ECOLOGY LETTERS

Ecology Letters, (2014) doi: 10.1111/ele.12282

LETTER

Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria

Albert Barberán,[1] Kelly S. Ramirez,[3] Jonathan W. Leff,[1,2] Mark A. Bradford,[4] Diana H. Wall[3] and Noah Fierer[1,2]*

[1]*Cooperative Institute for Research in Environmental Sciences University of Colorado Boulder, USA*
[2]*Department of Ecology and Evolutionary Biology University of Colorado Boulder, USA*
[3]*School of Global Environmental Sustainability and Department of Biology Colorado State University Fort Collins, USA*
[4]*School of Forestry and Environmental Studies Yale University New Haven, CT 06511, USA*

**Correspondence:*
E-mail: Noah.Fierer@colorado.edu

Abstract

Identifying the traits that determine spatial distributions can be challenging when studying organisms, like bacteria, for which phenotypic information is limited or non-existent. However, genomic data provide another means to infer traits and determine the ecological attributes that account for differences in distributions. We determined the spatial distributions of ~124 000 soil bacterial taxa across a 3.41 km^2 area to determine whether we could use phylogeny and/or genomic traits to explain differences in habitat breadth. We found that occupancy was strongly correlated with environmental range; taxa that were more ubiquitous were found across a broader range of soil conditions. Across the ~500 taxa for which genomic information was available, genomic traits were more useful than phylogeny alone in explaining the variation in habitat breadth; bacteria with larger genomes and more metabolic versatility were more likely to have larger environmental and geographical distributions. Just as trait-based approaches have proven to be so useful for understanding the distributions of animals and plants, we demonstrate that we can use genomic information to infer microbial traits that are difficult to measure directly and build trait-based predictions of the biogeographical patterns exhibited by microbes.

Keywords

Bacteria, functional traits, genome size, geographical distribution, habitat breadth, microbial ecology, phylogeny, soil.

Ecology Letters (2014)

INTRODUCTION

Like plants and animals, not all microbial taxa are found everywhere. Some microbial taxa can thrive, or at least tolerate, a broad range of environmental conditions and are more likely to be ubiquitous. In contrast, other taxa can only persist under a very specific set of environmental conditions and subsequently have far more restricted ranges and a high degree of endemicity. However, the factors that explain these differences in environmental and/or geographical distributions across microbial taxa remain unresolved. These types of questions have been asked for decades by ecologists seeking to understand which ecological, phylogenetic, or life-history attributes lead to some plant and animal taxa having larger distributions than others (Brown *et al*. 1996). The answers to these questions often remain elusive and dependent on the taxon in question, with recent work highlighting that the factors determining range size are often not well understood even for those organisms that have been well-studied (Lester *et al*. 2007).

In general, two factors might explain variation in geographical distributions across different taxa: shared evolutionary history or shared functional traits. Although these two factors are not necessarily independent (Losos 2008), we can determine the relative importance of each factor in predicting an ecological attribute of interest. If evolutionary history is more important, we would expect closely related taxa to have more similar distributions. We know from research on plants and animals that evolutionary history is often not a good predictor of observed differences in range sizes given that range size can vary considerably between closely related plant and animal species (Brown *et al*. 1996; Webb & Gaston 2003). We do not know if this holds true for bacteria where evolutionary history may be more useful for predicting bacterial distributions given that some key ecological attributes appear to be conserved across major bacterial lineages (e.g. Philippot *et al*. 2010). However, we would expect that the decoupling of ecological attributes from phylogeny to be more conspicuous in microbes for several reasons. First, since the domain *Bacteria* comprises high levels of phylogenetic diversity, taxa within the same genus or family may have very different ecological roles. Evolutionary processes have generated a remarkable amount of bacterial metabolic and functional diversity and thus, many optimal ecological strategies exist that could lead to some bacteria having larger environmental and geographical distributions than other bacteria. Second, microbes can rapidly gain and lose genes via horizontal gene transfer, thus introducing genetic connections among distantly related taxa (Cordero & Hogeweg 2009) and promoting a high degree of functional plasticity or functional convergence in disparate lineages (Tettelin *et al*. 2008).

Functional traits provide an alternative explanation for the variation in environmental and geographical breadth across taxa. Taxa that share similar range sizes may not have shared evolutionary histories, but they could share similar functional traits. Such functional trait-based explanations have been widely used by ecologists to identify the phenotypic characteristics of organisms that predict distribution patterns (McGill *et al*. 2006). For example, while dispersal ability is often claimed to be the main determinant of a species' range (Brown *et al*. 1996; Lester *et al*. 2007), other traits like body size or fecundity have also been shown to influence range size (Laube *et al*. 2013). Such trait-based approaches are less frequently

applied to microbial communities because their phenotypic characteristics often remain undetermined, but recent work has demonstrated how a trait-based framework can be used to understand bacterial community assembly (e.g. Burke *et al.* 2011; Barberán *et al.* 2012) and to explain shifts in microbial community composition across environmental gradients (Edwards *et al.* 2013). To apply a trait-based perspective to complex microbial communities, one must first identify which traits to investigate and how the relevant traits can be measured. These tasks are arguably far more difficult to do with microbes than with plants or animals for several reasons. First, sampling individual microorganisms to estimate intraspecific trait variation is often not feasible, particularly in environments like soil where cells cannot be readily separated from the surrounding matrix. For example, even simple traits like cell size are difficult to measure for individual cells of a given soil microbial taxon (Portillo *et al.* 2013). Moreover, measuring traits of uncultured microorganisms is also a challenging task given their extraordinary phenotypic diversity and the preponderance of novel metabolic pathways. Even for the small minority of microbial taxa that can be readily cultured and studied in the lab, the phenotypic traits they exhibit in culture may bear little resemblance to the traits they exhibit when growing in the environment (Lennon *et al.* 2012). Perhaps most importantly, microbial ecologists often lack a fundamental understanding of which characteristics or specific traits are most strongly linked to the performance of microbes (Green *et al.* 2008). Unlike plants, where a handful of leaf traits have proven useful for describing their ecological strategies (Wright *et al.* 2004), we often do not know what traits are important in differentiating microbial taxa or how to effectively measure those traits.

Genomic information provides a tractable starting point for identifying and quantifying relevant ecological traits of microorganisms when phenotypic information is missing. As some genomic traits are shared across the entire tree of life, they can be used to understand the trade-offs associated with particular strategies that contribute to eco-evolutionary adaptation (Gudelj *et al.* 2010; Verberk *et al.* 2013). Recent studies with microbial communities have shown a closer relationship between ecology and genomic functional potential than between ecology and phylogeny (Burke *et al.* 2011; Barberán *et al.* 2012). Furthermore, there are associations between genomic traits and general microbial lifestyles (Garcia *et al.* 2008), trophic strategies in marine bacteria (Lauro *et al.* 2009), microbial growth rates (Vieira-Silva & Rocha 2010) and the successional status of human gut symbionts (Lozupone *et al.* 2012).

Here we investigated how information on phylogeny or traits (inferred from genomic data) can be used to explain the degree of ubiquity (i.e. habitat breadth) of microbial taxa living in soil, an environment which houses large amounts of novel and understudied microbial diversity (Torsvik *et al.* 1990; Fierer *et al.* 2007). To do so, we collected 596 soil samples from a 3.41-km^2 landscape (Central Park in New York City) and deeply sequenced a portion of the 16S rRNA gene to identify bacterial taxa and measured the breadth of habitats across which each taxon could be found (measured as occupancy and environmental range size). We then matched the identified bacterial taxa to those with fully sequenced genomes to: (1) assess whether phylogenetic relatedness and/or genomic traits explain why some soil bacterial taxa can persist in a wide range of soil habitats, while others are far more restricted in their distributions; and (2) determine which genomic traits, if any, predict habitat breadth. Given that the degree of similarity between genomes is not independent from evolutionary history (Snel *et al.* 1999), we constructed a model to (3) estimate how phylogenetic relatedness and genomic traits can together be used to predict the environmental and geographical distribution of soil microorganisms.

MATERIALS AND METHODS

Data set, sampling and molecular analyses

Central Park (New York City, USA) was established in 1857 and is a 3.41 km^2, 0.80 km wide by 4.02 km long, urban park. The park is a useful site for our examination of microbial environmental and geographical breadth given that it has clearly defined boundaries yet the landscape within the urban park is highly heterogeneous, ranging from large lawns to dense forests. Although climatic conditions are essentially invariant across Central Park, soil edaphic characteristics are highly variable (Fig. S1). For example, soil pH values across Central Park range from 3.86 to 8.35, organic carbon concentrations (C) range from 1.19 mg g^{-1} soil to 139.1 mg g^{-1} soil and organic nitrogen concentrations (N) range from 0.06 mg g^{-1} soil to 4.91 mg g^{-1} soil, ranges in soil edaphic characteristics that match ranges observed across a broad array of soil types collected from across North and South America (Fierer & Jackson 2006).

We collected soils from 596 locations within the park with a sample collected nearly every 50 m^2 of land surface across the entire park. At each of the 596 sampling locations, four soil cores (2.54 cm diameter by 5 cm depth) were composited together to yield one soil sample per sampling location. Genomic DNA was extracted using the MoBio PowerSoil DNA extraction kit as described previously (Fierer *et al.* 2012) with the DNA from each soil sample amplified and sequenced on an Illumina HiSeq2000 following standard protocols (Caporaso *et al.* 2012). The V4–V5 region of the 16S rRNA gene was amplified in triplicate using the F515 and R806 primer set. This primer set has few biases against specific taxa (Walters *et al.* 2011) and the sequenced gene region has been shown to provide accurate taxonomic and phylogenetic information for bacteria (Liu *et al.* 2007).

16S rRNA gene analyses

The raw sequence data were processed using the UPARSE pipeline (Edgar 2013). Sequences were truncated to 150 bp and quality filtered sequences were clustered into operational taxonomic units (OTUs) using the consensus ≥97% 16S rRNA gene identity as a threshold, and unique sequences (i.e. singletons) were removed. Taxonomic assignment was carried out with the Ribosomal Database Project (RDP) classifier (Wang *et al.* 2007) against the March 2013 version of the Greengenes reference database (McDonald *et al.* 2012). As raw counts can vary

by orders of magnitude from the same sequencing run, communities were rarefied to 40 000 sequences per sample yielding a total of ~124 000 bacterial OTUs identified across the park. We calculated the occupancy (i.e. the number of sites out of 596 where a particular OTU was found) and the environmental range (i.e. the average of the range in soil edaphic factors reported in Fig. S1 standardised from 0 to 1) for each OTU.

Matching 16S rRNA genes to sequenced genomes

For most of the ~124 000 bacterial OTUs, no genomic information was available as there were no closely related taxa for which whole-genome sequence data are available. However, some of these OTUs were closely related to representatives with sequenced bacterial genomes. To identify those OTUs for which genomic information was available to determine genomic traits, we matched the 16S rRNA representative sequences from Central Park soils against a subset of the Ribosomal Database Project (RDP; Cole *et al.* 2009) derived from representatives of sequenced bacterial genomes (~2000 complete 16S rRNA genes) using UCLUST (Edgar 2010) at $\geq$ 99% identity (i.e. allowing just one base mismatch between subject and query sequences). For all the OTUs and for that subset of OTUs for which genomic information was available, the range of mean abundances, the range in occupancy levels and the relationship between abundance and occupancy were very similar (Fig. S2) indicating that the OTUs for which we had genomic trait data did not represent a biased subset of taxa as far as the variation in environmental and geographical distribution was concerned.

For the ~500 OTUs for which genomic information was available, we determined structural genomic characteristics (genome size, number of genes, guanine-cytosine (G+C) content, number of rRNA genes, % of coding DNA sequence, % of signal peptides, % of transmembrane proteins) and the abundance of different protein families' functional domains (~8000 Pfam domains in total; Finn *et al.* 2008) by downloading this information from the Integrated Microbial Genomes (IMG) database (Markowitz *et al.* 2012).

Phylogenetic analyses

To infer the evolutionary relationships between the ~124 000 OTUs identified from Central Park and the ~500 OTUs for which genomic information was available, we aligned the 16S rRNA gene sequences using PyNAST (Caporaso *et al.* 2010) with the Greengenes database (McDonald *et al.* 2012) as a template. The resulting multiple sequence alignment was subsequently trimmed to remove non-informative positions (i.e. positions which are gaps in every sequence). Phylogenies were reconstructed with the FastTree approximate maximum-likelihood algorithm (Price *et al.* 2010) using the mid-point method for rooting.

Functional (genomic) analyses

To identify which protein families (Pfam) tended to be over- or under-represented in genomes that had larger occupancy (more ubiquitous taxa) compared to those taxa with smaller range sizes, we used Mann–Whitney tests corrected for multiple comparisons by False Discovery Rate (FDR; Benjamini & Hochberg 1995). Pfam domains were mapped to KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology (KO) pathway groups, and then the KO pathways were grouped into more general functional categories using the BRITE hierarchy (Kanehisa *et al.* 2012). As the same Pfam domain can match several KO pathways, we used MinPath to yield a more conservative estimation of the biological pathways present (Ye & Doak 2009). Given a set of protein domains (Pfam) that can be mapped to one or more pathways (KO), the MinPath parsimony algorithm attempts to find the minimum number of pathways that can explain the presence of all protein domains.

Path model

We explored the relationships between evolutionary history (phylogenetic relatedness), genomic traits and habitat breadth (occupancy and environmental range) using Partial Least Squares Path Modeling (PLS-PM). PLS-PM is a statistical method for studying cause and effect relationships among observed and latent variables, and is particularly useful when a theoretical understanding of the relationships between variables is scarce because it does not impose any distributional assumption (Tenenhaus *et al.* 2005). Since PLS-PM does not rely on any distributional assumptions, we ran the path model using 1000 bootstraps to validate the estimates of path coefficients and the coefficients of determination (R^2). Path coefficients (i.e. standardised partial regression coefficients) represent the direction and strength of the linear relationships between variables (direct effects). Indirect effects are the multiplied path coefficients between a predictor and a response variable, adding the product of all possible paths excluding the direct effect.

We selected observed variables based on collinearity and predictability power. Five latent variables were used: phylogeny (the first two non-metric multidimensional axes of the cophenetic matrix derived from the phylogenetic tree constructed from complete sequences of 16S rRNA genes of sequenced genomes), functional potential (the number of different Pfam domains and the first non-metric multidimensional axis of the Pfam domain categories table), genome structure (genome size, % G+C content, % coding DNA sequence and % signal peptides), genomic traits (a combination of the first principal component analysis score for functional potential and genomic structure) and habitat breadth (occupancy and environmental range). Models with different structures were evaluated using the Goodness of Fit (GoF) statistic, a measure of their overall predictive power. The R package plspm (Sanchez & Trinchera 2012) was used to construct the model.

RESULTS AND DISCUSSION

Environmental range size correlates with occupancy

We observed a strong correlation between occupancy (the number of sites where a taxon is present) and environmental range (the breadth of environmental conditions where a taxon is present). This correlation between occupancy and environmental range was observed when we considered all of the OTUs (Spearman's $\rho = 0.95$, $P < 0.0001$) or just the subset of

OTUs for which genomic information was available ($\rho = 0.97$, $P < 0.0001$; Fig. 1). Essentially, those bacterial taxa that were found in more samples were also those taxa that persisted under a wider range of soil environmental conditions. For smaller organisms like bacteria, geographical range size can be more difficult to measure than the range sizes of plants and animals (Brown *et al*. 1996; Slatyer *et al*. 2013) given that it is often difficult to know the appropriate spatial scale(s) to sample and because the high diversity of bacterial communities make it difficult to determine whether a given taxon is truly absent from a sample. Nevertheless, given the close correspondence between occupancy and environmental range in our data, we can assume that occupancy is a reasonable proxy for habitat breadth (i.e. how generalised or specialised a biological entity is in its habitat requirements).

Evolutionary history cannot explain differences in occupancy

For both total OTUs and those OTUs with sequenced genomes, occupancy showed a similar highly skewed distribution with relatively few taxa present in most of the sites (Fig. 2a). For example, only 2% of all OTUs were found in >300 of the 596 soils that were collected from Central Park and 58% of all OTUs were restricted in their distribution (i.e. defined as being found in <10 of the soils). The differences observed in the distribution of occupancy values between all OTUs and for those with sequenced genomes (Fig. 2a) were likely a product of differences in sample size (~124 000 and ~500 respectively) rather than any underlying ecological process. Given these patterns, we then set out to determine why some taxa were found across a broader set of environmental conditions than other taxa. We started by determining if evolutionary patterns can explain the occupancy patterns shown in Fig. 2a. We expected that evolutionary history, expressed as phylogenetic relatedness, may explain differences in habitat breadth for soil bacteria because general ecological attributes are often conserved across major bacterial lineages (Fierer *et al*. 2007; Philippot *et al*. 2010). However, we did not observe any relationship between 16S rRNA phylogenetic distance and pairwise differences in occupancy for all OTUs and for that subset of OTUs with sequenced genomes (Mantel's $r_M = -0.05$ in both cases; Fig. 2b and c respectively). Furthermore, the non-association between phylogeny and occupancy is evident from the observation that OTUs affiliated with a given phylum typically had a wide range of occupancy values (Fig. 2a). Thus, we can conclude that phylogenetic relatedness, by itself, is not a good predictor of habitat breadth, despite expectations that environmental conditions may limit the types of microorganisms that inhabit a certain soil habitat (Fierer *et al*. 2007) and that some bacterial phyla have conserved ecological features (Philippot *et al*. 2010).

For plants and animals, geographical range sizes can vary considerably between closely related species (Brown *et al*. 1996), even though those morphological, physiological or life-history traits that might determine range size tend to be phylogenetically conserved (Webb & Gaston 2003). For microorganisms, this decoupling might be more prominent due to their high phylogenetic and metabolic diversity, and their high degree of functional plasticity (Tettelin *et al*. 2008). In addition, those ecological attributes that do tend to be conserved within bacterial phyla (e.g. oxygenic photosynthesis or methanogenesis; Martiny *et al*. 2013) may not necessarily be relevant for predicting occupancy patterns. Given that phylogeny does not help us predict occupancy patterns,

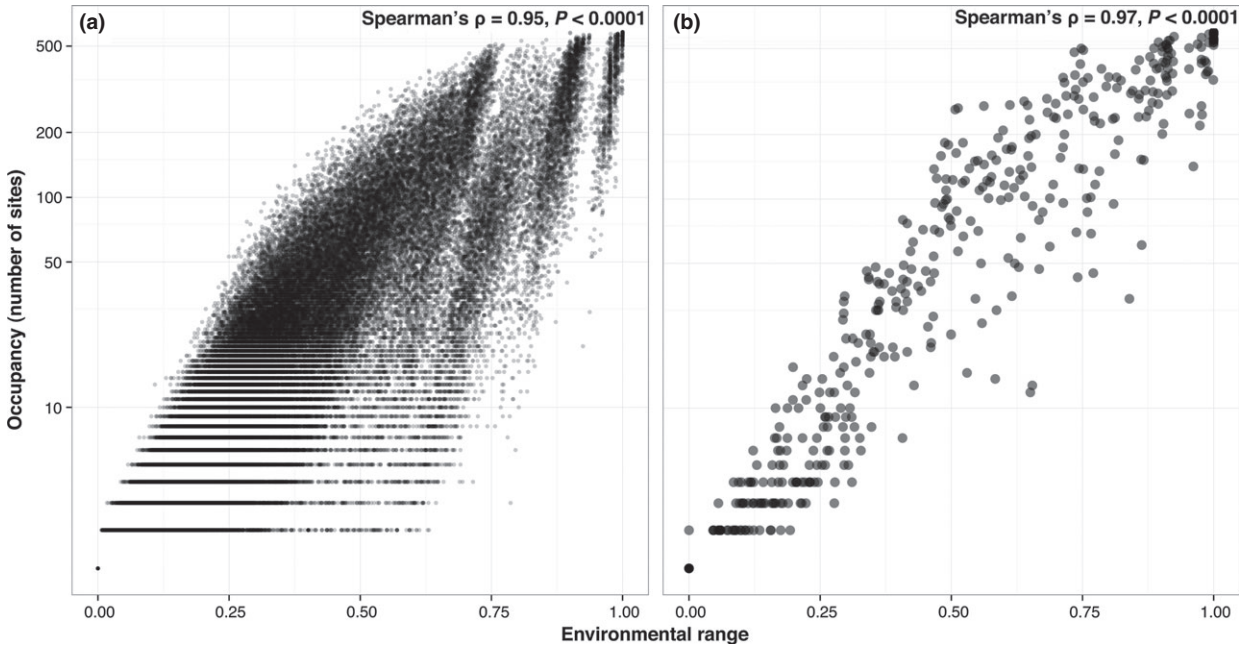

Figure 1 Correlation between occupancy and environmental range (where environmental range is calculated as the average of the range in soil edaphic factors standardised from 0 to 1) for the ~124 000 OTUs (a) and for that subset of OTUs (~500) for which genomic information was available (b).

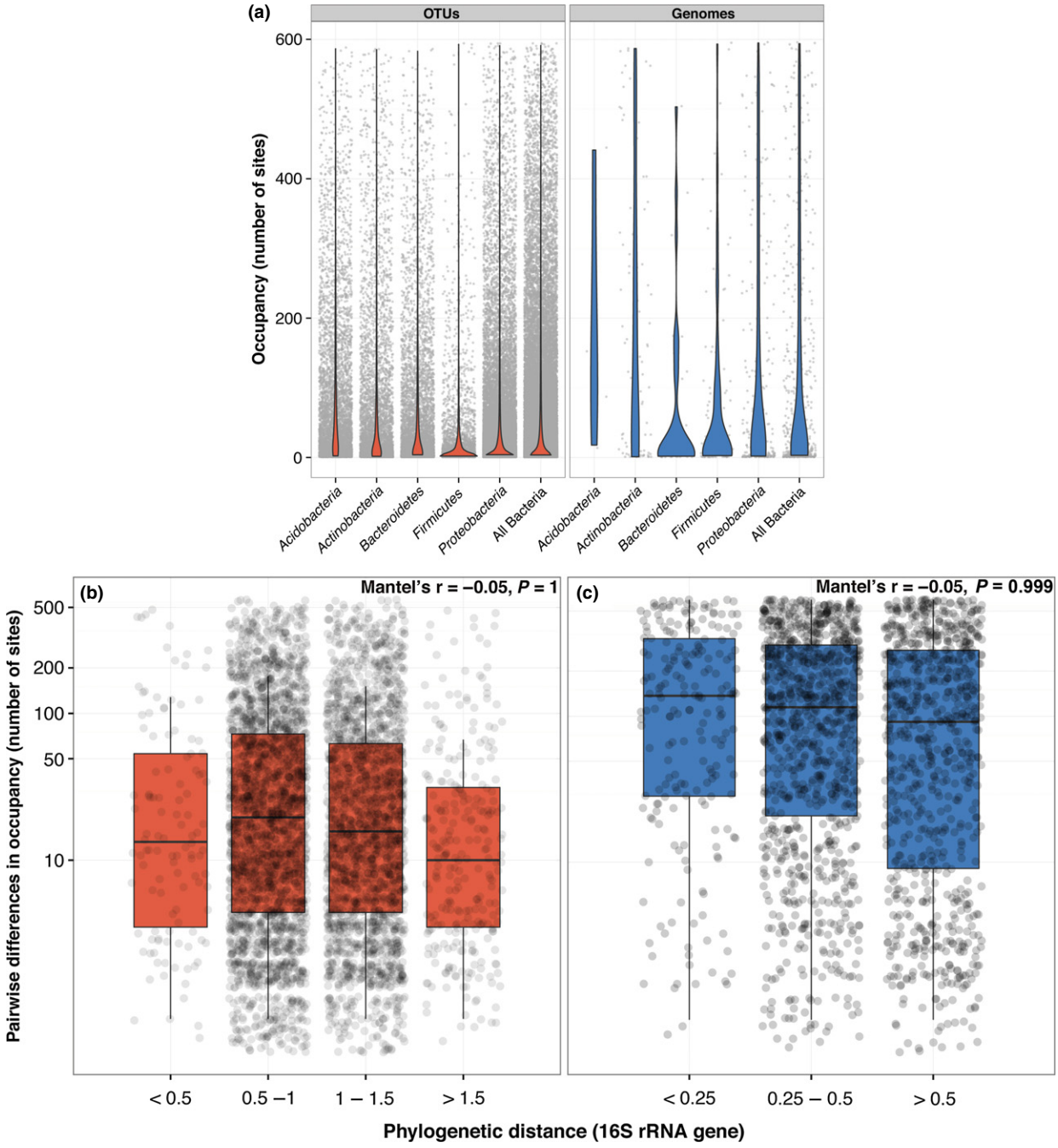

Figure 2 Distribution of occupancy for the total number of OTUs (in red) and the subset of sequences that matched sequenced bacterial genomes (in blue) (a). Relationship between 16S rRNA phylogenetic distance and pairwise differences in occupancy for all of the OTUs in red (b), and for that subset of OTUs for which genomic information was available in blue (c). Note that the range of phylogenetic distances for all of the OTUs is larger than for that subset of OTUs for which genome sequence data were available.

we then asked if we could use functional traits (gleaned from genomic information) to better explain the differences in habitat breadth.

Genomic traits can partially explain differences in occupancy

We hypothesised that habitat breadth of individual bacterial taxa would be better predicted by incorporating information on genomic traits instead of phylogeny alone, a poor predictor of range size (Fig. 2). Recent studies on microbial communities have shown that genomic functional potential rather than taxonomic composition tends to correlate better with biogeographical patterns (Burke *et al.* 2011; Barberán *et al.* 2012). For example, Barberán *et al.* (2012) showed that 16S rRNA gene-based taxonomic community structure failed to differentiate among marine habitats, while community patterns based on

genomic traits could better discriminate among those habitats. Just as trait-based analyses have been used by plant and animal ecologists to predict distribution patterns that are not well-explained by taxonomy or phylogeny alone (e.g. Laube *et al.* 2013), we hypothesised that genomic traits could be used to predict why some soil bacteria are more ubiquitous than others.

We observed significant relationships between individual genomic traits and occupancy (as one surrogate of habitat breadth, Fig. 1). Genome size (a basic structural characteristic) and the number of different protein family (Pfam) domains (a measure of the genomic functional richness) were both positively correlated with occupancy (Fig. 3). That is, bacteria with larger genomes and more metabolic and functional capabilities (both attributes that were highly co-correlated; $\rho = 0.80$, $P < 0.0001$) tended to occur across more of the samples than bacteria with smaller genomes. For microbes, genome size is a product of vertical inheritance, horizontal gene transfer and gene duplications and losses (Koonin & Wolf 2008). Thus, microbial genome size at least partially reflects an adaptation to the external complexity imposed by the organism's lifestyle and ecology (Konstantinidis & Tiedje 2004). That is, organisms inhabiting variable, heterogeneous environments, such as free-living microorganisms in soil, tend to have larger and more versatile genomes than organisms thriving in a constant or more stable habitat, such as parasites living inside a host (Guieysse & Wuertz 2012). In addition, larger genomes tend to be more susceptible to horizontal gene transfer from distantly related organisms, leading to the hypothesis that there is a correlation between genome size and environmental heterogeneity: more diverse communities living in complex environments increase the demand for larger gene repertoires, which are expanded by increased uptake of genes from phylogenetically distant organisms interacting in the same environment (Cordero & Hogeweg 2009). Although this hypothesis has not yet been validated, it is worth noting that many of the largest microbial genomes are from soil-associated microbes, including *Myxobacteria*, *Pseudomonas, Bradyrhizobium* and *Streptomyces* (Konstantinidis & Tiedje 2004), suggesting that environments like soil with a high degree of phylogenetic diversity (Torsvik *et al.* 1990) may select for organisms with larger genome sizes.

The presence of specific functional gene categories (based on Pfam domains) was significantly correlated with different levels of occupancy (Mann–Whitney test, $P < 0.001$ after FDR correction) (Fig. 4). For example, a number of protein families were more common in those taxa with large range sizes, including the phosphotransferase enzyme involved in antibiotic resistance (PF01636; Fig. 4a), the conserved domain containing the catalytic phosphorylation function of protein kinases (PF00069; Fig. 4b), the cytochrome P450 which catalyses the oxidation of organic substance such as xenobiotics (PF00067; Fig. 4c), and the transglycosylase domain found in secretion systems (PF01464; Fig. 4d). Overall, we observed that the functions present in more ubiquitous taxa tended to be involved in a higher proportion of pathways related to amino acid metabolism and xenobiotics biodegradation compared to the pathways predicted by pooling all the analysed taxa together (Fig. 4e, Figs S3 and S4 for the detailed cases of amino acid and xenobiotics metabolism). These functional differences in genome properties agree with our finding that ubiquitous soil bacteria tended to have bigger and more versatile genomes (Fig. 3). Konstantinidis & Tiedje (2004) showed that bacteria with larger genomes preferentially accumulated regulatory and secondary metabolism-related genes (Metabolism categories in Fig. 4e), while being depleted in informational and DNA metabolism-related genes (Genetic information categories in

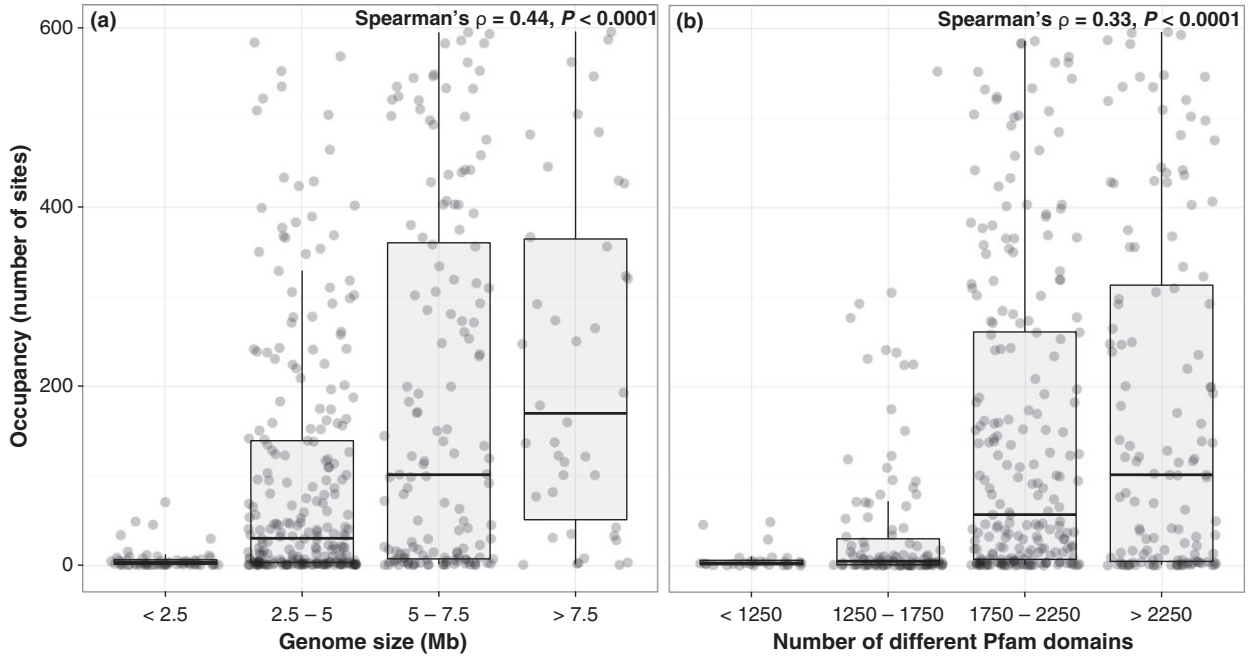

Figure 3 Relationship between occupancy and the genomic traits of genome size (a) and number of different protein families (Pfam) domains (b).

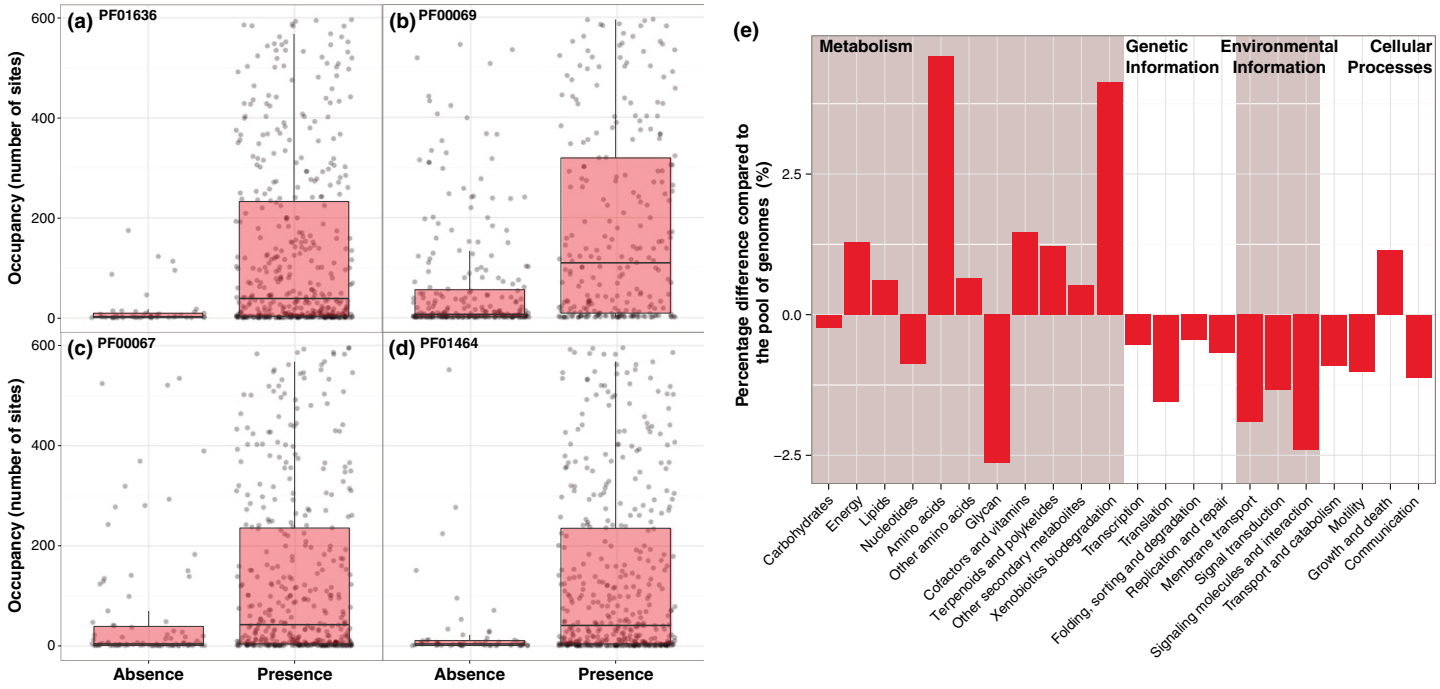

Figure 4 Differences in occupancy and the presence of different protein families (Pfam) domains: PF01636 (phosphotransferase) (a), PF00069 (protein kinase) (b), PF00067 (cytochrome P450) (c) and PF01464 (transglycosylase) (d). Barplot of the different abundance of functional categories from genomes with higher occupancy compared to all the analysed genomes pooled together (Mann–Whitney test; $P < 0.001$ after FDR correction) (e).

Fig. 4e). One explanation for this phenomenon is that larger genomes might be more adaptive in heterogeneous, but resource-scarce, environments like soil where the selection pressure for fast growth is relaxed and instead there is selection for oligotrophs able to cope with environmental stress (Konstantinidis & Tiedje 2004; Guieysse & Wuertz 2012). We extend their initial results and propose that larger, metabolically diverse and informationally depleted genomes are particularly common among those soil bacteria that are more ubiquitous and have larger environmental range sizes.

Integrating phylogenetic and genomic trait information to predict habitat breadth

As noted above, genomic traits were a far better predictor of occupancy than phylogeny. However, these two factors are not independent. The degree of similarity between genomes will often be a function of the time since divergence from the common ancestor. Because gene content and many metabolic traits are phylogenetically conserved (Snel *et al.* 1999; Martiny *et al.* 2013), it is not surprising that we also observed a significant relationship between 16S rRNA phylogenetic distance and functional distance based on the profile of protein families (Pfam) domains ($r_M = 0.70$, $P < 0.001$; Fig. S5). To better integrate these complex interrelationships, we constructed a partial least squares path model (PLS-PM) relating phylogeny and genomic traits to habitat breadth constructed from occupancy and environmental range (Fig. 5). Path models allow us to study cause and effect relationships among observed (indicators) and latent (constructs) variables (Tenenhaus *et al.* 2005). Instead of focusing on single trait relationships, our path model approach used latent variables as combinations of related traits. As highlighted in Verberk *et al.* (2013), environmental selection does not act independently on single traits, but rather, on individuals within species possessing a combination of interacting traits, and it is these combinations of traits that ultimately define ecological and life-history strategies.

In our path model (Fig. 5), genomic traits were defined as a composite of the latent variables genome structure (itself a combination of genome size, G + C content, % of coding DNA sequence, and % of signal peptides) and functional potential (a combination of protein families richness and composition). Genomic traits had an important (0.43 ± 0.03; Fig. S6) direct effect on the latent variable habitat breadth (a composite of occupancy and environmental range). In addition, genomic traits alone could explain $19 \pm 3\%$ of the variation in habitat breadth. Genomic structure had a larger indirect effect (0.38 ± 0.03) than functional potential (0.23 ± 0.02) on habitat breadth. We found an indirect effect (0.20 ± 0.02) of phylogeny on habitat breadth that was mediated by the genomic variables (Fig. 5). Thus, although phylogeny alone cannot explain habitat breadth patterns, patterns of common ancestry are often important for understanding how genomic traits are constrained by evolutionary history (Snel *et al.* 1999; Martiny *et al.* 2013). We acknowledge that the adaptive value of traits is context dependent. Although genomic traits could significantly explain differences in habitat breadth across soil bacteria, this association might change across different environments. Furthermore, the set of explanatory traits might be different in another habitat (e.g. marine systems) or if we were to examine other ecological attributes besides habitat breadth.

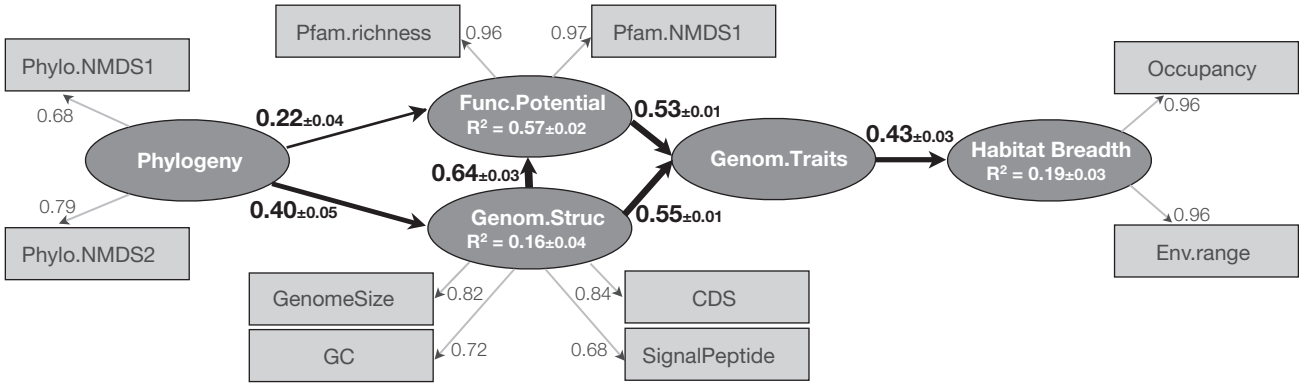

Figure 5 Directed graph of the Partial Least Squares Path Model (PLS-PM). Observed (i.e. measured) variables are represented in a rectangular form, while latent variables (i.e. constructs) are represented in an elliptical form. Indicated are the loadings (the correlations between a latent variable and its observed variables), the path coefficients and the coefficients of determination (R^2) calculated after 1000 bootstraps. Models with different structures were assessed using the Goodness of Fit (GoF) statistic, a measure of the overall prediction performance. For the best model represented here, the GoF was 0.59.

CONCLUDING REMARKS

The distributions of taxa across space and the ability of taxa to cope with different environmental conditions are fundamental ecological attributes of both microbial and 'macro'-bial organisms (Brown *et al.* 1996; Slatyer *et al.* 2013), yet the traits that determine them are not well understood even for organisms whose life-history characteristics are reasonably well known (Lester *et al.* 2007). For bacteria in the soil environment, we show that genomic traits can be used to predict why some microbial taxa are more ubiquitous than others. Habitat breadth was positively correlated with genome size and genes associated with specific functional capabilities, including secondary metabolic pathways. This ability to predict occupancy patterns from genomic traits is surprising given that it is the phenotypic traits that directly impact how bacteria interact with their environment. However, since we do not have phenotypic information for most soil bacteria, even for many of those bacterial taxa that have been cultured and sequenced, we have to rely on genomic traits as phenotypic information is often scarce or collected inconsistently across taxa, making direct phenotypic characterisation difficult. Therefore, this work highlights the potential for using genomic information to begin building a trait-based understanding of microbial ecology that has proven so useful for comprehending the distributions of animals and plants.

ACKNOWLEDGEMENTS

We thank Antonio Fernàndez-Guerra for advice on the use of MinPath, and Elise S. Gornish and the students from the 'Genomes and Traits' seminar at the University of Colorado for helpful comments. We also thank the Central Park Conservancy and the American Museum of Natural History for logistical assistance and Jessica Henley, Scott Bates, Jason Betley, Thomas Crowther, Eugene Kelly, Emily Oldfield, Ashley Shaw, and Chris Steenbock for their help with the sample collection and analyses. AB is supported by a James S. McDonnell (JSMF) Postdoctoral Fellowship. Funding for this study was provided to NF from the National Science Foundation (DEB0953331).

AUTHORSHIP

All authors were involved in the study design and contributed to the writing of the manuscript. JWL processed the 16S rRNA sequence data; KSR, MAB, DHW and NF led the sample collection; AB analysed the data and, along with NF, led the writing of the manuscript.

REFERENCES

Barberán, A., Fernández-Guerra, A., Bohannan, B.J.M. & Casamayor, E.O. (2012). Exploration of community traits as ecological markers in microbial metagenomes. *Mol. Ecol.*, 21, 1909–1917.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B*, 57, 289–300.

Brown, J.H., Stevens, G.C. & Kaufman, D.M. (1996). The geographic range: size, shape, boundaries, and internal structure. *Annu. Rev. Ecol. Syst.*, 27, 597–623.

Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S. & Thomas, T. (2011). Bacterial community assembly based on functional genes rather than species. *Proc. Natl. Acad. Sci. U S A*, 108, 14288–14293.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L. & Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26, 266–267.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N. *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.*, 6, 1621–1624.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J. *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, 37, D141–D145.

Cordero, O.X. & Hogeweg, P. (2009). The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci U S A*, 106, 21748–21753.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460–2461.

Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods*, 10, 996–998.

Edwards, K.F., Litchman, E. & Klausmeier, C.A. (2013). Functional traits explain phytoplankton responses to environmental gradients across lakes of the United States. *Ecology*, 94, 1626–1635.

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

Fierer, N., Bradford, M.A. & Jackson, R.B. (2007). Toward an ecological classification of soil bacteria. *Ecology*, 88, 1354–1364.
Fierer, N., Lauber, C.L., Ramirez, K.S., Zaneveld, J., Bradford, M.A. & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities. *ISME J.*, 6, 1007–1017.
Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R. *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res.*, 36, D281–D288.
Garcia, J.A.L., Bartumeus, F., Roche, D., Giraldo, J., Stanley, H.E. & Casamayor, E.O. (2008). Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses. *Genomics*, 91, 538–543.
Green, J.L., Bohannan, B.J.M. & Whitaker, R.J. (2008). Microbial biogeography: from taxonomy to traits. *Science*, 320, 1039–1043.
Gudelj, I., Weitz, J.S., Ferenci, T., Horner-Devine, M.C., Marx, C.J., Meyer, J.R. *et al.* (2010). An integrative approach to understanding microbial diversity: from intracellular mechanisms to community structure. *Ecol. Lett.*, 13, 1073–1084.
Guieysse, B. & Wuertz, S. (2012). Metabolically versatile large-genome prokaryotes. *Curr. Opin. Biotechnol.*, 23, 467–473.
Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40, D109–D114.
Konstantinidis, K.T. & Tiedje, J.M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A*, 101, 3160–3165.
Koonin, E.V. & Wolf, Y.I. (2008). Genomics of Bacteria and Archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, 36, 6688–6719.
Laube, I., Korntheuer, H., Schwager, M., Trautmann, S., Rahbek, C. & Böhning-Gaese, K. (2013). Towards a more mechanistic understanding of traits and range sizes. *Glob. Ecol. Biogeogr.*, 22, 233–241.
Lauro, F.M., McDougald, D., Thomas, T., Williams, T.J., Egan, S., Rice, S. *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci U S A*, 106, 15527–15533.
Lennon, J.T., Aanderud, Z.T., Lehmkuhl, B.K. & Schoolmaster, D.R. Jr (2012). Mapping the niche space of soil microorganisms using taxonomy and traits. *Ecology*, 93, 1867–1879.
Lester, S.E., Ruttenberg, B.I., Gaines, S.D. & Kinlan, B.P. (2007). The relationship between dispersal ability and geographic range size. *Ecol. Lett.*, 10, 745–758.
Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D. & Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res.*, 35, e120.
Losos, J.B. (2008). Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecol. Lett.*, 11, 995–1003.
Lozupone, C., Faust, K., Raes, J., Faith, J., Frank, D.N., Zaneveld, J. *et al.* (2012). Identifying genomic and metabolic features that can underlie early successional and opprtunistic lifestyles of human gut symbionts. *Genome Res.*, 22, 1974–1984.
Markowitz, V.M., Chen, I.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y. *et al.* (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.*, 40, D115–D122.
Martiny, A.C., Treseder, K. & Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *ISME J.*, 7, 830–838.
McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A. *et al.* (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, 6, 610–618.
McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends Ecol. Evol.*, 21, 178–185.
Philippot, L., Andersson, S.G.E., Battin, T.J., Prosser, J.I., Schimel, J.P., Whitman, W.B. *et al.* (2010). The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.*, 8, 523–529.
Portillo, M.C., Leff, J.W., Lauber, C.L. & Fierer, N. (2013). Cell size distributions of soil bacterial and archaeal taxa. *Appl. Environ. Microbiol.*, 79, 7610–7617.
Price, M.N., Dehal, P.S. & Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5, e9490.
Sanchez, G. & Trinchera, L. (2012). Plspm: partial least squares data analysis methods. *R package*. http://cran.r-project.org/web/packages/plspm
Slatyer, R.A., Hirst, M. & Sexton, J.P. (2013). Niche breadth predicts geographical range size: a general ecological pattern. *Ecol. Lett.*, 16, 1104–1114.
Snel, B., Bork, P. & Huynen, M.A. (1999). Genome phylogeny based on gene content. *Nat. Genet.*, 21, 108–110.
Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M. & Lauro, C. (2005). PLS path modeling. *Comput. Stat. Data Anal.*, 48, 159–205.
Tettelin, H., Riley, D., Cattuto, C. & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, 11, 472–477.
Torsvik, V., Goksoyr, J. & Daae, F.L. (1990). High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.*, 56, 782–787.
Verberk, W., van Noordwijk, C.G.E. & Hildrew, A.G. (2013). Delivering on a promise: integrating species traits to transform descriptive community ecology into a predictive science. *Freshwater Science*, 32, 531–547.
Vieira-Silva, S. & Rocha, E.P.C. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.*, 6, e1000808.
Walters, W.A., Caporaso, J.G., Lauber, C.L., Berg-Lyons, D., Fierer, N. & Knight, R. (2011). PrimerProspector: de nove design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics*, 27, 1159–1161.
Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007). Naïve Bayesian Classifier for rapid assessment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73, 5261–5267.
Webb, T.J. & Gaston, K.J. (2003). On the heritability of geographic range sizes. *Am. Nat.*, 161, 553–566.
Wright, I.J., Reich, P.B., Westoby, M., Ackerly, D.D., Baruch, Z., Bongers, F. *et al.* (2004). The worldwide leaf economics spectrum. *Nature*, 428, 821–827.
Ye, Y. & Doak, T.G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.*, 5, e1000465.

SUPPORTING INFORMATION

Additional Supporting Information may be downloaded via the online version of this article at Wiley Online Library (www.ecologyletters.com).

Editor, John Klironomos
Manuscript received 13 January 2014
First decision made 22 February 2014
Second decision made 12 March 2014
Third decision made 19 March 2014
Manuscript accepted 22 March 2014