DR. MALLORY J CHOUDOIR (Orcid ID : 0000-0002-9117-5150)

Article type : Articles

Running Head: Range size of microbial taxa

Variation in range size and dispersal capabilities of microbial taxa

Mallory J. Choudoir[1] (mjchoudoir@gmail.com)
Albert Barberán[2] (barberan@email.arizona.edu)
Holly L. Menninger[3] (holly.menninger@gmail.com)
Rob R. Dunn[3] (rrdunn@ncsu.edu)
*Noah Fierer[1,5] (noahfierer@gmail.com)

[1] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309 USA

[2] Department of Soil, Water, and Environmental Science, University of Arizona, Tucson, AZ 85721 USA
[3] Department of Applied Ecology North Carolina State University, Raleigh, NC 27695 USA
[4] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309 USA

***Corresponding author:**
Noah Fierer
215 CIRES, 216 UCB
University of Colorado Boulder
Boulder, CO 80309
noahfierer@gmail.com
303-492-5615

ABSTRACT

Geographic range size can span orders of magnitude for plant and animal species, with the study of why range sizes vary having preoccupied biogeographers for decades. In contrast, there have been few comparable studies of how range size varies across microbial taxa and what traits may be associated with this variation. We determined the range sizes of 74,134 bacterial and archaeal taxa found in settled dust collected from 1,065 locations across the United States. We found that most microorganisms have small ranges and few have large ranges, a pattern similar to the range

size distributions commonly observed for macrobes. However, contrary to expectations, those microbial taxa which were locally abundant did not necessarily have larger range sizes. The observed differences in microbial range sizes were generally predictable from taxonomic identity, phenotypic traits, genomic attributes, and habitat preferences, findings that provide insight into the factors shaping patterns of microbial biogeography.

Key Words: Geographic range size, biogeography, dispersal, microbiology, microbial dispersal, dust-associated microbes

INTRODUCTION

Not all microbes are everywhere all the time. Due to both dispersal constraints and habitat filtering, we know that many microbial taxa are restricted in their geographic and ecological distributions (Martiny *et al.* 2006; Hanson *et al.* 2012). Microbial endemism has been demonstrated across a range of habitats including geothermal hot springs (Papke *et al.* 2003; Whitaker *et al.* 2003), benthic ecosystems (Ruff *et al.* 2015), soil (Cho and Tiedje 2000; Vos and Velicer 2008; Andam *et al.* 2016), and marine waters (Boucher *et al.* 2011; Ghiglione *et al.* 2012; Sul *et al.* 2013). Perhaps the best evidence for restricted microbial distributions comes from decades of work on pathogens. Many pathogens of humans, domestic animals, and crops are restricted to certain geographic areas and regions with specific environmental conditions (Achtman 2008; Bebber *et al.* 2014; Just *et al.* 2014; Murray *et al.* 2015).

Like plants and animals, many microorganisms clearly have ranges – the geographic area where a given taxon is found – and range sizes are likely to vary across bacterial and archaeal taxa. The study of range size, and the factors that drive differences in range size and shape, have been

studied for more than a century by ecologists, biogeographers, and conservation biologists. However, there is surprisingly little explicit documentation of microbial geographic range size, taxonomic variation in range size, or the traits that might contribute to this variation.

Much of the variation in the geographic range size of plant and animal species is often predictable. For example, occupancy-abundance relationships are generally positive, and organisms that are locally abundant also often have large geographic ranges (Gaston *et al.* 2000; Holt *et al.* 2002; Roney *et al.* 2015), with causality likely flowing in both directions. Certain life history strategies also vary predictably with range size. For example, species with greater dispersal capabilities tend to have larger geographic ranges due to their ability to populate new regions and to maintain gene flow among regions, as is the case for certain insects (McCauley *et al.* 2014), birds (Laube *et al.* 2013), plants (Paul *et al.* 2009), and marine taxa (Macpherson 2003; Lester and Ruttenberg 2005; Lester *et al.* 2007). In addition, taxa able to live in many habitat types, whether because they are generalists or have a high degree of phenotypic plasticity, also tend to have larger geographic ranges (Pohlman *et al.* 2005; Pichancourt and van Klinken 2012; Morueta-Holme *et al.* 2013; Ofstad *et al.* 2016). Finally, closely related taxa often have similar geographic range sizes due to shared ecological attributes, as shown for species of birds (Mouillot and Gaston 2009; Herrera-Alsina and Villegas-Patraca 2014).

With their small cell size, massive population numbers, and diverse physiologies, microbial taxa have the potential for widespread dispersal and colonization, and consequently, large range sizes. Evidence suggests there are unifying theories of biodiversity and biogeography across all domains of life (Green and Bohannan 2006; Locey and Lennon 2016). Thus, we predict that many of the factors driving range size in plants and animals also influence microbial range size. For instance, we would expect that locally abundant microbial taxa would tend to have larger geographic ranges than rare species in concordance with previous work which has demonstrated a positive occupancy-abundance relationship for some microbial taxa living in specific environments (Nemergut *et al.* 2011; Ruff *et al.* 2015). We would also expect that closely related taxa should have more similar range size distributions due to a greater likelihood of sharing traits that govern capacity for dispersal and colonization. As for macrobes, taxa that disperse well and are able to tolerate a wide range of environmental conditions should have larger range sizes. We

predict that the relevant traits governing microbial dispersal may include dormancy or other strategies related to stress (e.g. UV radiation, desiccation, extreme temperature) tolerance. For example, endospore formation facilitates the dispersal of microbes through both time and hostile conditions. This trait allows *Bacillus* species to travel across continents in the upper atmosphere (Roberts and Cohan 1995) and thermophilic marine *Firmicutes* to persist in cold sediments (Müller *et al.* 2014). Similarly, we predict that traits associated with the ability to colonize and grow in diverse environments may include genomic characteristics related to phenotypic plasticity and habitat breadth. From previous work showing that genome size correlates with the ability of soil bacteria to persist in a broad range of habitats (Barberán *et al.* 2014a; Cobo-Simón and Tamames 2017), we would expect genome size to positively correlate with range size. Hence, the suite of phenotypic traits and genomic attributes that influence the ecological distribution of microbial taxa also likely influence range size.

To build a more comprehensive understanding of how and why microbial range sizes may vary, we determined the range sizes and shapes of 74,134 bacterial and archaeal taxa found in settled dust collected from outdoor building surfaces from 1,065 homes across the United States. We focus on settled dust because it is found everywhere and easy to sample consistently. Likewise, we know that those microbes found in settled dust were at one point airborne, allowing us to identify organisms that can be dispersed through the atmosphere. Also, the settled dust found on outdoor building surfaces is nutrient-limited and is unlikely to represent the original environmental source of the taxa found therein. In other words, by examining the range sizes of those microbes found in settled dust, we can more readily assess differences across taxa in their dispersal capabilities as opposed to differences related to colonization and establishment in a more suitable environment for growth.

We calculated range size using both the area of occupancy (AOO) and the extent of occurrence (EOO) approximations, both of which are commonly used in macroecology (Gaston and Fuller 2009). Simplified, AOO is akin to a ‘dot map’ of observations across a grid overlay that are summed together, while EOO is comparable to ‘connecting the dots’ and calculating the area of the resulting shape. We determined if the distribution of range sizes for these microbial taxa is similar to plant and animal species, and to what extent the occupancy-abundance relationship

explains variation in range size. Furthermore, we determined if taxonomy could predict differences in range size distributions. Finally, we mined the extensive wealth of information available in curated microbial databases to determine if phenotypic traits, genomic attributes, or habitat preferences can explain the measured variability in range size. This study represents one of the first comprehensive efforts to understand the variation in range size across a broad range of microbial taxa, whether microbes follow the same biogeographical patterns commonly observed for 'macrobes', and why some microbial taxa have larger range sizes than others.

METHODS

Sample collection and molecular analysis

Details of sample collection and molecular analysis have been described previously (Barberán *et al.* 2015). Briefly, outdoor dust samples were collected from the upper trim on the outside surface of an exterior door by participants of the Wild Life of Our Homes (http://robdunnlab.com/projects/wild-life-of-our-homes/) citizen science project. Bacterial and archaeal diversity was determined by sequencing the V4 hypervariable region of the 16S rRNA gene with primers 515-F (GTGCCAGCMGCCGCGGTAA) and 806-R (GGA-CTACHVGGGTWTCTAAT) (Fierer *et al.* 2012) using the direct PCR approach previously described (Flores *et al.* 2012). Sequencing was done on the Illumina HiSeq or MiSeq platforms with all reads trimmed to 100 bp. All reads were quality filtered (maximum e-value of 0.5), dereplicated, and clustered into phylotypes at a 97% similarity threshold with the UPARSE pipeline (Edgar 2013). Taxonomic identity was determined using the Ribosomal Database Project classifier (Wang *et al.* 2007) trained on the Greengenes 13_8 16S rRNA database (McDonald *et al.* 2012). All sequence data are accessible through the FigShare repository (https://doi.org/10.6084/m9.figshare.1270900.v8).

Eukaryotic sequences were removed, and those phylotypes present in >25% of negative control samples (including phylotypes classified as *Mycoplasma*, *Pseudomonas*, *Serratia*, and *Acinetobacter*) were also filtered prior to downstream analyses as they likely represent taxa originating from reagent or amplification contamination (Salter *et al.* 2014). To minimize amplicon sequencing biases between samples, low coverage samples (i.e. samples with <10,000 reads after quality filtering) were removed, and total sequence counts were normalized using a

cumulative-sum scaling approach (Paulson *et al.* 2013). We restricted our analyses to the contiguous United States. and hence removed samples originating from Hawaii and Alaska. Finally, we excluded rare phylotypes (i.e. phylotypes present in fewer than five samples, as at least five observations are required to calculate range size using the minimum convex polygon approach described below). In total, 74,134 phylotypes across 1,065 samples were included in all downstream analyses.

Range size and shape calculations

Latitude and longitude coordinates were inferred from sample locations (i.e. reported addresses), and these coordinates were transformed into the Lambert conformal conic projection (LCC) for all spatial analyses. Many plant species are dispersed by wind (Howe and Smallwood 1992; Clark *et al.* 2002), so given the potential similarities between plants and dust-associated microbes in their dispersal dynamics, we used approaches to calculate range sizes commonly employed by plant biogeographers. Range size was determined using both area of occupancy (AOO) (see Kolb *et al.* 2006; Kreft *et al.* 2006; Essl *et al.* 2009) and extent of occurrence (EOO) (see Sérgio *et al.* 2007; Brummit *et al.* 2015). To determine AOO, we overlaid a 100 x 100 km^2 grid that encompassed all sample locations and used the R package sp (Pebesma and Bivand 2005) to count the total number of grid cells in which each phylotype was observed. AOO range size (km^2) was calculated by summing the area of total occupied grid cells. To determine EOO, we used the R package adehabitatHR (Calenge 2006) to find the minimum convex polygon (MCP) after excluding 5% of the extreme points. EOO range size (km^2) was calculated from the area of the MCP circumscribing all observations for each phylotype. Range shape was determined by calculating the maximum longitudinal and latitudinal dimensions of occurrence for each phylotype. To control for biases introduced by uneven sampling intensity, we divided the U.S. into six regions, sub-sampled 70 locations from each of these regions, and repeated the range dimension analyses.

Taxonomic signal and phenotypic, genomic, and habitat trait-based analyses

Next, we assessed potential taxonomic determinants of range size. Phylotype range size was ranked by Phylum, and phyla with fewer than 25 representative phylotypes were excluded. For the most abundant phyla (i.e. *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes*),

phylotype range size was also ranked by Family, and families with fewer than 25 representative phylotypes were excluded.

Finally, we determined if differences in phenotypic traits, genomic attributes, or habitat preferences could further explain variation in range size. We inferred putative traits of dust phylotypes by matching their 16S rRNA gene sequences to those of reference strains from curated, publicly available databases. Representative partial 16S rRNA gene sequences from each phylotype were matched against full length 16S rRNA gene sequences from the IJSEM phenotypic database (Barberán *et al.* 2017) and from the Integrated Microbial Genomes (IMG) database (Markowitz *et al.* 2014). Matches were determined using BLASTn (Altschul *et al.* 1990) at ≥99% identity and ≥95% coverage. We restricted these analyses to the top four most abundant phyla, which included *Proteobacteria, Actinobacteria, Firmicutes,* and *Bacteroidetes*. We recognize that partial 16S rRNA gene sequences may not provide a level of resolution sufficient for accurately identifying the phenotypic and genomic traits of all taxa. However, the selected traits typically show a strong phylogenetic signal and are generally conserved across broader taxa and lineages (Barberán *et al.* 2017).

We were able to match a total of 1,461 16S rRNA gene sequences of dust phylotypes (including 518 *Proteobacteria,* 428 *Actinobacteria,* 293 *Firmicutes,* and 222 *Bacteroidetes*) to 2,487 unique full length 16S rRNA gene sequences in the IJSEM phenotypic database (Barberán *et al.* 2017). We assessed how the AOO varied in relation to the following phenotypic traits: oxygen tolerance, sporulation, pigmentation, Gram stain reaction, and source habitat. Here, source habitat refers to the reported isolation source of a given strain from the IJSEM phenotypic database (Barberán *et al.* 2017). We selected these traits because we expected that these traits may influence dispersal and colonization capabilities. For discrete traits, we excluded phylotypes with matches to multiple strains that had conflicting trait values.

We matched a total of 1,186 16S rRNA gene sequences of dust phylotypes (including 415 *Proteobacteria,* 276 *Actinobacteria,* 325 *Firmicutes,* and 170 *Bacteroidetes*) to 6,321 unique full length 16S rRNA gene sequences in the IMG database (Markowitz *et al.* 2014). We assessed how AOO varied with the following genomic attributes: G+C content, genome size, and 16S

rRNA operon copy number. We hypothesized that these attributes may influence phenotypic plasticity and habitat preferences. G+C content and genome size are highly correlated in bacteria, and large genomes are thought to confer broad niche breadth (Bently and Parkhill 2004). Multiple copies of the 16S rRNA operon are common in microbial genomes and are reflective of copiotrophic or oligotrophic life history strategies (Klappenbach *et al.* 2000), with those taxa capable of higher maximum growth rates generally having a larger number of rRNA operons. For these genomic attributes, we determined mean values for phylotypes with matches to multiple strains.

RESULTS

Microbial diversity and community composition

A total of 74,134 16S rRNA gene sequence phylotypes were observed across the 1,065 dust samples (Fig. 1a), with each sample harboring 4,850 phylotypes on average. A total of 50 bacterial and archaeal phyla were recovered, and the dominant phyla were *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes* with >75% of phylotypes assigned to these four phyla (Appendix S1: Fig S1). On average a given phylotype was observed in 70 and 27 different samples (mean and median, respectively). Nearly 24% of phylotypes were found in ≤10 samples, and only 35 phylotypes were observed in ≥90% of the samples (Fig. 1b). Community composition was highly variable across the samples, with geographic distance as well as environmental factors including soil pH, precipitation, primary productivity, and temperature being the best predictors of overall differences in community composition (see Barberán *et al.* 2015 for details).

Range size and shape

The frequency distribution in range size as measured using the area of occupancy (AOO) was highly right-skewed and best described using a log-normal distribution; many taxa have small ranges and few have very large ranges (Fig. 1c). Across all phylotypes, the mean and median estimated AOO range sizes were 3,984 and 2,200 km^2, respectively. Alternatively, using the extent of occurrence (EOO), we found that the frequency distribution of EOO range size is best described as irregular and bimodal (Fig. 1d). Across all phylotypes, the estimated mean and median EOO range sizes were 4.2 and 4.3 million km^2, respectively, which is approximately half

of the area of the contiguous U.S. We observed a strong positive correlation (Spearman's ρ = 0.89, $P < 0.0001$) between log AOO and EOO estimations of range size (Appendix S1: Fig S2). Given the high potential for cell dispersal and the wide distribution of suitable habitats, we expect that EOO is likely overestimating microbial range sizes here. Therefore, we used the more conservative AOO estimation for downstream analyses focused on determining what potential factors might explain range size variation.

We described range shape by calculating the maximum geographic spread in both longitudinal and latitudinal dimensions. The frequency distribution of the longitudinal range is highly left-skewed; most phylotypes were found on both the eastern and western coasts and have a mean and median east-west span of 3,869 km and 4,183 km, respectively (Appendix S1: Fig S3a). While also left-skewed, there was a greater variation in the latitudinal range with a mean and median north-south span of 1,855 km and 1,920 km, respectively (Appendix S1: Fig S3b). While phylotypes with greater longitudinal spread also tend to have greater latitudinal spread (Spearman's $\rho = 0.67$, $P < 0.0001$), range dimensions for most phylotypes (89.9%) are elongated east-west as opposed to north-south (Fig. 2), and this pattern persists after normalization for the irregular shape of the sampling region (i.e. the U.S. is larger east-west than north-south) and after correcting for differences in sampling intensities across different regions (Appendix S1: Fig S4). The bacterial and archaeal phylotypes are far more likely to have larger east-west distributions than north-south distributions.

One of the most widely observed correlates of range size is local density or abundance, and species that are more abundant tend to have larger geographic ranges than rare species (Gaston 1996a; Holt *et al*. 2002). Interestingly, we find little support for this relationship for dust-associated microbes. Instead, we found only a weak correlation between the local relative abundance of a phylotype and its range size (Spearman's $\rho = 0.14$, $P < 0.0001$) (Appendix S1: Fig S5).

Taxonomic differences in range size

Given this broad distribution of range sizes across dust phylotypes, we next sought to determine what additional factors could further explain this variation. To begin, we asked if range size

differed across taxonomic groups. We found that geographic range size has a strong taxonomic signal and varies significantly across phyla (one-way ANOVA; $F_{28, 70297} = 136.2$, $P < 0.0001$) (Fig. 3). For example, phylotypes within the phylum *Actinobacteria* tend to have range sizes that are approximately 13% larger than the range sizes of phylotypes within the phylum *Acidobacteria* (Tukey's test; $P = 0.00022$). Within the Archaea, range sizes of *Crenarchaeota* are approximately 69% larger than those of *Euryarachaeota* (Tukey's test; $P < 0.0001$).

At greater taxonomic resolution, the taxonomic signal for range size within the top four most abundant phyla was even more pronounced (Appendix S1: Fig S6). We observed significant differences across families of *Proteobacteria* (one-way ANOVA; $F_{45, 15833} = 84.9$, $P < 0.0001$) (Appendix S1: Fig S6a), *Actinobacteria* (one-way ANOVA; $F_{40, 8005} = 21.8$) (Appendix S1: Fig S6b), *Firmicutes* (one-way ANOVA; $F_{21, 7159} = 66.6$, $P < 0.0001$) (Appendix S1: Fig S6c) and *Bacteroidetes* (one-way ANOVA; $F_{15, 5359} = 70.1$, $P < 0.0001$) (Appendix S1: Fig S6d). For example, within the *Proteobacteria*, *Burkholderiaceae* share a similar range size with *Rhizobiaceae* (Tukey's test; $P = 0.98$) and *Bradyrhizobiaceae* (Tukey's test; $P = 0.052$), and all three of these families have larger range sizes than *Neisseriaceae* or *Legionellaceae* (Tukey's test; $P < 0.0001$) (Appendix S1: Fig S6a).

Phenotypic and genomic traits that vary with range size

Finally, we asked if certain phenotypic or genomic traits could predict variation in range size. We found that range size varies with oxygen tolerance (two-way ANOVA; $F_{3, 2324} = 67.5$, $P < 0.0001$); aerobes have geographic ranges approximately 63% larger than anaerobes (Tukey's test; $P < 0.0001$) (Fig. 4a). However, the strength of the relationship between oxygen tolerance and range size differs between phyla (two-way ANOVA; $F_{9, 2324} = 2.6$, $P = 0.0062$) (Appendix S1: Fig S7a). Unexpectedly, range sizes were approximately 19% smaller for those phylotypes inferred to be capable of spore formation, even after restricting the analysis to obligate aerobes to minimize potential biases incurred by many anaerobes being spore-formers (two-way ANOVA; $F_{1, 1643} = 36.0$, $P < 0.0001$) (Fig. 4b), although the strength of this relationship differed between phyla capable of spore formation (two-way ANOVA; $F_{1, 1643} = 5.0$, $P = 0.025$) (Appendix S1: Fig S7b). Taxa that are pigmented tended to have ranges that are approximately 39% larger than taxa that are not pigmented (two-way ANOVA; $F_{1, 1802} = 55.4$, $P < 0.0001$) (Fig. 4c), and this

pattern was independent of phylum identity (two-way ANOVA; $F_{3, 1802} = 1.5$, $P = 0.22$) (Appendix S1: Fig S7c). Range size also varied with Gram stain; taxa with Gram stain positive cell walls have approximately 17% larger ranges than taxa with Gram stain negative cell walls (two-way ANOVA; $F_{2, 31207} = 32.4$, $P < 0.0001$) (Fig. 4d). Finally, range size varied with source habitat (two-way ANOVA; $F_{7, 907} = 11.1$, $P < 0.0001$); taxa derived from soil and plants were more likely to have larger ranges compared to taxa associated with aquatic environments such as seawater or marine sediments (Tukey's test; $P < 0.005$) (Fig. 4e).

With regards to genomic attributes, we found that range size was positively correlated with G+C content (Pearson's $r = 0.45$, $P < 0.0001$) (Fig. 5a), but this relationship was largely driven by *Proteobacteria* (Pearson's $r = 0.39$, $P < 0.0001$) and *Actinobacteria* (Pearson's $r = 0.32$, $P < 0.0001$) (Appendix S1: Fig S8a). Range size and genome size were also positively correlated (Pearson's $r = 0.22$, $P < 0.0001$) (Fig. 5b), and this relationship was significant when we ran the analyses for each phylum individually (Appendix S1: Fig S8b). Finally, range size was negatively correlated with 16S rRNA operon copy number (Pearson's $r = -0.28$, $P < 0.0001$) (Fig. 5c), but the direction and significance of this relationship varied when these analyses were conducted within individual phyla (Appendix S1: Fig S8c).

DISCUSSION

Geographic range size is a cornerstone of biogeography, and studies of how range sizes vary across taxa have contributed to the development of key paradigms in conservation biology, evolutionary biology, and ecology. Despite decades of studies investigating range size and range size determinants in plants and animals, comparable studies are rarely conducted with microbial taxa. We addressed this knowledge gap by investigating the range sizes and the potential factors associated with range size variation across a broad breadth of bacterial and archaeal taxa (Appendix S1: Fig S1) identified in dust samples collected from across the United States (Fig. 1a).

The accurate evaluation of microbial range size distributions is challenging, and many of these challenges also apply to the accurate estimation of plant or animal range sizes. First, most microbial communities are highly diverse. Thus, adequate sampling depth is important, and it

remains challenging to determine with confidence whether a given taxon is truly absent in a community or simply below the level of detection. High-throughput culture-independent sequencing approaches, like the approach used here in which we identified microbial taxa in samples by analyzing a mean of 59,831 16S rRNA gene sequences per sample, can help to reduce the magnitude of this problem (Sogin *et al.* 2006; Lynch and Neufeld 2015). Even so, we are undoubtedly underestimating the full extent of microbial diversity in individual samples. Importantly, this problem of insufficient sampling depth, which limits our ability to confirm which taxa are 'truly absent' in a given sample versus those taxa that were simply not detected, also plagues plant and animal surveys (MacKenzie *et al.* 2002; Cunningham and Lindenmayer 2005). Second, accurate estimations of range size are best achieved through extensive population surveys across a broader geographic region of interest. While sampling efforts are inevitably constrained by logistics, more is always better, and we were able to collect samples from 1,065 locations across the contiguous U.S. (Fig. 1a). Third, range sizes will undoubtedly vary as a function of taxonomic resolution – the range sizes of sub-populations will likely be smaller than range sizes of the broader species or genus. Most studies of plant and animal range size focus on species or intra-species level resolutions. While the species definitions for plants and animals are often arbitrary and somewhat inconsistent, microbiologists continue to intensely debate the 'microbial species concept' and even the mere existence of species (Roselló-Mora and Amann 2001; Gevers *et al.* 2005; Achtman and Wager 2008; Doolittle 2012). To remedy this, microbial ecologists often define units of diversity, or phylotypes, based on similarity in marker gene sequences. Such an approach was used here as we defined phylotypes as those taxa which shared ≥97% similarity in their 16S rRNA gene sequences, a threshold that roughly corresponds to a bacterial 'species' (Stackebrandt and Goebel 1994; Kim *et al.* 2014). In short, the challenges associated with estimating microbial range sizes are not unique to microbial ecology, and we argue that robust investigation of microbial range size is possible with the sampling effort and methodologies used here.

The AOO range size frequency distribution for dust phylotypes was highly right-skewed (Fig. 1c); many microbial taxa have small geographic ranges and fewer have large ones. This distribution of geographic range sizes, described as a 'hollow curve' that is approximately log-normally distributed, is widely observed for many plant and animal species (Gaston 1996b;

Berry and Riina 2005; Orme *et al.* 2006; Agosta *et al*. 2013). In addition to range size, the shape of a species' range is also commonly studied by plant and animal biogeographers (Brown *et al.* 1996). For instance, range shape can be used to identify the environmental variables that determine patterns of range expansion (Pigot *et al.* 2010). Here, we described the range shapes of these microbial taxa by measuring the maximum east-west and north-south spread of each phylotype. We found that the north-south spread of taxa was more constrained that the east-west spread (Fig. 2, Appendix S1: Fig S3). To put simply, many taxa are found on both eastern and western coasts, but fewer are distributed across the southern and northern boundaries of the U.S. This results in an east-west elongated range for a majority of dust phylotypes (Fig. 2, Appendix S1: Fig S4), a pattern that is consistent with the east-west range elongation that is observed for many North American plant and animal species (Brown *et al.* 1996; Rosenfield 2002; Schlachter 2010). This pattern may be a product of dispersal driven by the prevailing winds, which predominately blow across North America from the west to the east. The migration of microbes through the atmosphere has been previously linked to wind patterns and weather dynamics (Yamagucki *et al.* 2012; Smith *et al.* 2013; Barberán *et al.* 2014b; Weil *et al.* 2017). This pattern also suggests that there are latitudinal limits to dispersal, which could be the result of climatic temperature constraints or historical biogeographical processes (Mittelbach *et al.* 2007). Latitudinal constraints to dispersal are well documented across diverse plant and animal species (Wiens *et al.* 2006; Svenning and Skov 2007; Salisbury *et al.* 2012), and more recently such constraints have been documented in terrestrial soil bacteria (Andam *et al.* 2016; Choudoir *et al.* 2016). We think that future work integrating information on weather systems and other climate variables to address mechanisms of microbial migration will be particularly insightful.

The frequency distribution in geographic range sizes and the spatial dimensions of range shape for these dust-associated microbes are qualitatively similar to what is commonly observed for plants and animals. In contrast, we find little support for the occupancy-abundance relationship for dust-associated bacteria (Appendix S1: Fig S5). This finding goes against expectations as the occupancy-abundance relationship has been widely observed for plants and animals (Gaston *et al.* 2000). Although this relationship may somewhat be inflated by the challenges associated with sampling rare taxa (Wenger and Freeman 2008; Sileshi *et al.* 2009), most bacterial phylotypes, regardless of their local abundance, had small ranges, while phylotypes with high local

abundance were nearly as likely to have large ranges as rare taxa (Appendix S1: Fig S5). Thus, abundance alone is not a useful predictor of microbial range sizes, and instead we expected that much of the observed variation in microbial range size is likely due to evolutionary or ecological traits affecting dispersal or habitat preferences.

Range size distributions varied across taxonomy, and mean range size differed significantly between phyla (Fig. 3). Importantly, this relationship was not just driven by the most abundant phyla. For example, range sizes for the *Crenarchaeota* and candidate phylum FBP are amongst the largest in the dataset, yet these phyla are not ranked among the top ten most abundant phyla (Appendix S1: Fig S1). We also see intra-group differences in range size distributions between phyla. For example, range size approximations for *Proteobacteria* and *Bacteroidetes* encompass values spanning the minimum and maximum of the entire dataset, while range size approximations for candidate phyla WPS-2 or *Chlorobi* have a much narrower size distribution (Fig. 3). Some of this variation in range size for *Proteobacteria* or *Bacteroidetes* is further explained by clear differences in range size at the Family level of taxonomic resolution (Appendix S1: Fig S6).

We identified a number of phenotypic traits, genomic attributes, and habitat preferences that varied predictably as a function of geographic range size (Fig. 4, Fig. 5). Some of these traits are consistent across phyla, while other traits explain more variation in range size within certain phyla (Appendix S1: Fig S7, Appendix S1: Fig S8). For instance, we found that anaerobes were more likely to have smaller range sizes than aerobes (Fig. 4a), potentially due to their inability to survive dispersal through the oxygen-rich atmosphere. Contrary to expectations, we found non-spore forming aerobes had larger range sizes than spore-formers (Fig. 4b). This pattern was consistent for *Actinobacteria* and *Firmicutes*, which are phyla with both spore-forming and non-spore forming members (Appendix S1: Fig S7b). Either there are other traits that are more important than spore formation in determining dispersal capabilities, or we are limited in our ability to accurately predict spore formation from the available *in vitro* data. Finally, we found that pigmentation was associated with larger geographic ranges (Fig. 4c), potentially due to pigment production offering UV protection to microbial cells during atmospheric dispersal. Pigments have been shown to protect *Bacillus* endospores from radiation (Moeller *et al.* 2005),

and carotenoid pigments are also shown to protect proteobacterial phytopathogens from UV (To *et al.* 1994; Mohammadi *et al.* 2013).

Additionally, we found that both genome G+C content and genome size increased with range size (Fig. 5a-b, Appendix S1: Fig S8a-b), although these genomic attributes are also positively correlated with each other (Nishida 2012). Greater G+C content has been associated with genome stability and thermal tolerance in some microbes (Nishio *et al.* 2003; Mann and Chen 2010). Larger genomes correspond to more genes and metabolic pathways that likely confer greater physiological versatility and ability to survive diverse environmental conditions (Bently and Parkhill 2004; Konstantinidis *et al.* 2006). Our findings are in line with recent studies showing that larger genomes are linked to ubiquity and greater environmental and spatial distributions (Barberán *et al.* 2014a; Cobo-Simón and Tamames 2017). Conversely, we observed a negative correlation between 16S rRNA gene copy number and range size (Fig. 5c), suggesting that oligotrophic life history strategies (see Klappenbach *et al.* 2000) are associated with greater range sizes within some phyla. Finally, we found that the inferred habitat preferences of microbes could explain some of the variation in range size. Soil and plant associated taxa had larger range size distributions than marine and aquatic habitat associated taxa (Fig. 4e). Not surprisingly, these results suggest that those taxa that are likely found in widespread source environments tend to have larger ranges. While we cannot explicitly determine the source origin for each taxon, phyla that are dominant in soil, including *Actinobacteria* and *Acidobacteria*, have some of the largest range size distributions (Fig. 3, Fig. 4e). Conversely, taxa from seawater and other aquatic habitats tend to have smaller ranges (Fig. 4e), a pattern that may result from these source habitats not being as widespread across the sampled region, limited aerosolization of microbial cells from these source environments, or a reduced capacity for these aquatic taxa to survive desiccation.

Together our results illustrate a wide variation in range size of diverse bacterial and archaeal taxa found in settled outdoor dust. The shape of the range size frequency distribution of these microbes is similar to many plants and animals, suggesting similar processes can drive observed biogeographical patterns. However, the canonical occupancy-abundance relationship explains little of the variation observed here. Instead, we found range size to vary between major phyla

and identified phenotypic traits and genomic attributes that also vary across taxonomy. These traits likely influence dispersal capabilities or the ability to colonize and establish in an environment following a dispersal event. Many dust-associated taxa are of ecological, agricultural, and medical importance, and integrating range size calculations and range size determinants into microbial ecology will advance our understanding of the spatial distributions of taxa of interest. Together, this work highlights the importance of both dispersal dynamics and habitat distribution in generating patterns in microbial biogeography.

ACKNOWLEDGEMENTS

We want to thank all of the volunteers who participated in the Wildlife of Our Homes project for collecting dust samples and members of the Fierer lab group for critical feedback on earlier drafts of this manuscript. Funding for this work was provided by grants from the U.S. Department of Defense ("Forensic Geolocation via Biological Signatures"), the U.S. Army Research Office ("Isolation and Characterization of Airborne Bacterial Strains That Produce Antimicrobial Compounds"), and the A.P. Sloan Foundation Microbiology of the Built Environment Program.

REFERENCES

Achtman, M. 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual Review of Microbiology* **62**:53–70.

Achtman, M., and M. Wagner. 2008. Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology* **6**: 431–440.

Agosta, S. J., J. Bernardo, G. Ceballos, and M. A. Steele. 2013. A macrophysiological analysis of energetic constraints on geographic range size in mammals. *PLoS ONE* **8**:e72731.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410.

Andam, C. P., J. R. Doroghazi, A. N. Campbell, P. J. Kelly, M. J. Choudoir, and D. H. Buckley. 2016. A latitudinal diversity gradient in terrestrial bacteria of the genus *Streptomyces*. *mBio* **7**:e02200–15.

Barberán, A., J. Ladau, J. W. Leff, K. S. Pollard, H. L. Menninger, R. R. Dunn, and N. Fierer. 2015. Continental-scale distributions of dust-associated bacteria and fungi. *Proceedings of the National Academy of Sciences* **112**:5756–5761.

Barberán, A., K. S. Ramirez, J. W. Leff, M. A. Bradford, D. H. Wall, and N. Fierer. 2014a. Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecology Letters* **17**:794–802.

Barberán, A., J. Henley, N. Fierer, and E. O. Casamayor. 2014b. Structure, inter-annual recurrence, and global-scale connectivity of airborne microbial communities. *The Science of the Total Environment* **487**:187–195.

Barberán, A., Velazques, H. C., Jones, S., and Fierer, N. 2017. Hiding in plain sight: mining bacterial species records for phenotypic trait information. *mSphere* **2**:e00237–17.

Bebber, D. P., T. Holmes, and S. J. Gurr. 2014. The global spread of crop pests and pathogens. *Global Ecology and Biogeography* **23**:1398–1407.

Bentley, S. D., and J. Parkhill. 2004. Comparative genomic structure of Prokaryotes. *Annual Review of Genetics* **38**:771–791.

Berry, P. E., and R. Riina. 2005. Insights into the diversity of the Pantepui flora and the biogeographic complexity of the Guayana Shield. *Biologiske Skrifter* **55**:145–167.

Boucher, Y., O. X. Cordero, A. Takemura, D. E. Hunt, K. Schliep, E. Bapteste, P. Lopez, C. L. Tarr, and M. F. Polz. 2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *mBio* **2**:e00335–10.

Brown, J. H., G. C. Stevens, and D. M. Kaufman. 1996. The geographic range: size, shape, boundaries, and internal structure. *Annual Review of Ecology and Systematics* **27**:597–623.

Brummitt, N., S. P. Bachman, E. Aletrari, H. Chadburn, J. Griffiths-Lee, M. Lutz, J. Moat, M. C. Rivers, M. M. Syfert, and E. M. Nic Lughadha. 2015. The Sampled Red List Index for Plants, phase II: ground-truthing specimen-based conservation assessments. *Philosophical Transactions of the Royal Society of London B, Biological sciences* **370**:20140015.

Calenge, C. 2006. The package “adehabitat” for the R software: a tool for the analysis of space and habitat use by animals. *Ecological Modelling* **197**:516–519.

Cho, J.-C., and J. M. Tiedje. 2000. Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. *Applied and Environmental Microbiology* **66**:5448–5456.

Choudoir, M. J., J. R. Doroghazi, and D. H. Buckley. 2016. Latitude delineates patterns of biogeography in terrestrial *Streptomyces*. *Environmental Microbiology* **18**:4931–4945.

Clark, J. S., Beckage, B., HilleRisLambers, J., Ibanez, I., LaDeau, S., McLachlan, J., Mohan, J., and M. Rocca. 2002. Plant dispersal and migration. Pages 81–93 in R. E. Munn, editor. *Encyclopedia of Global Change*. John Wiley & Sons, Ltd., Chichester, U.K.

Cobo-Simón, M., and J. Tamames. 2017. Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC genomics* **18**: 499.

Cunningham, R. B., and D. B. Lindenmayer. 2005. Modeling count data of rare species: some statistical issues. *Ecology* **86**:1135–1142.

Doolittle, F. 2012. Population genomics: how bacterial species form and why they don't exist. *Current Biology* **22**:R449–R451.

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

Essl, F., M. Staudinger, O. Stöhr, L. Schratt-Ehrendorfer, W. Rabitsch, and H. Niklfeld. 2009. Distribution patterns, range size and niche breadth of Austrian endemic plants. *Biological Conservation* **142**:2547–2558.

Fierer, N., J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, J. A. Gilbert, D. H. Wall, and J. G. Caporaso. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences* **109**:21390–21395.

Gaston, K. J., and R. A. Fuller. 2009. The sizes of species' geographic ranges. *The Journal of Applied Ecology* **46**:1–9.

Gaston, K. J. 1996a. The multiple forms of the interspecific abundance-distribution relationship. *Oikos* **76**:211.

Gaston, K. J., T. M. Blackburn, J. J. D. Greenwood, R. D. Gregory, R. M. Quinn, and J. H. Lawton. 2000. Abundance–occupancy relationships. *The Journal of Applied Ecology* **37**:39–59.

Gaston, K. J. 1996b. Species-range-size distributions: patterns, mechanisms and implications. *Trees* **11**:197–201.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F. L., and J. Swings. 2005. Re-evaluating prokaryotic species. *Nature Reviews Microbiology* **3**:733–739.

Ghiglione, J.-F., P. E. Galand, T. Pommier, C. Pedros-Alio, E. W. Maas, K. Bakker, S. Bertilson, D. L. Kirchman, C. Lovejoy, P. L. Yager, and A. E. Murray. 2012. Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Sciences* **109**:17633–17638.

Green, J., and B. J. M. Bohannan. 2006. Spatial scaling of microbial biodiversity. *Trends in Ecology and Evolution* **21**: 501–507.

Hanson, C. A., J. A. Fuhrman, M. C. Horner-Devine, and J. B. H. Martiny. 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*. **10**:497–506.

Herrera-Alsina, L., and R. Villegas-Patraca. 2014. Biologic interactions determining geographic range size: a one species response to phylogenetic community structure. *Ecology and Evolution* **4**:968–976.

Holt, A. R., K. J. Gaston, and F. He. 2002. Occupancy-abundance relationships and spatial distribution: a review. *Basic and Applied Ecology* **3**:1–13.

Howe, H. F., and J. Smallwood. 982. Ecology of seed dispersal. *Annual Review of Ecology and Systematics* **13**:201–228.

Just, M. G., J. F. Norton, A. L. Traud, T. Antonelli, A. S. Poteate, G. A. Backus, A. Snyder-Beattie, R. W. Sanders, and R. R. Dunn. 2014. Global biogeographic regions in a human-dominated world: the case of human diseases. *Ecosphere* **5**:art143.

Kim, M., H.-S. Oh, S.-C. Park, and J. Chun. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* **64**:346–351.

Klappenbach, J. A., J. M. Dunbar, and T. M. Schmidt. 2000. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology* **66**:1328–1333.

Kolb, A., F. Barsch, and M. Diekmann. 2006. Determinants of local abundance and range size in forest vascular plants. *Global Ecology and Biogeography* **15**:237–247.

Konstantinidis, K. T., A. Ramette, and J. M. Tiedje. 2006. The bacterial species definition in the genomic era. *Philosophical transactions of the Royal Society of London B, Biological Sciences* **361**:1929–1940.

Kreft, H., J. H. Sommer, and W. Barthlott. 2006. The significance of geographic range size for spatial diversity patterns in Neotropical palms. *Ecography* **29**:21–30.

Laube, I., H. Korntheuer, M. Schwager, S. Trautmann, C. Rahbek, and K. Böhning-Gaese. 2013. Towards a more mechanistic understanding of traits and range sizes: avian traits and range size. *Global Ecology and Biogeography* **22**:233–241.

Lester, S. E., and B. I. Ruttenberg. 2005. The relationship between pelagic larval duration and range size in tropical reef fishes: a synthetic analysis. *Proceedings of the Royal Society B, Biological Sciences* **272**:585–591.

Lester, S. E., B. I. Ruttenberg, S. D. Gaines, and B. P. Kinlan. 2007. The relationship between dispersal ability and geographic range size. *Ecology Letters* **10**:745–758.

Locey, K. J., and J. T. Lennon. 2016. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* **113**: 5970–5975.

Lynch, M. D. J., and J. D. Neufeld. 2015. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology* **13**:217–229.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. Andrew Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**:2248–2255.

Macpherson, E. 2003. Species range size distributions for some marine taxa in the Atlantic Ocean. Effect of latitude and depth. *Biological Journal of the Linnean Society* **80**:437–455.

Mann, S., and Y.-P. P. Chen. 2010. Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics* **95**:7–15.

Markowitz, V. M., I.-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, M. Pillay, A. Ratner, J. Huang, T. Woyke, M. Huntemann, I. Anderson, K. Billis, N. Varghese, K. Mavromatis, A. Pati, N. N. Ivanova, and N. C. Kyrpides. 2014. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research* **42**:D560–D567.

Martiny, J. B. H., B. J. M. Bohannan, J. H. Brown, R. K. Colwell, J. A. Fuhrman, J. L. Green, M. C. Horner-Devine, M. Kane, J. A. Krumins, C. R. Kuske, P. J. Morin, S. Naeem, L. Øvreås,

A.-L. Reysenbach, V. H. Smith, and J. T. Staley. 2006. Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology* **4**:102–112.

McCauley, S. J., C. J. Davis, E. E. Werner, and M. S. Robeson. 2014. Dispersal, niche breadth and population extinction: colonization ratios predict range size in North American dragonflies. *The Journal of Animal Ecology* **83**:858–865.

McDonald, D., M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal* **6**:610–618.

Mittelbach, G. G., D. W. Schemske, H. V. Cornell, A. P. Allen, J. M. Brown, M. B. Bush, S. P. Harrison, A. H. Hurlbert, N. Knowlton, H. A. Lessios, C. M. McCain, A. R. McCune, L. A. McDade, M. A. McPeek, T. J. Near, T. D. Price, R. E. Ricklefs, K. Roy, D. F. Sax, D. Schluter, J. M. Sobel, and M. Turelli. 2007. Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecology Letters* **10**:315–331.

Moeller, R., G. Horneck, R. Facius, and E. Stackebrandt. 2005. Role of pigmentation in protecting *Bacillus* sp. endospores against environmental UV radiation. *FEMS Microbiology Ecology* **51**:231–236.

Mohammadi, M., L. Burbank, and M. C. Roper. 2012. Biological role of pigment production for the bacterial phytopathogen *Pantoea stewartii* subsp. *stewartii*. *Applied and Environmental Microbiology* **78**:6859–6865.

Morueta-Holme, N., B. J. Enquist, B. J. McGill, B. Boyle, P. M. Jørgensen, J. E. Ott, R. K. Peet, I. Šímová, L. L. Sloat, B. Thiers, C. Violle, S. K. Wiser, S. Dolins, J. C. Donoghue, N. J. B. Kraft, J. Regetz, M. Schildhauer, N. Spencer, and J.-C. Svenning. 2013. Habitat area and climate stability determine geographical variation in plant species range sizes. *Ecology Letters* **16**:1446–1454.

Mouillot, D., and K. Gaston. 2009. Spatial overlap enhances geographic range size conservatism. *Ecography* **32**:671–675.

Murray, K. A., N. Preston, T. Allen, C. Zambrana-Torrelio, P. R. Hosseini, and P. Daszak. 2015. Global biogeography of human infectious diseases. *Proceedings of the National Academy of Sciences* 1**12**:12746–12751.

Müller, Albert L., de Rezende, Júlia Rosa, Hubert, Casey RJ, Kjeldsen, Kasper Urup, Lagkouvardos, Ilias, Berry, David, Jørgensen, Bo Barker, and Loy, Alexander. 2014. Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents. *The ISME Journal* **8**:1153–1165.

Nemergut, D. R., E. K. Costello, M. Hamady, C. Lozupone, L. Jiang, S. K. Schmidt, N. Fierer, A. R. Townsend, C. C. Cleveland, L. Stanish, and R. Knight. 2011. Global patterns in the biogeography of bacterial taxa. *Environmental Microbiology* **13**:135–144.

Nishida, H. 2012. Evolution of genome base composition and genome size in bacteria. *Frontiers in Microbiology* **3**:420.

Nishio, Yousuke, Nakamura, Yoji, Kawarabayasi, Yutaka, Usuda, Yoshihiro, Kimura, Eiichiro, Sugimoto, S, Matsui, K, Yamagishi, A, Kikuchi, H, Ikeo, K, and Gojobori, T. 2003. Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. *Genome Research* **13**:1572–1579.

Ofstad, E. G., I. Herfindal, E. J. Solberg, and B.-E. Sæther. 2016. Home ranges, habitat and body mass: simple correlates of home range size in ungulates. *Proceedings of the Royal Society Bm Biological Sciences* **283**:20161234.

Orme, C. D. L., R. G. Davies, V. A. Olson, G. H. Thomas, T.-S. Ding, P. C. Rasmussen, R. S. Ridgely, A. J. Stattersfield, P. M. Bennett, and I. P. F. Owens. 2006. Global patterns of geographic range size in birds. *PLoS Biology* **4**:e208.

Papke, R. T., N. B. Ramsing, M. M. Bateson, and D. M. Ward. 2003. Geographical isolation in hot spring cyanobacteria. *Environmental Microbiology* **5**:650–659.

Paul, J. R., C. Morton, C. M. Taylor, and S. J. Tonsor. 2009. Evolutionary time for dispersal limits the extent but not the occupancy of species' potential ranges in the tropical plant genus *Psychotria* (Rubiaceae). *The American Naturalist* **173**:188–199.

Paulson, J. N., O. C. Stine, H. C. Bravo, and M. Pop. 2013. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* **10**:1200–1202.

Pebesma, E., and R. S. Bivand. 2005. Classes and methods for spatial data in R. *R News* **5**.

Pichancourt, J.-B., and R. D. van Klinken. 2012. Phenotypic plasticity influences the size, shape and dynamics of the geographic distribution of an invasive plant. *PloS ONE* **7**:e32323.

Pigot, A. L., I. P. F. Owens, and C. D. L. Orme. 2010. The environmental limits to geographic range expansion in birds. *Ecology Letters* **13**:705–715.

Pohlman, C. L., A. B. Nicotra, and B. R. Murray. 2005. Geographic range size, seedling ecophysiology and phenotypic plasticity in Australian Acacia species. *Journal of Biogeography* **32**:341–351.

Roberts, M. S., and F. M. Cohan. 1995. Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. *Evolution* **49**:1081–1094.

Roney, N. E., A. Kuparinen, and J. A. Hutchings. 2015. Comparative analysis of abundance–occupancy relationships for species at risk at both broad taxonomic and spatial scales. *Canadian Journal of Zoology* **93**:515–519.

Roselló-Mora, R., and R. Amann. 2001. The species concept for prokaryotes. *FEMS Microbiology Reviews* **25**:39–67.

Rosenfield, J. A. 2002. Pattern and process in the geographical ranges of freshwater fishes. *Global Ecology and Biogeography* **11**:323–332.

Ruff, S. E., J. F. Biddle, A. P. Teske, K. Knittel, A. Boetius, and A. Ramette. 2015. Global dispersion and local diversification of the methane seep microbiome. *Proceedings of the National Academy of Sciences* **112**:4015–4020.

Salisbury, C. L., N. Seddon, C. R. Cooney, and J. A. Tobias. 2012. The latitudinal gradient in dispersal constraints: ecological specialisation drives diversification in tropical birds. *Ecology Letters* **15**:847–855.

Salter, S. J., M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. J. Loman, and A. W. Walker. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12**:87.

Schlachter, K. J. 2010. Range shape and range elongation of North American trees. *Physical Geography* **31**:40–57.

Sergio, C., R. Figueira, D. Draper, R. Menezes, and A. Sousa. 2007. Modelling bryophyte distribution based on ecological information for extent of occurrence assessment. *Biological Conservation* **135**:341–351.

Sileshi, G., G. Hailu, and G. I. Nyadzi. 2009. Traditional occupancy–abundance models are inadequate for zero-inflated ecological count data. *Ecological Modelling* **220**:1764–1775.

687 Smith, D. J., H. J. Timonen, D. A. Jaffe, D. W. Griffin, M. N. Birmele, K. D. Perry, P. D. Ward,
688 and M. S. Roberts. 2013. Intercontinental dispersal of bacteria and archaea by transpacific
689 winds. *Applied and Environmental Microbiology* **79**:1134–1139.
690 Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta,
691 and G. J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored "rare
692 biosphere." *Proceedings of the National Academy of Sciences* **103**:12115–12120.
693 Stackebrandt, E., and B. M. Goebel. 1994. Taxonomic note: a place for DNA-DNA reassociation
694 and 16S rRNA sequence analysis in the present species definition in bacteriology.
695 *International Journal of Systematic and Evolutionary Microbiology* **44**:846–849.
696 Sul, W. J., T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, and M. L. Sogin. 2013. Marine
697 bacteria exhibit a bipolar distribution. *Proceedings of the National Academy of Sciences*
698 **110**:2342–2347.
699 Svenning, J.-C., and F. Skov. 2007. Could the tree diversity pattern in Europe be generated by
700 postglacial dispersal limitation? *Ecology Letters* **10**:453–460.
701 To, K.-Y., E.-M. Lai, L.-Y. Lee, T.-P. Lin, C.-H. Hung, C.-L. Chen, Y.-S. Chang, and S.-T. Liu.
702 1994. Analysis of the gene cluster encoding carotenoid biosynthesis in *Erwinia herbicola*
703 Eho13. *Microbiology* **140**:331–339.
704 Vos, M., and G. J. Velicer. 2008. Isolation by distance in the spore-forming soil bacterium
705 *Myxococcus xanthus*. *Current Biology* **18**:386–391.
706 Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naive Bayesian classifier for rapid
707 assignment of rRNA sequences into the new bacterial taxonomy. *Applied and*
708 *Environmental Microbiology* **73**:5261–5267.
709 Weil, T., C. De Filippo, D. Albanese, C. Donati, M. Pindo, L. Pavarini, F. Carotenuto, M.
710 Pasqui, L. Poto, J. Gabrieli, C. Barbante, B. Sattler, D. Cavalieri, and F. Miglietta. 12/2017.
711 Legal immigrants: invasion of alien microbial communities during winter occurring desert
712 dust storms. *Microbiome* **5**:23.
713 Wenger, S. J., and M. C. Freeman. 2008. Estimating species occurrence, abundance, and
714 detection probability using zero-inflated distributions. *Ecology* **89**:2953–2959.
715 Whitaker, R. J. 2003. Geographic barriers isolate endemic populations of hyperthermophilic
716 Archaea. *Science* **301**:976–978.

Wiens, J. J., C. H. Graham, D. S. Moen, S. A. Smith, and T. W. Reeder. 2006. Evolutionary and ecological causes of the latitudinal diversity gradient in hylid frogs: treefrog trees unearth the roots of high tropical diversity. *The American Naturalist* **168**:579–596.

Yamaguchi, N., T. Ichijo, A. Sakotani, T. Baba, and M. Nasu. 2012. Global dispersion of bacterial cells on Asian dust. *Scientific Reports* **2**:535.

DATA AVAILABILITY

Data associated with this study are available on figshare:
https://doi.org/10.6084/m9.figshare.1270900.v8

FIGURE LEGENDS

Figure 1.

Map of the contiguous United States with the locations of the 1,065 outdoor dust samples shown with blue points (1a). Geographic range size was calculated for dust taxa using two approaches, the area of occupancy (AOO) and the extent of occurrence (EOO) approximations (see Methods). Kernel density distributions for occupancy (i.e. total observations across sample sites) (1b), area of occupancy (AOO) range estimations (1c), and extent of occurrence (EOO) range estimations (1d) for dust phylotypes.

Figure 2.

Points show the maximum longitudinal and corresponding latitudinal range for each phylotype. Phylotypes with greater east-west spread also tend to have greater north-south spread (Spearman's $\rho = 0.67$, $P < 0.0001$). Since the United States has greater east-west than north-south dimensions, the blue dashed line normalizes for this difference and depicts the ratio of possible maximum spread. Points above this line (10.1%) indicate ranges elongated north-south, and points below this line (89.9%) indicate ranges elongated east-west (see inset). See Appendix S1: Fig S3 for the density distributions of longitudinal and latitudinal ranges.

Figure 3.

Boxplots illustrating range size distributions for dust taxa ranked by Phylum. Log_{10} AOO range size estimations vary significantly between phyla (one-way ANOVA; $F_{28, 70297} = 136.2$, $P < 0.0001$).

Figure 4.

Phenotypic traits and source habitats of dust bacteria were inferred by matching representative partial 16S rRNA phylotype sequences to full length 16S rRNA sequences in the IJSEM phenotype database (see Methods). Boxplots illustrate the relationship between the AOO range size estimation and oxygen tolerance (4a), spore formation in obligate aerobes (4b), pigmentation (4c), Gram stain (4d), and habitat (4e) for the most abundant phyla including *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes*. Range size varies significantly with oxygen tolerance (ANOVA; $F_{3, 2324} = 67.5$, $P < 0.0001$) (4a), spore formation (ANOVA; $F_{1, 1643} = 5.0$, $P = 0.025$) (4b), pigmentation (ANOVA; $F_{1, 1802} = 55.4$, $P < 0.0001$) (5c), Gram stain (ANOVA; $F_{2, 31207} = 32.4$, $P < 0.0001$) (4d), and habitat (ANOVA; $F_{7, 907} = 11.1$, $P < 0.0001$). See Appendix S1: Fig S7 for phenotypic traits by phyla.

Figure 5.

Genomic attributes of dust taxa were inferred by matching representative partial 16S rRNA sequences to full length 16S rRNA sequences in the IMG database (see Methods). Panels depict the relationship between AOO range size estimation and mean G+C content (5a), genome size (5b), and log_{10} 16S rRNA copy number (5c) for the most abundant phyla including *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes*. Points depict the AOO range size estimations and the mean values of genomic traits. Blue lines show the linear regression with gray shading indicating 95% confidence intervals. Pearson's product-moment correlation r is reported. See Appendix S1: Fig S8 for genomic traits by phyla.

Figure 1.

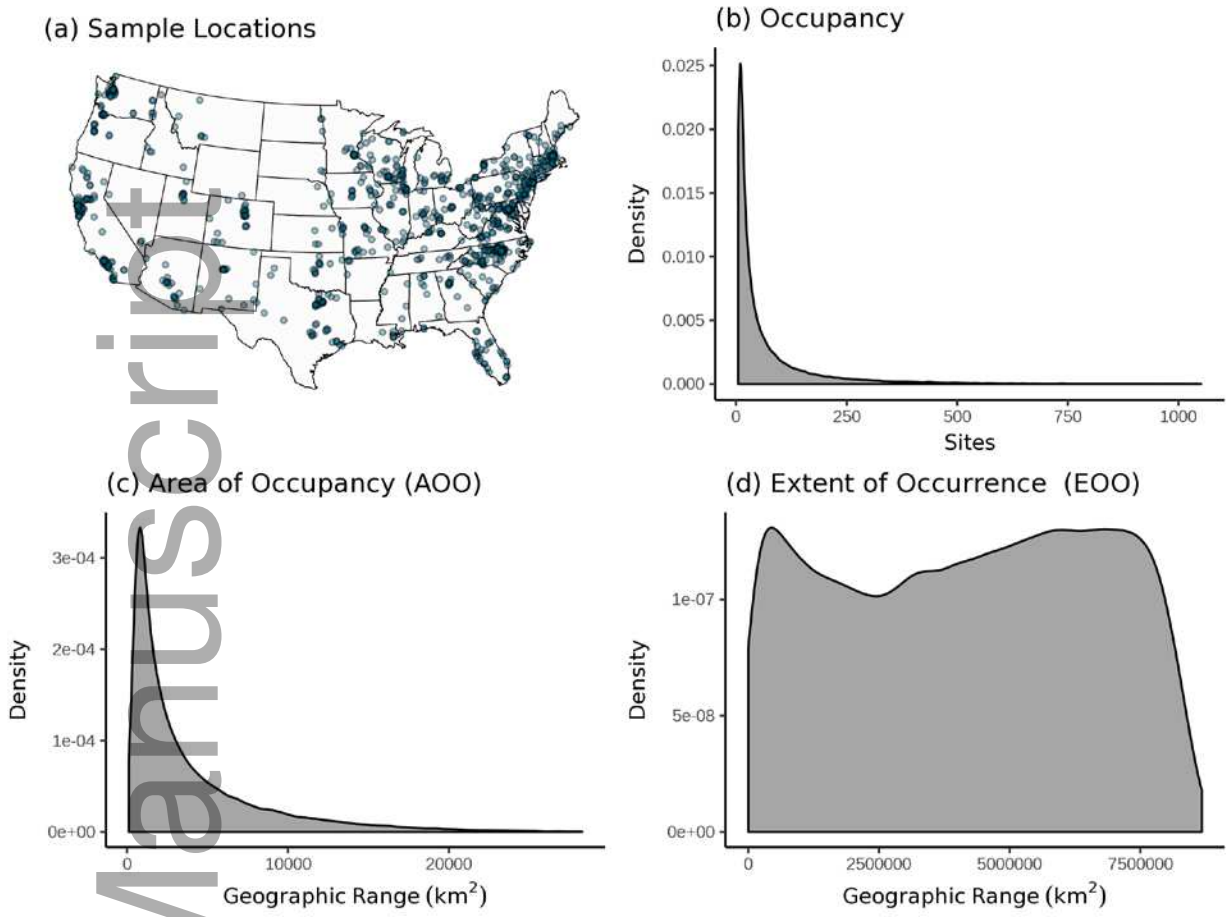

Figure 2.

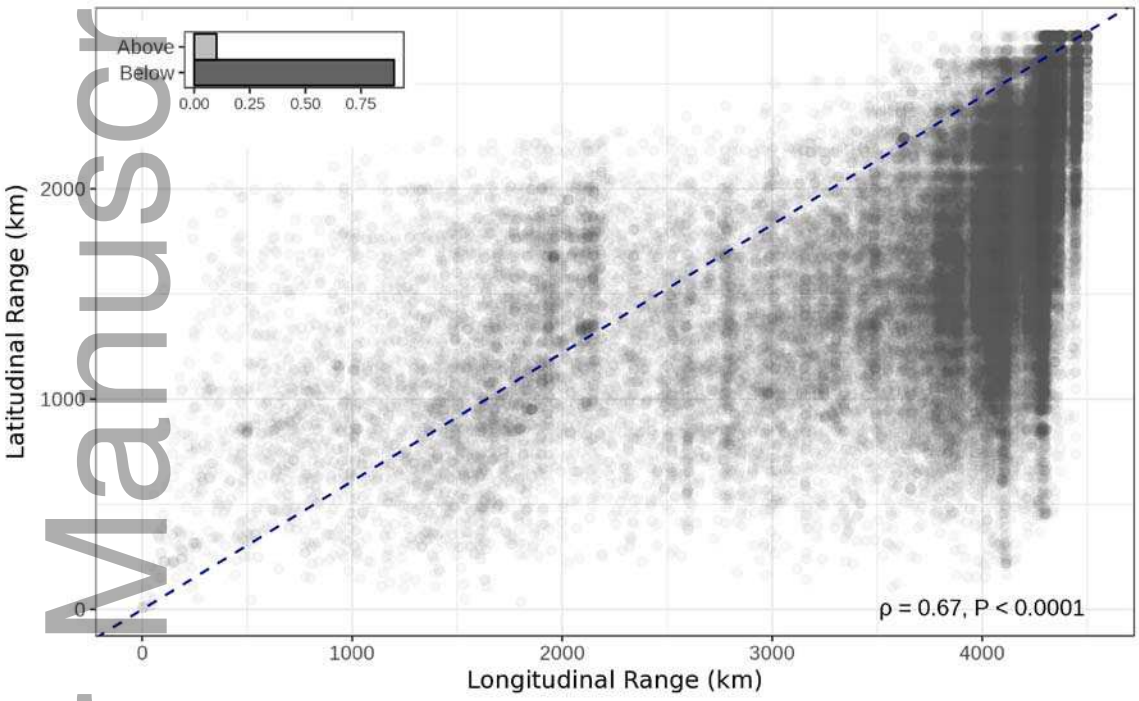

Figure 3.

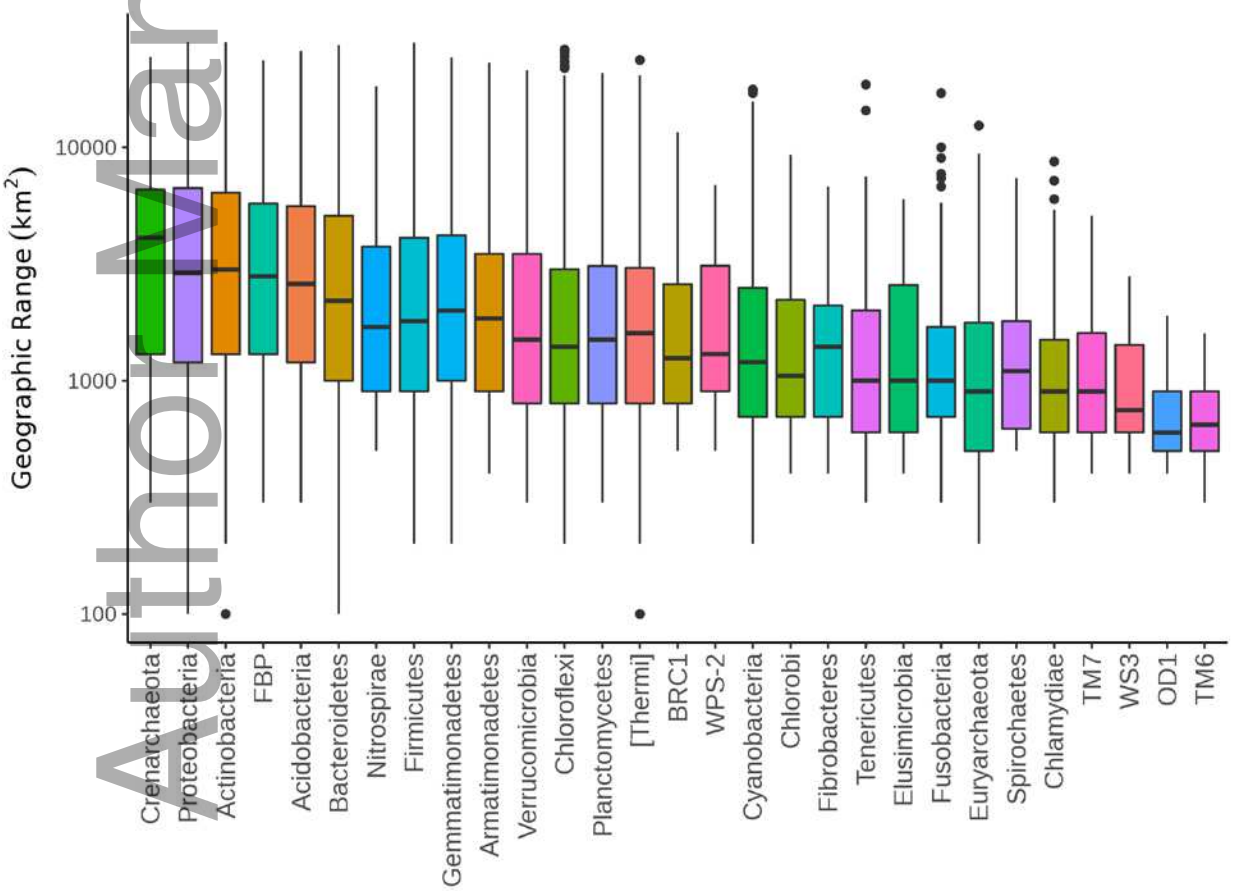

Figure 4.

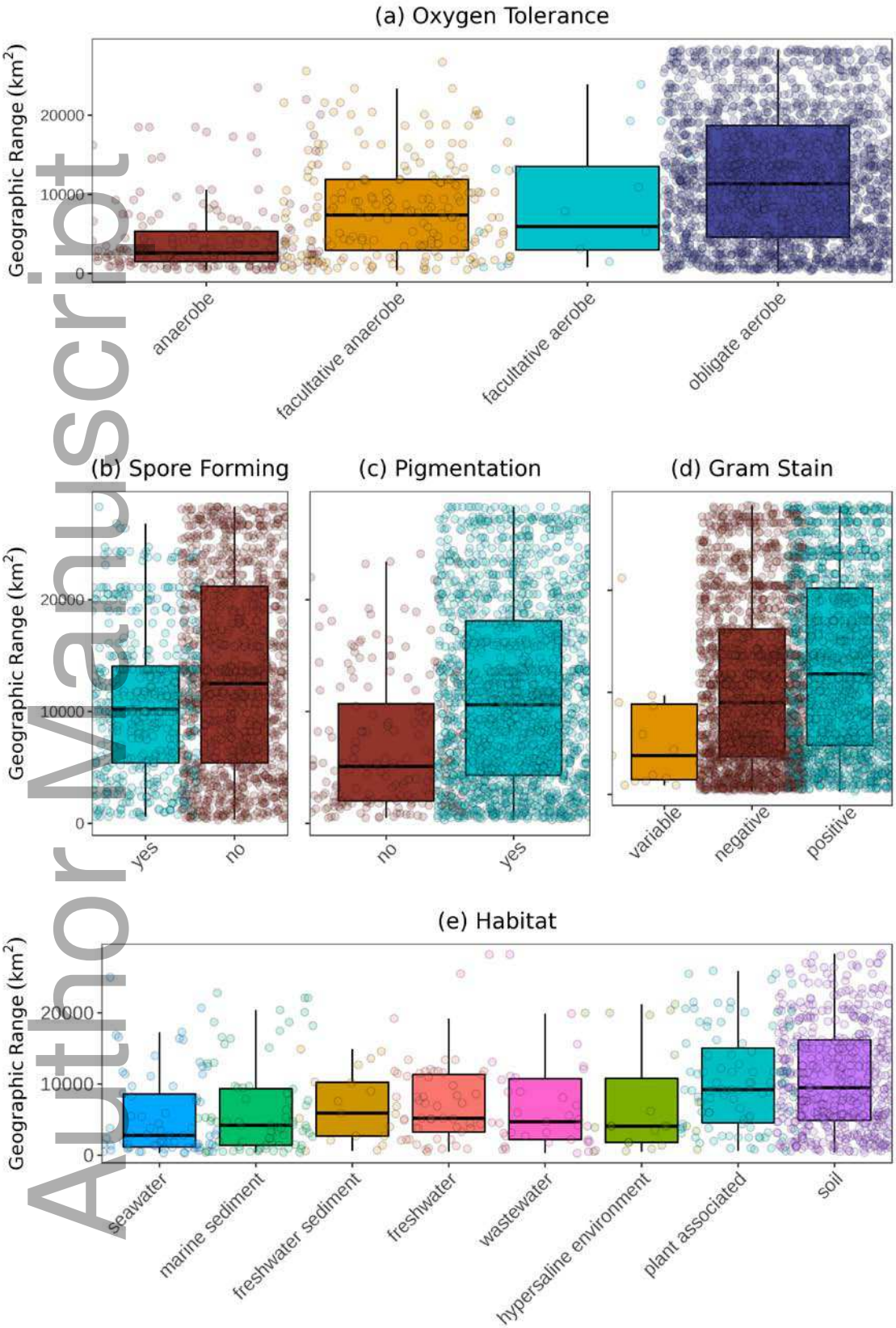

Figure 5.