Title: A global atlas of the dominant bacteria found in soil.

Authors: Manuel Delgado-Baquerizo[1,2*], Angela M. Oliverio[1,3], Tess E. Brewer[1,4], Alberto Benavent-González[5], David J. Eldridge[6], Richard D. Bardgett[7], Fernando T. Maestre[2], Brajesh K. Singh[8,9], Noah Fierer[1,3*].

Affiliations:

1. Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309.
2. Departamento de Biología y Geología, Física y Química Inorgánica, Escuela Superior de Ciencias Experimentales y Tecnología. Universidad Rey Juan Carlos, c/ Tulipán s/n, 28933 Móstoles, Spain.
3. Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309.
4. Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Colorado 80309.
5. Departamento de Biología Vegetal II, Fac. Farmacia, Universidad Complutense de Madrid, 28040 Madrid, Spain.
6. Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia.
7. School of Earth and Environmental Sciences, Michael Smith Building, The University of Manchester, Oxford Road, Manchester M13 9PT, UK.
8. Hawkesbury Institute for the Environment, Western Sydney University, Penrith, 2751, New South Wales, Australia.
9. Global Centre for Land-Based Innovation, Western Sydney University, Penrith South DC, NSW 2751, Australia.

*Authors for correspondence:

Manuel Delgado-Baquerizo. E-mail: M.DelgadoBaquerizo@gmail.com

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

Abstract

The immense diversity of soil bacterial communities has stymied efforts to characterize individual taxa and document their global distributions. We analyzed soils from 237 locations across six continents and found that only 2% of bacterial phylotypes (~500 phylotypes) consistently accounted for almost half of the soil bacterial communities worldwide. Despite the overwhelming diversity of bacterial communities, relatively few bacterial taxa are abundant in soils globally. We clustered these dominant taxa into ecological groups to build the first global atlas of soil bacterial taxa. Our study narrows down the immense number of bacterial taxa to a ‘most wanted’ list that will be fruitful targets for genomic and cultivation-based efforts aimed at improving our understanding of soil microbes and their contributions to ecosystem functioning.

One Sentence Summary: A few hundred bacterial taxa dominate soils globally, and we predict their ecological preferences and map their distributions.

Main text

Although soil bacteria have been studied for more than a century, most of the diversity of soil bacteria remains undescribed. This is unsurprising given that soil bacteria rank among the most abundant and diverse group of organisms on Earth (*1-4*), challenging our capacity to understand their specific contributions to ecosystem processes, including nutrient and carbon cycling, plant production, and greenhouse gas emissions (*1-3*). Put simply, characterizing the ecological attributes (environmental preferences and functional traits) of the thousands of bacterial taxa found in soil is unfeasible. Most soil bacteria do not match those found in pre-existing 16S rRNA gene databases (*5*), we have genomic information for relatively few of them (*5-7*), and the majority of soil bacteria have not been successfully cultivated *in vitro* (*6-7*). For these reasons, we lack a predictive understanding of the ecological attributes of most soil individual bacterial taxa, with their environmental preferences, traits, and metabolic capabilities remaining largely unknown.

Previous work has shown that only a small fraction of soil bacteria is typically shared between any pair of unique soil samples (*4,8-9*). However, we also know that, as with most 'macrobial' communities (*10*), not all bacterial taxa are equally abundant in soil. There are often sub-sets of soil bacterial taxa that are far more abundant than others. For example, the genus *Bradyrhizobium* has been found to be dominant in forest soils from North America (*11*). Similarly, a lineage within the class *Spartobacteria* was found to be highly abundant in undisturbed grassland soils (*12*). Perhaps more importantly, many individual taxa that are highly abundant in individual soil samples may also be abundant across distinct soil samples, even when those soil samples are from sites located far apart (e.g. *Candidatus Udaeobacter copiosus*) (*13*). Therefore, a critical and logical next step to advance our understanding of soil bacterial communities is to identify the dominant bacterial phylotypes that are abundant and ubiquitous across soils, and determine their ecological attributes.

From the large body of literature using marker gene sequencing to characterize soil bacterial communities, we know which major phyla tend to be more abundant in soil (*14*) and we have a growing understanding of how various factors, including soil properties (e.g. pH) (*15*), climate (*9,16*), vegetation type (*17*) and nutrient availability (*18*), structure the composition of soil bacterial communities worldwide. What is currently missing is a detailed ecological understanding of common soil bacterial species, which we refer to as phylotypes (as bacterial species definitions can be problematic) (*19*). Understanding the ecological attributes of dominant phylotypes will increase our ability to successfully cultivate them *in vitro,* and allow us to build a more predictive understanding of how soil bacterial communities vary across space, time, and in response to anthropogenic changes. For example, if we could identify those dominant phylotypes with strong preferences for a given set of environmental conditions (e.g. low or high pH), we could then use this information to predict their distributions and enrich for these dominant phylotypes *in vitro*. Ultimately, a better understanding of dominant soil bacterial taxa will improve our ability to actively manage soil bacterial communities to promote their functional capabilities.

Here we conducted a global analysis of the bacterial communities found in surface soils from 237 locations across six continents and eighteen countries (Fig. S1) to: (i) identify the most dominant (i.e. most abundant and ubiquitous) soil bacterial phylotypes worldwide; (ii) determine which of these dominant phylotypes tend to co-occur and share similar environmental preferences; (iii) map the abundances of these ecological clusters of dominant soil bacteria across the globe; and (iv) assess the genomic attributes that differentiate phylotypes with distinct environmental preferences. The soils included in this study were selected to span a wide range of vegetation types, edaphic characteristics, and bioclimatic regions (arid, temperate, tropical, continental and polar; *20*).

We first identified the most dominant bacterial phylotypes by 16S rRNA gene amplicon sequencing (*20*). Dominant phylotypes (taxa which share ≥97% sequence similarity across the amplified 16S rRNA gene region) include those that are highly abundant (top 10% most common phylotypes sorted by their % of 16S rRNA reads) (*21*) and ubiquitous (found in more than half of the 237 soil samples evaluated; *20*). Not surprisingly, our global dataset comprised bacterial communities that were highly variable with respect to their diversity and overall composition (Fig. S2). For example, observed phylotype richness ranged from 774 to 2869 phylotypes per sample and there was a large amount of variability in the relative abundances of major phyla across the studied sites (Fig. S2). Also, as expected, only a small fraction of phylotypes was found to be shared across soil samples and most phylotypes were relatively rare (Fig. S3). Based on our criteria, only 2% of the bacterial phylotypes (511 out of 25224 phylotypes) were dominant (Fig. 1A). However, this small number of phylotypes accounted for, on average, 41% of 16S rRNA gene sequences across all samples (Fig. 1A), although they collectively accounted for more than half of the bacterial communities in some environments (e.g. forests from arid environments; Fig. 1B). In other words, most soil bacterial phylotypes are rare and relatively few are abundant, but many of these are found across a wide range of soils.

Importantly, 85% of the dominant phylotypes identified from our dataset were also found to be dominant in the bacterial communities recovered from 123 global soils that were analyzed using a shotgun metagenomic approach (*20*). This cross-validation indicates that our list of dominant phylotypes is not biased by PCR amplification or by our choice of primers, as most of the identified dominant phylotypes were shared between two independent sets of soils analyzed using two different approaches (amplicon versus shotgun metagenomic sequencing). In addition, we compared the results from our sample set with those soils analyzed via amplicon sequencing as part of the Earth Microbiome Project (EMP, *22*). The majority of the dominant phylotypes in the EMP dataset (80%) –identified using the same criteria explained above– were included within our list of dominant taxa (>97% similarity) (*20*). Also, the top 511 phylotypes, comparable to our top 511 dominant taxa, accounted for 0.5% of all bacterial phylotypes and 41% of all 16S rRNA gene reads in the EMP dataset. Despite important methodological differences between the two datasets (*20*), this concordance between the results from EMP and our study reinforces our conclusion that a relatively small sub-set of bacterial phylotypes dominate soils across the globe.

On average, the dominant bacterial phylotypes identified from our dataset were highly abundant in soils across multiple continents, ecosystem types, and bioclimatic regions (Fig. 1B). The only exception was soil from tropical forests, where the dominant phylotypes accounted for only ~20% of 16S rRNA gene sequences, which is likely a product of soils from tropical forests being under-represented in our database and/or tropical forest bacterial communities being very distinct from those found in other ecosystem types (Fig. S4). Together, our results suggest that soil bacterial communities, like plant communities (*10*), are typically dominated by a relatively small subset of phylotypes. As such, we focus all downstream analyses on the 511 phylotypes found to be the most abundant and ubiquitous in soils from across the globe.

The identified dominant phylotypes accurately predicted overall patterns in β-diversity for the 'sub-dominant' component of the bacterial communities surveyed (98% of phylotypes; Figs. S2, S5 and Fig. 1C). In other words, patterns in the distribution of the dominant bacterial phylotypes across the globe closely mirrored those observed for the remaining 98% of bacterial phylotypes. The most abundant and ubiquitous of these 511 phylotypes included Alphaproteobacteria (*Bradyrhizobium sp.*, *Sphingomonas sp.*, *Rhodoplanes sp.*, *Devosia sp.* and *Kaistobacter sp.*), Betaproteobacteria (*Methylibium* sp. and *Ramlibacter* sp.), Actinobacteria (*Streptomyces* sp., *Salinibacterium* sp. and *Mycobacterium* sp.), Acidobacteria (*Candidatus Solibacter* sp. and order iii1-15), and Planctomycetes (order WD2101) (see Table S1 for a complete list). Remarkably, less than 18% of the 511 phylotypes we identified had a match to an available reference genome at the >97% 16S rRNA sequence similarity level, the level commonly used for delineating different bacterial species (*23*) (Fig. 2; Table S1). Approximately 42% of the dominant 511 phylotypes had no genome match even at the >90% 16S rRNA sequence similarity level, indicating that we do not have genomic information for taxa even within the same genus or family (Fig. 2A; Table S1). Further, only 45% of the identified 511 dominant phylotypes are related to cultivated isolates and <30% of the phylotypes have representative type strains at the >97% sequence similarity level (Fig. 2B, Table S1), which emphasizes the limited amount of phenotypic information we have available for these dominant phylotypes. Not surprisingly, phylotypes closely related to previously cultivated taxa tended to come from a few well-studied taxonomic groups, mostly *Proteobacteria* and *Actinobacteria*, with only a few representatives available from other phyla (Figs. 1C and 2B; Table S1), highlighting the well-known taxonomic biases of many pre-existing culture collections (*6*).

After identifying the dominant 511 phylotypes, we used Random Forest modeling (*24*) to identify habitat preferences for each phylotype (*20*). Our statistical models included 15 environmental factors: climate (aridity index, minimum and maximum temperature, precipitation seasonality and mean diurnal temperature range –MDR), UV radiation, net primary productivity, soil abiotic properties (soil texture, pH, total C, N and P concentrations, and C:N ratio), and dominant ecosystem type (forests and grasslands; *20*). We found that 53% (270) of the dominant 511 phylotypes had predictable habitat preferences (models explaining >30% of the variation; see ref. *20* and Table S1), with soil pH, climatic factors (aridity index, maximum temperature, and precipitation seasonality), and plant productivity consistently being the best predictors of their abundances across the globe (Fig. S6). These findings are in line with previous research demonstrating that climatic factors and soil pH are often highly correlated with observed differences in overall soil bacterial community composition (*4,8-9,15-16*), but, additionally, we found a strong link between microbial community composition and plant productivity (Fig. S7).

We were unable to identify a strong ecological preference for the remaining 241 of the 511 phylotypes, which included representatives from a wide range of phyla and sub-phyla (Fig. S8). Our inability to predict the distributions of these 241 phylotypes could be related to the absence of key, but hard to measure, environmental predictors (e.g. soil C availability) or the fact that our models did not take into account specific associations between the bacteria and plants, fungi, or animals (e.g. pathogen/host or predator/prey interactions), which may be driving their distribution patterns. Alternatively, we may not have been able to identify the habitat preferences of these phylotypes due to low variability in their abundances across the samples (Figs. S9 and S10). Indeed, the relative abundance of the group including all 241 undetermined phylotypes showed a much lower coefficient of variation than the relative abundance of those phylotypes for which we could identify their habitat preferences, as explained below (Fig. S9). This result suggests that the undetermined phylotypes, those with no clearly identifiable habitat preferences, represent a ‘core’ group of dominant phylotypes that are ubiquitous across global soils with proportional abundances that are relatively invariant.

We then used semi-partial correlations (Spearman) and clustering analyses (*20*) to identify groups of phylotypes with shared habitat preferences, restricting our analyses to those 270 phylotypes with predictable distribution patterns. We found that the phylotypes group into five reasonably well-defined ecological clusters sharing environmental preferences for: (i) high pH; (ii) low pH; (iii) drylands; (iv) low plant productivity; and (v) dry-forest environments (Figs. 2B, 3A and Fig. S11 and Table S1). These five clusters of phylotypes included 200 out of the 270 phylotypes for which we were able to identify their habitat preferences (Table S1). Each of the ecological clusters identified included phylotypes from multiple phyla, suggesting that habitat preferences are not linked to phylogeny at coarse levels of resolution (Fig. S8). The remaining 70 phylotypes were classified into three minor clusters including a small cluster consisting of six phylotypes (high pH-forest preference; Table S1 and Fig. 11) and two clusters that included phylotypes with preferences including warm-forests, sites with low seasonal variation in precipitation, mesic environments, and soils of low phosphorus content (Table S1 and Fig. 11). These results suggest that the dominant bacterial phylotypes can be clustered into predictable ecological groups that share similar habitat preferences. To cross-validate the ecological clusters, we used correlation network analyses (*20*,*25*) to investigate whether bacterial phylotypes sharing similar habitat and environmental preferences tend to co-occur (Fig. 3B).

Indeed, our network analyses indicated that bacterial phylotypes sharing a particular habitat preference (e.g. low pH) tend to co-occur with other phylotypes belonging to the same cluster more than we would expect by chance ($P < 0.001$ for all clusters; Fig. 3B; Fig. S12).

We next sought to determine if we could identify genomic attributes that delineate bacteria assigned to the individual ecological clusters. These analyses were restricted to the relatively small subset of bacterial phylotypes for which genomic data were available (>97% 16S rRNA sequence similarity to a reference genome). An insufficient number of representative unique genomes were available from phylotypes in four of the five major clusters identified (Fig. S13). However, we had genomic data for 10 unique genomes out of 25 phylotypes assigned to the 'drylands' cluster, including representatives of the *Proteobacteria* and *Actinobacteria* phyla (Fig. S13). We then identified functional genes that were over-represented in this 'drylands' cluster as compared to the genomes available for the other dominant taxa. A total of 72 genomes were included in this analysis, with 10 of these genomes belonging to the dryland cluster (*20*). We found that the genomes within this dryland cluster had significantly higher relative abundances of 18 genes (Fig. S14) compared to genomes representative of phylotypes assigned to other ecological clusters. Notably, Mnh/Mrp genes, which encode membrane transport proteins responsible for the proton-mediated efflux of monovalent cations (e.g. Na+, K+), were over-represented in the 'drylands' cluster (Fig. S14). These genes have frequently been linked to increased bacterial tolerance to alkaline or saline conditions, and, more generally, a greater capacity to tolerate external changes in the osmotic environment (*26*). These adaptations are likely to be important for bacteria living in arid soils, which are often saline, have high pH values, and experience prolonged periods of low moisture availability (*27*). Given the low number of reference genomes available, these findings are not conclusive and are simply a 'proof of concept'. Nevertheless, our results highlight that it is possible to identify genomic attributes that differentiate soil bacteria with distinct environmental preferences. They also emphasize the importance of acquiring new genomes to further understand the ecological attributes of dominant soil bacterial taxa. As such, our results pave the way for leveraging genomic data to predict the spatial distributions of soil bacterial taxa, efforts that will be improved as the collections of reference genomes from these microorganisms increase in size.

Together, our results suggest that there are predictable clusters of co-occurring dominant bacterial phylotypes in soils from across the globe. This finding indicates that commonly available environmental information could be used to build predictive maps of the global distributions of these bacterial clusters at a global scale. We did so for the four major ecological clusters (i.e., low pH, high pH, drylands and low productivity, Fig. 4; see Appendix S1 for details) using the prediction-oriented regression model Cubist (*28*) and information on 12 environmental variables for which we could acquire globally distributed information (*20*). Our models confirm that pH, aridity levels, and net primary productivity are major drivers of the low/high pH, dryland, and low productivity clusters observed, respectively (Appendix S1). Importantly, our maps (which accounted for 36-64% of the spatial variation in these clusters, Fig. 4) provide estimates of the regions where we would expect the groups of dominant soil bacterial phylotypes to be most abundant (Fig. 4). As expected, the dryland and low productivity clusters were relatively abundant in dryland and low productivity regions across the globe, and the low and high pH clusters were particularly abundant in areas known for their low or high pH soils, respectively.

This global inventory of dominant soil bacterial phylotypes represents a small subset of phylotypes that account for almost half of the 16S rRNA sequences recovered from soils. We show that we can predict the environmental preferences for over half of these dominant phylotypes, making it possible to predict how future environmental change will affect the spatial distribution of these taxa. Following Grime's mass ratio hypothesis (*10*), we would expect that identifying the physiological attributes of these dominant taxa will be critical for improving our understanding of the microbial controls on some key soil processes, including those that regulate soil C and nutrient cycling (*1-3,29*). Also, given the strong links between the distribution of bacterial phylotypes and their functional attributes across the globe (*8,12*), and the observed associations between dominant and sub-dominant phylotypes (Fig. S5), we expect that these dominant bacteria will be critical drivers, or indicators, of key soil processes worldwide. We also found that habitat preferences were not predictable from phylum-level identity alone, given that all of the ecological clusters included phylotypes from multiple phyla. This suggests that phylotypes from diverse taxa share some phenotypic traits (e.g. osmoregulatory capabilities) or life history strategies (*29-30*) that allow them to survive under particular environmental conditions. By narrowing down the number of phylotypes to be targeted in future studies from tens of thousands to a few hundred, our study paves the way for a more predictive understanding of soil bacterial communities, which is critical for accurately forecasting the ecological consequences of ongoing global environmental change.

References

1. J.M. Tiedje *et al., Appl. Soil Ecol.* **13,** 109-122 (1999)
2. R.D. Bardgett, W.H. van der Putten, *Nature* **515,** 505–511 (2014).
3. P.L.E. Bodelier, *Front. Microbiol.* **2,** 80 (2011).
4. K.S. Ramirez *et al., Proc R Soc Lond [Biol]* **281,** 20141988 (2014).
5. M Land *et al., Funct. Integr. Genomics.* **15,** 141–161 (2015).
6. P.D. Schloss *et al., mBio* **7,** e00201-16 (2016)
7. C. Lok, *Nature* **522**, 270-3 (2015)
8. N. Fierer *et al., Proc. Natl. Acad. Sci. USA.* **109,** 21390-21395 (2012).
9. F.T. Maestre *et al., Proc Natl Acad Sci U.S.A* **112,** 15684–15689 (2015).
10. J.P. Grime, *J Ecol* **86**, 902–910 (1998).
11. D. VanInsberghe *et al., The ISME J* **9,** 2435–2441 (2015)
12. Fierer N. *et al., Science* **342,** 621-624 (2013)
13. T.E. Brewer *et al., Nat. Microbiol.* **2,** 16198 (2016).
14. P.H. Janssen, *Appl Environ Microbiol* **72,** 1719-28 (2006).
15. C.L. Lauber *et al., Appl Environ Microbiol* **75**, 5111-5120 (2009).
16. J. Zhou et al., *Nat Commun* **7,** 12083 (2016).
17. S.M. Prober *et al., Ecol Lett* **18,** 85–95 (2015).
18. J.W. Leff *et al., Proc. Natl. Acad. Sci. U.S.A.* **112,** 10967-72 (2015).
19. K.T. Konstantinidis *et al., Philos Trans R Soc Lond B Biol Sci* **361,** 1929–1940 (2006).
20. Materials and methods are available as supplementary materials.
21. S. Soliveres *et al., Phil. Trans. R. Soc. B* **371,** 20150269 (2016).
22. L.R. Thompson *et al., Nature* 551, 457–463 (2017).
23. E. Stackebrandt, B.M. Goebel, *Int J Syst Bacteriol* **44,** 846–849 (1994).
24. L. Breiman, *Machine Learning* **45,** 5 (2001).
25. A.B. Menezes *et al., Environ. Microbiol.* **17**, 2677-2689 (2015).
26. T.H. Swartz *et al., Extremophiles.* **9,** 345-54 (2005).
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 Publishers, San Mateo, California, 1993).
29. A. Barberán *et al., mSphere* **2,** e00237-17 (2017).
30. N. Fierer *et al., Ecology* **88,** 2162-2173 (2007).
31. J.G. Caporaso *et al., Nat Method* **7,** 335 (2010).
32. R.C. Edgar, *Bioinformatics* **26**, 2460 (2010).
33. R.G. Edgar, *Nature Meth* **10,** 996-998 (2013).
34. J.R. Cole *et al., Nucleic Acids Res*. **33,** 294–296 (2005).
35. T.Z. DeSantis *et al., Appl Environ Micro*. **72,** 5069-72 (2006).
36. T.A. Kettler et al., *Soil Sci Soc Am J* **65,** 849 (2001).
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 *et al., Int J Climatol* **25,** 1965-1978 (2005).
39. R.J. Zomer *et al.,* Agric Ecosyst Envir **126,** 67-80 (2008).
40. P.A. Newman, R.L. McKenzie, *Photochem Photobiol Sci* **10,** 1152–1160 (2011).
41. N. Pettorelli *et al., Trend Ecol Evol* **20,** 503 (2005).
42. M. Delgado-Baquerizo *et al., Nat Commun* **28,** 10541 (2016).
43. S. Mukherjee *et al., Nat Biotechnol* **35,** 676-683 (2017).
44. R. Ranjan *et al., Biochem Biophys Res Commun*. **22**, 967-77 (2016).
45. J. Bengtsson-Palme *et al., Antimicrob Agents Chemother* **59,** 6551-60 (2015).
46. A.T. Moles *et al., J Ecol* **100,** 116-127 (2012).
47. E. Archer, rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics. R package version 1.5.2 (2016).
48. S. Kim, *Commun Statist Applic Meth* **22,** 665-674 (2015).
49. R.M. Warner, *Applied statistics: From bivariate through multivariate techniques* (Sage Publications, Inc, Thousand Oaks, California, 2012).
50. E. Pruesse, *et al., Bioinformatics* **28,** 1823–1829 (2012).
51. S. Capella-Gutiérrez *et al., Bioinformatics* **25,** 1972–1973 (2009).
52. F. Asnicar *et al., PeerJ* **3,** e1029 (2015).
53. A. Barberán, *et al., The ISME J* **6,** 343-351 (2012).

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 *et al., PLoS ONE* **12,** e0176751 (2017).

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

57. T. Hengl *et al., PLoS ONE* **12,** e0169748 (2017).

58. M. Delgado-Baquerizo *et al., Funct Ecol* **29,** 1087–1098 (2015).

Acknowledgements

We would like to thank Victoria Ochoa and Beatriz Gozalo for their help with soil analyses. We also would like to thank Dr. Senani Karunaratne for his advice on mapping analyses, and Tony Hartshorn and Ben Turner for sending us soil samples from Montana and Panama. M.D-B. acknowledge support from the Marie Sklodowska-Curie Actions of the Horizon 2020 Framework Programme H2020-MSCA-IF-2016 under REA grant agreement n° 702057. N.F was supported by grants from the U.S. National Science Foundation (EAR 1331828 and DEB 1554415). B.K.S. and M.D-B were supported by the Australian Research Council project (DP170104634). D.J.E. was supported by the Hermon Slade Foundation and R.D.B. was supported by the UK Department of Environment, Food and Rural Affairs (DEFRA) project number BD5003 and a BBSRC International Exchange Grant (BB/L026406/1). AB-G was supported by a FPI PhD fellowship (BES-2013-062945) and two research grants (CTM2015-64728-C2-1-R and CTM2012-38222-CO2-01) from the Spanish Ministerio de Economía y Competitividad. The work of FTM and the global drylands database were supported by the European Research Council (ERC Grant Agreements 242658 [BIOCOM] and 647038 [BIODESERT]) and by the Spanish Ministry of Economy and Competitiveness (BIOMOD project, ref. CGL2013-44661-R).

Author contributions

M.D-B. and N.F. designed this study. Field sampling was conducted by N.F., M.D-B., F.T.M., A.B-G., D.J.E., and R.D.B. Lab analyses were done by F.T.M. and B.K.S. Bioinformatic analyses were done by T.E.B. Statistical modeling and mapping were done by M.D-B. Network analyses were done by A.M.O. The manuscript was written by M.D-B and N.F. with contributions from all co-authors.

Data accessibility

All data used in this study are publicly available in Figshare (https://figshare.com/s/82a2d3f5d38ace925492; DOI: 10.6084/m9.figshare.5611321).

Supplementary Materials

Material and Methods

Table S1

Appendix S1

Figs. S1-S14

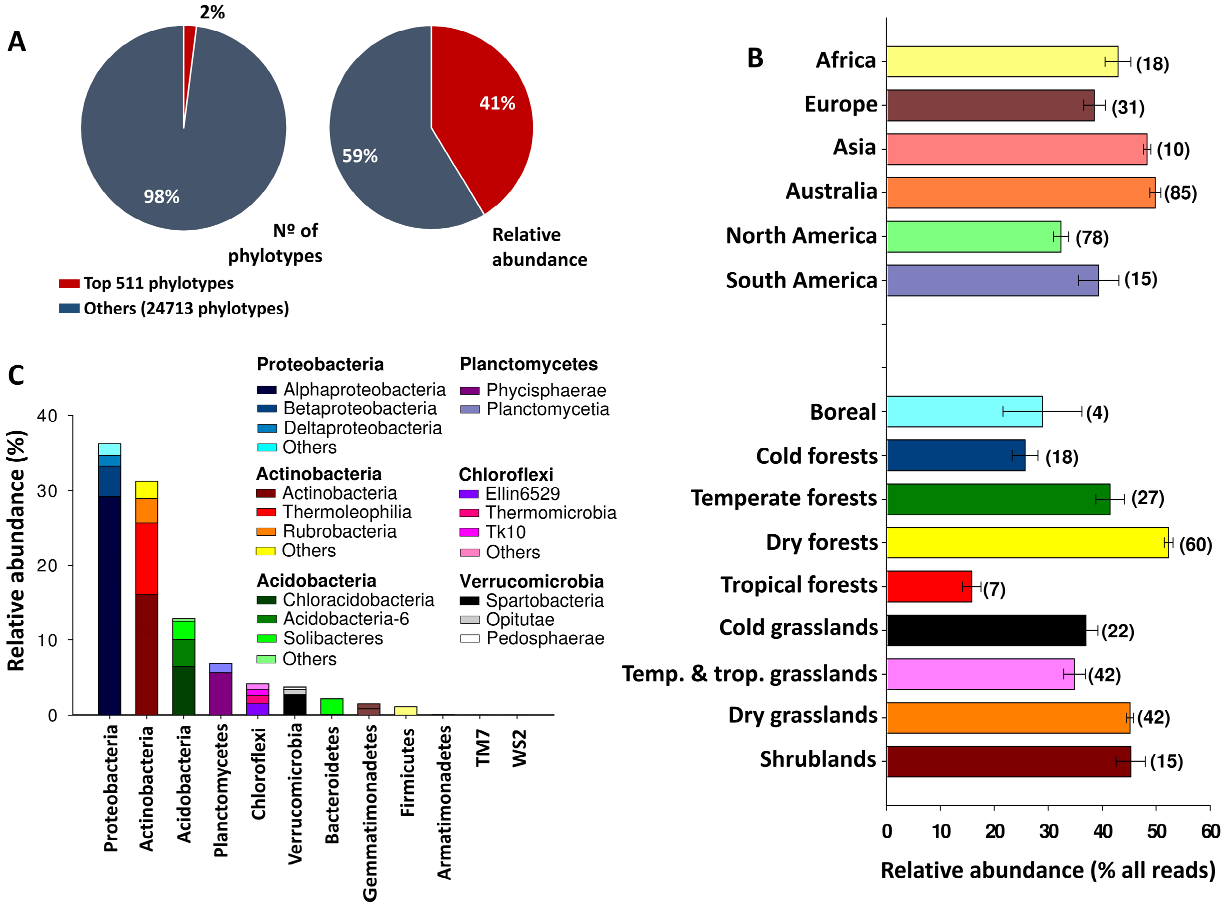


Figure 1. **Abundance and composition of dominant soil bacterial phylotypes across the globe**. Percentage of phylotypes and relative abundance of 16S rRNA genes representing the dominant versus the remaining bacterial phylotypes (A). Relative abundance (mean ± SE) of dominant phylotypes across continents and ecosystem types (B). Ecosystem type classification followed the Köppen climate classification and the major vegetation types found in our database. Grasslands include both tropical and temperate grasslands. Shrublands include polar, temperate and tropical shrublands. The number of samples in each category is indicated in brackets. The taxonomic composition of the dominant phylotypes is shown in (C). The phylotypes assigned to the least abundant phyla are not shown (including *Armatimonadetes* = 0.08%, TM7 = 0.05% and WS2 = 0.03%). Details on the top 511 dominant phylotypes are shown in Table S1.

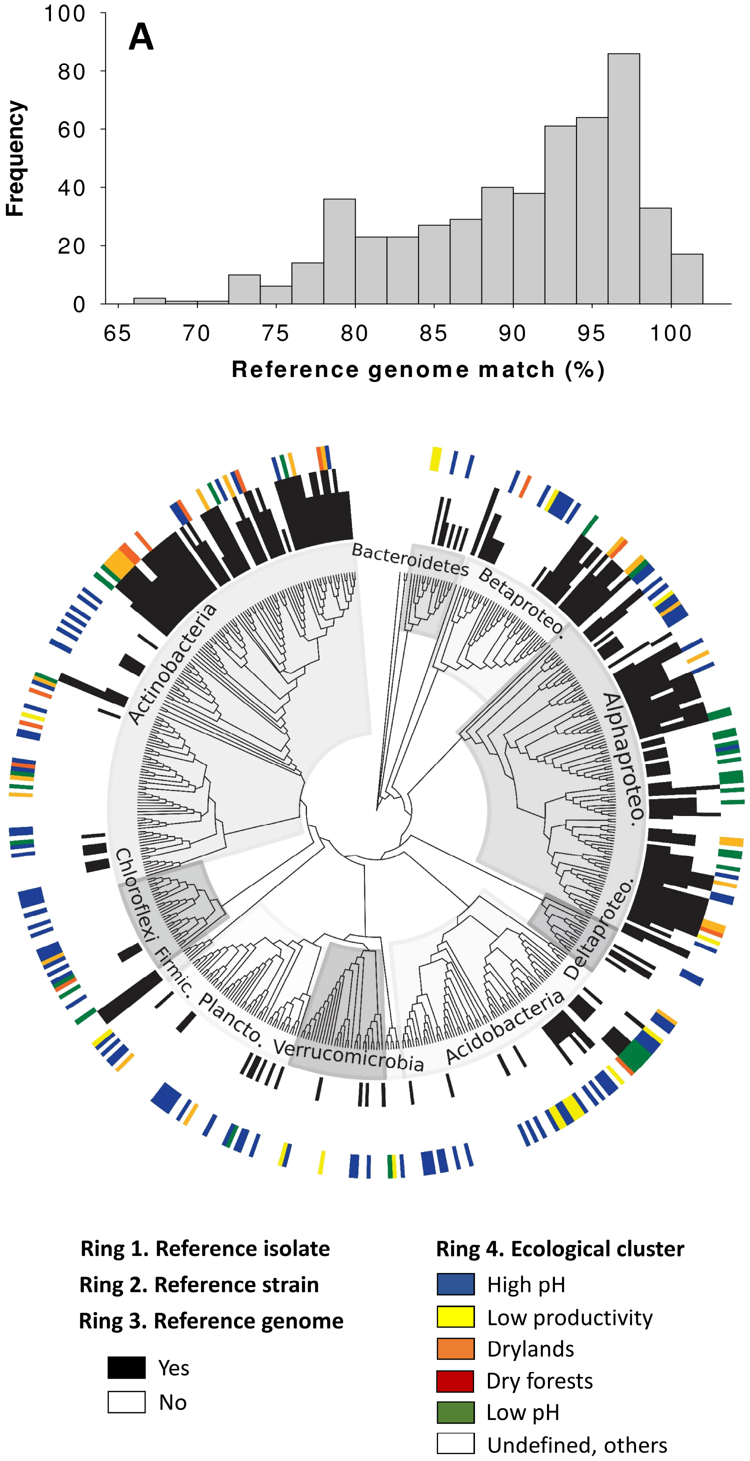


Figure 2. **Phylogenetic tree including the taxonomic information on dominant soil bacterial phylotypes**. Histogram showing the percentage 16S rRNA gene sequence similarity between the 511 dominant phylotypes and the most closely related available reference genome for each phylotype (A). Phylogenetic distribution of the 511 dominant phylotypes (B). Black shading on

the innermost and middle rings indicate, for each phylotype, whether there is a representative isolate and a genome match at the ≥97% 16S rRNA gene sequence similarity level. The coloring on the outermost ring highlights the distribution of environmental preferences for all phylotypes (n = 511). For the few phylotypes where taxonomic assignment did not correspond to tree topology, no manual corrections were made. Betaproteo. = Betaproteobacteria; Alphaproteo. = Alphaproteobacteria; Deltaproteo. = Deltaproteobacteria; Plancto. = Planctomycetes; Firmic. = Firmicutes.

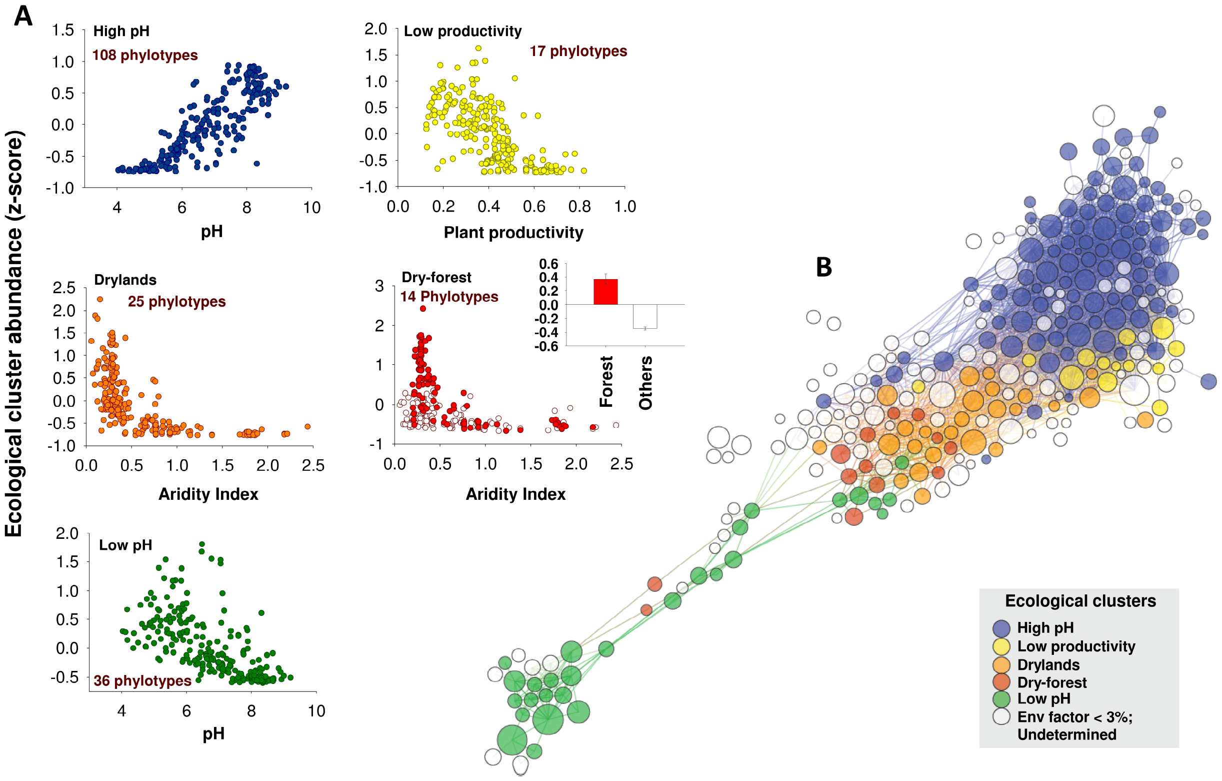


Figure 3. **Identified habitat preferences for dominant soil bacterial phylotypes**. Relationships between the relative abundance of the phylotypes assigned to each ecological cluster and their major environmental predictors (A, statistical analyses and identity of phylotypes within each cluster are presented in Table S1). Network diagram with nodes (bacterial phylotypes) colored by each of the five major ecological clusters that were identified, highlighting that the phylotypes within each ecological cluster tend to co-occur more than expected by chance (B, statistical analyses presented in Fig. S12).

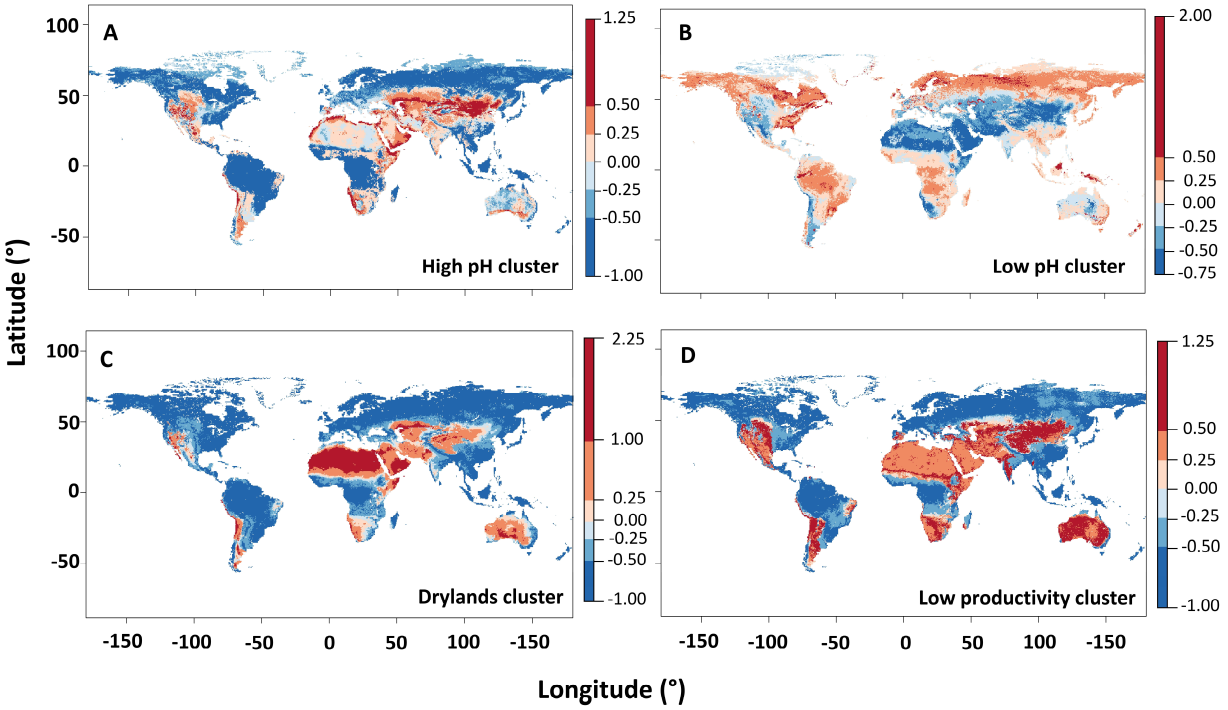


Figure 4. **A global atlas of the dominant bacteria found in soil.** Predicted global distribution of the relative abundances of the four major ecological clusters of bacterial phylotypes sharing habitat preferences for high pH, low pH, drylands and low plant productivity. R^2 (percentage of variation explained by the models) as follows: (1) High pH cluster, $R^2 = 0.53$, $P < 0.001$; (2) Low pH cluster, $R^2 = 0.36$, $P < 0.001$; (3) Drylands cluster, $R^2 = 0.64$, $P < 0.001$; and (4) Low productivity cluster, $R^2 = 0.40$, $P < 0.001$. The scale bar represents the standardized abundance (z-score) of each ecological cluster. An independent cross-validation for these maps is available in ref. *20*.

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.

Authors for correspondence:

Manuel Delgado-Baquerizo. E-mail: M.DelgadoBaquerizo@gmail.com

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

This PDF file includes:

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%).

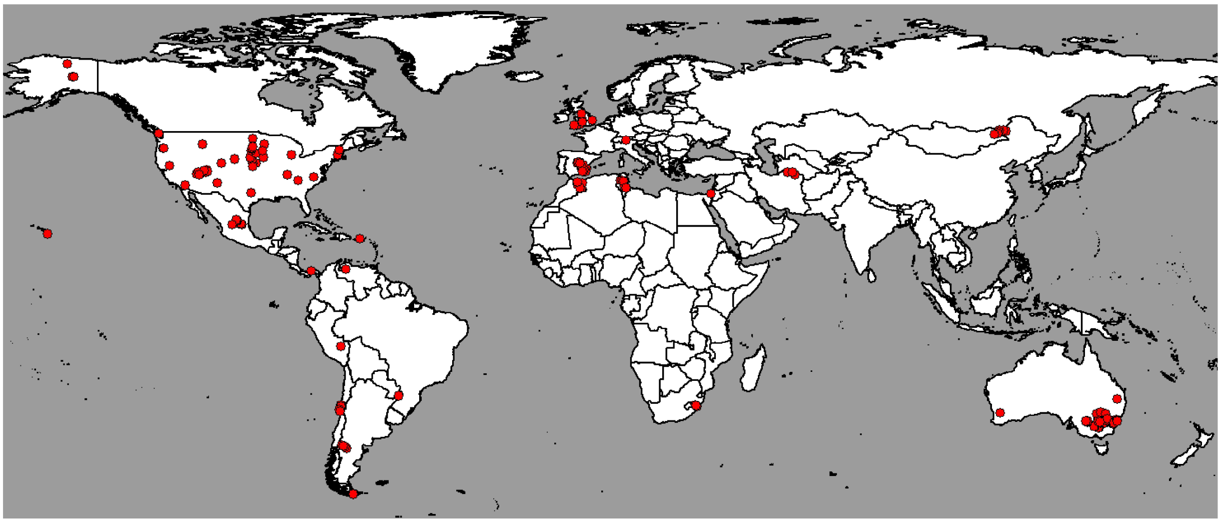


Figure S1. Locations of the 237 soil sampling sites included in this study.

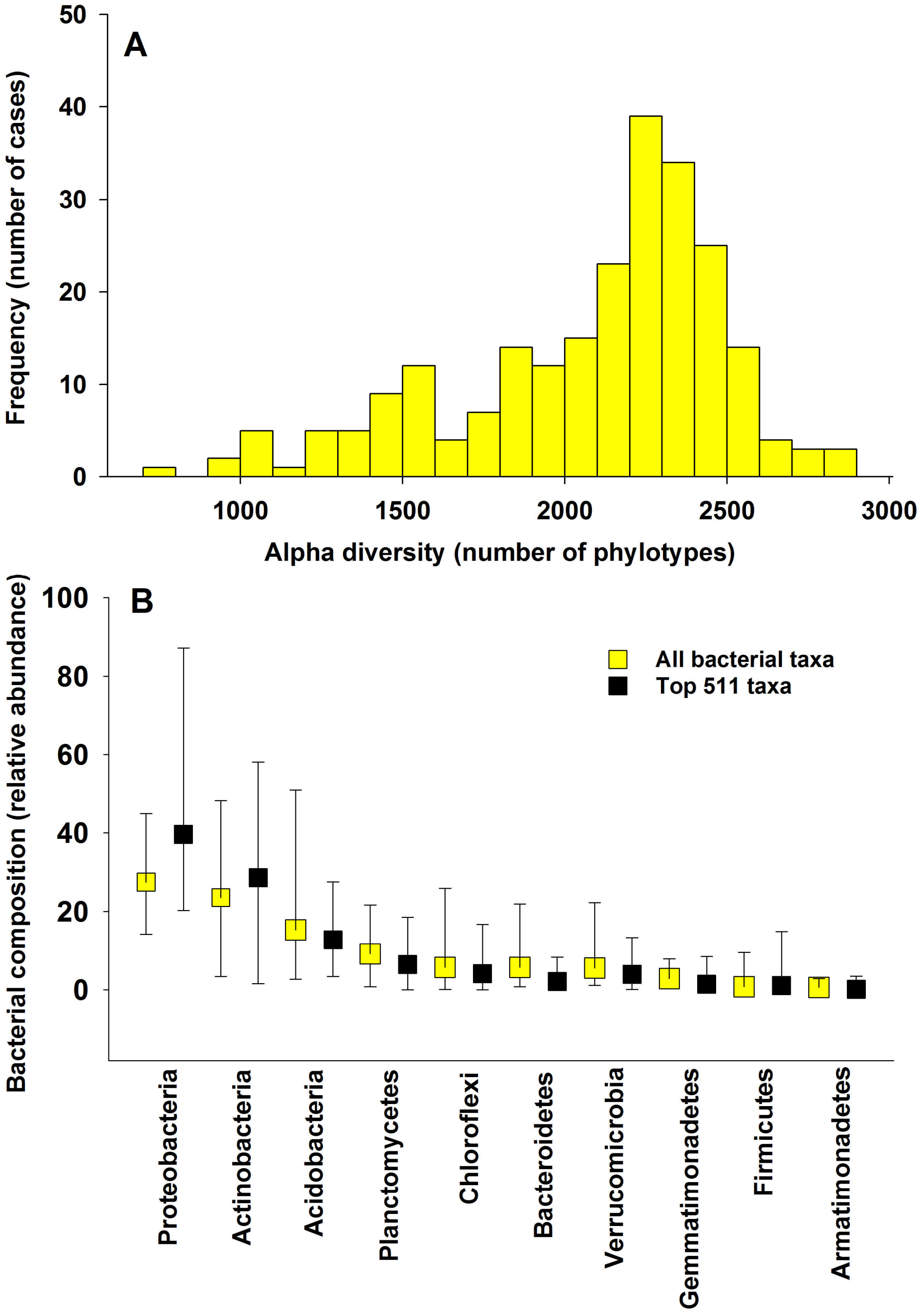


Figure S2. Richness and composition of the bacterial communities across the 237 soil samples included in this study. (A) Distribution of per-sample bacterial richness across the globe at a rarefied sequencing depth of 10,000 16S rRNA gene reads per sample. (B) Relative abundances (mean ± maximum/minimum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

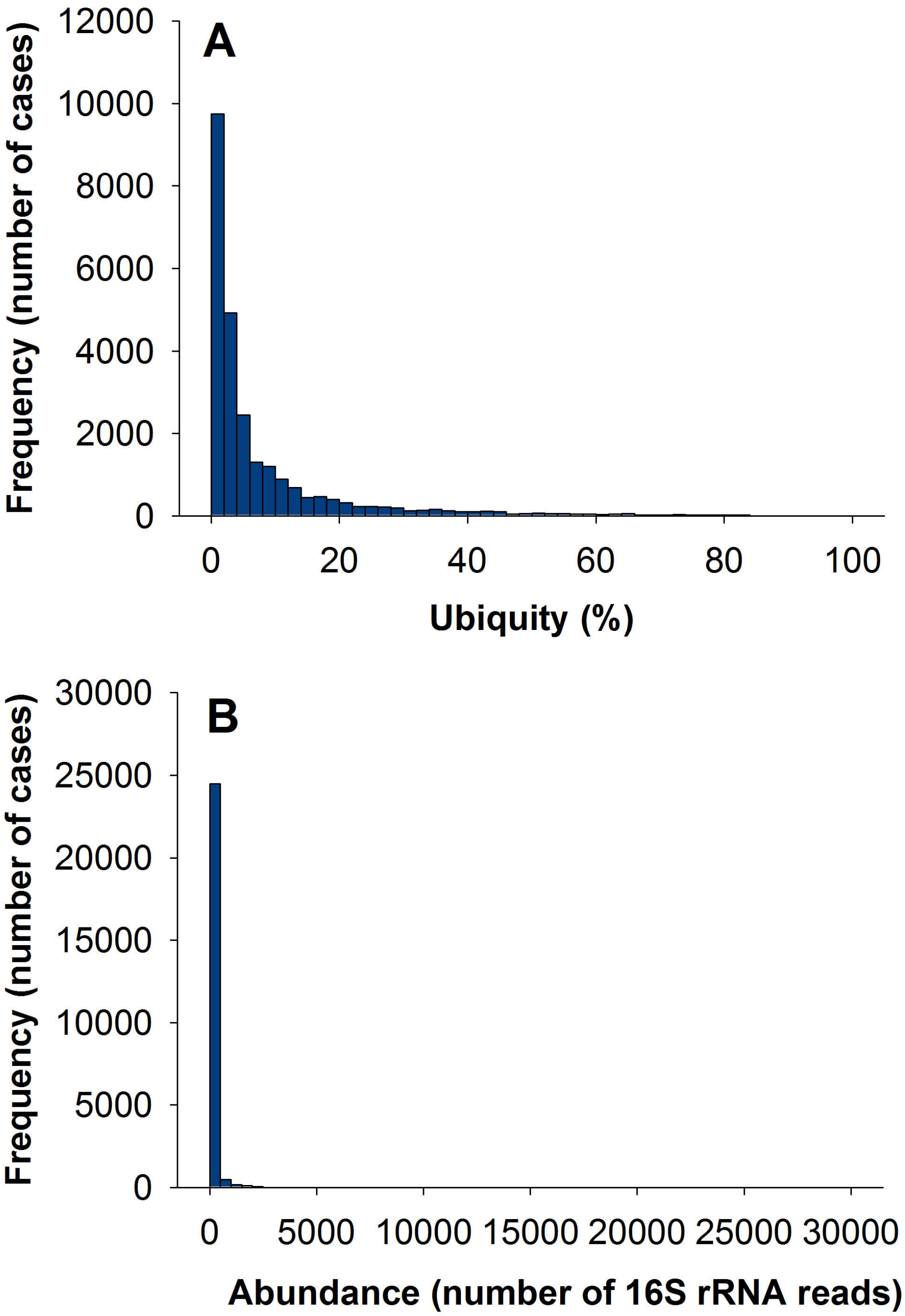


Figure S3. Histograms reporting the distributions for ubiquity (A) and the relative abundances of 25224 taxa (B) in the 237 soil samples from across the globe.

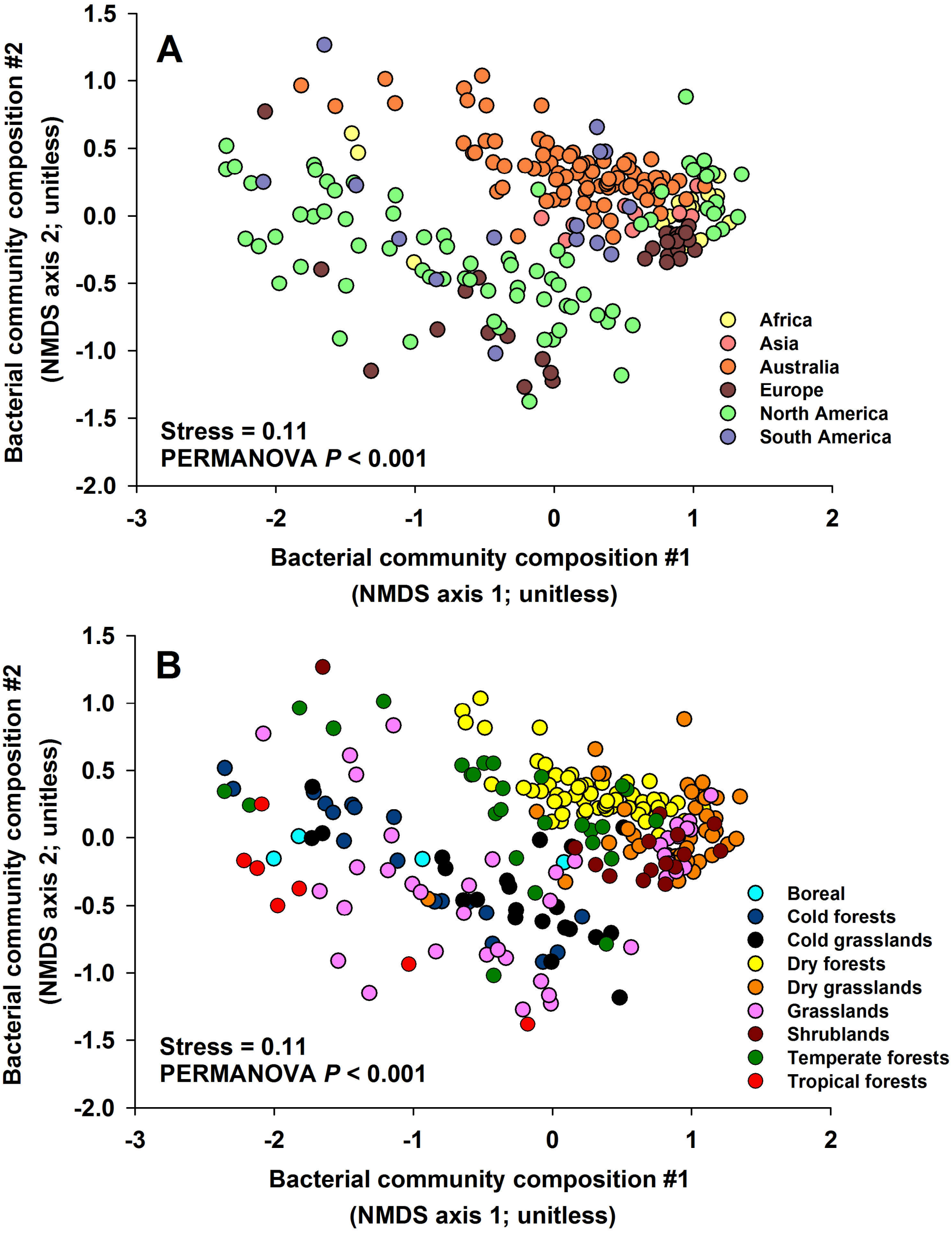


Figure S4. NMDS ordination summarizing the dissimilarity in community composition of bacteria across the globe for different continents (A) and ecosystem types (B). Grasslands include both tropical and temperate grasslands. Shrublands include polar, temperate and tropical shrublands.

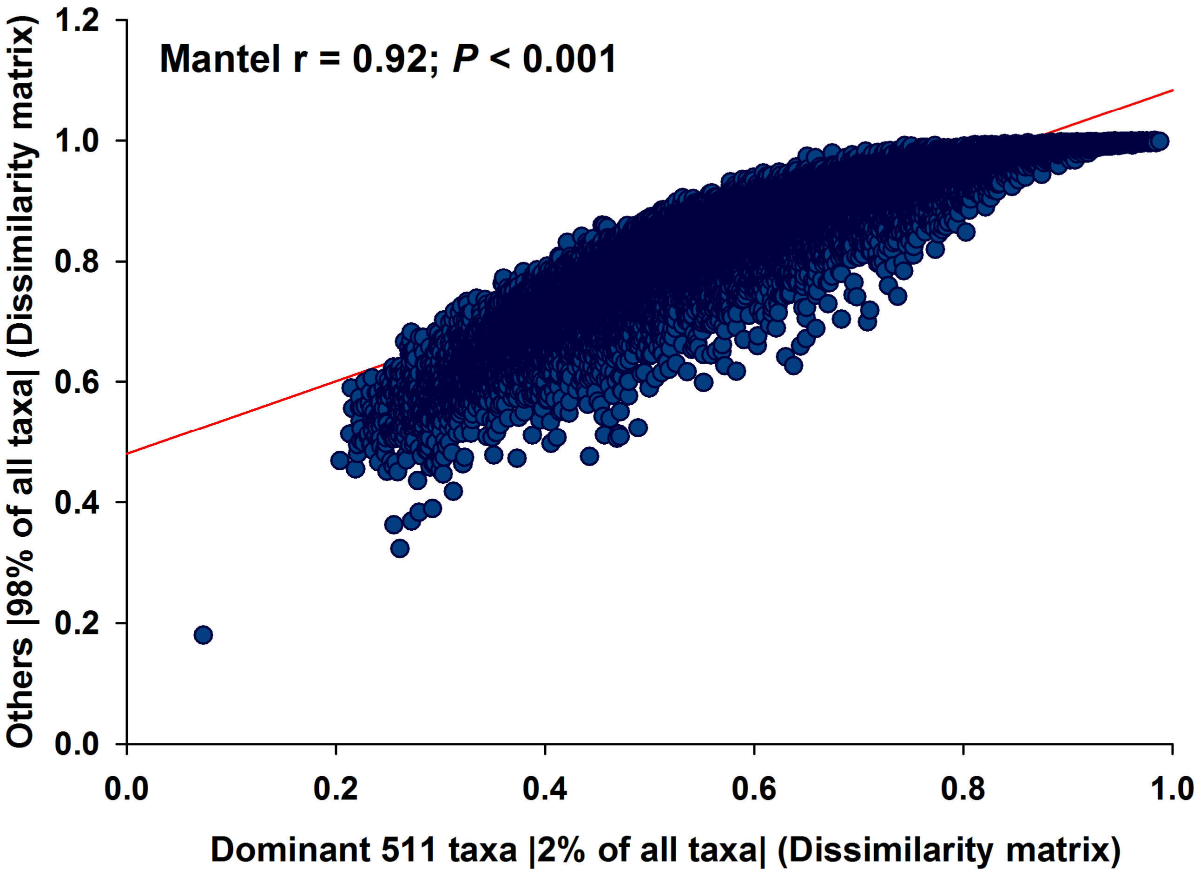


Figure S5. Relationship between beta diversity (community dissimilarity) based on Bray-Curtis distance for the dominant (511 phylotypes) and the remaining 24713 bacterial phylotypes. Correlation was done using the Mantel test.

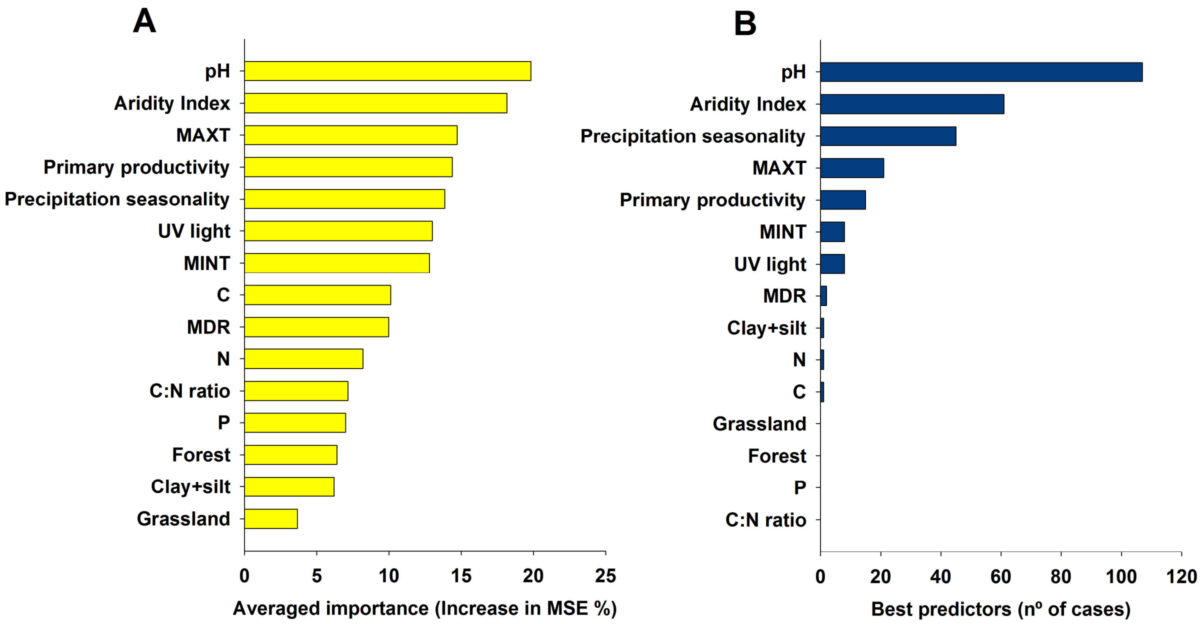


Figure S6. Major predictors of the distribution of dominant bacterial taxa across the globe. Averaged importance of environmental factors (across 270 Random Forest models) in predicting the relative abundance of dominant bacterial taxa (A). Number of cases (out of 270 Random Forest models) for which 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.

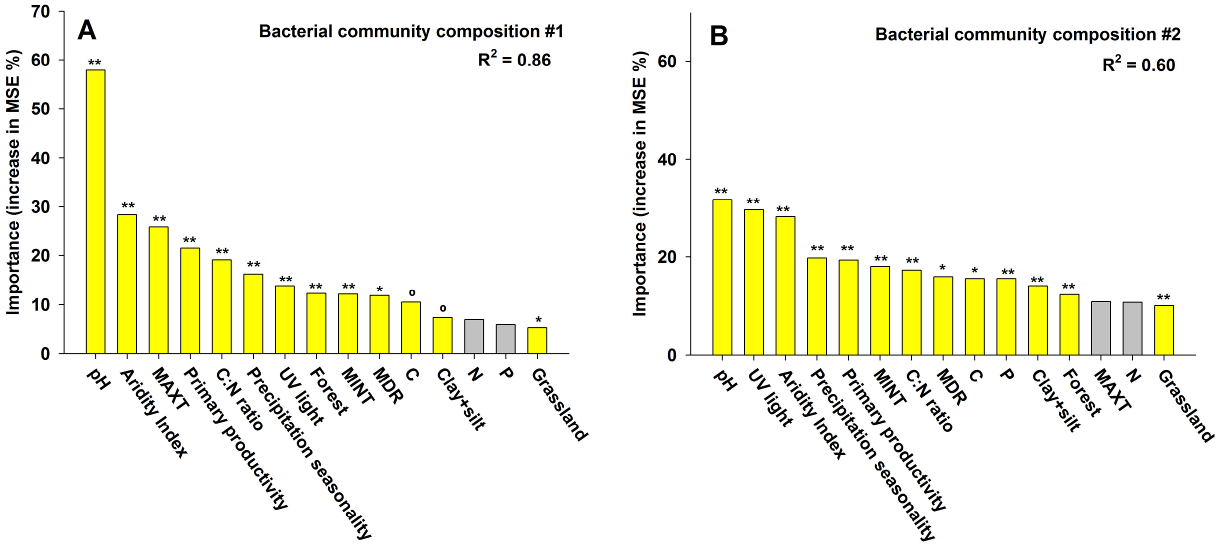


Figure S7. Major predictors of the distribution of bacterial communities across the globe. Panels (A) and (B) include the importance of environmental factors in predicting the relative abundance of bacterial community composition (two axes from a NMDS summarizing information on the overall community 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.

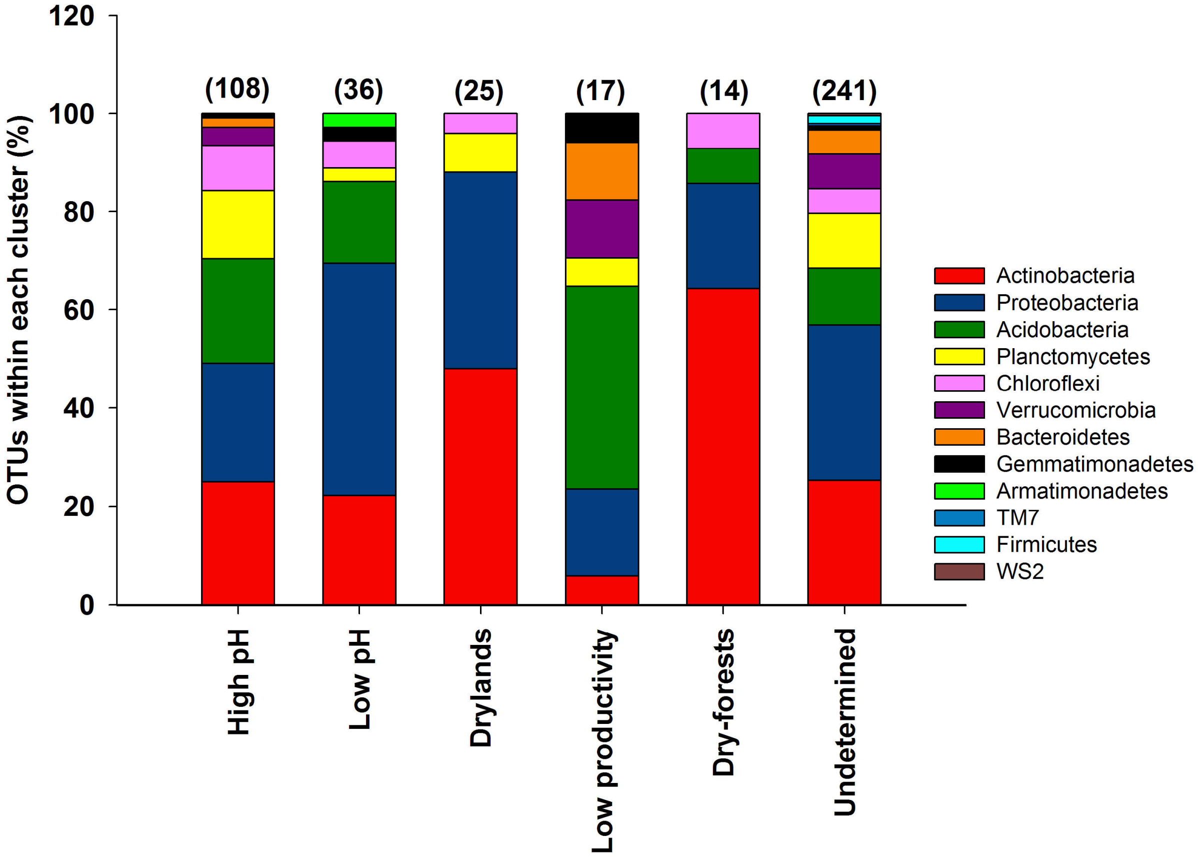


Figure S8. Taxonomic composition (% of phylotypes (OTUs) within each cluster) for five well-defined ecological clusters of bacterial phylotypes sharing habitat preferences and also for those phylotypes for which we were not able to identify their niche model (undetermined phylotypes). The total number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

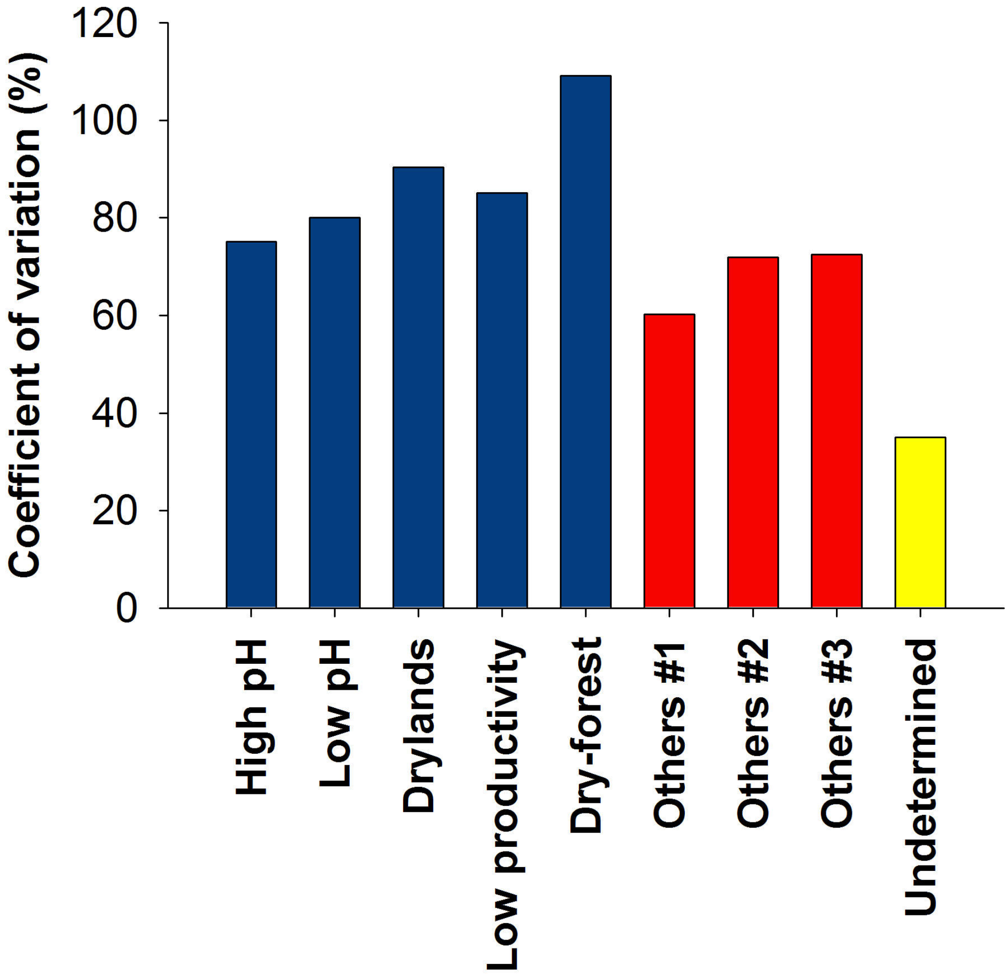


Figure S9. Coefficients of variation in the relative abundances of dominant bacterial phylotypes assigned to each of the five major ecological clusters and those phylotypes that fell within an 'undetermined' group (those dominant bacterial phylotypes with no identifiable habitat or environmental preferences).

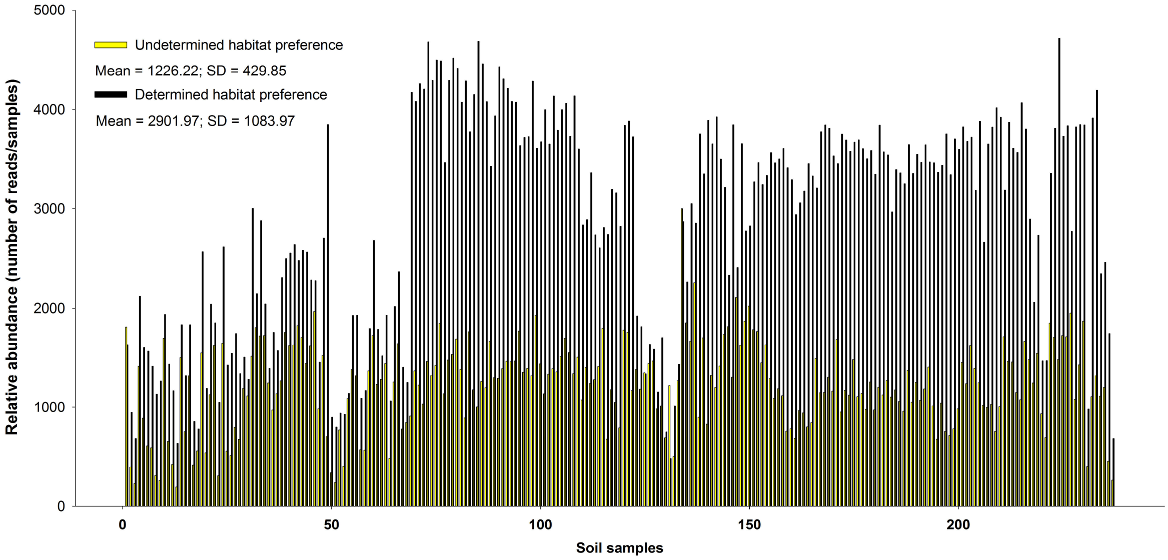


Figure S10. Sum of the relative abundances (per sample) of taxa with defined and undefined habitat preferences for the 511 dominant bacterial phylotypes. SD = standard deviation.

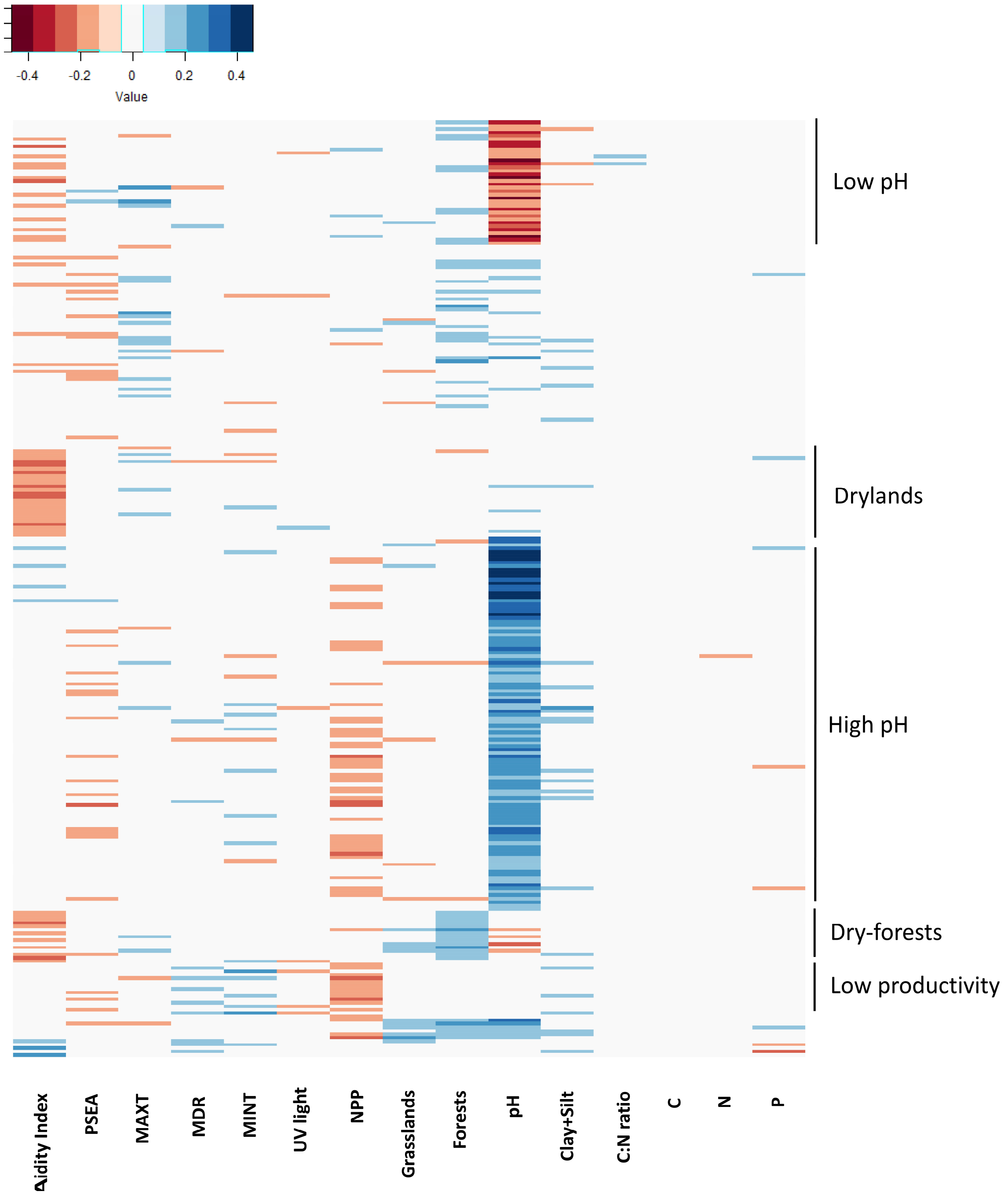


Figure S11. Heatmap including coefficients of correlation from semi-partial correlations between the relative abundance of each bacterial taxon (out of 270 phylotypes) with multiple environmental predictors. Data were sorted using the ecological cluster information provided in Table S1. MINT = minimum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

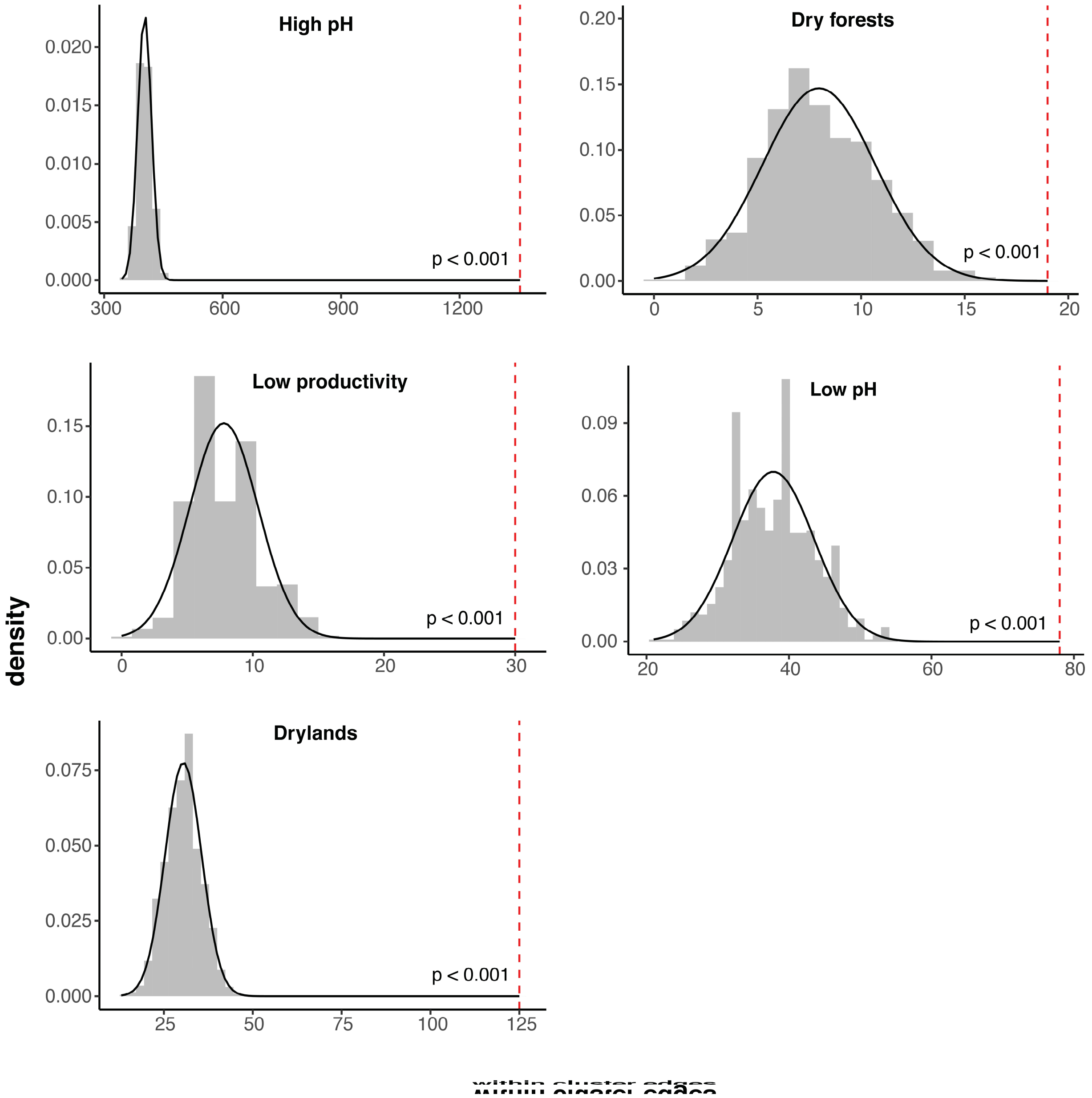


Figure S12. **Distributions of cluster environmental assembly across observed data and null models** (dashed line). For each major cluster, including high pH, low pH cluster, low productivity, drylands and dry-forest a histogram displays the expected distribution of expected number of edges between taxa 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.

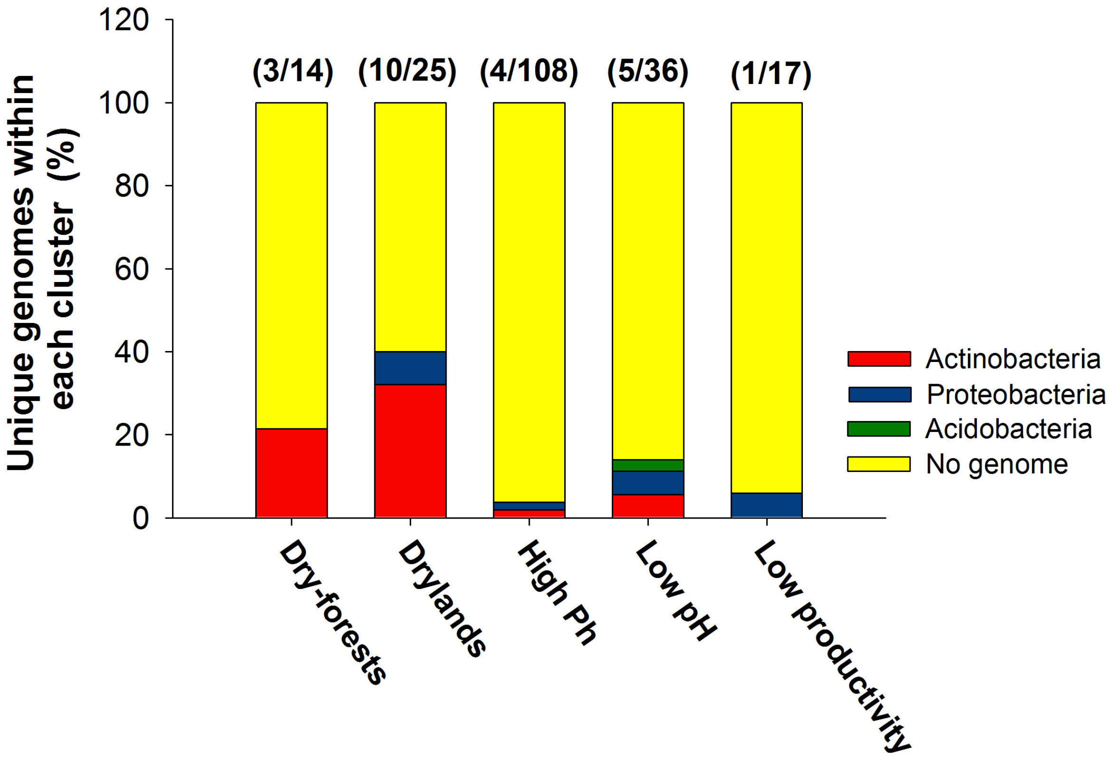
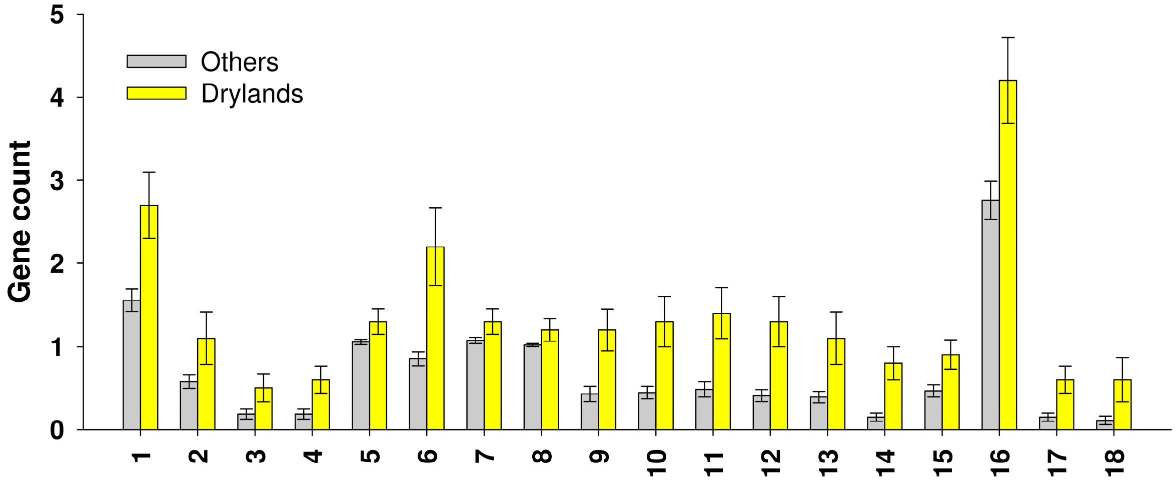
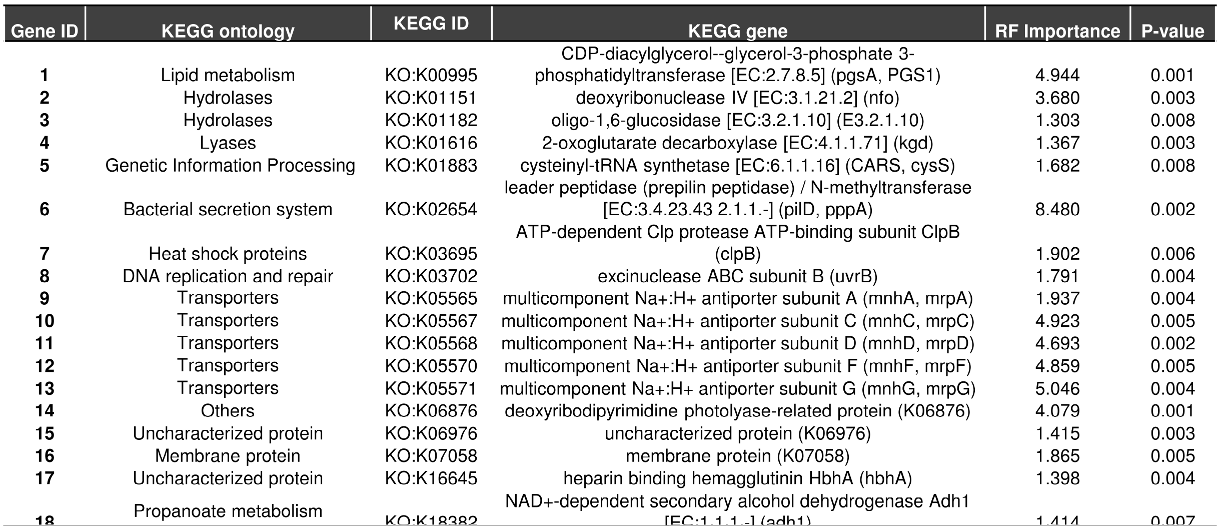


Figure S13. Percentage of genomes within each cluster for five well-defined ecological clusters of bacterial phylotypes with shared habitat preferences. The total number of phyloypes for which representative genomic data are available per cluster are indicated in brackets.





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.