). Because the way that we sample microorganisms (e.g. 1 g of soil) is far too coarse to permit the examination of single communities, we are actually examining the composition of many communities within a single assemblage (Grundmann, 2004). Indeed, high 'regional' distribution patterns may cause high abundance values within a sampled assemblage, thus the lognormal-shaped species abundance curves observed within a single community (Curtis *et al*., 2002) may actually reflect the same phenomena as the distribution patterns that we observed between assemblages (Figs 1–4).

## Conclusion

We emphasize that our results apply to the most abundant organisms that are detected by contemporary sequencing technology. New technological advances are on the horizon, and it is unknown how the patterns that we observed may change when more assemblages can be completely sampled and analysed. Additionally, methods to assay the entire genomic complement of individual assemblages will become easier in the near future, enabling the analysis of 'functional biogeography' (Green *et al*., 2008) to better understand the process-level implications of the observed patterns of bacterial distribution.

Despite these caveats, our results suggest that while macro-scale environmental factors structure the ecological distribution of bacterial taxa, most bacteria demonstrate a limited distribution within habitat types. We also show a positive relationship between the abundance and distribution of soil bacteria within habitat types. Given the high degree of genetic differentiation between even very closely related lineages of bacteria living in close proximity (Thompson *et al*., 2005), our results add to the growing body of data that support that the diversity of the overall bacterial metagenome is enormous. The mechanisms governing microbial distribution remain poorly known, but our analyses provide a framework with which to test the importance of macro-ecological environmental gradients, relative abundance, the ecological strategies of individual taxa and neutral processes in structuring microbial communities.

## Experimental procedures

### *Data sets*

To examine the distribution of different bacterial taxa within and between different habitat types, we expanded upon the 16S rRNA gene clone library data set compiled by Lozupone and Knight (2007) so that it now includes 28 115 16S rRNA sequences, derived from 238 samples taken across 14 different habitat types (Table S1). This data set was assembled from studies examining the microbial communities of natural environments using 16S rRNA gene cloning and sequencing targeted towards all bacteria. Specifically, sequences were identified from the ENV database of GenBank, reference information was extracted for each record, the sequences that had the same title were grouped, and studies that were associated with the most sequences were selected. Since a single study can report sequences from different assemblages and different habitat types, the sequences were divided into assemblages and habitat types using annotations from the associated publications. Here, habitat types were defined at the macroscale and ranged from soil and

© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, *Environmental Microbiology*

seawater to insect and sponge-associated assemblages (Table S1). These clone libraries contained an average of 118 16S rRNA gene sequences per assemblage with a minimum of 20 and a maximum of 836 sequences.

Next, to determine the distribution of bacteria between relatively well-sampled assemblages within soils as well as to examine the relationship between relative abundance and distribution, we examined two pyrosequencing-generated 16S rRNA gene data sets from soils. The first was taken from Lauber and colleagues (2009) and featured nucleotides 27–338 (*Escherichia coli* numbering) of the 16S rRNA gene (regions V1 and V2). In this study, 88 different soils from across North and South America that were first described by Fierer and Jackson (2006) were analysed. An average of 1501 classifiable sequences per soil was obtained with a maximum of 2167 and a minimum of 1047 sequences. In addition, we introduce a new pyrosequencing-based data set of 16S rRNA genes from 27 soil samples obtained from within a single hectare of lowland tropical rainforest soil from the Osa Peninsula in Costa Rica [see Cleveland and Townsend (2006) for site description]. Samples were taken from litter removal, litter augmentation, and precipitation exclusion manipulations as well as from control plots (Wieder *et al*., 2009). Control plots were sampled in April, June and October of 2008, while plots subjected to experimental manipulations were sampled in June and October of 2008. For each treatment at each time point, three replicate plots were obtained for a total of 27 soil samples. For each plot, the top 5 cm of soil was aseptically collected and DNA was extracted using PowerSoil DNA Isolation kits (MO BIO, Carlsbad, CA, USA). Error-corrected bar-coded pyrosequencing was performed as described by Fierer and colleagues (2008) with the sequencing performed at the Environmental Genomics Core Facility at the University of South Carolina on a Roche FLX 454 pyrosequencing machine. Data were processed as described by Fierer and colleagues (2008) and Hamady and colleagues (2008) with an average of 1384 sequences obtained per soil (range of 1087–2030).

### *Data analysis*

Taxonomy was assigned to clone library sequences using BLAST with a minimum *e*-value cut-off of $1e^{-12}$, minimum identity of 88%, and a word size of 38 against the Greengenes database and the Hugenholtz taxonomic nomenclature (DeSantis *et al*., 2006). For the pyrosequencing data, sequences were removed from the analysis if they were $< 200$ or $> 400$ nt, had a quality score $< 25$, contained ambiguous characters, contained an uncorrectable barcode, or did not contain the primer sequence. For the clone library data set, OTUs were selected at four different levels of minimum sequence identity: 98%, 95%, 92% and 89% using cd-hit (Li and Godzik, 2006). The total number of OTUs at each level of phylogenetic resolution was 14 627 (98%), 9479 (95%), 6348 (92%) and 4383 (89%). For the pyrosequencing data sets we only classified OTUs at the 97% similarity level because of difficulties with consistency when examining shorter pyrosequenced fragments at different levels of phylogenetic resolution as compared with full-length 16S rRNA genes (Elshahed *et al*., 2008). For the Lauber and colleagues (2009) data set, 50 891 OTUs were examined while 10 374 OTUs were obtained from the tropical forest data set.

For each data set (clone library data and both of the pyrosequencing data sets) an assemblage by OTU presence–absence matrix was created. For the pyrosequencing data sets, matrices containing the relative sequence abundances of different OTUs in different assemblages were also created. Distribution analysis within and between assemblages and habitats was performed in Microsoft Excel and in MySQL using phpMyAdmin as a graphical user interface. We then tested the significance of macro-scale habitat in structuring the presence/absence of bacteria at all levels of phylogenetic resolution in the clone library data using PERMANOVA analyses in PRIMER v6 (Anderson, 2001).

## Acknowledgements

The authors recognize support from the National Science Foundation Grants DEB-0136957 and DEB-0852916. Antonio Peña-Gonzales provided support and advice implementing MySQL. The authors also wish to thank an anonymous reviewer for providing a thorough and thoughtful critique of the manuscript.

## References

Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecol* **26:** 32–46.

Ashby, M.N., Rine, J., Mongodin, E.F., Nelson, K.E., and Dimster-Denk, D. (2007) Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities. *Appl Environ Microbiol* **73:** 4532–4542.

Auguet, J.-C., Barberan, A., and Casamayor, E.O. (2010) Global ecological patterns in uncultured Archaea. *ISME J* **4:** 1–9.

Brown, J.H. (1984) On the relationship between abundance and distribution of species. *Am Nat* **124:** 255–279.

Brown, J.H., and Lomolino, M.V. (1998) *Biogeography*, 2nd edn. Sunderland, MA, USA: Sinauer.

Burtscher, M.M., Zibuschka, F., Mach, R.L., Lindner, G., and Farnleitner, A.H. (2009) Heterotrophic plate count vs. *in situ* bacterial 16S rRNA gene amplicon profiles from drinking water reveal completely different communities with distinct spatial and temporal allocations in a distribution net. *Water SA* **35:** 495–504.

Cho, J.C., and Tiedje, J.M. (2000) Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. *Appl Environ Microbiol* **66:** 5448–5456.

Cleveland, C.C., and Townsend, A.R. (2006) Nutrient additions to a tropical rain forest drive substantial soil carbon dioxide losses to the atmosphere. *Proc Natl Acad Sci USA* **103:** 10316–10321.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009) Bacterial community variation in human body habitats across space and time. *Science* **326:** 1694–1697.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99:** 10494–10499.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., *et al*. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72:** 5069–5072.

Dojka, M.A., Harris, J.K., and Pace, N.R. (2000) Expanding the known diversity and environmental distribution of an uncultured phylogenetic division of bacteria. *Appl Environ Microbiol* **66:** 1617–1621.

Elshahed, M.S., Youssef, N.H., Spain, A.M., Sheik, C., Najar, F.Z., Sukharnikov, L.O., *et al*. (2008) Novelty and uniqueness patterns of rare members of the soil biosphere. *Appl Environ Microbiol* **74:** 5422–5428.

Fenchel, T., and Finlay, B.J. (2004) The ubiquity of small species: patterns of local and global diversity. *Bioscience* **54:** 777–784.

Fierer, N., and Jackson, R.B. (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103:** 626–631.

Fierer, N., Hamady, M., Lauber, C.L., and Knight, R. (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci USA* **105:** 17994–17999.

Fulthorpe, R.R., Roesch, L.F.W., Riva, A., and Triplett, E.W. (2008) Distantly sampled soils carry few species in common. *ISME J* **2:** 901–910.

Green, J., and Bohannan, B.J.M. (2006) Spatial scaling of microbial biodiversity. *Trends Ecol Evol* **21:** 501–507.

Green, J.L., Bohannan, B.J.M., and Whitaker, R.J. (2008) Microbial biogeography: from taxonomy to traits. *Science* **320:** 1039–1043.

Grundmann, G.L. (2004) Spatial scales of soil bacterial diversity – the size of a clone. *FEMS Microbiol Ecol* **48:** 119–127.

Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J., and Knight, R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5:** 235–237.

Horner-Devine, M.C., Lage, M., Hughes, J.B., and Bohannan, B.J.M. (2004) A taxa-area relationship for bacteria. *Nature* **432:** 750–753.

Horner-Devine, M.C., Silver, J.M., Leibold, M.A., Bohannan, B.J.M., Colwell, R.K., Fuhrman, J.A., *et al*. (2007) A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* **88:** 1345–1353.

Hugenholtz, P., Goebel, B.M., and Pace, N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180:** 4765–4774.

Janssen, P.H. (2006) Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* **72:** 1719–1728.

Lauber, C.L., Hamady, M., Knight, R., and 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:** 5111–5120.

Li, W.Z., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658–1659.

Lozupone, C.A., and Knight, R. (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104:** 11436–11440.

Madigan, M., Martinko, J., and Parker, J. (2000) *Brock Biology of Microorganisms*. Upper Saddle River, NJ, USA: Prentice-Hall.

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., *et al*. (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4:** 102–112.

Newton, R.J., Jones, S.E., Helmus, M.R., and McMahon, K.D. (2007) Phylogenetic ecology of the freshwater *Actinobacteria* acI lineage. *Appl Environ Microbiol* **73:** 7169–7176.

Noguez, A.M., Arita, H.T., Escalante, A.E., Forney, L.J., Garcia-Oliva, F., and Souza, V. (2005) Microbial macroecology: highly structured prokaryotic soil assemblages in a tropical deciduous forest. *Glob Ecol Biogeogr* **14:** 241–248.

Pommier, T., Canback, B., Riemann, L., Bostrom, K.H., Simu, K., Lundberg, P., *et al*. (2007) Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16:** 867–880.

Preston, F.W. (1948) The commonness, and rarity, of species. *Ecology* **29:** 254–283.

Prinzing, A., Ozinga, W.A., and Durka, W. (2004) The relationship between global and regional distribution diminishes among phylogenetically basal species. *Evolution* **58:** 2622–2633.

Prosser, J.I., Bohannan, B.J.M., Curtis, T.P., Ellis, R.J., Firestone, M.K., Freckleton, R.P., *et al*. (2007) Essay – the role of ecological theory in microbial ecology. *Nat Rev Microbiol* **5:** 384–392.

Roesch, L.F., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K.M., Kent, A.D., *et al*. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1:** 283–290.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al*. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5:** 398–431.

Schloss, P.D., and Handelsman, J. (2004) Status of the microbial census. *Microbiol Mol Biol Rev* **68:** 686–691.

Sloan, W.T., Lunn, M., Woodcock, S., Head, I.M., Nee, S., and Curtis, T.P (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol* **8:** 732–740.

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

Spain, A.M., Krumholz, L.R., and Elshahed, M.S. (2009) Abundance, composition, diversity and novelty of soil *Proteobacteria*. *ISME J* **3:** 992–1000.

Stackebrandt, E., and Goebel, B.M. (1994) A place for DNA–DNA reassociation and 16S ribosomal-RNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44:** 846–849.

Tanner, M.A., Goebel, B.M., Dojka, M.A., and Pace, N.R. (1998) Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Appl Environ Microbiol* **64:** 3110–3113.

Thompson, J.R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D.E., Benoit, J., *et al*. (2005) Genotypic diversity

© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, *Environmental Microbiology*

within a natural coastal bacterioplankton population. *Science* **307:** 1311–1313.

Van der Gucht, K., Cottenie, K., Muylaert, K., Vloemans, N., Cousin, S., Declerck, S., *et al*. (2007) The power of species sorting: local factors drive bacterial community composition over a wide range of spatial scales. *Proc Natl Acad Sci USA* **104:** 20404–20409.

Wawrik, B., Kudiev, D., Abdivasievna, U.A., Kukor, J.J., Zystra, G.J., and Kerkhof, L. (2007) Biogeography of actinomycete communities and type II polyketide synthase genes in soils collected in New Jersey and Central Asia. *Appl Environ Microbiol* **73:** 2982–2989.

Whitaker, R.J., Grogan, D.W., and Taylor, J.W. (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* **301:** 976–978.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95:** 6578–6583.

Wieder, W.R., Cleveland, C.C., and Townsend, A.R. (2009) Controls over leaf litter decomposition in wet tropical forests. *Ecology* **90:** 3333–3341.

## Supporting information

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

**Table S1.** Information on the clone library-based studies used in this work.

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

© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd, *Environmental Microbiology*