www.sciencemag.org/content/359/6373/320/suppl/DC1

Supplementary Materials for

A global atlas of the dominant bacteria found in soil

Manuel Delgado-Baquerizo,* Angela M. Oliverio, Tess E. Brewer, Alberto Benavent-González, David J. Eldridge, Richard D. Bardgett, Fernando T. Maestre, Brajesh K. Singh, Noah Fierer*

*Corresponding author. Email: m.delgadobaquerizo@gmail.com (M.D.-B.); noah.fierer@colorado.edu (N.F.)

Published 19 January 2018, *Science* **359**, 320 (2018)
DOI: 10.1126/science.aap9516

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:
(available at www.sciencemag.org/content/359/6373/320/suppl/DC1)

Table S1 (Excel)

Material and Methods

Field survey and soil sample collection. Soils were collected from 237 locations across eighteen countries and six continents (Fig. S1). These sites include a wide range of ecosystem types (forests, grasslands, and shrublands) and climatic regions (arid, temperate, tropical, continental, and polar ecosystems). Mean annual precipitation and temperature in these locations ranged from 67 to 3085mm and -11.4º to 26.5ºC, respectively. Soil sample collection took place between 2003 and 2015. The coordinates of each site were recorded *in situ* with a portable GPS, and the ecosystem type (grassland, shrubland, or forest) of each location recorded. At each site, a composite soil sample (top ~7.5cm depth) was collected under the most common vegetation. After field collection, each soil sample was separated into two sub-samples - one sub-sample was immediately frozen at -20 ºC for molecular analyses while the other sub-sample was air-dried for chemical analyses.

PCR-based 16S rRNA gene analyses. Soil DNA was extracted using the Powersoil® DNA Isolation Kit (MoBio Laboratories, Carlsbad, CA, USA) according to the manufacturer's instructions. The extracted DNA samples were frozen and shipped to the Next Generation Genome Sequencing Facility of the University of Western Sydney (Australia), where a portion of the bacterial 16S rRNA gene (V3-V4 region) was sequenced using the Illumina MiSeq platform and the 341F/805R primer set. Bioinformatic processing was performed using a combination of QIIME (*31*), USEARCH (*32*) and UPARSE (*33*). Raw data were processed by trimming 20 nucleotides off the beginning and end of each sequence, then merged using the usearch7 command with a fastq_maxee of 1. Sequences were next dereplicated, and phylotypes were identified at the ≥97% identity level using UCLUST (*32*). Taxonomy was assigned using the Ribosomal Database Project classifier (*34*) and the Greengenes 13_8 database (*35*). The resulting phylotype tables were rarefied to 10000 sequences per sample. We further removed phylotypes that were represented by only a single read across all samples. In addition, we removed any archaeal, chloroplasts and mitochondria phylotypes, which together accounted for 0.8% of all phylotypes (204 of 25,424 phylotypes).

Soil and site characteristics. To avoid biases associated with having multiple laboratories analyzing soils from different sites, and to facilitate the comparison of results between them, all dried soil samples were shipped to the Universidad Rey Juan Carlos (Spain) for laboratory analyses. For all soil samples, we measured pH, texture, total organic carbon (soil C), total nitrogen (soil N) and total phosphorus (soil P) concentrations using standard laboratory methods. pH was measured in all the soil samples with a pH meter, in a 1: 2.5 mass: volume soil and water suspension. Texture (% of fine fractions: clay + silt) was determined according to ref. *36*. The concentration of soil total organic carbon (C) was determined using a wet chemistry method described in ref. *37*. Soil total N was measured with a CN analyzer (Leco CHN628 Series, LECO Corporation, St Joseph, MI, USA) and total phosphorus (P) was measured using a SKALAR San++

Analyzer (Skalar, Breda, The Netherlands) after digestion with sulphuric acid. The collected soils represent a wide range in soil properties. In brief, soil pH ranged from 4.04 to 9.21, soil C from 0.15 to 34.77%, soil N from to 0.02 to 1.57%, soil P from 75.10 to 4111.04 mg P Kg-1 soil, C:N ratio from 2.12 to 67.52 and fine texture fraction (% clay+silt) from 1.40 to 92.00%.

We obtained information on maximum and minimum temperature, precipitation seasonality, and mean diurnal temperature range (MDR) for all sampling locations from the Worldclim database (www.worldclim.org), which has a 1 km resolution (*38*). In addition, for each site we estimated the Aridity Index (Precipitation/evapotranspiration) from the Global Potential Evapotranspiration database (*39*), which is based on interpolations provided by WorldClim (*38*). We used the Aridity Index rather than mean annual precipitation because Aridity Index includes both mean annual precipitation and potential evapotranspiration, and is therefore a better measure of the long-term water availability at each site. We obtained information on annual ultraviolet index (UV index) from the NASA's Aura satellite (https://neo.sci.gsfc.nasa.gov) (*40*), which has a 50 km resolution. The UV index is a measure of the intensity of UV radiation ranging from 0 (minimal UV exposure risk) to 16 (extreme risk).

We used the Normalized Difference Vegetation Index (NDVI) as our proxy for net plant primary productivity (*41-42*). This index provides a global measure of the "greenness" of vegetation across Earth's landscapes for a given composite period, and thus acts as a proxy of photosynthetic activity and large-scale vegetation distribution (*41-42*). NDVI data were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Terra satellites (http://neo.sci.gsfc.nasa.gov/) as described in ref. *42*. We calculated the monthly average value for this variable between the 2003-2015 period (~10km resolution), when all soil sampling was conducted.

Identification of the dominant bacterial phylotypes. We identified the most common and ubiquitous phylotypes across our global dataset following two criteria. First, we identified the top 10% most abundant phylotypes based on total number of reads across all samples (as described in ref. *21*). Abundance is widely accepted as a metric of how common or rare species (here 'phylotypes') are in their environment, therefore is a useful metric to identify dominant phylotypes *(21*). Second, we only kept those phylotypes that were also found in more than half of the samples (i.e., > 55% of samples). These phylotypes were considered to be widely present across soil samples and therefore to be reasonably ubiquitous.

For the isolate and reference strain data, we matched our amplicon sequences to appropriate databases maintained by RDP and counted hits of 97% similarity as matches. For the reference genomes, we matched our amplicon sequences to the Integrated Microbial Genomes & Microbiomes (IMG/M) database (https://img.jgi.doe.gov). We took into consideration those new genomes from ref. *43*.

Dominant taxa cross-validation #1: shotgun sequencing data. We validated the ubiquity of identified phylotypes with an independent previously published shotgun metagenomic dataset (*18*) that included a total of 123 soils collected from a broad range of locations to confirm that the same phylotypes are also dominant when community composition is assessed using a PCR-free approach (*44*). Using Metaxa2 (*45*), we extracted 16S rRNA gene sequences from these shotgun datasets, then matched the 16S rRNA gene sequences to the Greengenes database (*35*) using the usearch7 command -usearch_global at ≥97% identity. We used these matches to obtain longer sequences that would uniformly contain the specific hyper-variable region covered by the primer pair 341F/805R. We then used the same usearch command to compare the representative sequences of the dominant phylotypes to the Greengenes reference database. We counted sequences as present in both shotgun and amplicon datasets if they had at least 97% similarity to each other. Note that, unlike 16S rRNA amplicon sequencing, shotgun metagenomic sequencing can include all DNA present in a given sample, not just 16S rRNA genes from bacteria (with 16S rRNA genes representing ~0.04% of metagenomic reads, on average). As we were only able to recover a relatively small number of 16S rRNA genes in each shotgun metagenome, we assumed that the bacterial 16S rRNA genes identified using this approach represent those phylotypes that are highly abundant in soil.

Dominant taxa cross-validation #2: the Earth Microbiome Project (EMP). We used data from the EMP (*22*) to further validate our results. Note that any comparisons between the EMP dataset and our dataset need to be considered carefully given methodological differences in the primer sets used (here 341F/805R vs. 515F/806R for the EMP), read lengths (here 400bp/sequence vs. <150bp, but mostly <100bp, for the EMP) and lack of standardization in the EMP soil sampling protocols and metadata collection. We selected all soil samples from the EMP that were comparable with those in our dataset (soil samples from <10cm depth). We used the subset of 2,004 EMP samples (100bp) from soil (<10cm depth), rarefying this dataset to 10,000 reads/sample (as done in our original analyses). Using the same approach explained above, we identified the dominant taxa across the 2,004 EMP samples, i.e. the top 10% most common phylotypes found in more than half (>55%) of the soil samples. After conducting these analyses we found that 97 phylotypes were dominant in the subset of the EMP data used here (vs. our top dominant 511 phylotypes). The majority of the dominant phylotypes in the EMP data (80%) were included within our dominant taxa (>97% similarity). We also repeated the analyses included in Fig. 1A of our manuscript and found, that for the EMP data used, the top 511 dominant phylotypes accounted for ~41% of all reads, but they only represented 0.5% of all the bacterial phylotypes (>35% ubiquity). Therefore, we found very similar results to those reported from our dataset (511 phylotypes accounting for 41% of all reads). Considering that our study used different methods and given the aforementioned limitations of the EMP dataset, we believe that the similarity between our results and the results obtained by re-analyzing the EMP dataset are compelling.

Importantly, in both datasets we find that a few hundred taxa account for an enormous proportion of the soil bacterial communities found across the globe.

Identifying groups of dominant phylotypes with shared habitat preferences. We used Random Forest analysis (*24*) as explained in ref. *42* to identify the environmental preferences of each of the dominant bacterial phylotypes across the globe. We considered that we were able to identify the environmental preferences for a given phylotype when the Random Forest model explained >30% of variation in the distribution of this phylotype, which is considered to be a high level of variation explained in the context of large scale studies (*46*). Our models included 15 environmental predictors: climate variables (Aridity Index, minimum and maximum temperature, precipitation seasonality and mean diurnal temperature range –MDR), UV radiation, net primary productivity (NDVI index), soil properties (texture [% of clay + silt], soil pH, total C, N and P concentrations and C: N ratio) and dominant ecosystem types in our dataset (forest and grasslands). Ecosystem types were coded as categorical variables with two levels: 1 (a given ecosystem type) and 0 (remaining ecosystem types). This approach allowed us to compare the effect of a particular ecosystem type on the relative abundance of each phylotype compared with the average of the remaining ecosystem types. Note that minority ecosystem types in this dataset (i.e., shrublands) were selected as our baseline condition (i.e. procedural control), and thus were not explicitly included in our model. These analyses were conducted using the rfPermute package (*47*).

We next clustered the phylotypes with known environmental preferences (% variation explained from Random Forest > 30%) into different ecological groups. To do this, we conducted semi-partial correlations (Spearman) using the ppcor package (*48*) to further identify the unique contribution of each predictor in explaining the distribution of a given phylotype. Unlike regular correlations, semi-partial correlations allow us to identify the variance from a given response variable (here dominant bacterial phylotypes) that is uniquely predictable from a given predictor, controlling for all other predictors simultaneously (*49*). Information on semi-partial correlations (significant $P < 0.05$ correlation coefficients) was then used to cluster our dominant bacterial phylotypes in different ecological clusters with hierarchical cluster analysis (as implemented in the "hclust" function in the R package "stats"). We used a heatmap (heatmap.2 function in the R package gplots) to visualize our ecological clusters (Fig. S11). We then computed the relative abundance of each cluster per sample by averaging the standardized (z-score) relative abundance of the phylotypes that belong to each ecological cluster. Using this approach, each phylotype contributed equally to the final relative abundance of each ecological cluster.

Phylogenetic analyses. We built a phylogenetic tree for the 511 dominant phylotypes to visualize the extent to which environmental preferences, reference isolates, and reference genomes were phylogenetically clustered. To obtain a more robust phylogeny, we first identified the nearest neighbor for each sequence at

the 98% cutoff with the "search and classify" function with the Silva Incremental Aligner (SINA v1.2.11) (*50*). We then aligned those representative sequences and the remaining original sequences (those without a 98% match) using SINA with default parameters (*50*). After aligning, gaps were trimmed with trimAl (threshold = 0.2) (*51*). We then built a tree with FastTree using a GTR model of nucleotide evolution and visualized the tree with GraPhlAn (*52*).

Network analyses. We used correlation network analyses to evaluate whether dominant bacterial phylotypes within a particular ecological cluster were found to co-occur more often than expected by chance. To build the co-occurrence network, we first calculated pairwise Spearman's rank correlations (ρ) between all dominant bacterial phylotypes. We focused exclusively on positive correlations, as they provide information on microbial phylotypes that may respond similarly to environmental conditions. We considered a co-occurrence to be robust if the Spearman's correlation coefficient (ρ) was > 0.65 and $P < 0.00001$ (*53*). The network we recovered included 270 nodes with 3646 edges. This network was visualized with the interactive platform gephi (*54*). We then investigated whether microbial phylotypes tend to co-occur with others in the same ecological cluster (as identified with the Random Forest and clustering analyses). To do this, we generated 1,000 random graphs with the same number of nodes and edges as the derived network, under the Erdős–Rényi model. This allowed us to estimate null distributions (in a method similar to that described in ref. *55*) for the likelihood of co-occurrences across versus within ecological clusters, providing a metric for the robustness of each ecological cluster. We conducted these analyses using the igraph package (v1.0.1) and custom R functions (available from https://github.com/amoliverio/rnetworks).

Mapping of ecological groups across the globe. We used the prediction-oriented regression model Cubist (*28*) to predict the distribution of the four major ecological clusters (i.e., low pH, high pH, drylands and low productivity clusters) across the globe. The Cubist algorithm uses a regression tree analysis to generate a set of hierarchical rules using information on environmental covariates (*56*), which are later used for spatial prediction (*56*). Covariates in our models include 12 out of the 15 environmental predictors evaluated: soil properties (soil C, soil pH and texture), climate (MDR, Aridity Index, maximum and minimum temperature, precipitation seasonality), net primary productivity, UV radiation and major vegetation types (forests and grasslands). We did not include soil total N, P and C:N ratio in these analyses because (1) they were not selected as major drivers of dominant phylotypes (Fig. S6) and (2) high-resolution information on these variables is not available at the global scale. Global predictions on the distribution of major clusters were done on a 25km resolution grid. Global information on soil properties for this grid was obtained using the ISRIC (global gridded soil information) Soil Grids (https://soilgrids.org/#!/?layer=geonode:taxnwrb_250m). Similarly, global information on the major

vegetation types in this study (grasslands and forests) was obtained using the Globcover2009 map from the European Space Agency (http://due.esrin.esa.int/page_globcover.php) (*57*). Global information on climate, UV radiation and net primary productivity were obtained from the WorldClim database (www.worldclim.org) (*38*) and NASA satellites (https://neo.sci.gsfc.nasa.gov), as explained above. We used the package Cubist in R to conduct these analyses (*56*).

Cross-validation of maps using data from the Earth Microbiome Project (EMP). We cross-validated our maps using the selected soil samples from the EMP dataset used above (*22*). We focused this cross-validation on the top two clusters identified in this study (Low and High pH) which included the largest number of phylotypes (Fig. 3). The EMP high and low pH clusters included the dominant phylotypes from the EMP (as defined in our study), which were highly related (>97% similarity) to the phylotypes within the high pH and low pH cluster from our study. The relative abundance of the High and Low pH cluster in the EMP dataset was calculated as the average standardized abundance (z-score) of EMP phylotypes assigned to these two ecological clusters, as explained for our dataset above. Then, using the spatial information (latitude and longitude) for the selected 2004 soil samples from the EMP, and the information derived from our maps in Fig. 4, we extracted the predicted abundances of the high and low pH clusters for these selected EMP locations. Finally, we correlated the relative abundance of these two High and Low pH clusters based on our map predictions with that from the same clusters calculated for the EMP. We found strong positive and significant correlations between information based on our maps and that from the EMP data: High pH cluster Pearson´s r = 0.41 ($P < 0.001$) and Low pH cluster Pearson´s r = 0.32 ($P < 0.001$). This concordance between our predictions and independent results obtained from the EMP data is compelling given the local scale variation in soil properties and the fact that our data and the EMP data were independently generated using different methods (see above). Therefore, these results strongly support the validity of our maps as representations of the distribution of ecological clusters of dominant taxa across the globe.

Identifying genomic attributes within ecological groups. We next identified the genomic attributes within ecological groups – in particular within the drylands cluster, as there were a sufficient number of unique reference genomes for phylotypes included in this cluster (Fig. S13). To do this, we obtained information on ~20000 genes characterizing the genomic attributes for all unique genomes in our dataset (Table S1). We obtained this information from the Kyoto Encyclopedia of Genes and Genomes database (www.genome.jp/kegg/) using the Integrated Microbial Genomes & Microbiomes (IMG/M) system (https://img.jgi.doe.gov). We only included in our analyses those genomes that matched >97% a reference genome and were over ~90% complete. A total of 72 genomes were included in this analysis, with 10 of these genomes belonging to the dryland cluster. We further filtered our gene database to maintain those

genes that had >5 gene counts across all genomes. Finally, we used Random Forest analyses (*24*) as described in ref. *58* to identify the main genes characterizing genomes within the dryland clusters versus those genomes representing phylotypes assigned to other clusters. In this respect, our response variable in this analysis is a categorical variable including “drylands” and “others” and our predictor variables are the genomic attributes.

Appendix S1.

Extended results regarding the mapping of ecological clusters. According to the results from the Random Forest analyses and semi-partial correlations, the Cubist model found the following variables to be the most important predictors of the following ecological clusters (values inside the parenthesis indicate the model usage of those environmental covariates for mapping): **(1) High pH**: pH (100%), net primary productivity (60%), maximum temperature (30%), MDR (38%), UV radiation (32%), precipitation seasonality (30%) and minimum temperature (30%). **(2) Low pH**: pH (100%), precipitation seasonality (58%), minimum temperature (22%), MDR (72%), Aridity Index (41%), UV radiation (37%), net primary productivity (35%), maximum temperature (25%) and soil C (8%). **(3) Drylands**: Aridity Index (100%), precipitation seasonality (25%), UV radiation (100%), pH (96%), forests (96%), clay+silt (59%), C (35%), net primary productivity (29%) and MDR (25%). (4) **Low productivity**: net primary productivity (100%), soil C (100%), Aridity Index (63%), pH (45%), precipitation seasonality (45%), maximum temperature (35%), minimum temperature (35%), MDR (35%) and clay+silt (20%).

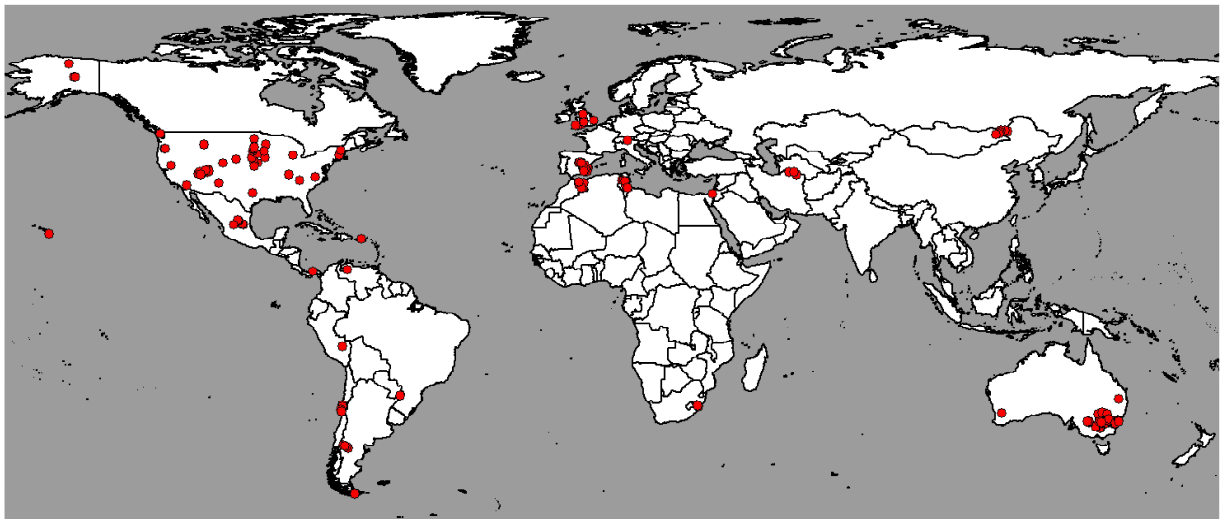


Figure S1. Locations of the 237 soil sampling sites included in this study.

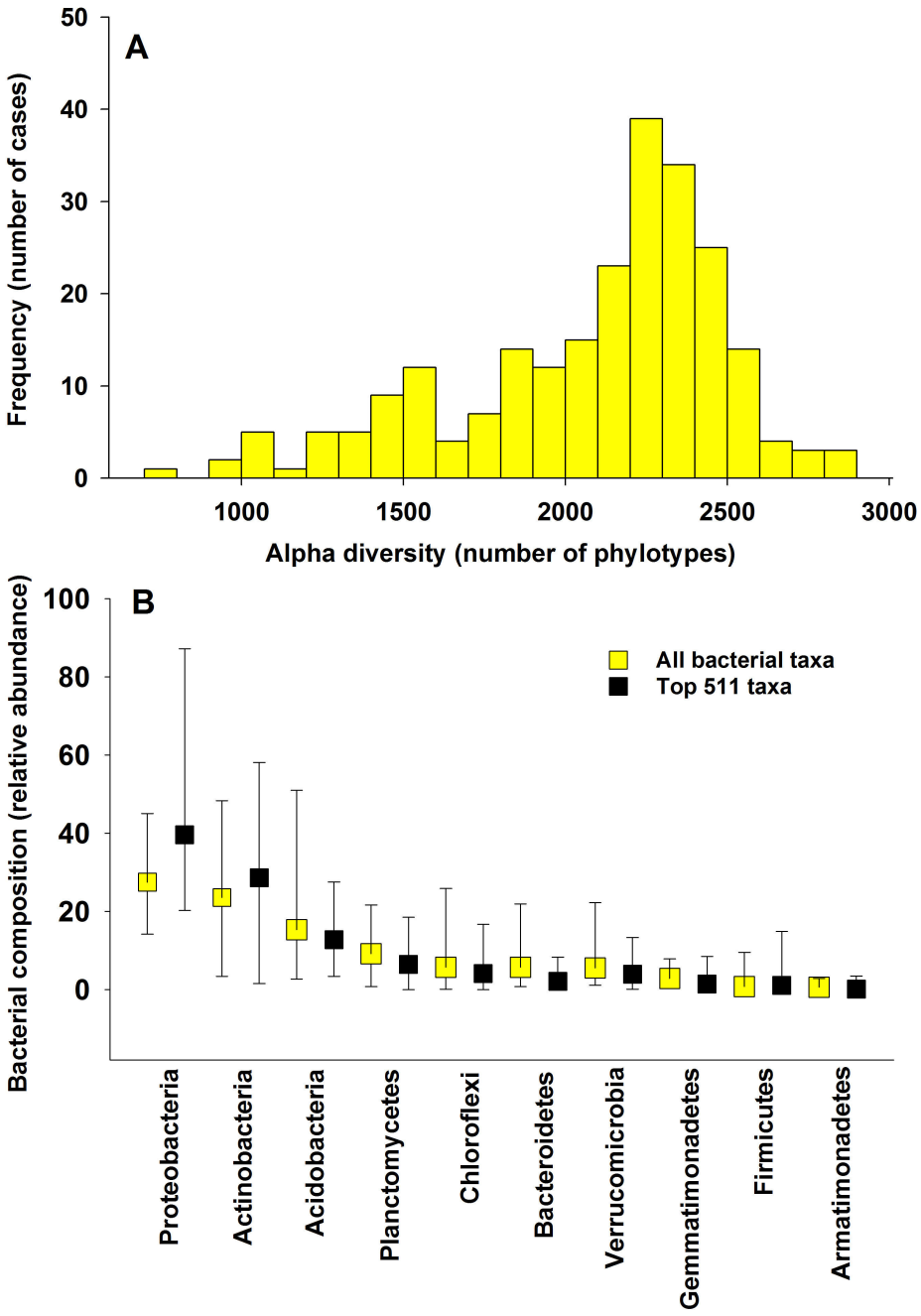


Figure S2. Richness and composition of the bacterial communities across the 237 soil samples included in this study. (A) Distribution of per-sample bacterial richness across the globe at a rarefied sequencing depth of 10,000 16S rRNA gene reads per sample. (B) Relative abundances (mean ± maximum/minimum values) of major groups of bacteria for the entire bacterial community and for the subset identified as dominant (2% of bacterial phylotypes).

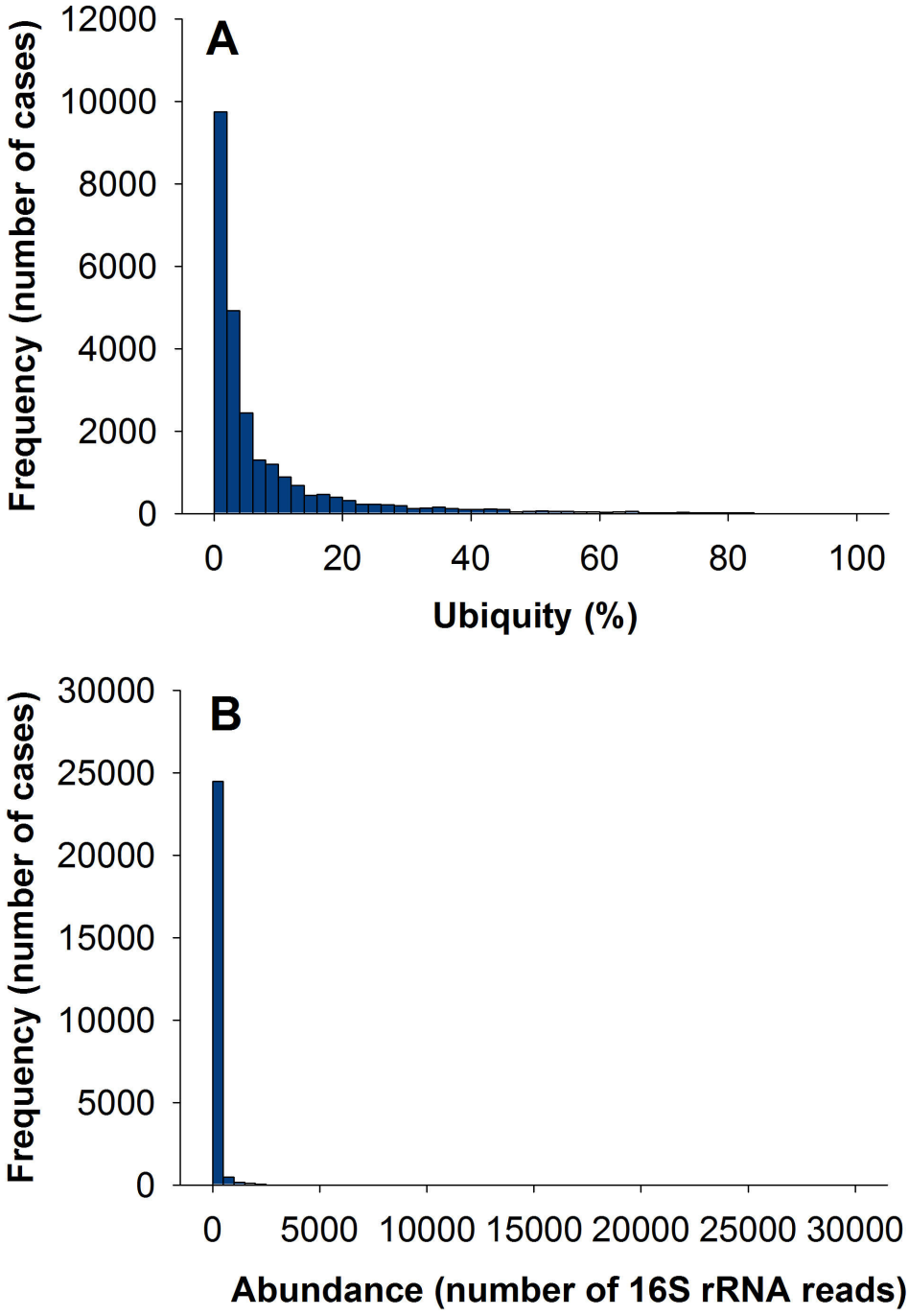


Figure S3. Histograms reporting the distributions for ubiquity (A) and the relative abundances of 25224 taxa (B) in the 237 soil samples from across the globe.

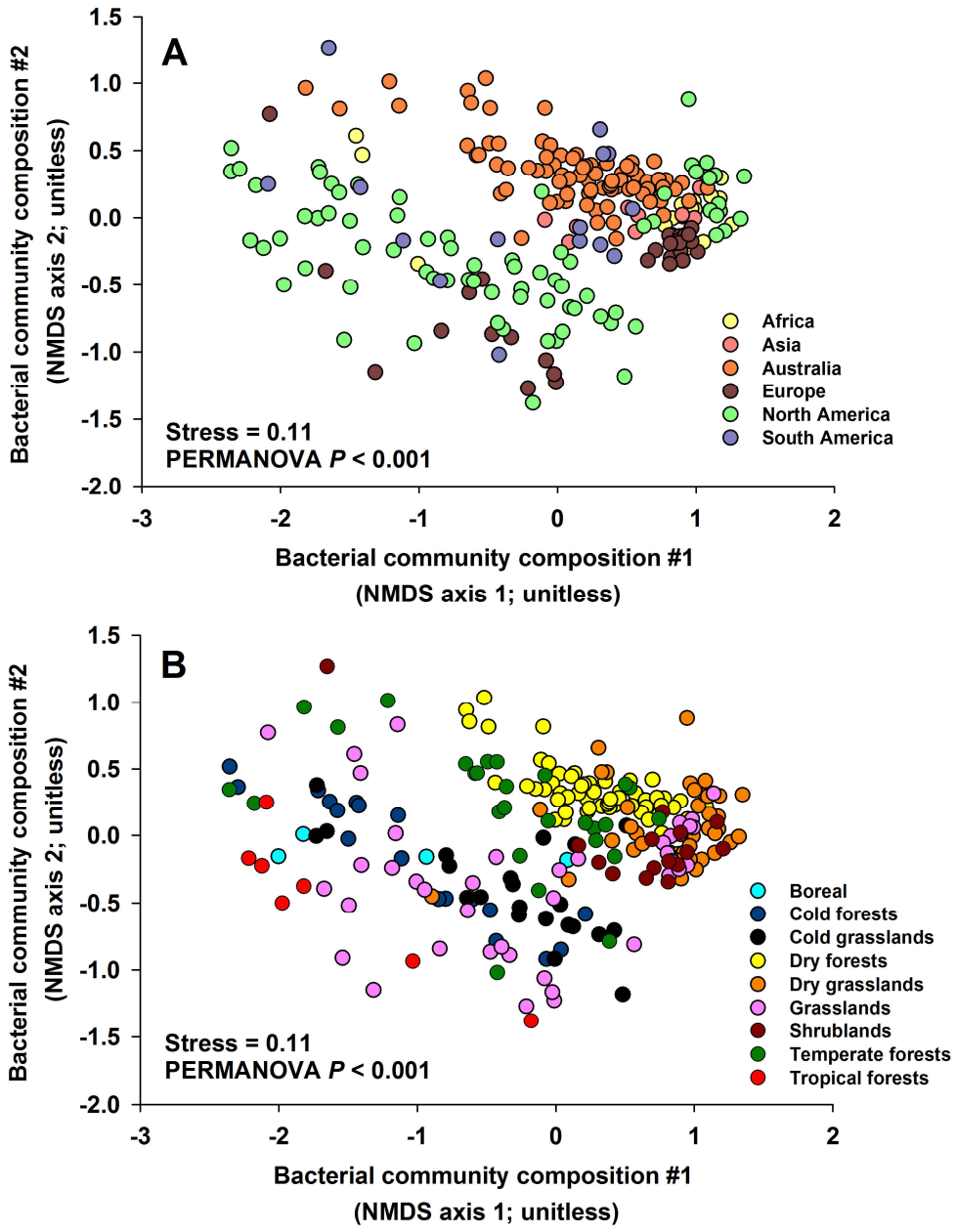


Figure S4. NMDS ordination summarizing the dissimilarity in community composition of bacteria across the globe for different continents (A) and ecosystem types (B). Grasslands include both tropical and temperate grasslands. Shrublands include polar, temperate and tropical shrublands.

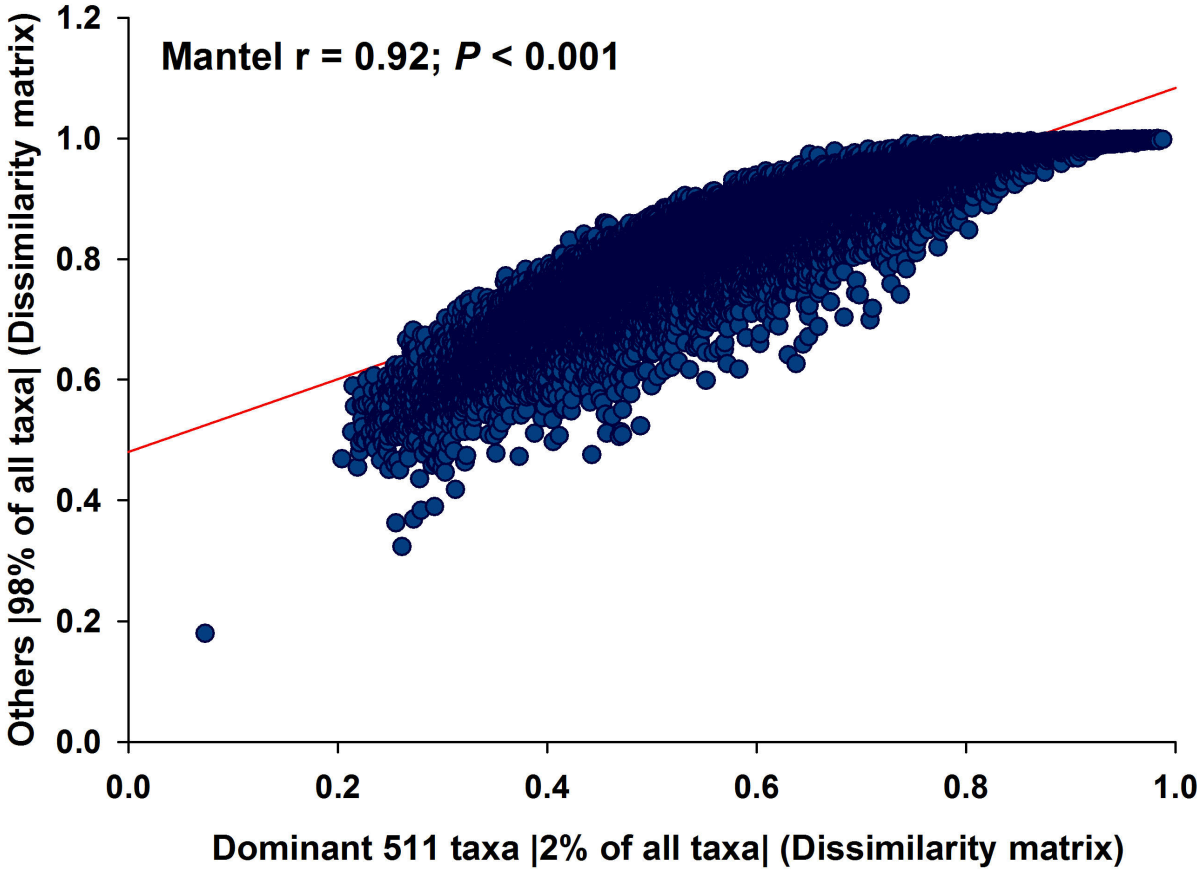


Figure S5. Relationship between beta diversity (community dissimilarity) based on Bray-Curtis distance for the dominant (511 phylotypes) and the remaining 24713 bacterial phylotypes. Correlation was done using the Mantel test.

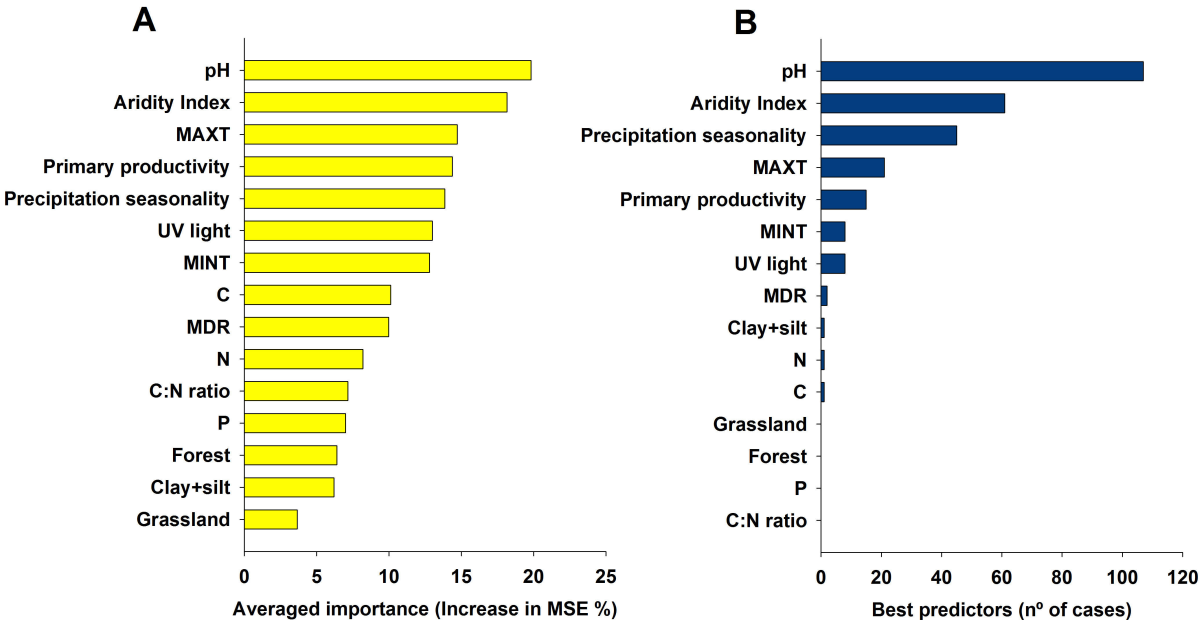


Figure S6. Major predictors of the distribution of dominant bacterial taxa across the globe. Averaged importance of environmental factors (across 270 Random Forest models) in predicting the relative abundance of dominant bacterial taxa (A). Number of cases (out of 270 Random Forest models) for which a particular environmental factor is the best predictor for the dominant bacterial taxa (B). MINT = minimum temperature; MAXT = maximum temperature; MDR = Mean diurnal temperature range. Primary productivity = net primary productivity.

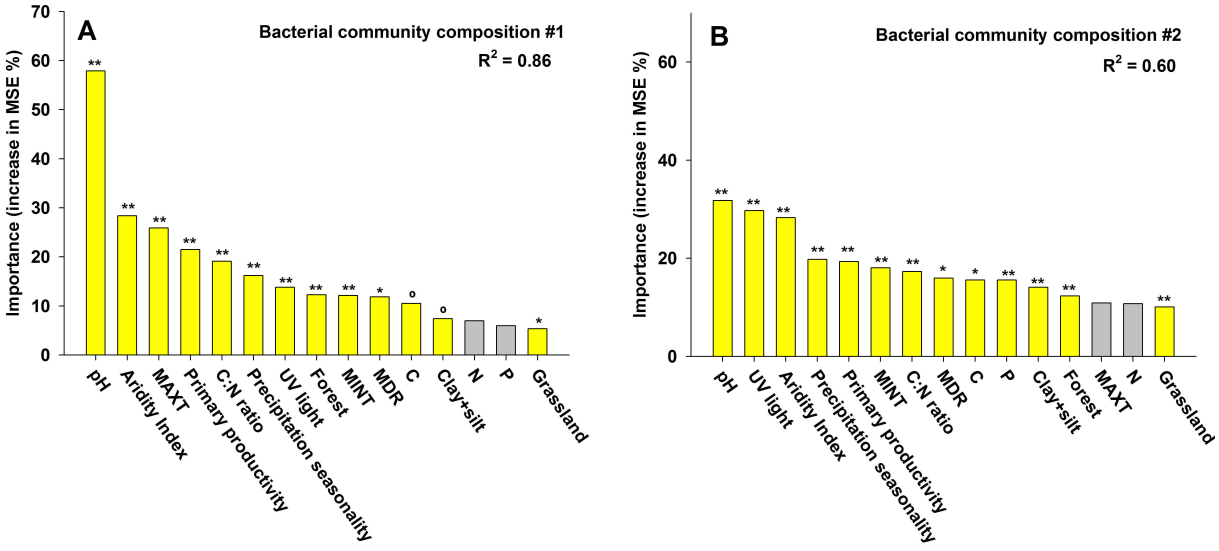


Figure S7. Major predictors of the distribution of bacterial communities across the globe. Panels (A) and (B) include the importance of environmental factors in predicting the relative abundance of bacterial community composition (two axes from a NMDS summarizing information on the overall community composition of bacteria at the phylotype level). MINT = minimum temperature; MAXT = maximum temperature; MDR = Mean diurnal temperature range. Primary productivity = net plant primary productivity. Significant levels are: **P < 0.01, *P < 0.05 and °P < 0.10.

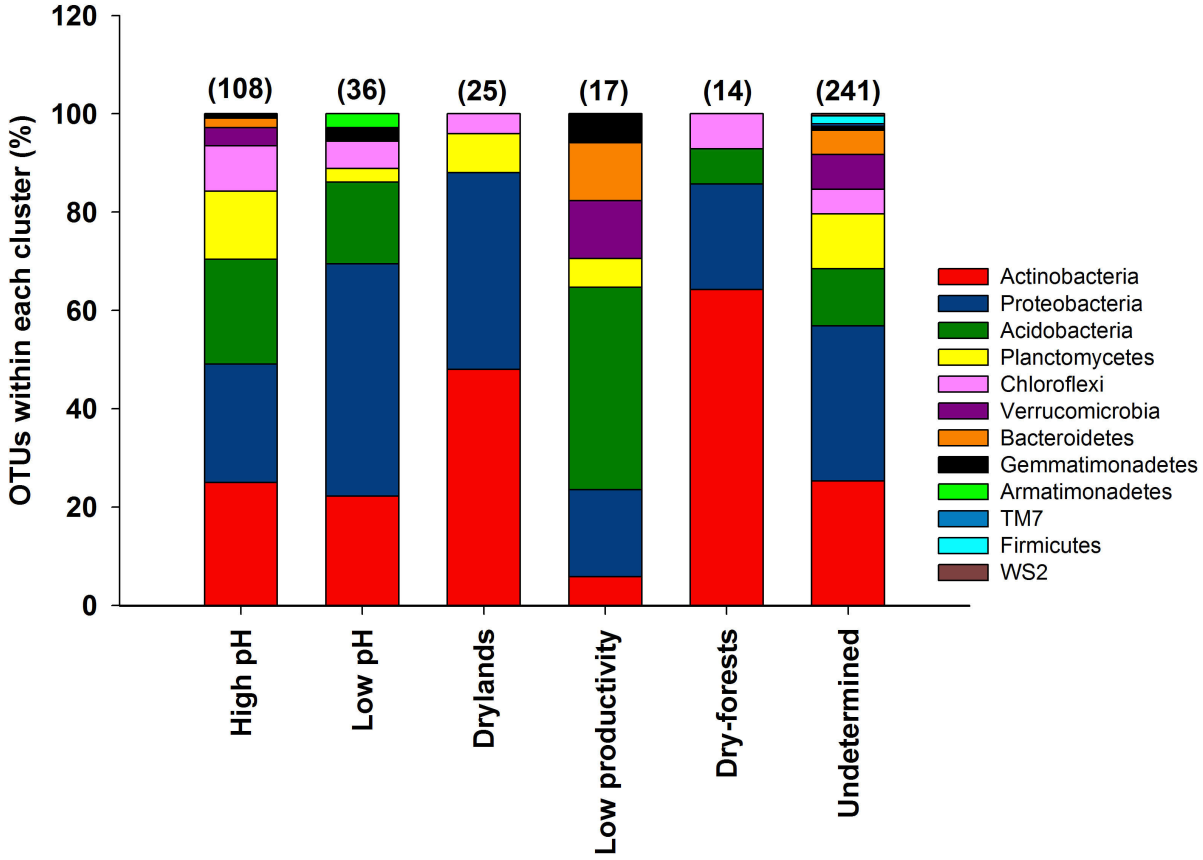


Figure S8. Taxonomic composition (% of phylotypes (OTUs) within each cluster) for five well-defined ecological clusters of bacterial phylotypes sharing habitat preferences and also for those phylotypes for which we were not able to identify their niche model (undetermined phylotypes). The total number of phylotypes per cluster is indicated in parentheses.

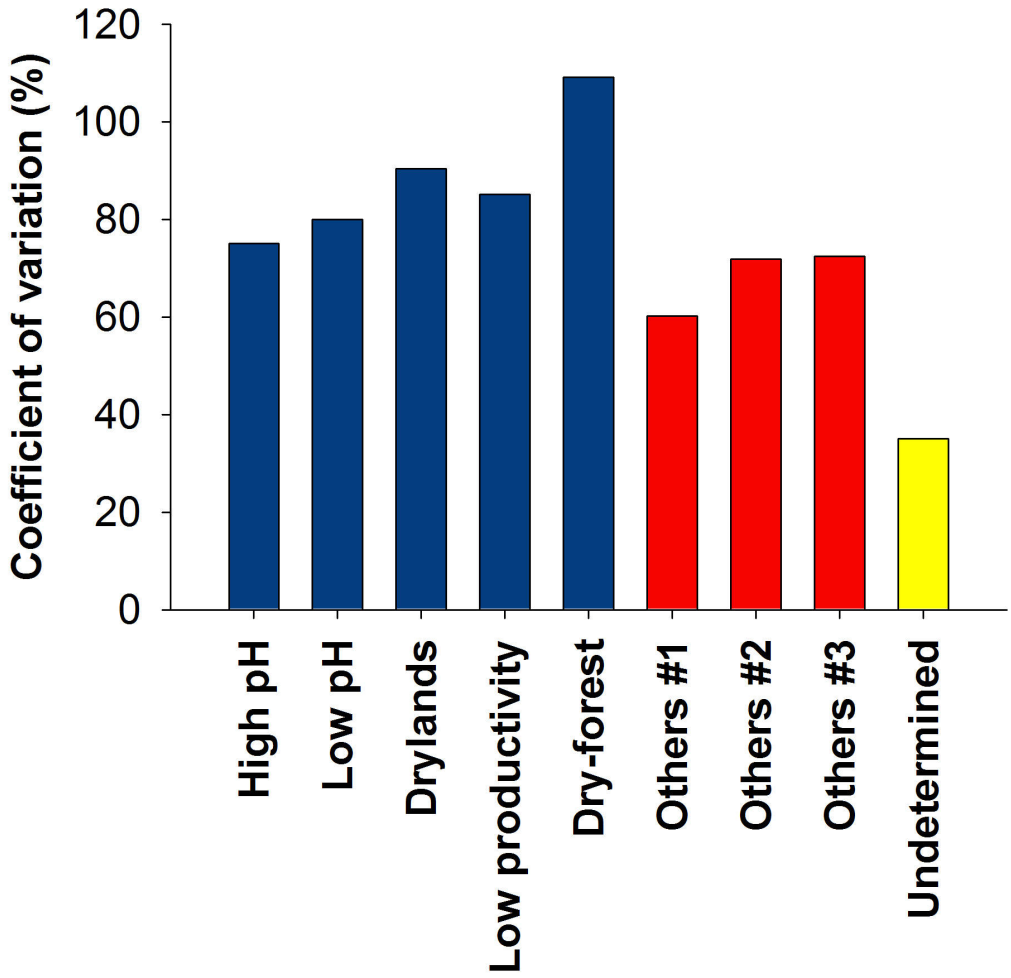


Figure S9. Coefficients of variation in the relative abundances of dominant bacterial phylotypes assigned to each of the five major ecological clusters and those phylotypes that fell within an 'undetermined' group (those dominant bacterial phylotypes with no identifiable habitat or environmental preferences).

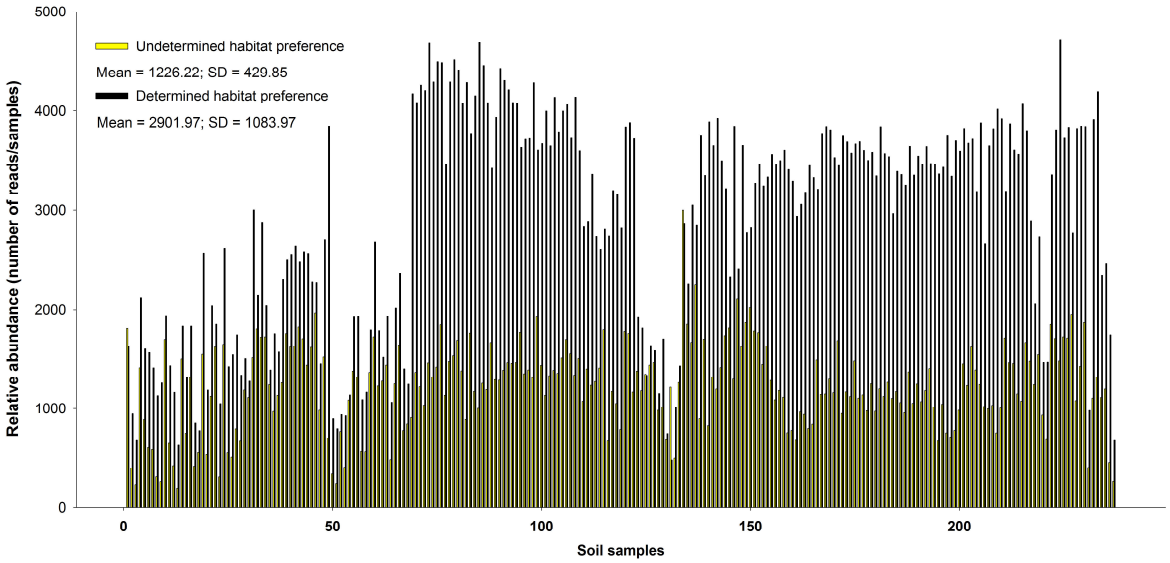


Figure S10. Sum of the relative abundances (per sample) of taxa with defined and undefined habitat preferences for the 511 dominant bacterial phylotypes. SD = standard deviation.

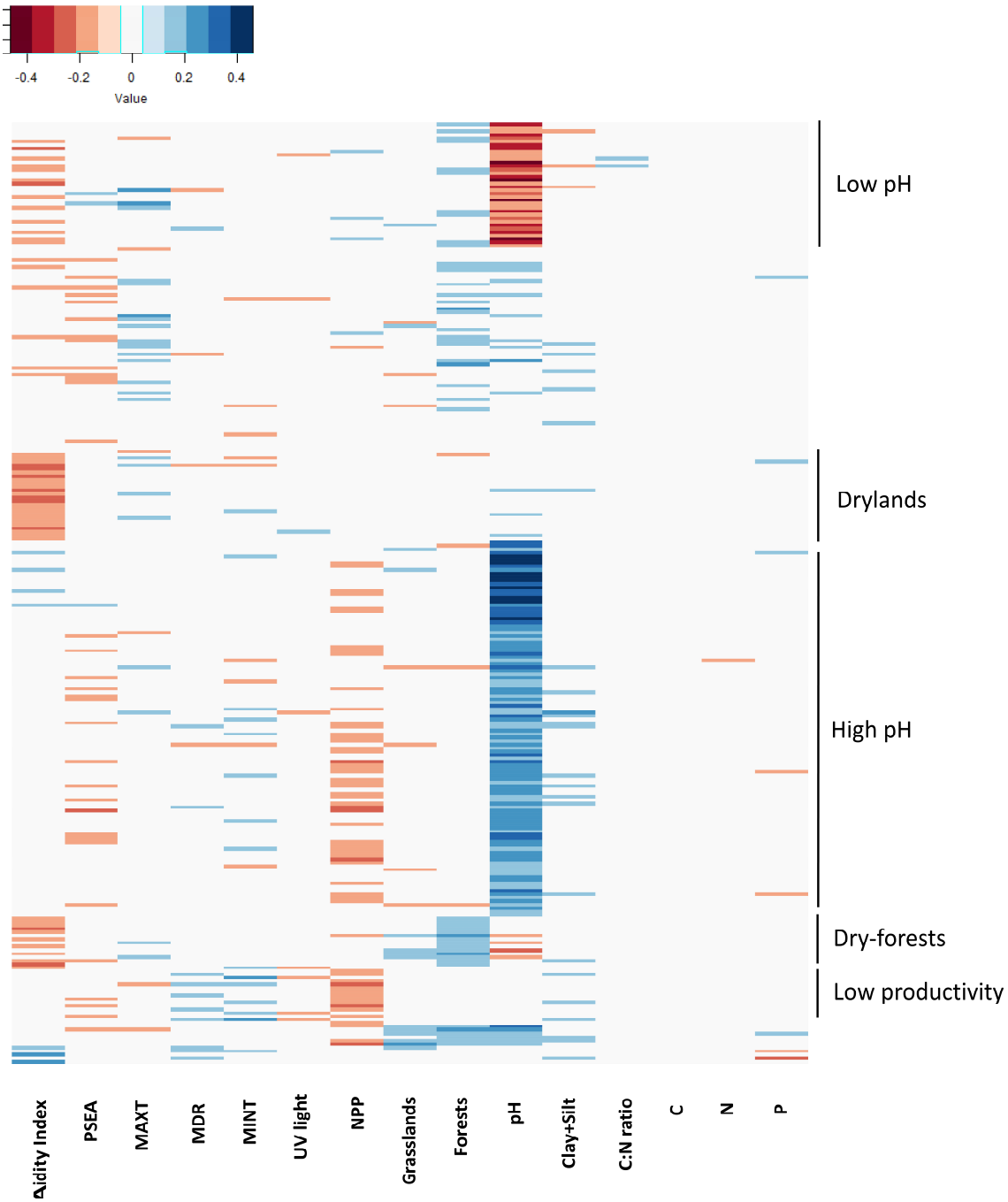


Figure S11. Heatmap including coefficients of correlation from semi-partial correlations between the relative abundance of each bacterial taxon (out of 270 phylotypes) with multiple environmental predictors. Data were sorted using the ecological cluster information provided in Table S1. MINT = minimum temperature; MAXT = maximum temperature; MDR = Mean diurnal temperature range. NPP = net primary productivity. PSEA = Precipitation seasonality.

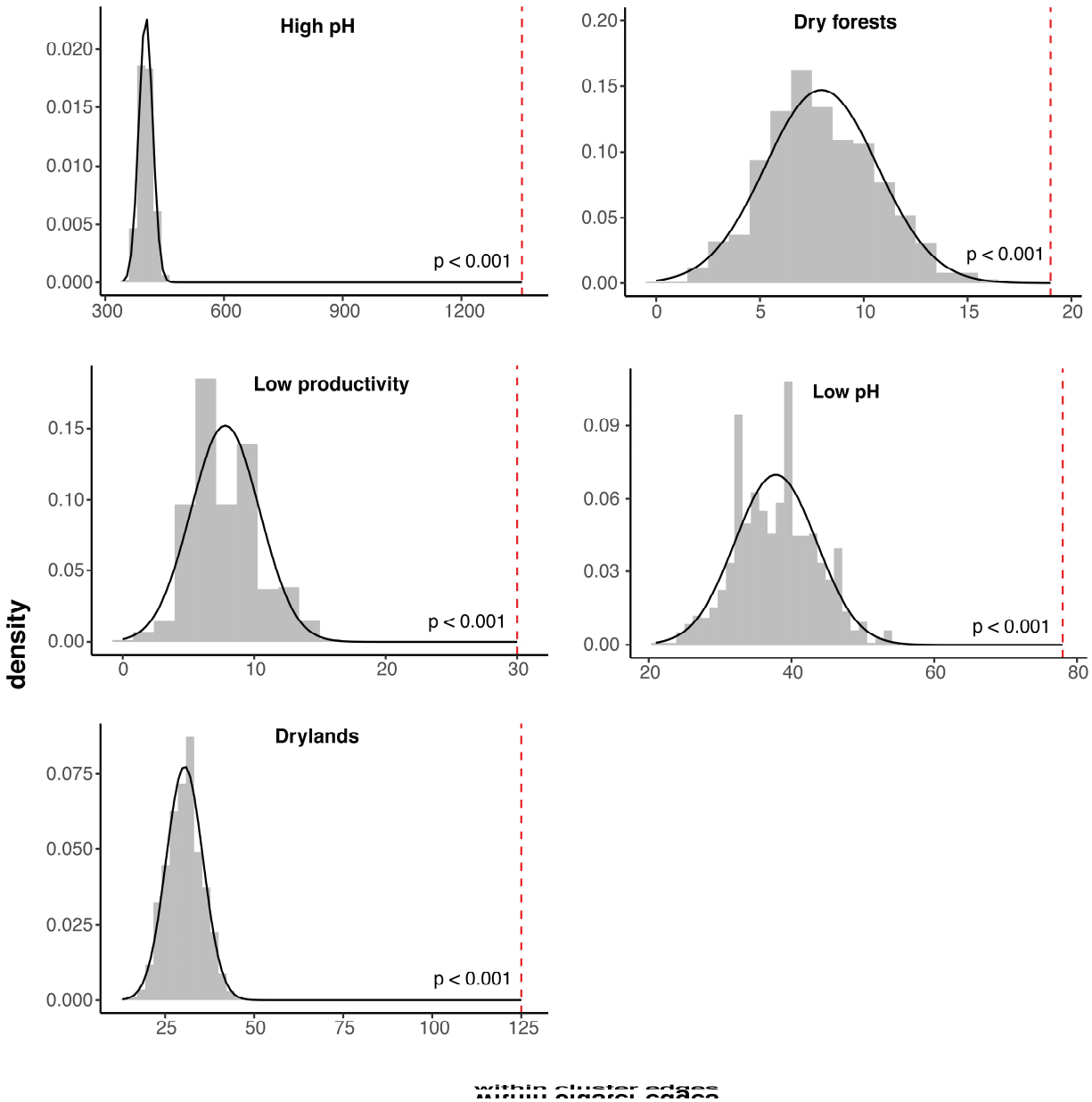


Figure S12. **Distributions of cluster environmental assembly across observed data and null models** (dashed line). For each major cluster, including high pH, low pH cluster, low productivity, drylands and dry-forest a histogram displays the expected distribution of expected number of edges between taxa that share an environmental cluster based on 1,000 random graphs under the Erdős–Rényi model if there were no structuring of co-occurrence patterns by environmental cluster (e.g. the null model). The dashed line indicates the observed number of edges between taxa that share an environmental cluster. For all environmental clusters, the *P*-value of expected versus observed is less than 0.001.

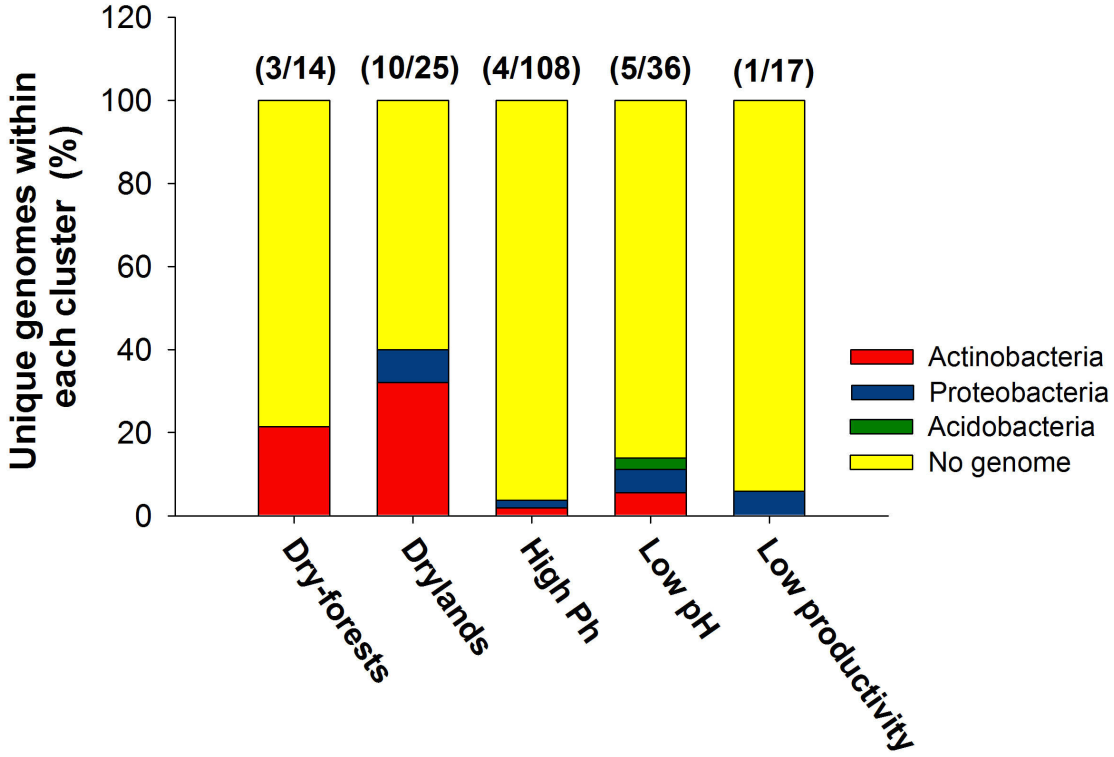
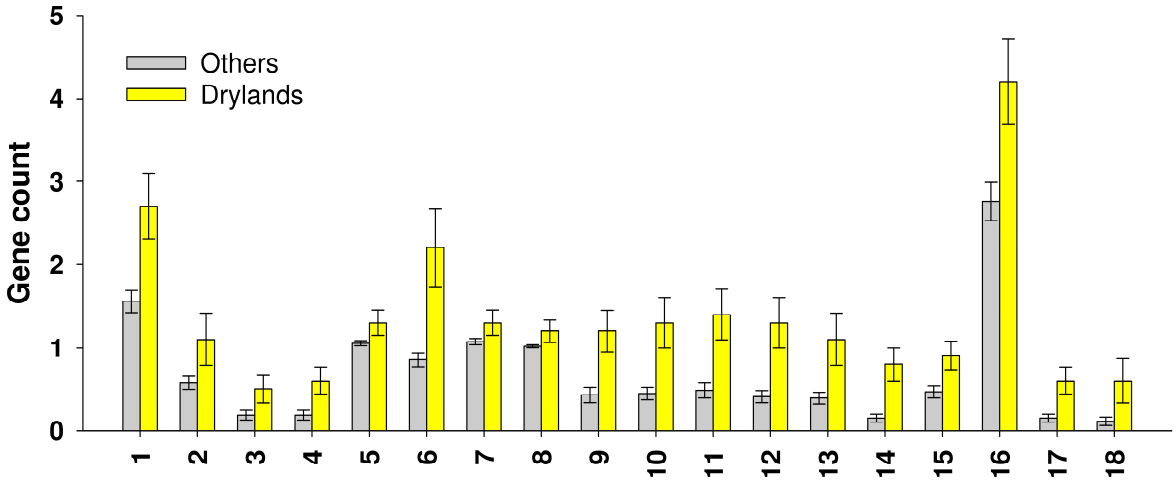


Figure S13. Percentage of genomes within each cluster for five well-defined ecological clusters of bacterial phylotypes with shared habitat preferences. The total number of phyloypes for which representative genomic data are available per cluster are indicated in brackets.



Gene ID	KEGG ontology	KEGG ID	KEGG gene	RF Importance	P-value
1	Lipid metabolism	KO:K00995	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase [EC:2.7.8.5] (pgsA, PGS1)	4.944	0.001
2	Hydrolases	KO:K01151	deoxyribonuclease IV [EC:3.1.21.2] (nfo)	3.680	0.003
3	Hydrolases	KO:K01182	oligo-1,6-glucosidase [EC:3.2.1.10] (E3.2.1.10)	1.303	0.008
4	Lyases	KO:K01616	2-oxoglutarate decarboxylase [EC:4.1.1.71] (kgd)	1.367	0.003
5	Genetic Information Processing	KO:K01883	cysteinyl-tRNA synthetase [EC:6.1.1.16] (CARS, cysS)	1.682	0.008
6	Bacterial secretion system	KO:K02654	leader peptidase (prepilin peptidase) / N-methyltransferase [EC:3.4.23.43 2.1.1.-] (pilD, pppA)	8.480	0.002
7	Heat shock proteins	KO:K03695	ATP-dependent Clp protease ATP-binding subunit ClpB (clpB)	1.902	0.006
8	DNA replication and repair	KO:K03702	excinuclease ABC subunit B (uvrB)	1.791	0.004
9	Transporters	KO:K05565	multicomponent Na+:H+ antiporter subunit A (mnhA, mrpA)	1.937	0.004
10	Transporters	KO:K05567	multicomponent Na+:H+ antiporter subunit C (mnhC, mrpC)	4.923	0.005
11	Transporters	KO:K05568	multicomponent Na+:H+ antiporter subunit D (mnhD, mrpD)	4.693	0.002
12	Transporters	KO:K05570	multicomponent Na+:H+ antiporter subunit F (mnhF, mrpF)	4.859	0.005
13	Transporters	KO:K05571	multicomponent Na+:H+ antiporter subunit G (mnhG, mrpG)	5.046	0.004
14	Others	KO:K06876	deoxyribodipyrimidine photolyase-related protein (K06876)	4.079	0.001
15	Uncharacterized protein	KO:K06976	uncharacterized protein (K06976)	1.415	0.003
16	Membrane protein	KO:K07058	membrane protein (K07058)	1.865	0.005
17	Uncharacterized protein	KO:K16645	heparin binding hemagglutinin HbhA (hbhA)	1.398	0.004
18	Propanoate metabolism	KO:K18382	NAD+-dependent secondary alcohol dehydrogenase Adh1 [EC:1.1.1.-] (adh1)	1.414	0.007

Figure S14. Gene count (mean ± 1 SE) for selected genes from Random Forest (RF) analyses of those genomes matching phylotypes in the drylands cluster versus those genomes representing phylotypes assigned to other clusters. RF Importance = Increase in % mean square error. Only predictors from RF with a $P < 0.01$ are selected for this analyses.

Table S1. List of identified dominant bacterial phylotypes from soils across the globe. This list contains information on the taxonomic identity of each phylotype, the ecological cluster it was assigned to, and the most closely related reference genome, cultivated strain and isolate.

Table S1 is available online as a Separate .XLS file under the Supporting Materials for this article.

References and Notes

1. J. M. Tiedje, S. Asuming-Brempong, K. Nüsslein, T. L. Marsh, S. J. Flynn, Opening the black box of soil microbial diversity. *Appl. Soil Ecol.* **13**, 109–122 (1999). doi:10.1016/S0929-1393(99)00026-8

2. R. D. Bardgett, W. H. van der Putten, Belowground biodiversity and ecosystem functioning. *Nature* **515**, 505–511 (2014). doi:10.1038/nature13855 Medline

3. P. L. E. Bodelier, Toward understanding, managing, and protecting microbial ecosystems. *Front. Microbiol.* **2**, 80 (2011). doi:10.3389/fmicb.2011.00080 Medline

4. K. S. Ramirez, J. W. Leff, A. Barberán, S. T. Bates, J. Betley, T. W. Crowther, E. F. Kelly, E. E. Oldfield, E. A. Shaw, C. Steenbock, M. A. Bradford, D. H. Wall, N. Fierer, Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proc. R. Soc. B* **281**, 20141988 (2014). doi:10.1098/rspb.2014.1988 Medline

5. M. Land, L. Hauser, S.-R. Jun, I. Nookaew, M. R. Leuze, T.-H. Ahn, T. Karpinets, O. Lund, G. Kora, T. Wassenaar, S. Poudel, D. W. Ussery, Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15**, 141–161 (2015). doi:10.1007/s10142-015-0433-4 Medline

6. P. D. Schloss, R. A. Girard, T. Martin, J. Edwards, J. C. Thrash, Status of the archaeal and bacterial census: An update. *MBio* **7**, e00201–e00216 (2016). doi:10.1128/mBio.00201-16 Medline

7. C. Lok, Mining the microbial dark matter. *Nature* **522**, 270–273 (2015). doi:10.1038/522270a Medline

8. N. Fierer, J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, J. A. Gilbert, D. H. Wall, J. G. Caporaso, Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 21390–21395 (2012). doi:10.1073/pnas.1215210110 Medline

9. F. T. Maestre, M. Delgado-Baquerizo, T. C. Jeffries, D. J. Eldridge, V. Ochoa, B. Gozalo, J. L. Quero, M. García-Gómez, A. Gallardo, W. Ulrich, M. A. Bowker, T. Arredondo, C. Barraza-Zepeda, D. Bran, A. Florentino, J. Gaitán, J. R. Gutiérrez, E. Huber-Sannwald, M. Jankju, R. L. Mau, M. Miriti, K. Naseri, A. Ospina, I. Stavi, D. Wang, N. N. Woods, X. Yuan, E. Zaady, B. K. Singh, Increasing aridity reduces soil microbial diversity and abundance in global drylands. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15684–15689 (2015). Medline

10. J. P. Grime, Benefits of plant diversity to ecosystems: Immediate, filter and founder effects. *J. Ecol.* **86**, 902–910 (1998). doi:10.1046/j.1365-2745.1998.00306.x

11. D. VanInsberghe, K. R. Maas, E. Cardenas, C. R. Strachan, S. J. Hallam, W. W. Mohn, Non-symbiotic Bradyrhizobium ecotypes dominate North American forest soils. *ISME J.* **9**, 2435–2441 (2015). doi:10.1038/ismej.2015.54 Medline

12. N. Fierer, J. Ladau, J. C. Clemente, J. W. Leff, S. M. Owens, K. S. Pollard, R. Knight, J. A. Gilbert, R. L. McCulley, Reconstructing the microbial diversity and function of pre-

agricultural tallgrass prairie soils in the United States. *Science* **342**, 621–624 (2013). doi:10.1126/science.1243768 Medline

13. T. E. Brewer, K. M. Handley, P. Carini, J. A. Gilbert, N. Fierer, Genome reduction in an abundant and ubiquitous soil bacterium '*Candidatus* Udaeobacter copiosus'. *Nat. Microbiol.* **2**, 16198 (2016). doi:10.1038/nmicrobiol.2016.198 Medline

14. P. H. Janssen, Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl. Environ. Microbiol.* **72**, 1719–1728 (2006). doi:10.1128/AEM.72.3.1719-1728.2006 Medline

15. C. L. Lauber, M. Hamady, R. Knight, N. Fierer, Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009). doi:10.1128/AEM.00335-09 Medline

16. J. Zhou, Y. Deng, L. Shen, C. Wen, Q. Yan, D. Ning, Y. Qin, K. Xue, L. Wu, Z. He, J. W. Voordeckers, J. D. V. Nostrand, V. Buzzard, S. T. Michaletz, B. J. Enquist, M. D. Weiser, M. Kaspari, R. Waide, Y. Yang, J. H. Brown, Temperature mediates continental-scale diversity of microbes in forest soils. *Nat. Commun.* **7**, 12083 (2016). doi:10.1038/ncomms12083 Medline

17. S. M. Prober, J. W. Leff, S. T. Bates, E. T. Borer, J. Firn, W. S. Harpole, E. M. Lind, E. W. Seabloom, P. B. Adler, J. D. Bakker, E. E. Cleland, N. M. DeCrappeo, E. DeLorenze, N. Hagenah, Y. Hautier, K. S. Hofmockel, K. P. Kirkman, J. M. H. Knops, K. J. La Pierre, A. S. MacDougall, R. L. McCulley, C. E. Mitchell, A. C. Risch, M. Schuetz, C. J. Stevens, R. J. Williams, N. Fierer, Plant diversity predicts beta but not alpha diversity of soil microbes across grasslands worldwide. *Ecol. Lett.* **18**, 85–95 (2015). doi:10.1111/ele.12381 Medline

18. J. W. Leff, S. E. Jones, S. M. Prober, A. Barberán, E. T. Borer, J. L. Firn, W. S. Harpole, S. E. Hobbie, K. S. Hofmockel, J. M. H. Knops, R. L. McCulley, K. La Pierre, A. C. Risch, E. W. Seabloom, M. Schütz, C. Steenbock, C. J. Stevens, N. Fierer, Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10967–10972 (2015). doi:10.1073/pnas.1508382112 Medline

19. K. T. Konstantinidis, A. Ramette, J. M. Tiedje, The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1929–1940 (2006). doi:10.1098/rstb.2006.1920 Medline

20. Materials and methods are available as supplementary materials.

21. S. Soliveres, P. Manning, D. Prati, M. M. Gossner, F. Alt, H. Arndt, V. Baumgartner, J. Binkenstein, K. Birkhofer, S. Blaser, N. Blüthgen, S. Boch, S. Böhm, C. Börschig, F. Buscot, T. Diekötter, J. Heinze, N. Hölzel, K. Jung, V. H. Klaus, A.-M. Klein, T. Kleinebecker, S. Klemmer, J. Krauss, M. Lange, E. K. Morris, J. Müller, Y. Oelmann, J. Overmann, E. Pašalić, S. C. Renner, M. C. Rillig, H. M. Schaefer, M. Schloter, B. Schmitt, I. Schöning, M. Schrumpf, J. Sikorski, S. A. Socher, E. F. Solly, I. Sonnemann, E. Sorkau, J. Steckel, I. Steffan-Dewenter, B. Stempfhuber, M. Tschapka, M. Türke, P. Venter, C. N. Weiner, W. W. Weisser, M. Werner, C. Westphal, W. Wilcke, V. Wolters, T. Wubet, S. Wurst, M. Fischer, E. Allan, Locally rare species influence grassland

ecosystem multifunctionality. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20150269 (2016). doi:10.1098/rstb.2015.0269 Medline

22. C. Thompson, Human embryos: Collect reliable data on embryo selection. *Nature* **551**, 33 (2017). doi:10.1038/551033a Medline

23. E. Stackebrandt, B. M. Goebel, Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Bacteriol.* **44**, 846–849 (1994). doi:10.1099/00207713-44-4-846

24. L. Breiman, *Mach. Learn.* **45**, 5–32 (2001). doi:10.1023/A:1010933404324

25. A. B. de Menezes, M. T. Prendergast-Miller, A. E. Richardson, P. Toscas, M. Farrell, L. M. Macdonald, G. Baker, T. Wark, P. H. Thrall, Network analysis reveals that bacteria and fungi form modules that correlate independently with soil parameters. *Environ. Microbiol.* **17**, 2677–2689 (2015). doi:10.1111/1462-2920.12559 Medline

26. T. H. Swartz, S. Ikewada, O. Ishikawa, M. Ito, T. A. Krulwich, The Mrp system: A giant among monovalent cation/proton antiporters? *Extremophiles* **9**, 345–354 (2005). doi:10.1007/s00792-005-0451-6 Medline

27. W. G. Whitford, *Ecology of Desert Systems* (Academic Press, San Diego, CA, 2002).

28. J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, CA, 1993).

29. A. Barberán, H. Caceres Velazquez, S. Jones, N. Fierer, Hiding in plain sight: Mining bacterial species records for phenotypic trait information. *MSphere* **2**, e00237–e17 (2017). doi:10.1128/mSphere.00237-17 Medline

30. N. Fierer, J. L. Morse, S. T. Berthrong, E. S. Bernhardt, R. B. Jackson, Environmental controls on the landscape-scale biogeography of stream bacterial communities. *Ecology* **88**, 2162–2173 (2007). doi:10.1890/06-1746.1 Medline

31. J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010). doi:10.1038/nmeth.f.303 Medline

32. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010). doi:10.1093/bioinformatics/btq461 Medline

33. R. C. Edgar, UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10**, 996–998 (2013). doi:10.1038/nmeth.2604 Medline

34. J. R. Cole, B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, J. M. Tiedje, The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33**, D294–D296 (2005). doi:10.1093/nar/gki038 Medline

35. T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, G. L. Andersen, Greengenes, a chimera-checked 16S rRNA gene database

and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006). doi:10.1128/AEM.03006-05 Medline

36. T. A. Kettler, J. W. Doran, T. L. Gilbert, Simplified Method for Soil Particle-Size Determination to Accompany Soil-Quality Analyses. *Soil Sci. Soc. Am. J.* **65**, 849 (2001). doi:10.2136/sssaj2001.653849x

37. J. M. Anderson, J.S.I., Ingramm, *Tropical Soil Biology and Fertility: A Handbook of Methods* (CABI, Wallingford, UK, ed. 2, 1993).

38. R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005). doi:10.1002/joc.1276

39. R. J. Zomer, A. Trabucco, D. A. Bossio, L. V. Verchot, Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agric. Ecosyst. Environ.* **126**, 67–80 (2008). doi:10.1016/j.agee.2008.01.014

40. P. A. Newman, R. McKenzie, UV impacts avoided by the Montreal Protocol. *Photochem. Photobiol. Sci.* **10**, 1152–1160 (2011). doi:10.1039/c0pp00387e Medline

41. N. Pettorelli, J. O. Vik, A. Mysterud, J.-M. Gaillard, C. J. Tucker, N. C. Stenseth, Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends Ecol. Evol.* **20**, 503–510 (2005). doi:10.1016/j.tree.2005.05.011 Medline

42. M. Delgado-Baquerizo, F. T. Maestre, P. B. Reich, T. C. Jeffries, J. J. Gaitan, D. Encinar, M. Berdugo, C. D. Campbell, B. K. Singh, Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nat. Commun.* **7**, 10541 (2016). doi:10.1038/ncomms10541 Medline

43. S. Mukherjee, R. Seshadri, N. J. Varghese, E. A. Eloe-Fadrosh, J. P. Meier-Kolthoff, M. Göker, R. C. Coates, M. Hadjithomas, G. A. Pavlopoulos, D. Paez-Espino, Y. Yoshikuni, A. Visel, W. B. Whitman, G. M. Garrity, J. A. Eisen, P. Hugenholtz, A. Pati, N. N. Ivanova, T. Woyke, H.-P. Klenk, N. C. Kyrpides, 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017). doi:10.1038/nbt.3886 Medline

44. R. Ranjan, A. Rani, A. Metwally, H. S. McGee, D. L. Perkins, Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469**, 967–977 (2016). doi:10.1016/j.bbrc.2015.12.083 Medline

45. J. Bengtsson-Palme, M. Angelin, M. Huss, S. Kjellqvist, E. Kristiansson, H. Palmgren, D. G. J. Larsson, A. Johansson, The human gut microbiome as a transporter of antibiotic resistance genes between continents. *Antimicrob. Agents Chemother.* **59**, 6551–6560 (2015). doi:10.1128/AAC.00933-15 Medline

46. A. T. Moles, H. Flores-Moreno, S. P. Bonser, D. I. Warton, A. Helm, L. Warman, D. J. Eldridge, E. Jurado, F. A. Hemmings, P. B. Reich, J. Cavender-Bares, E. W. Seabloom, M. M. Mayfield, D. Sheil, J. C. Djietror, P. L. Peri, L. Enrico, M. R. Cabido, S. A. Setterfield, C. E. R. Lehmann, F. J. Thomson, Invasions: The trail behind, the path ahead, and a test of a disturbing idea. *J. Ecol.* **100**, 116–127 (2012). doi:10.1111/j.1365-2745.2011.01915.x

47. E. Archer, rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics. R package version 1.5.2 (2016).

48. S. Kim, ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun. Stat. Appl. Methods* **22**, 665–674 (2015). doi:10.5351/CSAM.2015.22.6.665 Medline

49. R. M. Warner, *Applied statistics: From bivariate through multivariate techniques* (Sage Publications, Inc, Thousand Oaks, California, 2012).

50. E. Pruesse, J. Peplies, F. O. Glöckner, SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012). doi:10.1093/bioinformatics/bts252 Medline

51. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009). doi:10.1093/bioinformatics/btp348 Medline

52. F. Asnicar, G. Weingart, T. L. Tickle, C. Huttenhower, N. Segata, Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015). doi:10.7717/peerj.1029 Medline

53. A. Barberán, S. T. Bates, E. O. Casamayor, N. Fierer, Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* **6**, 343–351 (2012). doi:10.1038/ismej.2011.119 Medline

54. M. Bastian *et al*., *Gephi: An Open Source Software for Exploring and Manipulating Networks*, AAAI Publications, Third International AAAI Conference on Weblogs and Social Media (2009).

55. N. Connor, A. Barberán, A. Clauset, Using null models to infer microbial co-occurrence networks. *PLOS ONE* **12**, e0176751 (2017). doi:10.1371/journal.pone.0176751 Medline

56. M. Kuhn *et al.,* Cubist: Rule- And Instance-Based Regression Modeling. R package version 0.0.19 (2016).

57. T. Hengl, J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, B. Kempen, SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* **12**, e0169748 (2017). doi:10.1371/journal.pone.0169748 Medline

58. M. Delgado-Baquerizo, A. Gallardo, F. Covelo, A. Prado-Comesaña, V. Ochoa, F. T. Maestre, Differences in thallus chemistry are related to species-specific effects of biocrust-forming lichens on soil nutrients and microbial communities. *Funct. Ecol.* **29**, 1087–1098 (2015). doi:10.1111/1365-2435.12403