

Structure and function of the global topsoil microbiome

Mohammad Bahram^{1,2,3,21*}, Falk Hildebrand^{4,21}, Sofia K. Forslund^{4,16,17}, Jennifer L. Anderson², Nadejda A. Soudzilovskaia⁵, Peter M. Bodegom⁵, Johan Bengtsson-Palme^{6,7,18}, Sten Anslan^{1,8}, Luis Pedro Coelho⁴, Helery Harend¹, Jaime Huerta-Cepas^{4,19}, Marnix H. Medema⁹, Mia R. Maltz¹⁰, Sunil Mundra¹¹, Pål Axel Olsson¹², Mari Pent¹, Sergei Pöhlme¹, Shinichi Sunagawa^{4,20}, Martin Ryberg², Leho Tedersoo^{13*} & Peer Bork^{4,14,15*}

Soils harbour some of the most diverse microbiomes on Earth and are essential for both nutrient cycling and carbon storage. To understand soil functioning, it is necessary to model the global distribution patterns and functional gene repertoires of soil microorganisms, as well as the biotic and environmental associations between the diversity and structure of both bacterial and fungal soil communities^{1–4}. Here we show, by leveraging metagenomics and metabarcoding of global topsoil samples (189 sites, 7,560 subsamples), that bacterial, but not fungal, genetic diversity is highest in temperate habitats and that microbial gene composition varies more strongly with environmental variables than with geographic distance. We demonstrate that fungi and bacteria show global niche differentiation that is associated with contrasting diversity responses to precipitation and soil pH. Furthermore, we provide evidence for strong bacterial–fungal antagonism, inferred from antibiotic-resistance genes, in topsoil and ocean habitats, indicating the substantial role of biotic interactions in shaping microbial communities. Our results suggest that both competition and environmental filtering affect the abundance, composition and encoded gene functions of bacterial and fungal communities, indicating that the relative contributions of these microorganisms to global nutrient cycling varies spatially.

Bacteria and fungi dominate terrestrial soil habitats in terms of biodiversity, biomass and their influence over essential soil processes⁵. Specific roles of microbial communities in biogeochemical processes are reflected by their taxonomic composition, biotic interactions and gene functional potential^{1–4}. Although microbial-biogeography studies have focused largely on single taxonomic groups, and on how their diversity and composition respond to local abiotic soil factors (for example, pH^{6,7}), global patterns and the impact of biotic interactions on microbial biogeography remain relatively unexplored. In addition to constraints imposed by environmental factors, biotic interactions may strongly influence bacterial communities. For example, to outcompete bacteria, many fungal taxa secrete substantial amounts of antimicrobial compounds⁸, which may select for antibiotic-resistant bacteria and effectively increase the relative abundance of antibiotic-resistance genes (ARGs). Here we used metagenomics and DNA metabarcoding (16S, 18S and internal transcribed spacer (ITS) rRNA gene markers), soil chemistry and biomass assessments (phospholipid fatty acids analyses (PLFAs)) to determine the relationships among genetic (functional

potential), phylogenetic and taxonomic diversity and abundance in response to biotic and abiotic factors in 189 topsoil samples, covering all terrestrial regions and biomes of the world⁹ (Extended Data Fig. 1a, Supplementary Table 1). Altogether, 58,000 topsoil subsamples were collected from 0.25-ha plots from 1,450 sites (40 subsamples per site), harbouring homogeneous vegetation that were minimally affected by humans. We minimized biases and shortcomings in sampling¹⁰ as well as technical variation, including batch effects¹¹, by using highly standardized collection and processing protocols. From the total collection, 189 representative sites were selected for this analysis. We validated our main findings in external datasets, including an independent soil dataset (145 topsoil samples; Supplementary Table 1) that followed the same sampling and sequencing protocol.

Using metagenomics, we constructed a gene catalogue for soils, by combining our newly generated data with published soil metagenomes ($n=859$, Supplementary Table 1) and identified 159,907,547 unique genes (or fragments thereof). Only 0.51% of these 160 million genes overlapped with those from published genomes and large gut¹² and ocean¹³ gene catalogues that are much closer to saturation (Supplementary Table 2), indicating that the functional potential of soil microbiomes is enormously vast and undersampled. For functional analysis, we annotated genes and functional modules via orthologous groups using the eggNOG database¹⁴. For each sample, we also constructed taxonomic profiles at the class and phylum levels for both bacteria and fungi from relative abundance of rRNA genes in metagenomic datasets (miTags¹⁵), complemented by operational taxonomic units (OTUs) that were based on clustering 18S rRNA and ITS¹⁶ genes for soil fungi and 16S rRNA genes for soil bacteria at 97% similarity threshold (see Methods ‘Metagenomics and metabarcoding analyses’). In total, 34,522 16S-based bacterial, 2,086 18S-based and 33,476 ITS-based fungal OTUs were analysed in the context of geographic space and 16 edaphic and climatic parameters were determined for each sampling site (see Methods ‘Statistical analyses’). Archaea were poorly represented in our metabarcoding data (less than 1% of OTUs) and metagenomics data (less than 1% of miTags) and hence are excluded from most analyses.

We examined whether the latitudinal diversity gradient (LDG), a trend of increasing diversity from the poles to the tropics seen in many macroscopic organisms, especially plants¹⁷, applies to microbial global distribution patterns¹⁰. We found that, contrary to the typical LDG,

¹Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia. ²Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden. ³Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden. ⁴Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. ⁵Environmental Biology Department, Institute of Environmental Sciences, CML, Leiden University, Leiden, The Netherlands. ⁶Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Göteborg, Göteborg, Sweden. ⁷Centre for Antibiotic Resistance research (CARe), University of Göteborg, Göteborg, Sweden. ⁸Braunschweig University of Technology, Zoological Institute, Braunschweig, Germany. ⁹Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. ¹⁰Center for Conservation Biology, University of California, Riverside, Riverside, CA, USA. ¹¹Section for Genetics and Evolutionary Biology (Evogene), Department of Biosciences, University of Oslo, Oslo, Norway. ¹²Biodiversity Unit, Department of Biology, Ecology building, Lund University, Lund, Sweden. ¹³Natural History Museum, University of Tartu, Tartu, Estonia. ¹⁴Max Delbrück Centre for Molecular Medicine, Berlin, Germany. ¹⁵Department of Bioinformatics, University of Würzburg, Würzburg, Germany. ¹⁶Present address: Experimental and Clinical Research Center, a cooperation of Charité-Universitätsmedizin and the Max-Delbrück Center, Berlin, Germany. ¹⁷Present address: Max Delbrück Centre for Molecular Medicine, Berlin, Germany. ¹⁸Present address: Wisconsin Institute of Discovery, University of Wisconsin-Madison, Madison, WI, USA. ¹⁹Present address: Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain. ²⁰Present address: Department of Biology, Institute of Microbiology, ETH Zurich, Zurich, Switzerland. ²¹These authors contributed equally: Mohammad Bahram, Falk Hildebrand. *e-mail: bahram@ut.ee; leho.tedersoo@ut.ee; bork@embl.de

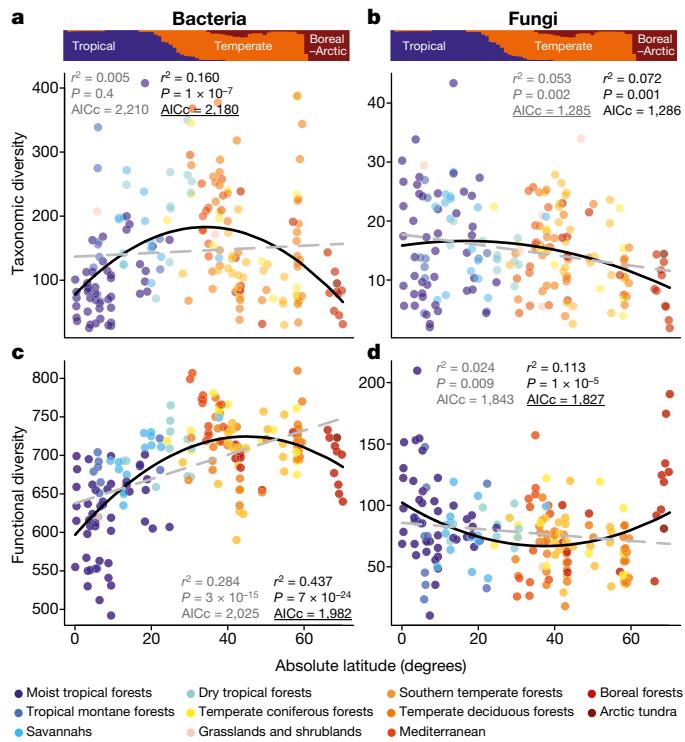


Fig. 1 | Fungal and bacterial diversity exhibit contrasting patterns across the latitudinal gradient. **a–d,** Latitudinal distributions of bacterial (**a, c**) and fungal (**b, d**) taxonomic (**a, b; n = 188** biologically independent samples) and gene functional (**c, d; n = 189** biologically independent samples) diversity in global soil samples. First- and second-order polynomial fits are shown in grey and black, respectively. The best polynomial fit was determined (as underlined) on the basis of the corrected Akaike Information Criterion (AICc; see Methods ‘Statistical analyses’) of the first and second order polynomial models (ANOVA: **a**, $F = 34.28$, $P < 10^{-7}$; **b**, $F = 3.84$, $P = 0.052$; **c**, $F = 50.48$, $P < 10^{-10}$; **d**, $F = 18.55$, $P < 10^{-4}$). Grey dashed and black solid lines are the first and second order polynomial regression lines, respectively. Diversity was measured using inverse Simpson index (these trends were robust to the choice of index, see Extended Data Fig. 2b, c). The latitudinal distribution of the high-level biome (tropical, temperate and boreal-arctic) is given at the top of **a** and **b**.

both taxonomic and gene functional diversity of bacteria peaked at mid-latitudes and declined towards the poles and the equator, as is also seen in the global ocean¹³, although the pattern was relatively weak for taxonomic diversity herein (Fig. 1a, c, Extended Data Figs. 1b, 2). The deviation of several bacterial phyla (5 out of 20) from the general trends may be explained by responses to edaphic and climate factors weakly related to latitude (Extended Data Fig. 1b) or contrasting effects at lower taxonomic levels (Supplementary Discussion). By contrast, the LDG does apply to overall fungal taxonomic diversity, and to three out of five fungal phyla when examined separately, but not to fungal functional diversity, which was lowest in temperate biomes and exhibited an inverse unimodal relationship with latitude (Fig. 1b, d, Extended Data Fig. 2c). The LDG was negligible for oceanic fungi (regression analysis, $P > 0.05$)¹³, possibly owing to their lower dispersal limitation and the paucity of plant associations. Although fungal taxonomic diversity decreased poleward, the total fungal biomass (inferred from PLFA markers) and the fungal/bacterial biomass ratio increased poleward, partly owing to a decline in bacterial biomass with increasing latitude (Extended Data Fig. 3a–c).

We tested the extent to which deterministic processes (such as competition and environmental filtering; that is, the niche theory) versus neutral processes (dispersal and drift; the neutral theory) explain the distributions of fungal and bacterial taxa and functions¹⁸. In bacteria, environmental variation correlated strongly with taxonomic

composition (partial Mantel test accounting for geographic distance between samples: $r_{\text{Env}|\text{Geo}} = 0.729$, $P = 0.001$) and moderately with gene functional composition ($r_{\text{Env}|\text{Geo}} = 0.100$, $P = 0.001$), whereas the overall effect of geographic distance among samples was negligible ($P > 0.05$). The weak correlation between geographic and taxonomic as well as functional composition suggests that environmental variables are more important than dispersal capacity in determining global distributions of soil bacteria and their encoded functions, as previously suggested¹⁹ and observed for oceanic prokaryotes¹³.

For fungi, both geographic distance and environmental parameters were correlated with taxonomic composition (ITS data: $r_{\text{Geo}|\text{Env}} = 0.307$, $P = 0.001$; $r_{\text{Env}|\text{Geo}} = 0.208$, $P = 0.001$; 18S data: $r_{\text{Geo}|\text{Env}} = 0.193$, $P = 0.001$; $r_{\text{Env}|\text{Geo}} = 0.333$, $P = 0.001$). Environmental distance (but not geographic distance) correlated with composition of fungal functional genes ($r_{\text{Env}|\text{Geo}} = 0.197$, $P = 0.001$), as was also observed for bacteria. The relatively weaker correlation of fungi with environmental variation is consistent with results from local scales⁷. Thus, at both global and local scales, different processes appear to underlie community assembly of fungi and bacteria.

To more specifically investigate the association between environmental parameters and the distribution of taxa and gene functions on a global scale, we used multiple regression modelling (see Methods ‘Statistical analyses’). We found that bacterial taxonomic diversity, composition, richness and biomass as well as the relative abundance of major bacterial phyla can be explained by soil pH, nutrient concentration and to a lesser extent by climatic variables (Extended Data Figs. 4, 5, Supplementary Table 4). The composition of bacterial communities responded most strongly to soil pH, followed by climatic variables, particularly mean annual precipitation (MAP; Extended Data Figs. 4, 5). This predominant role of pH agrees with studies from local to continental scales⁶, and may be ascribed to the direct effect of pH or related variables such as the concentration of calcium and other cations⁶. The relative abundance of genes that encode several metabolic and transport pathways were strongly increased with pH (Extended Data Fig. 4c), suggesting that there may be greater metabolic demand for these functions for bacteria in high-nutrient and alkaline conditions.

Compared to temperate biomes, tropical and boreal habitats contained more closely related taxa at the tip of phylogenetic trees, but from more distantly related clades (Extended Data Fig. 2d), indicating a deeper evolutionary niche specialization in bacteria²⁰. Together with global biomass patterns (Extended Data Fig. 2a), these results suggest that soil bacterial communities in the tropics and at high latitudes are subjected to stronger environmental filtering and include a relatively greater proportion of edaphic-niche specialists, possibly rendering these communities more vulnerable to global change. By contrast, phylogenetic overdispersion in temperate bacterial communities, may result from greater competitive pressure²⁰ or nutrient availability as predicted by the niche theory²¹.

In contrast to the strong association between bacterial taxonomic diversity and soil pH, diversity of bacterial gene functions was more strongly correlated with MAP (Extended Data Fig. 5a–h). The steeper LDG in gene functions than in taxa (Fig. 1a, c) may thus relate to the stronger association of specific metabolic functions to climate than to local soil conditions. Although soil and climate variables exhibited comparable correlations with fungal taxa, the soil carbon-to-nitrogen ratio (C/N) was the major predictor for fungal biomass and relative abundance and composition of gene functions (Extended Data Figs. 3g, 4b, d, Supplementary Table 4). We hypothesize that, compared to bacteria, the global distribution of fungi is more limited by resource availability owing to specialization for the use of specific compounds as substrates and greater energy demand.

We interpret opposing biogeographic trends for bacteria and fungi as niche segregation, driven by differential responses of bacteria and fungi to environmental factors⁷ and their direct competition. Gene functional diversity of both bacteria and fungi responded to MAP and soil pH, albeit in opposite directions (Extended Data Fig. 5c, d, g, h, Supplementary Table 3). This may partly explain the observed inverse

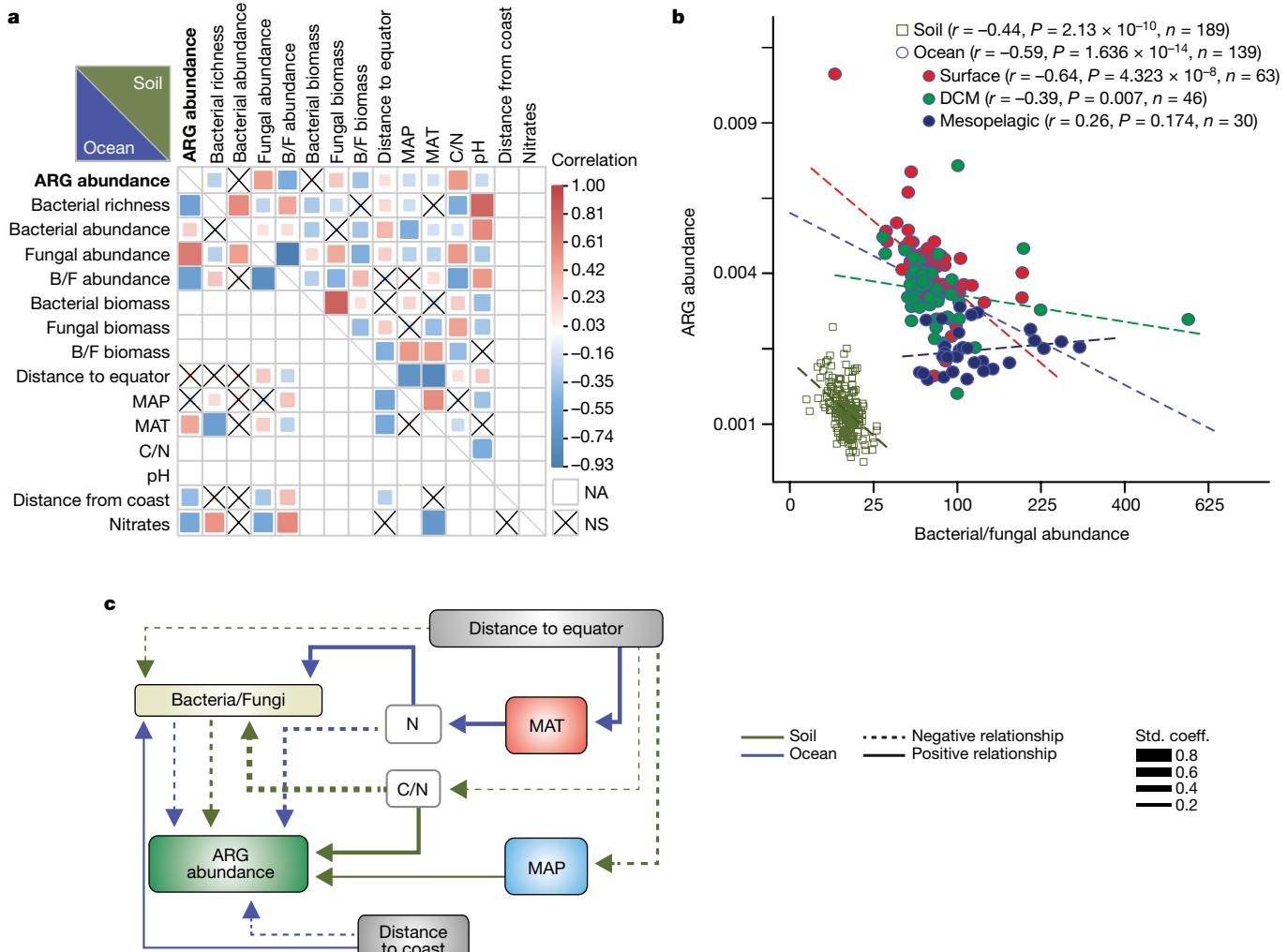


Fig. 2 | Global relative abundance of ARGs can be explained by a combination of biotic and abiotic factors. **a**, Pairwise Spearman's correlation matrix of the main biotic and abiotic determinants of the relative abundance of ARGs. **b**, Bacterial/fungal abundance ratio significantly correlated with the relative abundance of ARGs on a global scale. **c**, Structural equation modelling (SEM) of the relative abundance of ARGs in the soil (green) and ocean (blue) datasets (explaining 44% and 51% of variation, respectively; Supplementary Table 5). The goodness of fit was acceptable (soil dataset: root mean square error of estimation (RMSEA) = 0.00, P value for a test of close fit (P_{CLOSE}) = 0.989, n = 189

pattern of gene functional diversity across the latitudinal gradient, that is, niche differentiation, between bacteria and fungi (Fig. 1, Extended Data Fig. 2). Although increasing precipitation seems to favour higher fungal diversity, it is associated with higher bacterial/fungal biomass and abundance ratios (Extended Data Figs. 3d, g, 5f, h). The increasing proportion of fungi towards higher latitudes may be explained by competitive advantages, perhaps owing to a greater tolerance to nutrient and water limitation associated with potential long-distance transport by hyphae.

A role of inter-kingdom biotic interactions in determining the distributions of functional diversity and biomass in fungi and bacteria has been suggested previously²². As competition for resources affects the biomass of fungi and bacteria^{22,23}, we hypothesized that the bacterial/fungal biomass ratio is related to the prevalence of fungi and bacterial antibiotic-resistance capacity, because of broader activities of fungi than bacteria in using complex carbon substrates²⁴ as well as increased antibiotic production of fungi in high C/N environments²⁵. Consistent with this hypothesis, we found that both fungal biomass and the bacterial/fungal biomass ratio correlated with the relative abundance of ARGs (Extended Data Fig. 6) and that most fungal orthologous group

biologically independent samples; ocean dataset: $\text{RMSEA} = 0.059$, $P_{\text{CLOSE}} = 0.302$, $n = 139$ biologically independent samples). Abundance, relative abundance of miTags determined as fungi (including fungus-like protists) or bacteria; B/F, bacterial/fungal abundance or biomass ratio; bacterial richness, bacterial OTU (>97% similarity) richness on the basis of the metabarcoding dataset; biomass (nmol g^{-1}), absolute biomass on the basis of PLFA analysis; DCM, deep chlorophyll maximum; MAT, mean annual temperature; N, nitrates; NA, not applicable; NS, not significant ($P > 0.05$, $q > 0.1$); Std. coeff., standardized coefficients.

subcategories, particularly those involved in biosynthesis of antibiotic and reactive oxygen species, increased with soil C/N (Supplementary Table 4; Supplementary Discussion). We also found that the relative abundance of ARGs in topsoil is more strongly related to fungal relative abundance ($r = 0.435$, $P < 10^{-9}$) and bacterial/fungal abundance ratio ($r = -0.445$, $P < 10^{-12}$; Fig. 2b) than to bacterial relative abundance ($r = 0.232$, $P = 0.002$, on the basis of miTags), which is supported by our external validation dataset (fungal relative abundance $r = 0.637$, $P < 10^{-15}$; bacterial/fungal abundance ratio $r = -0.621$, $P < 10^{-15}$; bacterial relative abundance $r = 0.174$, $P = 0.036$). In addition, the relative abundance of ARGs in topsoil was significantly negatively correlated with bacterial phylogenetic diversity and OTU richness on the basis of the 16S rRNA gene (Spearman correlation, $P < 0.01$; Extended Data Figs. 7a, c, 8a), further supporting a role for biotic interactions in shaping microbial communities.

We also tested possible direct and indirect relationships between ARGs and 16 environmental predictors using structural equation modelling (SEM; Supplementary Table 5). The optimized model suggests that the soil C/N ratio and moisture, rather than pH—the predominant driver of bacterial diversity (Extended Data Fig. 3g, Supplementary

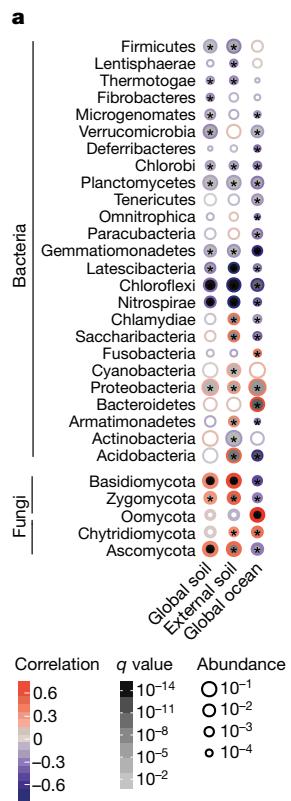


Fig. 3 | Fungi are the main determinants of the relative abundance of ARGs in soils and oceans. **a**, The association between the relative abundance of ARGs and major bacterial and fungal (including fungal-like protist) phyla in metagenomic samples from soils and oceans. Outer circle colour corresponds to the Pearson's correlation coefficient. Circle fill colour corresponds to the significance after adjustment for multiple testing (q value), as indicated in the legend. **b, c**, Relationships (non-parametric correlations) between the relative abundances of the most correlated fungal groups with ARGs in soil metagenomes from this study (**b**) and ocean metagenomes (**c**). For statistical details and significance, see Supplementary Table 8. Asterisks denote significance after Benjamini–Hochberg correction for multiple testing; * $q < 0.1$. See also Supplementary Discussion and Supplementary Table 8 for analogous results as in **a** but at the class level, and in other habitats besides soil and ocean including published non-forest and agricultural soil as well as human skin and gut samples.

Discussion)—affect the bacterial/fungal abundance ratio, which in turn affects the relative abundance of ARGs at the global scale (Fig. 2c). In line with increased production of antibiotics in high-competition environments, the soil C/N ratio was the best predictor for richness of fungal functional genes ($r^2 = 0.331, P < 10^{-15}$; Supplementary Table 3) and bacterial carbohydrate active enzyme (CAZyme) genes involved in degrading fungal carbohydrates ($r = 0.501, P < 10^{-12}$). The relative abundance of ARGs was also strongly correlated with C/N in the external validation dataset ($r = 0.505, P < 10^{-10}$).

Although the concomitant increase in antibiotic-resistance potential and the relative abundance of bacteria (as potential ARG carriers) was expected, the strong correlation of fungal relative abundance with the relative abundance of ARGs and in turn bacterial phylogenetic diversity may be explained by selection against bacteria that lack ARGs, such that bacteria surviving fungal antagonism are enriched for ARGs. Among all studied phyla, the relative abundance of Chloroflexi, Nitrospirae, and Gemmatimonadetes bacteria (on the basis of miTags), taxa with relatively low genomic ARG content (Supplementary Table 6) were most strongly negatively correlated with ARG relative abundance (Fig. 3a). By contrast, ARGs were strongly positively correlated with the relative abundance of Proteobacteria, which have the greatest average number of ARGs per genome²⁶ among bacteria (Supplementary Table 6), and the fungal phyla Ascomycota and Zygomycota sensu lato

(including Zoopagomycota and Mucoromycota) in both the global soil and the external validation datasets (Fig. 3a, b, Extended Data Fig. 9a, c, Supplementary Table 7). More specifically, ITS metabarcoding revealed increasing relative abundances of ARGs with numerous fungal OTUs (Supplementary Table 8), particularly those belonging to *Oidiodendron* (Myxotrichaceae, Ascomycota) and *Penicillium* (Aspergillaceae, Ascomycota), which are known antibiotic producers^{27,28} (Supplementary Discussion). Among bacterial ARGs, the relative abundance of efflux pumps and β -lactamases, which act specifically on fungal-derived antibiotics, were significantly correlated to the relative abundance of Ascomycota (Extended Data Fig. 10a, Supplementary Table 7). Actinobacteria, encompassing antibiotic-producing *Streptomyces*, also significantly correlated to ARG diversity in topsoil (Supplementary Table 6). Together these results suggest that relationships between organismal and ARG abundances are probably the result of selective and/or suppressive actions of antibiotics on bacteria.

Consistent with our observations in topsoil, we found evidence for antagonism between fungi and bacteria in oceans by reanalysing the distribution of ARGs in 139 water samples from the global *Tara* Oceans project¹³ (see Methods ‘External metagenomic datasets’; Supplementary Table 1, Extended Data Fig. 8a): the fungus-like stramenopile class Oomycetes (water moulds) and the fungal phylum Chytridiomycota constituted the groups most strongly associated with the relative abundance of bacterial ARGs (Fig. 3a, c, Extended Data Figs. 9b, d, 10b, d). Although there is little direct evidence that oomycetes produce antibiotics, their high antagonistic activity can induce bacteria²⁹ and other organisms, including fungi³⁰, to produce antibiotics (Supplementary Discussion). As in topsoil, bacterial phylogenetic diversity was significantly negatively correlated with the relative abundance of ARGs in ocean samples (Extended Data Fig. 7b, c). In addition, the relative abundance of ARGs declined with increasing distance from the nearest coast in ocean samples (Extended Data Fig. 8b), which may reflect the effect of a decreasing nutrient gradient along distance from the coast on the pattern of bacterial and fungal abundance and in turn the abundance of ARGs. The agreement of results from these disparate habitats suggests that competition for resources related to nutrient availability and climate factors drive a eukaryotic–bacterial antagonism in both terrestrial and oceanic ecosystems.

Our results indicate that both environmental filtering and niche differentiation determine global soil microbial composition, with a minor role of dispersal limitation at this scale (for limitations, see Methods ‘Metagenomics and metabarcoding analyses’). In particular, the global distributions of soil bacteria and fungi were most strongly associated with soil pH and precipitation, respectively. Our data further indicate that inter-kingdom antagonism, as reflected in the association of bacterial ARGs with fungal relative abundance, is also important in structuring microbial communities. Although further studies are needed to explicitly address the interplay between the bacterial/fungal abundance ratio and the abundance of ARGs, our data suggest that environmental variables that affect the bacterial/fungal abundance ratio may have consequences for microbial interactions and may favour fungi- or bacteria-driven soil nutrient cycling. This unprecedented view of the global patterns of microbial distributions indicates that global climate change may differentially affect bacterial and fungal community composition and their functional potential, because acidification, nitrogen pollution and shifts in precipitation all have contrasting effects on topsoil bacterial and fungal abundance, diversity and functioning.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0386-6>.

Received: 7 March 2017; Accepted: 13 June 2018;
Published online 1 August 2018.

1. Green, J. L., Bohannan, B. J. & Whitaker, R. J. Microbial biogeography: from taxonomy to traits. *Science* **320**, 1039–1043 (2008).
2. Reed, D. C., Algar, C. K., Huber, J. A. & Dick, G. J. Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proc. Natl Acad. Sci. USA* **111**, 1879–1884 (2014).
3. Maynard, D. S., Crowther, T. W. & Bradford, M. A. Fungal interactions reduce carbon use efficiency. *Ecol. Lett.* **20**, 1034–1042 (2017).
4. de Menezes, A. B., Richardson, A. E. & Thrall, P. H. Linking fungal–bacterial co-occurrences to soil ecosystem function. *Curr. Opin. Microbiol.* **37**, 135–141 (2017).
5. Bardgett, R. D. & van der Putten, W. H. Belowground biodiversity and ecosystem functioning. *Nature* **515**, 505–511 (2014).
6. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. 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).
7. Rousk, J. et al. Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* **4**, 1340–1351 (2010).
8. de Boer, W., Folman, L. B., Summerbell, R. C. & Boddy, L. Living in a fungal world: impact of fungi on soil bacterial niche development. *FEMS Microbiol. Rev.* **29**, 795–811 (2005).
9. Olson, D. M. et al. Terrestrial ecoregions of the world: a new map of life on earth: a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience* **51**, 933–938 (2001).
10. Green, J. & Bohannan, B. J. Spatial scaling of microbial biodiversity. *Trends Ecol. Evol.* **21**, 501–507 (2006).
11. Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS ONE* **7**, e33865 (2012).
12. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
13. Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
14. Huerta-Cepas, J. et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).
15. Logares, R. et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16**, 2659–2671 (2014).
16. Tedersoo, L. et al. Global diversity and geography of soil fungi. *Science* **346**, 1256688 (2014).
17. Willig, M. R., Kaufman, D. & Stevens, R. Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis. *Annu. Rev. Ecol. Syst.* **34**, 273–309 (2003).
18. Martiny, J. B. H. et al. Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).
19. Baas-Becking, L. G. M. *Geobiologie; Of Inleiding tot de Milieukunde* (W. P. Van Stockum & Zoon NV, The Hague, 1934).
20. Webb, C. O., Ackerly, D. D., McPeek, M. A. & Donoghue, M. J. Phylogenies and community ecology. *Annu. Rev. Ecol. Syst.* **33**, 475–505 (2002).
21. Bryant, J. A., Stewart, F. J., Eppley, J. M. & DeLong, E. F. Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone. *Ecology* **93**, 1659–1673 (2012).
22. Frey-Klett, P. et al. Bacterial–fungal interactions: hyphens between agricultural, clinical, environmental, and food microbiologists. *Microbiol. Mol. Biol. Rev.* **75**, 583–609 (2011).
23. Mille-Lindblom, C., Fischer, H. & Tranvik, J. L. Antagonism between bacteria and fungi: substrate competition and a possible tradeoff between fungal growth and tolerance towards bacteria. *Oikos* **113**, 233–242 (2006).
24. Koranda, M. et al. Fungal and bacterial utilization of organic substrates depends on substrate complexity and N availability. *FEMS Microbiol. Ecol.* **87**, 142–152 (2014).
25. Platas, G., Pelaez, F., Collado, J., Villuendas, G. & Diez, M. Screening of antimicrobial activities by aquatic hyphomycetes cultivated on various nutrient sources. *Cryptogam. Mycol.* **19**, 33–43 (1998).
26. Mende, D. R. et al. proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res.* **45**, D529–D534 (2017).
27. Bérdy, J. Thoughts and facts about antibiotics: where we are now and where we are heading. *J. Antibiot. (Tokyo)* **65**, 385–395 (2012).
28. Andersen, N. R. & Rasmussen, P. The constitution of clerodin a new antibiotic isolated from *Oidiodendron truncatum*. *Tetrahedron Lett.* **25**, 465–468 (1984).
29. Zhao, Y., Qian, G., Chen, Y., Du, L. & Liu, F. Transcriptional and antagonistic responses of biocontrol strain *Lysobacter enzymogenes* OH11 to the plant pathogenic oomycete *Pythium aphanidermatum*. *Front. Microbiol.* **8**, 1025 (2017).
30. Takahashi, K. et al. Cladomarine, a new anti-saprophytic compound isolated from the deep-sea fungus, *Penicillium coralligerum* YK-247. *J. Antibiot. (Tokyo)* **70**, 911–914 (2017).

Acknowledgements The authors thank I. Liiv for technical and laboratory assistance; S. Waszak for comments on the manuscript; Y. P. Yuan and A. Glazek for bioinformatics support and A. Holm Viborg for help in retrieving the CAZY database. We also thank V. Benes, R. Hercog and other members of the EMBL GeneCore (Heidelberg), who provided assistance and facilities for sequencing. This study was funded by the Estonian Research Council (grants PUT171, PUT1317, PUT1399, IUT20-30, MOBERC, KIK, RMK, ECOLCHANGE), the Swedish Research Council (VR grant 2017-05019), Royal Swedish Academy of Sciences, Helge Ax:son Johnsons Stiftelse, EU COST Action FP1305 Biolink (STSM grant), Netherlands Organization for Scientific research (vidi grant 016.161.318), EMBL European Union’s Horizon 2020 Research and Innovation Programme (#686070; DD-DeDaF) and Marie Skłodowska-Curie (600375).

Reviewer information *Nature* thanks S. Tringe and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.B., L.T. and P.B. conceived the project. L.T. supervised DNA extraction and sequencing. M.B., F.H., S.K.F., J.L.A., M.R. and P.M.B. designed and supervised the data analyses. F.H. designed and performed bioinformatics analysis. N.A.S. and P.A.O. performed biomass analysis. S.K.F., S.M., M.P., S.A., H.H., S.P., M.R.M., S.S. and L.T. contributed data. M.B., F.H., S.K.F., J.L.A., P.M.B., S.A., J.B.-P., M.H.M., L.P.C. and J.H.-C. performed the data analyses. M.B. wrote the first draft of the manuscript with input from F.H., S.F., J.L.A., L.T. and P.B. All authors contributed to data interpretation and editing of the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0386-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0386-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.B., L.T. or P.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Soil-sample preparation. Composite soil samples from 1,450 sites worldwide were collected using highly standardized protocols¹⁶. The sampling was conducted broadly across the most influential known environmental gradient (that is, the latitude) taking advantage of a global ‘natural laboratory’ to study the impact of climate on diversity across vegetation, biome and soil types and to enable testing of the effects of environmental parameters, spatial distance and biotic interactions in structuring microbial communities. We carefully selected representative sites for different vegetation types separated by spatial distances that were sufficient to minimize spatial autocorrelation and to cover most areas of the globe. Total DNA was extracted from 2 g of soil from each sample using the PowerMax Soil DNA Isolation kit (MoBio). A subset of 189 high-quality DNA samples representing different ecoregions spanning multiple forest, grassland and tundra biomes (Supplementary Table 1) were chosen for prokaryote and eukaryote metabarcoding (ribosomal rRNA genes) and whole metagenome analysis. Samples from desert ($n=8$; G4010, G4034, S357, S359, S411, S414, S418 and S421) and mangrove ($n=1$; G4023) biomes yielded sufficient DNA for metabarcoding, but not for metagenomics sequencing, thus these samples were used for global mapping of taxonomic diversity but excluded from all comparisons between functional and taxonomic diversity. One sample (S017) contained no 16S sequences; thus, altogether, 189 and 197 samples were used for metagenomics and metabarcoding analyses, respectively.

To determine the functional gene composition of each sample, 5 µg total soil DNA (300–400 bp fragments) was ligated to Illumina adaptors using the TruSeq Nano DNA HT Library Prep Kit (Illumina) and shotgun-sequenced in three runs of the Illumina HiSeq 2500 platform (2 × 250 bp paired-end chemistry, rapid run mode)³¹ in the Estonian Genomics Center (Tartu, Estonia). Taxonomic composition was estimated from the same DNA samples using ribosomal DNA metabarcoding for bacteria (16S V4 subregion) and eukaryotes (18S V9 subregion). For amplification of prokaryotes and eukaryotes, universal prokaryote primers 515F and 806RB³² (although this pair may discriminate against certain groups of Archaea and Bacteria such as Crenarchaeota/Thaumarchaeota (and SAR11³³) and eukaryote primers 1389f and 1510r³⁴ were used. Although the resolution of 16S rRNA sequencing is limited to assignments to the level of genus (and higher), it is currently a standard approach in profiling bacterial communities and thus enabled us at least to explore patterns at coarse phylogenetic resolution.

Each primer was tagged with a 10–12-base identifier barcode¹⁶. DNA samples were amplified using the following PCR conditions: 95 °C for 15 min, and then 30 cycles of 95 °C for 30 s, 50 °C for 45 s and 72 °C for 1 min with a final extension step at 72 °C for 10 min. The 25 µl PCR mix consisted of 16 µl sterilized H₂O, 5 µl 5 × HOT FIREPol Blend MasterMix (Solis Biodyne, Tartu, Estonia), 0.5 µl each primer (200 nM) and 3 µl template DNA. PCR products from three technical replicates were pooled and their relative quantity was evaluated after electrophoresis on an agarose gel. DNA samples producing no visible band or an overly strong band were amplified using 35 or 25 cycles, respectively. The amplicons were purified (FavourPrep Gel/PCR Purification Kit; Favourgen), checked for quality (ND 1000 spectrophotometer; NanoDrop Technologies), and quantified (Qubit dsDNA HS Assay Kit; Life Technologies). Quality and concentration of 16S amplicon pools were verified using Bioanalyzer HS DNA Analysis Kit (Agilent) and Qubit 2.0 Fluorometer with dsDNA HS Assay Kit (Thermo Fisher Scientific), respectively. Sequencing was performed on an Illumina MiSeq at the EMBL GeneCore facility (Heidelberg, Germany) using a v2 500 cycle kit, adjusting the read length to 300 and 200 bp for read1 and read2, respectively. 18S amplicon pools were quality checked using Bioanalyzer HS DNA Analysis Kit (Agilent), quantified using Qubit 2.0 Fluorometer with dsDNA HS Assay Kit (Thermo Fisher Scientific) and sequenced on an Illumina HiSeq at Estonian Genomics Center (Tartu, Estonia). Sequences resulting from potential contamination and tag switching were identified and discarded on the basis of two negative and positive control samples per sequencing run.

Soil chemical analysis and biomass analysis. All topsoil samples were subjected to chemical analysis of pH_{KCl}, P_{total} (total phosphorus), K, Ca and Mg; the content of ¹²C, ¹³C, ¹⁴N and ¹⁵N was determined using an elemental analyser (Eurovector) coupled with an isotope-ratio mass spectrometer³⁵.

To calculate the absolute abundance of bacteria and fungi using an independent approach, bacterial and fungal biomass were estimated from PLFAs³⁶ in nmol g⁻¹ as follows. Lipids were extracted from 2 g freeze dried soil in a one-phase solution of chloroform, methanol and citrate buffer³⁷. Chloroform and citrate buffer was added to split the collected extract into one lipophilic phase, and one hydrophilic phase. The lipid phase was collected and applied on a pre-packed silica column³⁷. The lipids were separated into neutral lipids, intermediate lipids and polar lipids (containing the phospholipids) by subsequent elution with chloroform, acetone

and methanol. The neutral and phospholipids were dried using a speed vac. Methyl nonadecanoic acid (Me19:0) was added as an internal standard. The lipids were subjected to a mild alkaline methanolysis, in which fatty acids were derivatized to fatty acid methyl esters (FAMEs). The FAMEs from neutral (NLFFAs) and phospholipids (PLFFAs) were dried, using speed vac, and then dissolved in hexane before analysis on a gas chromatograph as described³⁸. Fungal biomass was estimated as the concentration of PLFA 18:2ω6,9 and bacterial biomass from the sum of nine PLFAs (115:0, 116:0, 117:0, a15:0, a17:0, cy17:0, cy19:0, 10Me17:0 and 10Me18:0)³⁷. The nomenclature of fatty acids was according to previously published work³⁸.

Acquisition of metadata from public databases. Climate data including monthly temperature and precipitation were obtained from the WorldClim database (<http://www.worldclim.org>). In addition, estimates of soil carbon, moisture, pH, potential evapotranspiration (PET) and net primary productivity (NPP) at 30 arc minute resolution were obtained from the Atlas of the Biosphere (<https://nelson.wisc.edu/sage/data-and-models/atlas/maps.php>). Samples were categorized into 11 biomes⁹, with all grassland biomes being categorized as ‘grasslands’. Thus, the following biomes were considered and summarized to three global levels: moist tropical forests, tropical montane forests and dry tropical forests, savannahs as tropical; Mediterranean, grasslands and shrublands, southern temperate forests, coniferous temperate forests and deciduous temperate forests as temperate; and boreal forests and arctic tundra as boreal–arctic. The time from the last fire disturbance was estimated on the basis of enquiries to local authorities or collaborators and evidence from the field.

Metagenomics and metabarcoding analyses. *Processing of metagenomics sequence data.* Most soil microorganisms are uncultured, making their identification difficult. Metagenomics analysis has emerged as a way around this to capture both genetic and phylogenetic diversity. As such, it can only directly reveal the potential for functions through determining and tracing gene family abundances (as opposed to realized protein activity), which may be involved in various functional pathways³⁹, but we can safely assume a strong correspondence between gene functional potential and the resulting ecosystem functioning⁴⁰ or enzyme activities⁴¹.

Reads obtained from the shotgun metagenome sequencing of topsoil samples were quality-filtered, if the estimated accumulated error exceeded 2.5 with a probability of $\geq 0.01^{42}$, or > 1 ambiguous position. Reads were trimmed if base quality dropped below 20 in a window of 15 bases at the 3' end, or if the accumulated error exceeded 2 using the sdm read filtering software⁴³. After this, all reads shorter than 70% of the maximum expected read length (250 bp unless noted otherwise for external datasets) were removed. This resulted in retention of 894,017,558 out of 1,307,037,136 reads in total (Supplementary Table 1). We implemented a direct mapping approach to estimate the functional gene composition of each sample. First, the quality-filtered read pairs were merged using FLASH⁴⁴. The merged and unmerged reads were then mapped against functional reference sequence databases (see below) using DIAMOND v.0.8.10 in blastx mode using ‘-k 5 -e 1e-4 --sensitive’ options. The mapping scores of two unmerged query reads that mapped to the same target were combined to avoid double counting. In this case, the hit scores were combined by selecting the lower of the two *e* values and the sum of the bit scores from the two hits. The best hit for a given query was based on the highest bit score, longest alignment length and highest percentage identity to the subject sequence. Finally, aligned reads were filtered to those that had an alignment percentage identity $> 50\%$ and $e < 1 \times 10^{-9}$ (see ‘Parameterization and validation of metagenomics approach’ for parameter choice).

The functional databases to which metagenomic reads were mapped included gene categories related to ROS sources (peroxidases genes databases^{45,46}, KEGG⁴⁷ (Kyoto Encyclopedia of Genes and Genomes) and CAZyme genes (<http://www.cazy.org>, accessed 22 November 2015)⁴⁸). To facilitate the interpretation of the results, the relative abundance of CAZyme genes were summed on the basis of the substrates for each gene family. Substrate utilization information for CAZyme families was obtained from previously published work^{49,50} as well as the CAZypedia (http://www.cazypedia.org/index.php?title=Carbohydrate-binding_modules&oldid=9411). On the basis of the KEGG orthologue abundance matrices we calculated SEED functional module abundances. For functional annotations of metagenomic reads, we used *in silico* annotation on the basis of a curated database of the orthologous gene family resource eggNOG 4.5¹⁴.

For all databases that included taxonomic information (eggNOG, KEGG, CAZy), reads were mapped competitively against all kingdoms and assigned into prokaryotic and eukaryotic groups, on the basis of best bit score in the alignment and the taxonomic annotation provided within the database at kingdom level. All functional abundance matrices were normalized to the total number of reads used for mapping in the statistical analysis, unless mentioned otherwise (for example, rarefied in the case of diversity analysis, see ‘Statistical analyses’). This normalization better takes into account differences in library size as it has the advantage of including the fraction of unmapped (that is functionally unclassified) reads. Although there are limitations to using relative abundance of genes, our analysis shows which potential functions are relatively more important. Without any

normalization, such analyses cannot be performed. It is currently difficult to test the absolute numbers, owing to limitations of reliably quantifying soil DNA resulting from differences in extraction efficiency and the level of degradation.

To identify ARGs in our metagenome samples, the merged and unmerged reads were mapped to a homology expansion⁵¹ of the Antibiotic Resistance Gene Data Base (ARDB). Only hits that passed the minimum sequence identity values as listed in the ARDB for each family were taken further into account. Although newer ARG databases exist, only the ARDB presently has curated family inclusion thresholds that directly allow application to our topsoil dataset; as soil microbial diversity is so large, unlike for gut datasets, high-fidelity gene catalogue construction will not be possible until many more samples are available. Therefore, direct mapping of reads to the gene family databases becomes necessary for our analysis, in turn necessitating ARG inclusion thresholds that are well-defined for single reads, not merely for full-length genes. Thus, the cut-offs curated by ResFams⁵² or CARD⁵³, for example, are inappropriate, as they are defined in the length-dependent bit-score space. The ARDB cut-offs, however, are defined as sequence identities and thus in principle are applicable to sequences shorter than full length. Because of these technical limitations, we used a soil-gene catalogue to determine CARD-based ARG abundance matrices (see 'Gene catalogue construction').

It is important to note that measurements of functional genes, including ARGs, represent relative proportions of different gene families, because the absolute amount of DNA differs among samples. This necessitates the use of statistical tests that do not assume absolute measurements, and centres analysis of this type on comparisons across the set of samples.

Estimation of taxa abundance using miTag. We used a miTag approach¹⁵ to determine bacterial and fungal community composition from metagenome sequence data. First, SortMeRNA⁵⁴ was used to extract and blast search rRNA genes against the SILVA LSU/SSU database. Reads approximately matching these databases with $e < 10^{-1}$ were further filtered with custom Perl and C++ scripts, using FLASH to attempt to merge all matched read pairs. In case read pairs could not be merged, which happens when the overlap between read pairs is too small, the reads were interleaved such that the second read pair was reverse complemented and then sequentially added to the first read. To fine-match candidate interleaved or merged reads to the Silva LSU/SSU databases, lambda⁵⁵ was used. Using the lowest common ancestor (LCA) algorithm adapted from LotuS (v.1.462)⁴³, we determined the identity of filtered reads on the basis of lambda hits. This included a filtering step, in which queries were only assigned to phyla and classes if they had at least 88% and 91% similarity to the best database hit, respectively. The taxon-by-sample matrices were normalized to the total number of reads per sample to minimize the effects of uneven sequencing depth. The average of SSU and LSU matrices was used for calculating the relative abundance of phyla or classes. The abundance of miTag sequences matching bacteria and fungi was used to determine the bacterial/fungal abundance ratio. Although LSU/SSU assessments refer to the number of fungal cells rather than the number of discrete multicellular fungi (as this can apply to all samples equally), it is not systematically biased for comparing the trends of bacterial to fungal abundance across samples.

External metagenomic datasets. To validate and compare the global trends at smaller scales, we used a regional scale dataset of 145 topsoil samples that was generated and processed using the same protocol as our global dataset (Supplementary Table 1).

In addition, to compare patterns of ARG diversity in soils and oceans on a global scale, we re-analysed the metagenomics datasets of the *Tara Oceans*¹³, including all size fractions (Supplementary Table 1). After quality filtering, 41,790,928,650 out of 43,076,016,494 reads were retained from the *Tara Oceans* dataset.

The quality-filtered reads from all datasets were mapped to the corresponding databases using DIAMOND, with the exception that no merging of read pairs was attempted, because the chances of finding overlapping reads were too low (with a read length of 100 bp and insert size of 300 bp (*Tara Oceans*)). Sequences for SSU/LSU miTags were extracted from these metagenomics datasets as described above. ARG abundance matrices were also obtained from the *Tara Oceans* project on the basis of the published gene catalogues annotated using a similar approach as in the current study.

Gene catalogue construction. To create a gene catalogue, we first searched for complete reference genes that matched to read pairs in our collection using Bowtie2⁵⁶ with the options '--no-unal --end-to-end'. The resulting bam files were sorted and indexed using samtools 1.3.1⁵⁷ and the jgi_summarize_bam_contig_depths provided with MetaBat⁵⁸ was used to create a depth profile of genes from the reference databases that were covered with $\geq 95\%$ nucleotide identity. This cut-off is commonly used in constructing gene catalogues^{13,59} and chosen to delineate genes belonging to the same species. Using the coverage information, we extracted all genes that had at least 200 bp with $\geq 1\times$ coverage by reads from our topsoil metagenomes. The reference databases included an ocean microbial gene catalogue¹³, a gut microbial gene catalogue¹², as well as all genes extracted from 25,038 published bacterial genomes²⁶. Altogether, 273,723, 2,376 and 8,642 genes

from proGenomes, IGC and Tara database, respectively, could be matched to soil reads and were used in the gene catalogue.

The majority of genes in our catalogue were assembled from the topsoil samples presented here. To reduce the likelihood of chimaeric reads, each sample was assembled separately using Spades 3.7-0 (development version obtained from the authors)⁶⁰ in metagenomic mode with the parameters '-only-assembler -m 500 --meta -k 21,33,67,111,127'. Only sdm-filtered⁴³ paired reads were used in the assembly, with the same read-filtering parameters as described above. Resulting assemblies had an average N50 of 469 bases (total of all assemblies 21,538 Mb). The low N50 reflects difficulties in the assembly of soil metagenomes, which probably reflects the vast microbial genetic diversity of these ecosystems. We further de novo assembled reads from two other deep sequencing soil⁶¹ and sediment studies⁶², using the same procedure and parameters, except that the Spades parameter '-k 21,33,67,77' was adjusted to a shorter read length. Furthermore, we included publicly available data from the European Nucleotide Archive (ENA). The ENA was queried to identify all projects with publicly available metagenomes and whose metadata contained the keyword 'soil'. The initial set of hits was then manually curated to select relevant project and/or samples that were assembled as described above. Additionally, we integrated gene predictions from soil metagenomes downloaded from MG-RAST⁶³ (Supplementary Table 1). Assembly was not attempted for these samples owing to the absence of paired-end reads, and relatively low read depth; rather, only long reads or assemblies directly uploaded to MG-RAST with ≥ 400 bp length were retained. Therefore, only scaffolds and long reads, with at least 400 bp length, were used for analysis. On these filtered sequences, genes were de novo predicted using prodigal 2.6.1⁶⁴ in metagenomic mode. Finally, we merged the predicted genes from assemblies, long reads, gene catalogues and references genomes to construct a comprehensive soil gene catalogue.

Thus, 53,294,555,100 reads were processed, of which 31,015,827,636 (58.20%) passed our stringent quality control. The initial gene set predicted on the soil assemblies and long reads was separated into 17,114,295 complete genes and 111,875,596 incomplete genes. A non-redundant gene catalogue was built by comparing all genes to each other. This operation was performed initially in amino-acid space using DIAMOND⁶⁵. Subsequently, any reported hits were checked in nucleotide space. Any gene that covered at least 90% of another one (with at least 95% identity over the covered area) was considered to be a potential representative of it (genes are also potential representatives of themselves). The final set was chosen by greedily picking the genes that were representative of the highest number of input genes until all genes in the original input have at least one representative in the output. This resulted in a gene catalogue with a total of 159,907,547 non-redundant genes at 95% nucleotide identity cut-off. We mapped reads from our experiment onto the gene catalogue with bwa⁶⁶, requiring >45 nt overlap and $>95\%$ identity. The average mapping rate was $26.2 \pm 7.4\%$. Although the gene catalogue is an invaluable resource for future explorations of the soil microbiome, we decided to rely on using the direct mapping approach to gene functional composition, owing to the low overall mapping rate. Furthermore, using minimap2⁶⁷ to find genes at 95% similarity threshold, we compared the soil gene catalogue with the *Tara Oceans* gene catalogue¹³, human gut gene catalogue¹² and the proGenomes prokaryotic database²⁶. The gene catalogue nucleotide and amino acid sequences and abundance matrix estimates from rtk⁶⁸ have been deposited at http://vm-lux.embl.de/~hildebra/Soil_gene_cat/.

Estimation of ARG abundance using CARD. CARD abundances in topsoil samples were estimated by annotating the soil gene catalogue using a DIAMOND search of the predicted amino acid sequences against the CARD database and filtering hits to the specified bit-score cut-offs in the CARD database. On the basis of gene abundances in each sample, we estimated the abundance of different CARD categories per metagenomic sample. Despite qualitative similarities in overall trends of ARDB and CARD abundance matrices, CARD abundance estimation is limited by being based on the gene catalogue (only a $26.2 \pm 7.4\%$ of all metagenomic reads could be mapped to the gene catalogue).

Processing of metabarcoding sequence data. The LotuS pipeline was used for bacterial 16S rRNA amplicon sequence processing. Reads were demultiplexed with modified quality-filtering settings for MiSeq reads, increasing strictness to avoid false positive OTUs. These modified options were the requirement of correctly detected forward 16S primer, trimming of reads after an accumulated error of 1 and rejecting reads below 28 average quality or, exceeding an estimated accumulated error >2.5 with a probability of ≥ 0.01 ⁴². Furthermore, we required each unique read (reads preclustered at 100% identity) to be present eight or more times in at least one sample, four or more times in at least two samples, or three or more times in at least three samples. In total 27,883,607 read pairs were quality-filtered and clustered with uparse⁶⁹ at 97% identity. Chimeric OTUs were detected and removed on the basis of both reference-based and de novo chimaera checking algorithms, using the RDP reference database (http://drive5.com/uchime/rdp_gold.fa) in uchime⁶⁹, resulting in 13,070,436 high-quality read pairs to generate and estimate the abundance of bacterial OTUs. The seed sequence for each OTU cluster was

selected from all read pairs assigned to that OTU, selecting the read pair with the highest overall quality and closest to the OTU centroid. Selected OTU seed read pairs were merged with FLASH⁴⁴ and a taxonomic identity was assigned to each OTU by aligning full-length sequences with lambda to the SILVA v.123 database⁷⁰ and the LotuS least common ancestor (LCA) algorithm. This was performed using the following LotuS command line options: ‘-p miSeq -derepMin 8:1,4:2,3:3 --simBasedTaxo 2 --refDB SLV -thr 8’. OTU abundances per sample were summed to class and phylum level per sample, according to their taxonomic classification, to obtain taxa abundance matrices. However, the choice of clustering method (for example, Swarm) and identity threshold had little effect on retrieved OTU richness (comparison with 99% threshold: $r = 0.977$, $P < 10^{-15}$; comparison with Swarm clustering: $r = 0.979$, $P < 10^{-15}$).

For eukaryotic 18S rRNA genes, we used the same options in LotuS, except that reads were rejected if they did not occur at least six times each in a minimum of two samples or at least four times each in a minimum of three samples. This was done to account for lower sequencing depth in 18S rRNA compared to 16S rRNA dataset. Furthermore, the database to annotate fungal taxonomy was extended to include general annotations of SILVA and information from unicellular eukaryotes (PR² database⁷¹). Of 7,462,813 reads, 2,890,093 passed quality filtering. The fungal ITS metabarcoding dataset¹⁶ was downloaded and used in addition to 18S data in specific analyses, such as finding fungal OTUs associated with ARG relative abundance. The resulting taxon abundance matrix was further filtered to remove sequences of chloroplast origin for all three metabarcoding experiments.

Full-length sequences representing OTUs were aligned using the SILVA reference alignment as a template in mothur⁷². A phylogenetic tree was constructed using FastTree²³ with the maximum-likelihood method using default settings. This program uses the Jukes–Cantor models to correct for multiple substitutions. *Parameterization and validation of metagenomics approach.* Although we used state-of-art molecular approaches, there are several potential limitations regarding our analyses related to the technologies used. All metagenomics and amplicon-based analyses are affected by taxonomic biases in sequence databases, whereas (PCR-free) miTag as well as amplicon sequencing are biased owing to differential ribosomal gene copy number across taxonomic groups. Amplicon-based metabarcoding, specifically, is affected by both primer PCR artefacts and PCR biases that may affect estimates of absolute organism abundance. These biases are inherent to all metagenomics and metabarcoding studies. However, all these biases affect different samples equally (same rRNA gene copy numbers, same PCR biases per species, same database bias per taxa) and thus we estimate that our results are robust to these methodological shortcomings. Shotgun-based metagenomics is affected by reference bias, in which human pathogens or Proteobacteria are overrepresented. The necessity for lenient thresholds becomes obvious from annotating phylogenetic profiles with MetaPhlAn2⁷⁴ using standard parameters: whereas we observed that most fungal phyla are present abundantly in our samples, MetaPhlAn2 detected Ascomycota in only 2 out of 189 samples. In 48 out of 189 samples, no organism (bacteria/archaea/eukaryotes) was detected, and the most abundant phylum was Proteobacteria (55%). As these results are clearly deviating from our miTag, 16S, 18S and ITS analyses, specific database cut-off thresholds were required for this project.

To optimize the analysis pipeline and identify suitable e values for filtering blastx results, we used metagenomic simulations of four reference genomes for which CAZy assignments in the CAZy database were available. Simulated reads were created as 250-bp paired reads with 400 bp insert at differing sequence abundances from the four reference genomes in each simulated metagenome, using iMessi⁷⁵. For this simulated dataset, we used the pipeline described above to derive CAZy functional profiles. We found that querying short reads processed as above against databases results in the retrieval of most genes at relative abundances consistent with expectations on the basis of the reference genomes at $e < 10^{-9}$ ($r = 0.95 \pm 0.01$, $P < 0.001$). Furthermore, we simulated 200 metagenomes from 18 bacterial genomes, five bacterial plasmids, one fungal mitochondrion and two fungal genomes at differing relative proportions in each of these simulated metagenomes (Supplementary Table 11). We subsequently simulated 1,000,000 reads of 250-bp and 400-bp insert size using iMessi, and mapped these against reference databases and retained hits that fulfilled the following arbitrary criteria (used in all subsequent analyses): e value cut-off of 10^{-9} , alignment length ≥ 20 amino acids, and similarity $\geq 50\%$ amino acids to the target sequence. From these, we generated functional profiles and found a strong correlation of simulated to expected functional metagenomic composition on the basis of mixed fungal and bacterial genomes ($r = 0.94 \pm 0.05$, $P < 0.001$).

Estimating fungal antibiotics production. We also specifically screened for fungal gene clusters directly associated with antibiotic activity, on the basis of a compiled database of MiBiG (minimum information about a biosynthetic gene cluster, <https://mibig.secondarymetabolites.org>) repository entries that describe gene clusters for which the products have been shown experimentally to display antimicrobial activities (Supplementary Table 12). To extend the range of genes

that can be associated with the validated, antibiotic-producing, MiBiG protein domains, we downloaded all published non-redundant fungal genomes deposited in JGI (Supplementary Table 14) as well as all non-redundant fungal genes deposited in NCBI. The set of MiBiG and fungus-derived genes was screened with custom hidden Markov models for domains from secondary metabolite production (specifically these were drmat, AMP-binding, Condensation, PKS_KS and Terpene synthesis domains). All identified domains were aligned together with the MiBiG domains using Clustal Omega⁷⁶ and a tree was constructed with FastTree2. Phylogenetic trees were rooted to midpoint and automatically scanned to identify highly supported clades (aLRT branch support ≥ 0.99) in which antibiotic-producing MiBiG domains were monophyletically grouped. The average nucleotide identity within each such group was subsequently used as identity cut-off in the mapping step. All metagenomic reads were mapped with DIAMOND in blastx mode to the newly created database, using the previously mentioned sequence identity cut-offs and rejecting domains of reads that were mapping to bacterial non-supervised orthologous groups.

Statistical analyses. *Data normalization and diversity estimates.* All statistical analyses were performed using specific packages in R (v.3.3.2) unless otherwise noted. Diversity parameters were estimated from OTU and functional gene matrices that were rarefied to an equal number per sample to reduce the effect of variation in sequencing depth using the function rrarefy in vegan (v.2.2.1; <https://cran.r-project.org/web/packages/vegan/index.html>). ARG matrices were normalized to the total number of merged and singleton reads. The total abundance of ARGs per sample was estimated by summing the abundance of all individual ARGs per sample. ARG diversity measures indicate the variety and their proportions produced.

From the rarefied matrices we calculated OTU, orthologous group and CAZyme gene richness (function specnumber) and diversity (function diversity on the basis of the inverse Simpson index). The latter measure accounts for both richness and evenness, and it gives more weight to abundant groups compared to the Shannon index. Our results were robust to choice of index, and the various diversity indices were highly correlated in the present dataset (for example, bacterial taxonomic diversities calculated using inverse Simpson versus using Shannon diversity were highly correlated: $r = 0.888$, $P < 10^{-15}$; for a comparison of richness and diversity trends, see Extended Data Fig. 2b, c). As evenness and richness were highly correlated in all datasets, we report the results that, on the basis of the diversity index, represent both richness and evenness. The rarefaction process was repeated for calculating taxonomic and gene functional diversity and richness on the basis of the average of 100 rarefied datasets.

Phylogenetic diversity was calculated on the basis of Faith's Phylogenetic Diversity (PD) metric⁷⁷ in the Picante package (v.1.6-2; <https://cran.r-project.org/web/packages/picante/index.html>). In addition, to assess phylogenetic clustering and overdispersion, nearest relative index (NRI) and nearest taxon index (NTI) were calculated in Picante. Although both measures are closely related, NRI is more sensitive to phylogenetic diversity at deep nodes, whereas NTI is more sensitive to phylogenetic clustering towards tips. A null model of shuffling taxon labels (100 times) was used to randomize phylogenetic relationships among OTUs.

Correlating environmental parameters to taxa and functions. To identify the main determinants of taxonomic and gene functional composition or diversity and relative abundance of phyla and classes, we used a series of statistical tests. We included all prominent environmental variables that we expected to have a significant effect on microbial diversity on the basis of previous studies, and which were feasible to collect. These included soil pH, carbon and nutrient levels and factors that can affect these, such as fire, assuming soil as the major resource for microbial nutrition. We also included isotope ratios of nitrogen ($\delta^{15}\text{N}$) and carbon ($\delta^{13}\text{C}$) as these provide principal components for carbon and nitrogen cycling. To avoid overfitting and to ensure model simplicity, we excluded the variables that had no significant effect on fungal or bacterial diversity, such as altitude, age of vegetation, plant diversity and community (the first two principal component analysis axes of plant community variation at both genus and family level) and basal areas of trees. Thus, for univariate regression modelling, 16 variables (Supplementary Table 14) were included.

To understand which factors explain the orthologous group- and OTU-based community composition, variable selection was performed in the Forward.sel function of Packfor (v.0.0-8/r109; https://r-forge.r-project.org/R/?group_id=195) according to the coefficient of determination (threshold, $r^2 = 0.01$). All functional and taxonomic compositional matrices were transformed using Hellinger transformation before statistical analysis. Furthermore, Mantel tests and partial Mantel tests were used to test the effects of geographical versus environmental distances on the compositional similarity of OTUs and orthologous groups as implemented in vegan. Mantel tests allow testing of the correlation between two distance matrices, partial Mantel tests are similar but also control for variation in a third distance matrix. In our analysis, we controlled for the effect of geographic distance while testing the correlation of environmental variation and functional or taxonomic composition variation. The importance of biome type in explaining functional gene and taxonomic composition was tested in permutational multivariate

analysis of variance (PERMANOVA) using the Adonis function of vegan (using 10^3 permutation for calculating pseudo-*F* test statistic and its statistical significance). For constructing orthologous group and OTU distance matrices, the Bray–Curtis dissimilarity was calculated between each pair of samples. Great-circle distance was used to calculate a geographic distance matrix between samples on the basis of geographical coordinates. This test compares the intragroup distances to intergroup distances in a permutation scheme and from this assesses significance. PERMANOVA post hoc *P* values were corrected for multiple testing using the Benjamini–Hochberg correction. We visualized taxonomic (OTU) and functional (orthologous group) composition of bacteria using global nonmetric multidimensional scaling (GNMDS) in vegan with the following options: two dimensions, initial configurations = 100, maximum iterations = 200 and minimum stress improvement in each iteration = 10^{-7} . The main environmental drivers of the relative abundance of major taxonomic groups and main functional categories were recovered by random forest analysis⁷⁸ using the R package randomForest (v.4.6–10; <https://cran.r-project.org/web/packages/randomForest/index.html>).

To examine latitudinal gradients of diversity at phylum level (Fig. 2), the diversity of OTUs assigned to each phylum was calculated on the basis of inverse Simpson index. Diversity values were modelled in response to environmental variables and predicted values were extracted, which were used in a clustering and bootstrapping analysis to depict the similarities of phyla environmental associations using pvclust (v.1.3–2; <https://cran.r-project.org/web/packages/pvclust/index.html>) with 1,000 iterations. To model latitudinal gradients and environmental associations of diversity and biomass (Fig. 1, Extended Data Fig. 3), we compared the goodness-of-fit estimates between first and second order polynomial models on the basis of the corrected Akaike information criterion (AICc) using analysis of variance (ANOVA). AICc reflects both goodness of fit and parsimony of the models.

For univariate regression modelling of diversity and biomass measures, ordinary least squares (OLS) or generalized least squares (GLS) regression models were used depending on the importance of the spatial component in the nlme package (v.3.1–120; <https://cran.r-project.org/web/packages/nlme/index.html>). The model variance structure (Gaussian, exponential, spherical and linear) was evaluated on the basis of AICc. After selection of variance structure, variables were combined in a set of models with specified variance structure (that is, the number of tested models = $2^{\text{number of variables}}$). The resulting models were sorted according to AICc values to reveal the best model. Lists of the five best-fitting models for each response variable are given in Supplementary Table 4. Prior to model selection, all variables were evaluated for linearity, normality, and multicollinearity (excluded if the variance inflation factor was >5). The degree of polynomial functions (linear, quadratic, cubic) was chosen on the basis of the lowest AIC values. Because of nonlinear relationships with response variables, a quadratic term for pH was also included in the model selection procedure. The accuracy of the final models was evaluated using tenfold ‘leave-one-out’ cross-validation. For this, we used 1,000 randomly sampled 90%-data subsets for model training and predicting the withheld data. To minimize biases owing to the partitioning of the data and potential overfitting, the average of 1,000 resulting determination coefficients are reported as cross-validated r^2 (r^2_{cv}) for each regression model.

Correlating biotic interactions to taxa and functions. To test the associations of biotic variables on ARG relative abundance, we used a sparse partial least squares (sPLS) analysis, which reduces dimensionality by projecting predictor variables onto latent components to identify the 16S/18S lineages (phyla and classes) and the ITS OTUs most strongly associated with ARG relative abundance, as implemented in the mixOmics (v.5.0–4; <https://cran.r-project.org/web/packages/mixOmics/index.html>) package. ARG composition and taxonomic community matrices (miTags classes and phyla and ITS OTUs) were normalized to library size using Hellinger transformation. Significance of associations was examined by bootstrap tests of subsets of each dataset. We subsequently used partial least squares (PLS) analysis to predict ARG relative abundance on the basis of significantly correlated lineages, which allows the dimensionality of multivariate data to be reduced into PLS components. Optimal numbers of PLS components for prediction of the relative ARG abundance were selected on the basis of leave-one-out cross-validation. To confirm the results of PLS analysis, we further used a cross-validated LASSO model to simultaneously perform variable selection and model fitting, as implemented in glmnet (v.2.0–2; <https://cran.r-project.org/web/packages/glmnet/index.html>). First the lambda shrinkage parameter was determined from a cross-validated LASSO-penalized logistic regression classifier. Using this shrinkage parameter, a new logistic regression classifier was fit to the data to predict ARG relative abundance.

To further test direct and indirect effects of geographic and environmental variables on microbial distributions, we built SEM models in the AMOS software (SPSS) by including predictors of the best GLS model. In a priori models, all indirect and direct links between variables were established on the basis of their pairwise correlations. We subsequently removed non-significant links and variables or created new links between error terms until a significant model fit was achieved. Goodness of fit was assessed on the basis of a χ^2 test to evaluate

the difference between the observed and estimated by-model covariance matrices (a non-significant value indicates that the model fits the observed data). We also used RMSEA and *P*_{CLOSE} to assess the discrepancy between the observed data and model per degree of freedom, which is less sensitive to sample size compared to the χ^2 test (RMSEA < 0.08 and *P*_{CLOSE} > 0.05 show a good fit). Observed correlations between diversity and environmental values can serve as the first step towards understanding the structure and function of global topsoil microbiome, however, they are not proof of causations and mechanism. Despite the fact that we used SEM modelling to infer indirect links, we cannot preclude the possibility of other biotic or soil variables confounded with climate variables that we did not include in our models. Further laboratory experiments may be able to address the causality of relationships reported in this study.

Differences between univariate variables such as taxonomic and functional richness were tested using a non-parametric Wilcoxon rank-sum test, with Benjamini–Hochberg multiple testing correction. Post hoc statistical testing for significant differences between all combinations of two groups was conducted only for taxa with *P* < 0.2 in the Kruskal–Wallis test. For this, Wilcoxon rank-sum tests were calculated for all possible group combinations and corrected for multiple testing using Benjamini–Hochberg multiple testing correction.

Geographic coordinates were plotted on a world map transformed to a Winkler projection, using the maptools package (v.0.8–36; <https://cran.r-project.org/web/packages/maptools/index.html>).

Limitations of statistical modelling on a global scale. Although we performed cross-validations to test the accuracy of most of our statistical models, predictions might be limited by the vast diversity in soil microbiomes. For example, strong local variation in soil pH may lead to deviation from general patterns, which is a common limitation in environmental sciences. Given the large spatial scale and strong environmental gradient in our sampling design, and long-term persistence of DNA in soil⁷⁹, seasonal variation in soils is expected to have a minor impact⁸⁰ (in contrast to the oceans). In addition, the vast majority of our samples were collected during the growing season, further reducing possible seasonal biases. We nevertheless tested the effect of sampling month and seasons and found no significant effect of seasonality on diversity indices (*P* > 0.05). We also compared the effect of seasons and years in a time series study in two of our sites, which revealed no seasonal effects on richness and composition (unpublished data). In particular, the relationship between bacterial phylogenetic diversity and pH, are strongly consistent with studies performed at the local to continental scales and within a single season^{6,7,81}, which indicates the robustness of our results. Nonetheless, validation of the proposed models needs to be performed by other researchers with more data or an independent dataset, particularly by including samples from under-sampled regions (Extended Data Fig. 1a) and from different seasons (to account for seasonality). Under-sampled regions in our dataset (for example, North Asia) lowered precision of our models for those regions. Unfortunately, there are no published global datasets with comparable sampling protocols that could be directly compared and used for model validation, and we encourage future studies that will make this possible.

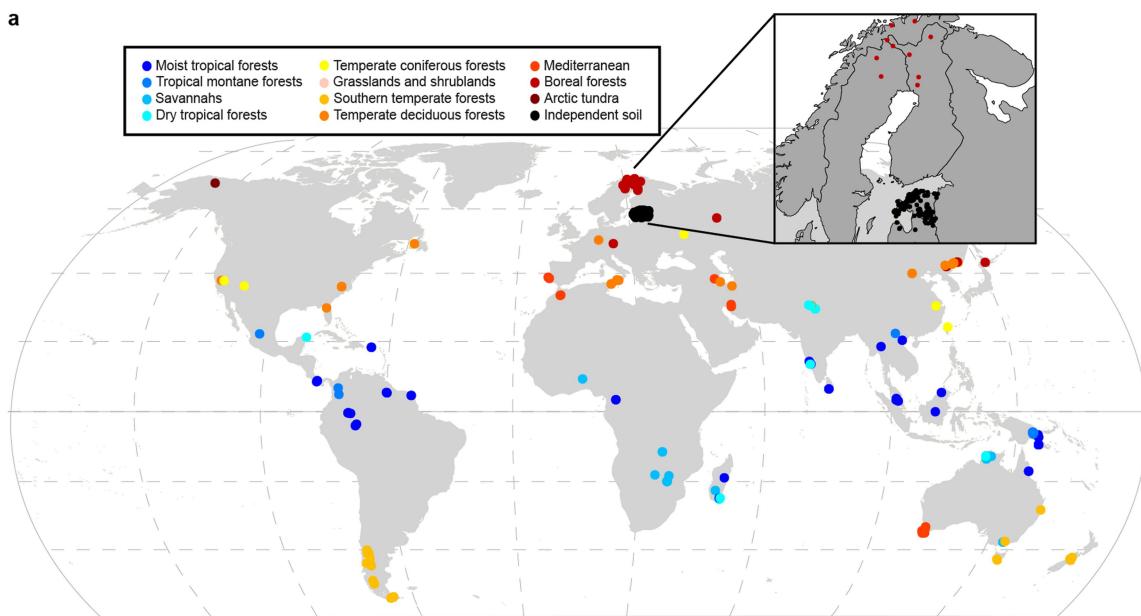
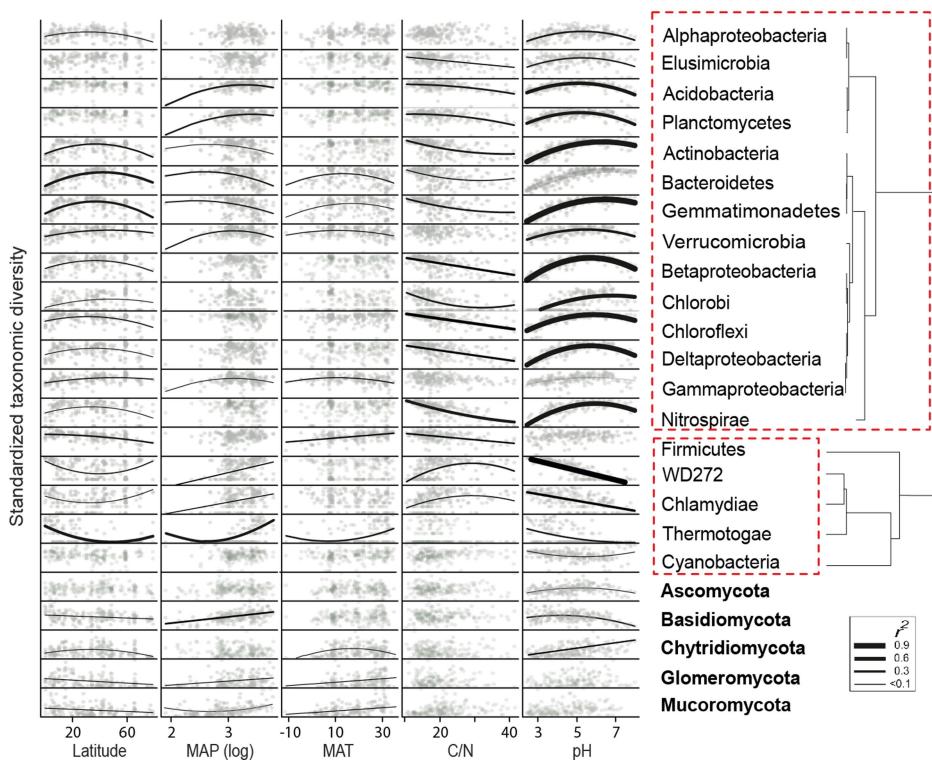
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The pipeline to process metabarcoding samples is available under <http://psbweb05.psb.ugent.be/lotus/>. The pipeline to process shotgun metagenomic samples is available under <https://github.com/hildebra/MATAFILER> and <https://github.com/hildebra/Rarefaction>.

Data availability. All metagenomics and metabarcoding sequences have been deposited in the European Bioinformatics Institute Sequence Read Archive database: Estonian forest and grassland topsoil samples, accession numbers PRJEB24121 (ERP0105926); 16S metabarcoding data of global soil samples, accession numbers PRJEB19856 (ERP021922); 18S metabarcoding data of global soil samples, accession numbers PRJEB19855 (ERP021921); Global analysis of soil microbiomes, accession numbers PRJEB18701 (ERP020652). The soil gene catalogue and dataset are available at http://vm-lux.embl.de/~hildebra/Soil_gene_cat/. The Tara Oceans data are available at <http://ocean-microbiome.embl.de/companion.html>. All other data that support the findings of this study are available from the corresponding authors upon request.

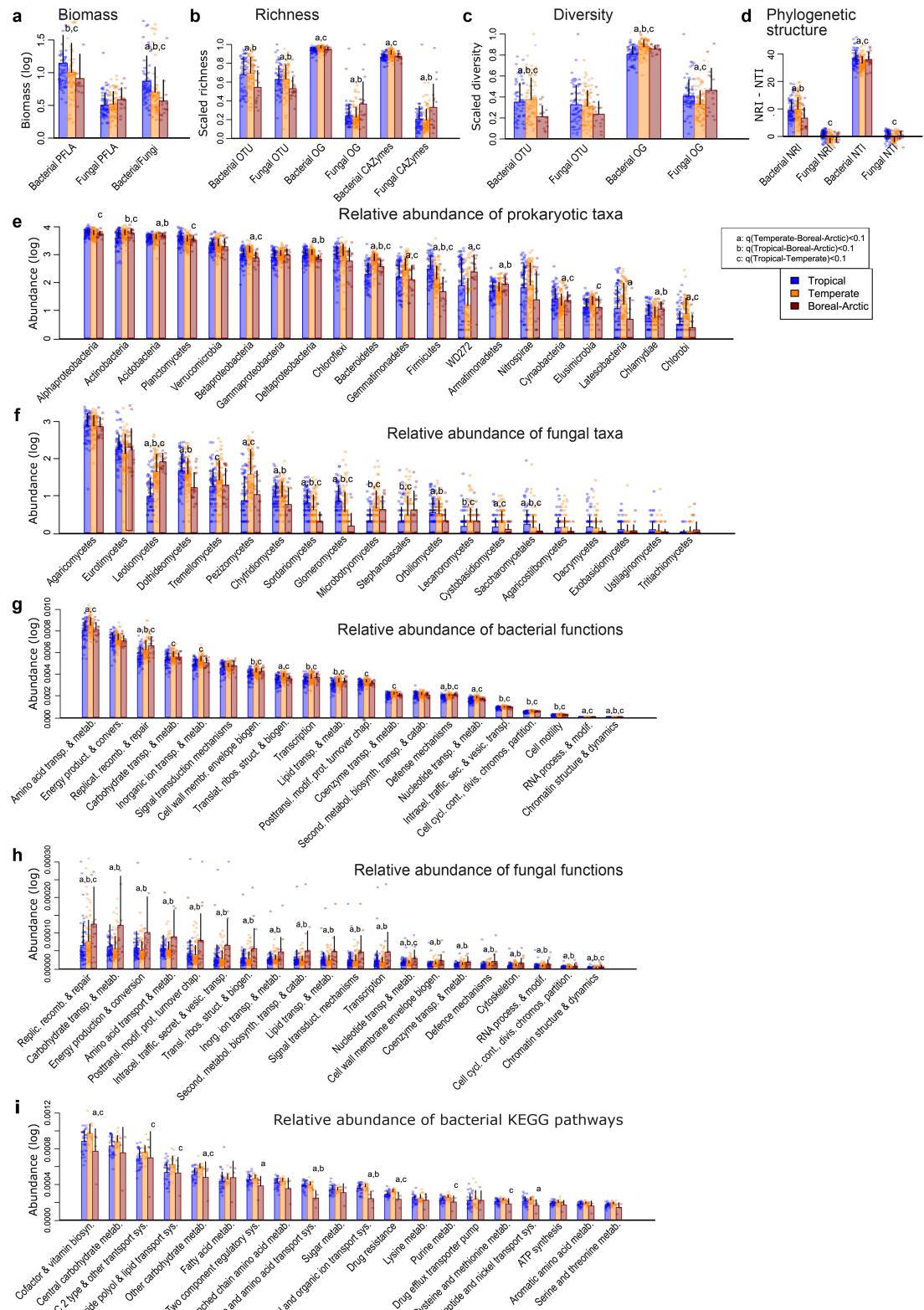
31. Tedersoo, L. et al. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *Mycobanks* **10**, 1–43 (2015).
32. Caporaso, J. G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA* **108**, 4516–4522 (2011).
33. Walters, W. et al. Improved bacterial 16S rRNA gene (V4 and V4–5) and fungal internal transcribed spacer marker gene primers for microbial community survey. *mSystems* **1**, e00009–e00015 (2015).
34. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4**, e6372 (2009).

35. Tedersoo, L. et al. Enzymatic activities and stable isotope patterns of ectomycorrhizal fungi in relation to phylogeny and exploration types in an afrotropical rain forest. *New Phytol.* **195**, 832–843 (2012).
36. Frostegård, Å., Tunlid, A. & Bäath, E. Use and misuse of PLFA measurements in soils. *Soil Biol. Biochem.* **43**, 1621–1625 (2011).
37. van Aarle, I. M. & Olsson, P. A. Fungal lipid accumulation and development of mycelial structures by two arbuscular mycorrhizal fungi. *Appl. Environ. Microbiol.* **69**, 6762–6767 (2003).
38. Frostegård, Å., Tunlid, A. & Bäath, E. Phospholipid fatty acid composition, biomass, and activity of microbial communities from two soil types experimentally exposed to different heavy metals. *Appl. Environ. Microbiol.* **59**, 3605–3617 (1993).
39. Prosser, J. I. Dispelling misconceptions and identifying opportunities for the use of ‘omics’ in soil microbial ecology. *Nat. Rev. Microbiol.* **13**, 439–446 (2015).
40. Salles, J. F., Le Roux, X. & Poly, F. Relating phylogenetic and functional diversity among denitrifiers and quantifying their capacity to predict community functioning. *Front. Microbiol.* **3**, 209 (2012).
41. Trivedi, P. et al. Microbial regulation of the soil carbon cycle: evidence from gene–enzyme relationships. *ISME J.* **10**, 2593–2604 (2016).
42. Puente-Sánchez, F., Aguirre, J. & Parro, V. A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. *Nucleic Acids Res.* **44**, e40 (2016).
43. Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P. & Raes, J. LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* **2**, 30 (2014).
44. Magoć, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
45. Choi, J. et al. fPoxDB: fungal peroxidase database for comparative genomics. *BMC Microbiol.* **14**, 117 (2014).
46. Fawal, N. et al. PeroxiBase: a database for large-scale evolutionary analysis of peroxidases. *Nucleic Acids Res.* **41**, D441–D444 (2012).
47. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
48. Cantarel, B. L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009).
49. Cantarel, B. L., Lombard, V. & Henrissat, B. Complex carbohydrate utilization by the healthy human microbiome. *PLoS ONE* **7**, e28742 (2012).
50. Cardenas, E. et al. Forest harvesting reduces the soil metagenomic potential for biomass decomposition. *ISME J.* **9**, 2465–2476 (2015).
51. Forslund, K. et al. Country-specific antibiotic use practices impact the human gut resistome. *Genome Res.* **23**, 1163–1169 (2013).
52. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).
53. McArthur, A. G. et al. The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57**, 3348–3357 (2013).
54. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
55. Hauswedell, H., Singer, J. & Reinert, K. Lambda: the local aligner for massive biological data. *Bioinformatics* **30**, i349–i355 (2014).
56. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
57. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
58. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
59. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
60. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
61. Howe, A. C. et al. Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl Acad. Sci. USA* **111**, 4904–4909 (2014).
62. Sharon, I. et al. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res.* **25**, 534–543 (2015).
63. Meyer, F. et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
64. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
65. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
66. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
67. Li, H. Minimap2: fast pairwise alignment for long DNA sequences. Preprint at <https://arxiv.org/abs/1708.01492> (2017).
68. Saary, P., Forslund, K., Bork, P. & Hildebrand, F. RTK: efficient rarefaction analysis of large datasets. *Bioinformatics* **33**, 2594–2595 (2017).
69. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10**, 996–998 (2013).
70. Pruesse, E. et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
71. Guillou, L. et al. The Protist Ribosomal Reference database (PR²): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2012).
72. Schloss, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
73. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
74. Truong, D. T. et al. MetaPhiAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
75. Mende, D. R. et al. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* **7**, e31386 (2012).
76. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
77. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10 (1992).
78. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
79. Carini, P. et al. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat. Microbiol.* **2**, 16242 (2016).
80. Žifáková, L., Větrovský, T., Howe, A. & Baldrian, P. Microbial activity in forest soil reflects the changes in ecosystem properties between summer and winter. *Environ. Microbiol.* **18**, 288–301 (2016).
81. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc. Natl Acad. Sci. USA* **103**, 626–631 (2006).

a**b**

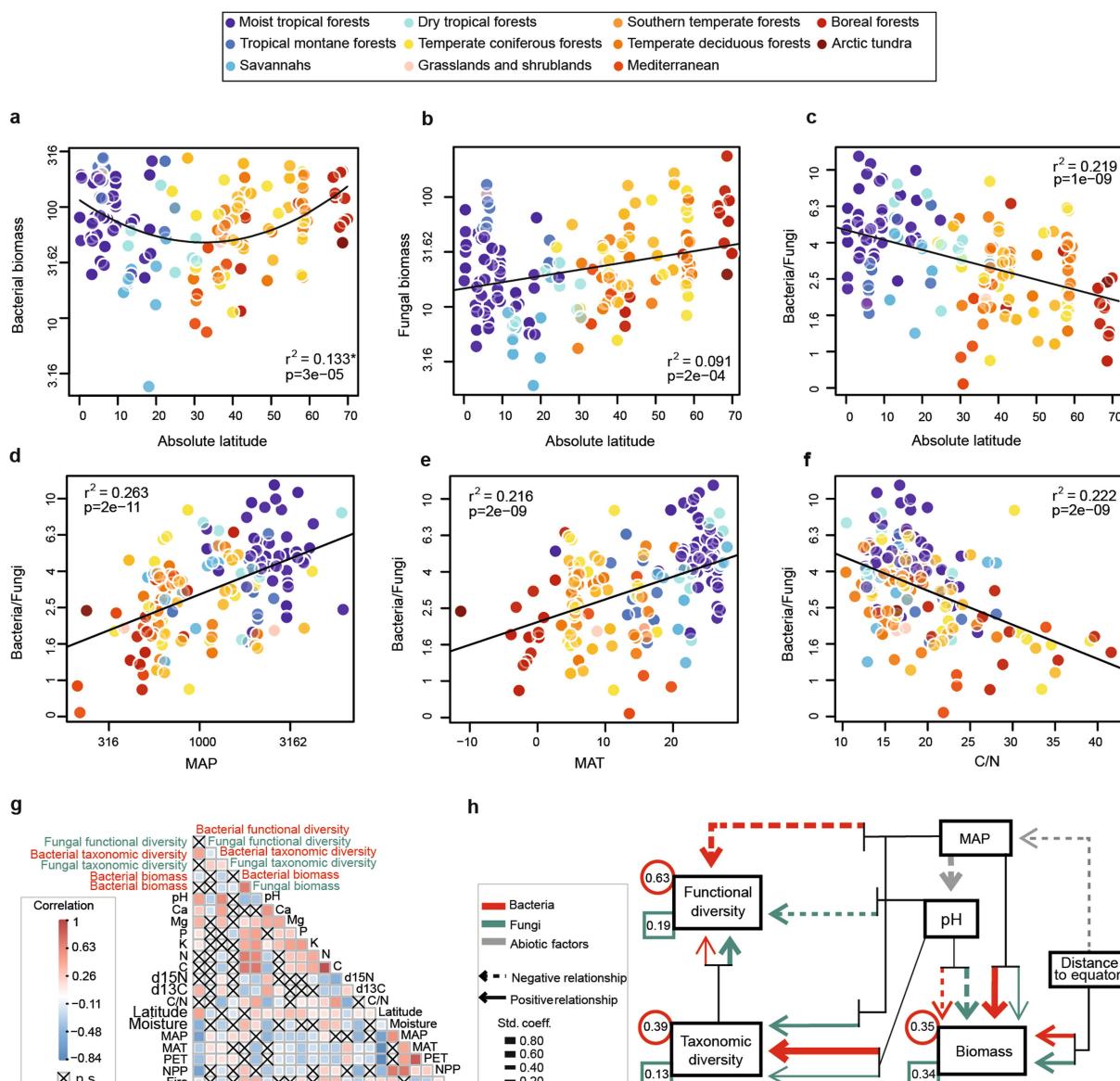
Extended Data Fig. 1 | Distribution of topsoil samples and diversity patterns of phyla. **a**, A map of samples used for metagenomic and metabarcoding analysis. Colours indicate biomes as shown in the legend. Desert samples were only used in metabarcoding analysis and were excluded in comparative analysis of functional and taxonomic patterns. Black symbols refer to samples from an independent soil dataset (145 topsoil samples; Supplementary Table 1) that were used for validation of

our results. **b**, Relationship between the diversity of major microbial phyla (classes for Proteobacteria) and environmental variables across the global soil samples ($n = 197$ biologically independent samples). Only regression lines for significant relationships after Bonferroni correction are shown. Diversity was measured using Hellinger-transformed matrices on the basis of inverse Simpson index. Latitude, absolute latitude.



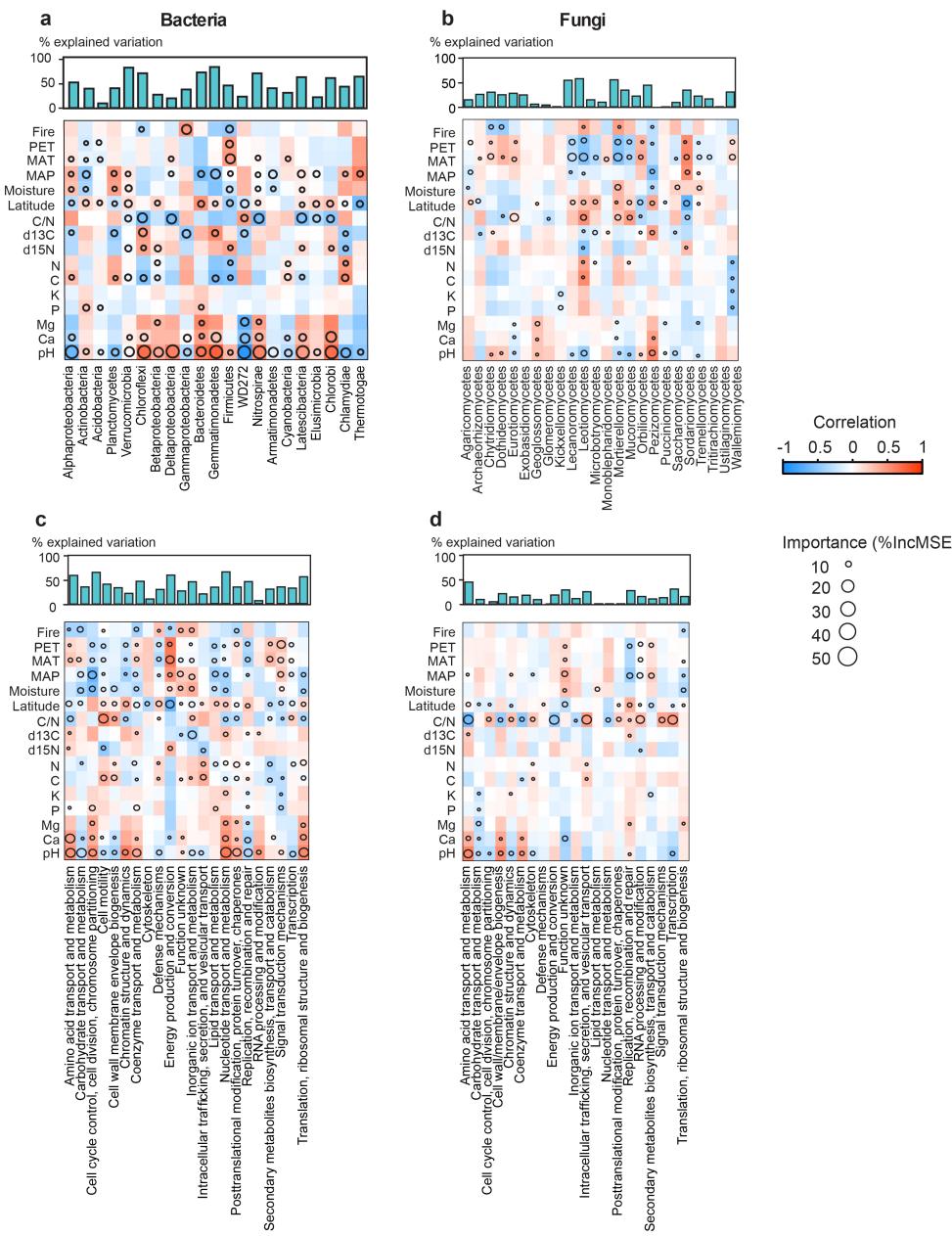
Extended Data Fig. 2 | Contrasting microbial structure and function in major terrestrial biomes. **a–d**, The average total biomass normalized to organic carbon (**a**, $n = 152$ biologically independent samples) as well as richness (**b**), diversity (**c**) and phylogenetic structure including NRI and NTI (**d**) ($n = 188$ biologically independent samples) of fungi and bacteria across samples categorized into major terrestrial biomes, including tropical (moist and dry tropical forests and savannahs), temperate (coniferous and deciduous forests, grasslands and shrublands, and Mediterranean biomes) and boreal–arctic ecosystems ($n = 188$ biologically independent samples).

($n = 189$ biologically independent samples) across biomes: bacterial phyla (classes for Proteobacteria) and archaea (**e**); fungal classes (**f**); functional categories of bacteria (**g**); functional categories of fungi (**h**); bacterial KEGG metabolic pathways (**i**). Biomass was measured on the basis of PLFA analysis. Different letters denote significant differences between groups (shown in the legend) at the 0.05 probability level on the basis of Kruskal–Wallis tests corrected for multiple testing. Additional details for these comparisons are presented in Supplementary Table 14. Taxonomic and gene functional diversity indices were calculated on the basis of inverse Simpson index. Data are mean \pm s.d.



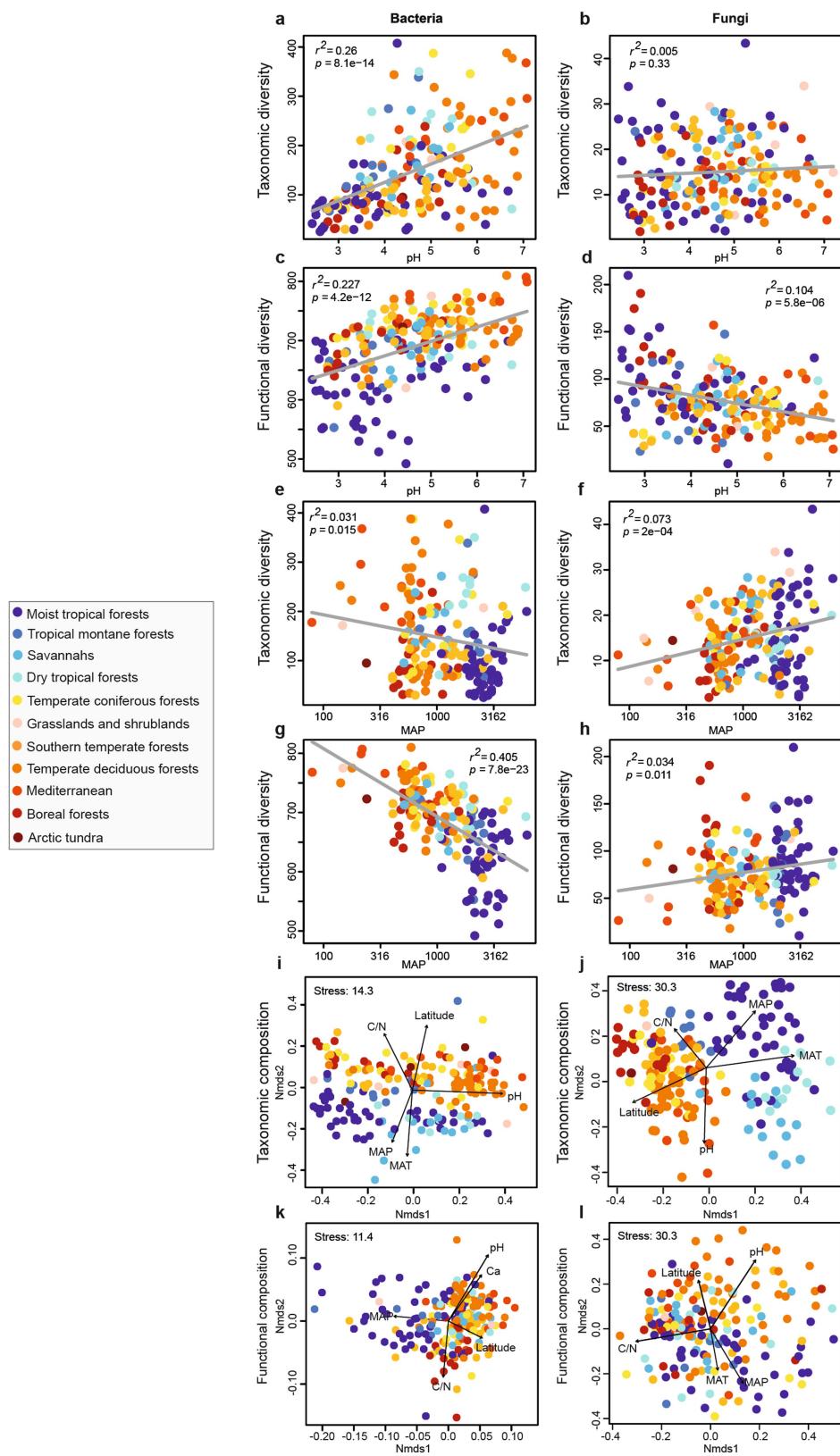
Extended Data Fig. 3 | The significant decrease in the bacterial/fungal biomass ratio with increasing latitude is driven by the joint effect of climate and soil fertility. **a**, The second order polynomial relationship between absolute latitude and the total biomass of bacteria ($n = 152$ biologically independent samples). **b**, The relationship between absolute latitude and the total biomass of fungi. **c**, The relationship between absolute latitude and the bacterial/fungal biomass ratio. **d–f**, The relationship between bacterial/fungal biomass ratio and MAP, MAT and C/N, as the main correlated environmental variables with bacterial/fungal biomass ratio. Linear regression analysis (Pearson's correlation) was used in **b–f** ($n = 152$ biologically independent samples). **g**, Pairwise Spearman's correlation matrix of biotic and abiotic variables in soil. **h**, Direct and

indirect relationships and directionality between variables determined from best-fitting structural equation model. Determination coefficients (R^2) are given for biomass and diversity factors (see Supplementary Table 5 for more details). Goodness of fit: bacteria, $\chi^2 = 15.37$, degrees of freedom = 11, $P = 0.166$; RMSEA = 0.041, $P_{CLOSE} = 0.573$, $n = 189$; fungi, $\chi^2 = 7.74$, degrees of freedom = 12, $P = 0.805$; RMSEA = 0.00, $P_{CLOSE} = 0.970$, $n = 189$). Biomass (nmol g^{-1}) was measured on the basis of PLFA analysis. pH, soil pH representing soil pH and its quadratic term; $\delta^{15}\text{N}$, nitrogen stable isotope signature; $\delta^{13}\text{C}$, carbon stable isotope signature; PET, potential of evapotranspiration; Fire, time from the last fire disturbance; NPP, net primary productivity.



Extended Data Fig. 4 | The environment has a stronger effect on bacterial taxa and functions than on those of fungi. Correlation and best random forest model for major taxonomic (**a, b**; $n = 188$ biologically independent samples) and functional (**c, d**; $n = 189$ biologically independent samples) categories of bacteria (**a, c**) and fungi (**b, d**) in the global soil samples ($n = 189$ biologically independent samples). **a**, Relative abundance of major 16S-based bacterial phyla (class for Proteobacteria).

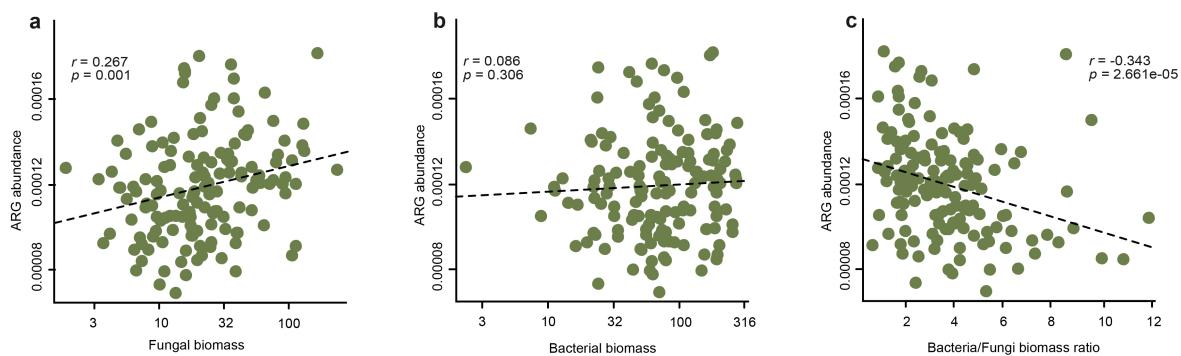
b, Relative abundance of ITS-based fungal classes. **c, d**, Major orthologous gene categories of bacteria (**c**) and fungi (**d**). For variable selection and estimating predictability, the random forest machine-learning algorithm was used. Circle size represents the variable importance (that is, decrease in the prediction accuracy (estimated with out-of-bag cross-validation)) as a result of the permutation of a given variable. Colours represent Spearman correlations. pH, soil pH.



Extended Data Fig. 5 | See next page for caption.

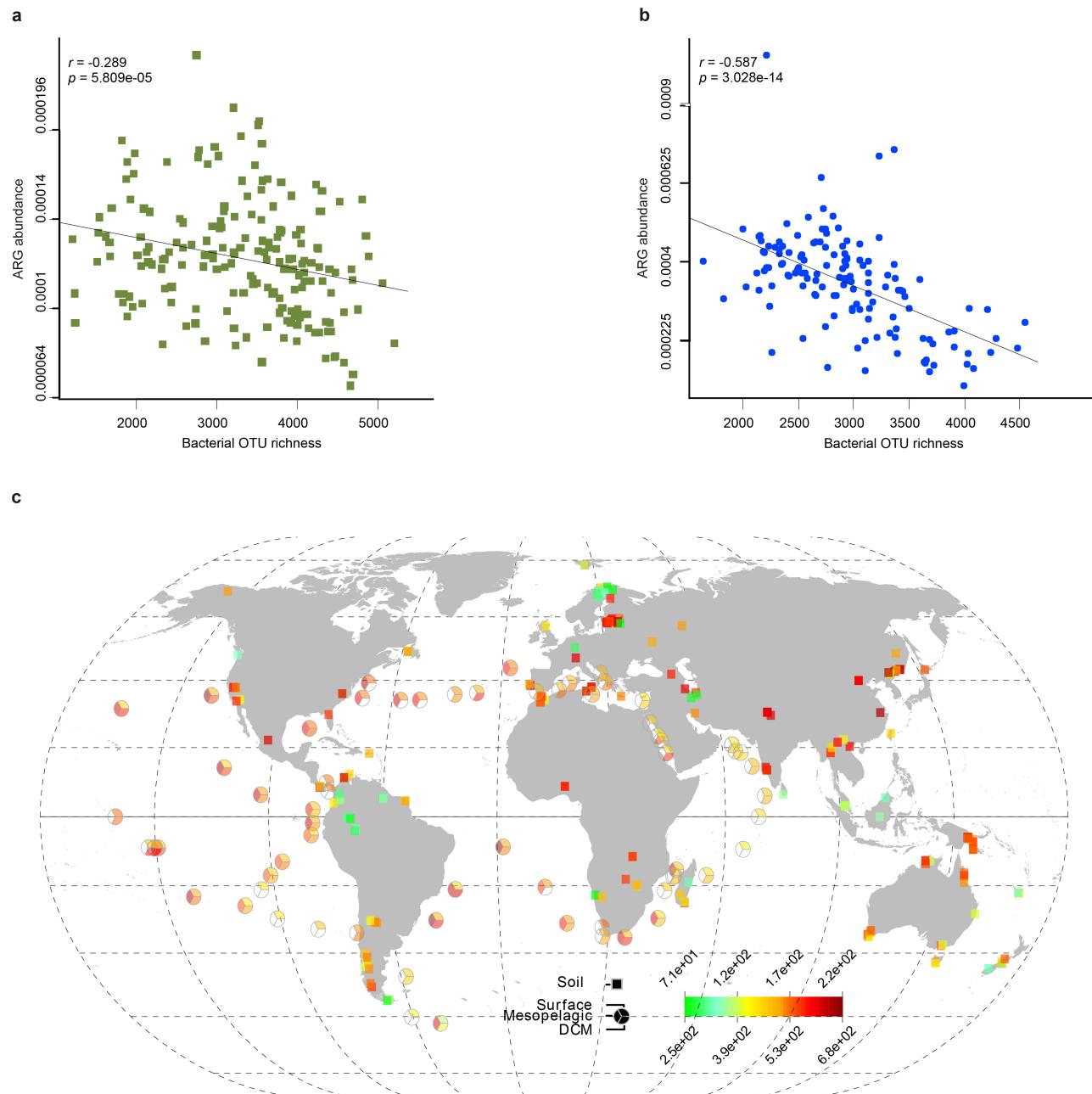
Extended Data Fig. 5 | Niche differentiation between bacteria and fungi is probably related to precipitation and soil pH. Contrasting effect of pH and MAP on bacterial (16S; left column) and fungal (18S; right column) taxonomic ($n=188$ biologically independent samples) and gene functional ($n=189$ biologically independent samples) diversity in the global soil samples. **a, b,** Relationship between soil pH and taxonomic diversity of bacteria and fungi. **c, d,** Relationship between soil pH and gene functional diversity of bacteria and fungi. **e, f,** Relationship between MAP and taxonomic diversity of bacteria and fungi. **g, h,** Relationship between MAP and gene functional diversity of bacteria and fungi. Lines represent regression lines of best fit. The choice of degree of polynomial was

determined by a goodness of fit. Colours denote biomes as indicated in the legend. Taxonomic and gene functional diversity indices were calculated on the basis of inverse Simpson index. **i–l,** NMDS plots of trends in taxonomic (16S and 18S datasets) and gene functional composition (orthologous groups from metagenomes) of bacteria and fungi on the basis of Bray–Curtis dissimilarity. **i,** Taxonomic composition of bacteria (16S). **j,** Taxonomic composition of fungi (18S). **k,** Gene functional composition of bacteria. **l,** Gene functional composition of fungi. **i,** Colours denote biomes as indicated in the legend. Vectors are the prominent environmental drivers fitted onto ordination.



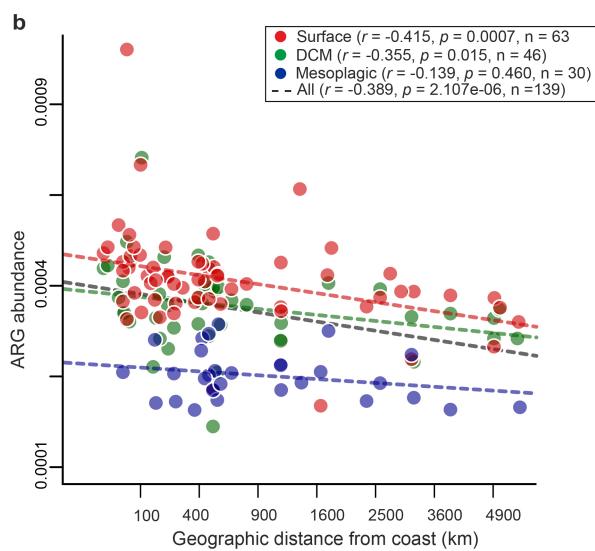
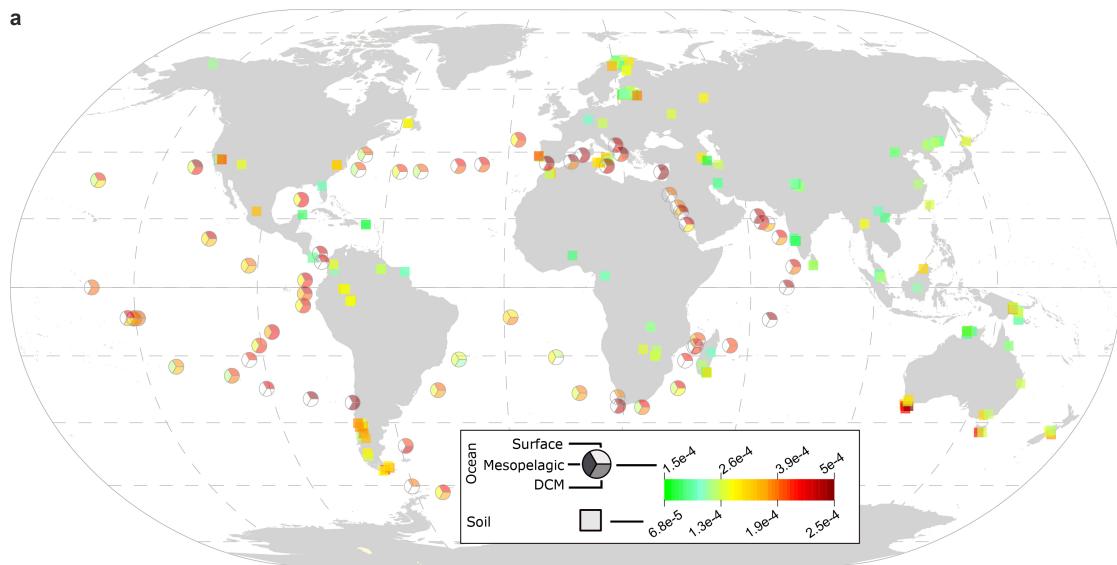
Extended Data Fig. 6 | Fungal biomass is significantly related to the relative abundance of ARGs. **a**, Increase in fungal biomass is related to ARG relative abundance. **b**, Bacterial biomass is unrelated to the relative abundance of ARGs. **c**, ARG relative abundance is inversely correlated

with the bacterial/fungal biomass ratio. Biomass (nmol g^{-1}) was measured on the basis of PLFA analysis. Spearman's correlation was used ($n = 152$ biologically independent samples).



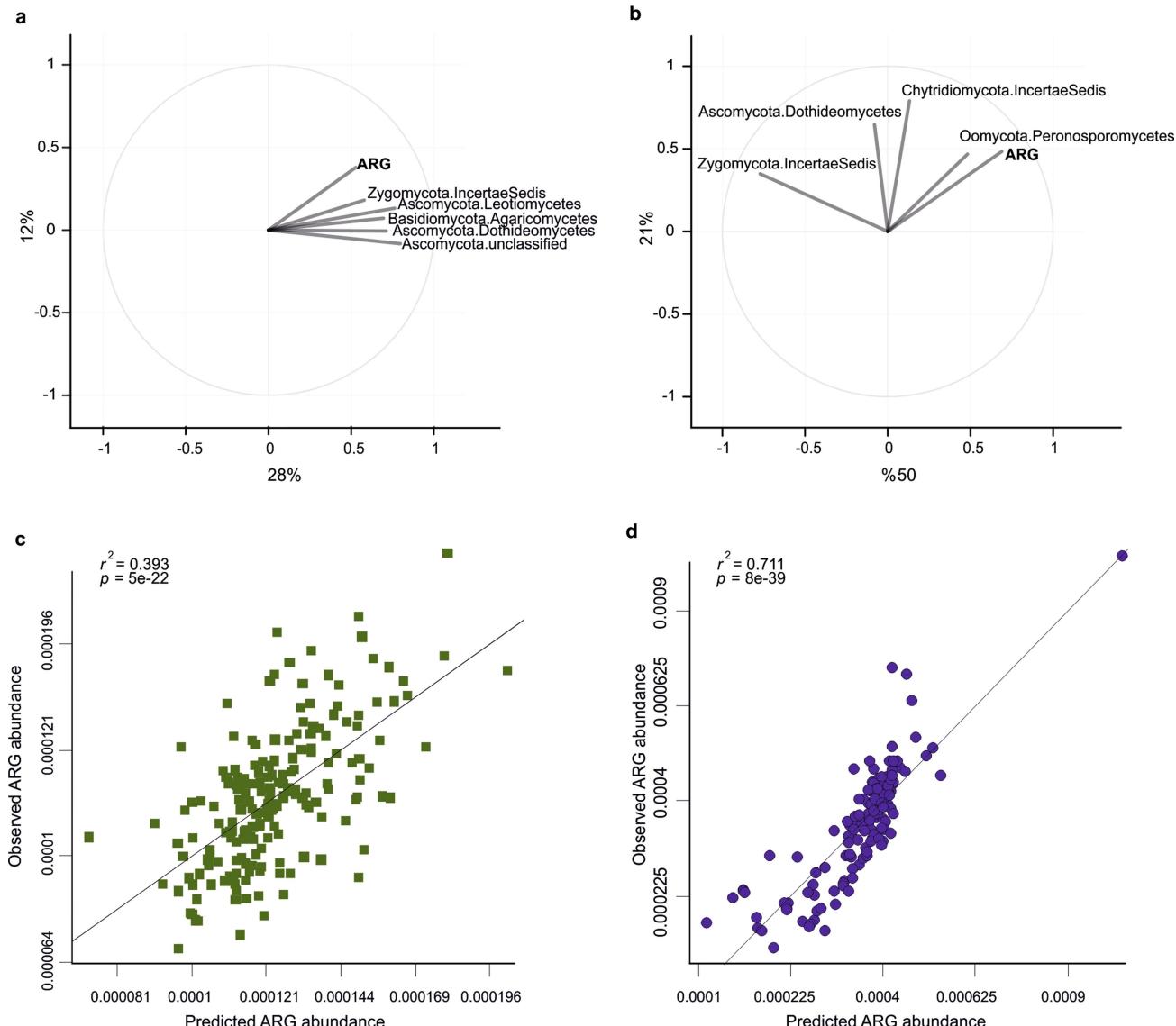
Extended Data Fig. 7 | Topsoil and ocean bacterial phylogenetic diversity is negatively correlated with the abundance of ARGs.
a, b, Spearman's correlation between the relative abundance of ARGs and bacterial phylogenetic diversity (Faith's index) in soil (**a**, $n = 188$ biologically independent samples) and the oceans (**b**, $n = 139$ biologically

independent samples) at the global scale. Similar trends were observed for richness ($r = -0.219, P = 0.007$ and $r = -0.659, P < 10^{-15}$ in soil and ocean, respectively). **c**, Global map of observed bacterial phylogenetic diversity (Faith's index) at the sampled sites. Note that hotspots of bacterial diversity do not correspond to ARG hotspots (See Extended Data Fig. 8).



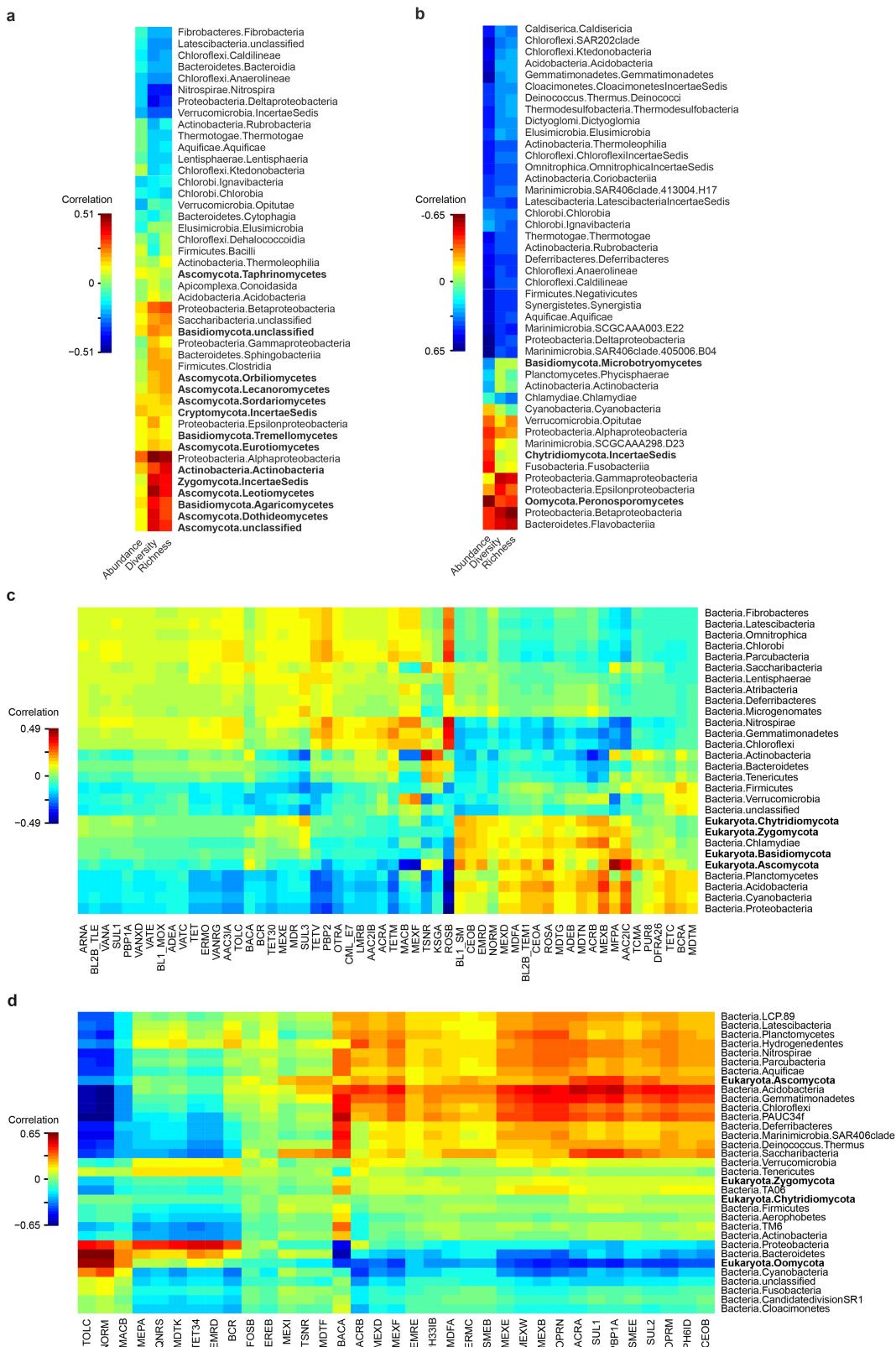
Extended Data Fig. 8 | Relative abundance of ARGs within and between terrestrial and oceanic ecosystems. **a**, Heat map of the observed relative abundance of ARGs at the global scale. Squares and circles correspond to soil and to ocean samples, respectively. ARG abundance is given on three relative scales for these three datasets. **b**, Relative abundance of ARGs in ocean samples (across depths) declines with the distance from land ($n = 139$ biologically independent samples), a pattern that was significant

at two water depths, including surface (red) and deep chlorophyll maximum (DCM; green), but not at mesopelagic (blue). Spearman's correlation statistics for specified comparisons are given in the legends. Dotted lines display Spearman's correlations across the whole dataset and within the three depth categories, respectively. n , number of biologically independent samples.



Extended Data Fig. 9 | Relative abundance of ARGs in both ocean and topsoil samples can be modelled by the relative abundance of fungi and fungus-like protists. **a, b,** Correlation circle indicating the relationships among fungal classes and the relative abundance of ARGs as well as the first two PLS components in soil (**a**) and ocean (**b**). Length and direction of vectors indicate the strength and direction of correlations. Percentages show the variation explained by each PLS component. **c, d,** Linear (Pearson) correlations between observed and modelled ARG relative abundance on the basis of the relative abundance of fungal taxa

in soil (**c**) and ocean (**d**). The two principal axes were chosen on the basis of leave-one-out cross-validation (LOOCV) and explained 40% (LOOCV: $R^2 = 0.381$) and 71% (LOOCV: $r^2 = 0.684$) of the variation of the relative abundance of ARGs in soil and the oceans, respectively. Only taxa significantly associated with the relative abundance of ARGs are shown. Cross-validation and LASSO regression confirmed this result. Soil dataset: $r = 0.619$, RMSE = 10^{-9} , $n = 189$ biologically independent samples; ocean dataset, $r = 0.832$, RMSE = 10^{-9} , $n = 139$ biologically independent samples.



Extended Data Fig. 10 | Fungal classes are among the main taxa associated with the relative abundance, diversity and richness of ARGs in different habitats. **a, b,** Heat map derived from sPLS analysis showing correlation of total relative abundance, richness and diversity of ARGs to that of the main taxonomic classes in soil (**a**) and ocean (**b**) metagenomes (see also the Supplementary Discussion for analogous results in previously published soil (from grasslands, deserts agricultural soils) as well as

human skin and gut samples). For statistical details and significance, see Supplementary Table 8. **c, d,** Heat maps showing correlation of total relative abundance of ARGs to that of the main eukaryotic and prokaryotic taxa in soil (**c**) and the ocean (**d**) on the basis of sPLS regression analysis. All matrices were normalized to library size and Hellinger transformation. Fungal and fungal-like classes are shown in bold text. See Supplementary Table 15 for ARG gene letter abbreviations.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

For DNA sequencing, we used RTA Version 1.18.54 and bcl2fastq v1.8.4 from illumina.

Data analysis

For statistical analysis we used R version 3.3.2. SEM models were built using AMOS software (SPSS, Chicago, IL). The bioinformatic pipelines used to process metagenomes are available at <https://github.com/hildebra/MATAFILER> and <https://github.com/hildebra/Rarefaction>. The pipeline to process amplicon sequences is available at <http://psbweb05.psb.ugent.be/lotus/>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability All metagenomics and metabarcoding sequences have been deposited in the European Bioinformatics Institute-Sequence Read Archive database,

under accession numbers PRJEB24121 (ERP105926): Estonian forest and grassland topsoil samples; PRJEB19856 (ERP021922): 16S metabarcoding data of global soil samples; PRJEB19855 (ERP021921): 18S metabarcoding data of global soil samples; PRJEB18701 (ERP020652): Global analysis of soil microbiomes. The soil gene catalogue nucleotide and amino acid sequences as well as abundance matrix estimates are available at http://vm-lux.embl.de/~hildebra/Soil_gene_cat/. The Tara Oceans data are available at <http://ocean-microbiome.embl.de/companion.html>. All other data that support the findings of this study are available from the corresponding authors upon request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

This study uses high-throughput sequencing methods, soil chemistry and biomass analysis to characterize soil microbiomes across diverse locations and biomes in relation to environmental variables on a global scale.

Research sample

We used shotgun sequencing and metabarcoding of global soil samples to investigate microorganisms present in samples based on their genetic signature, focusing on dominant soil microbial groups such as Bacteria and Fungi. All soil samples used for the main text were collected by us. All ocean samples are from the Tara Oceans expedition and available at EBI under the project identifiers PRJEB402 and PRJEB7988. All public soil samples (only mentioned in Supplement) are publicly available on MG-RAST or EBI, as specified in the Methods.

Sampling strategy

We collected composite soil samples from 1450 sites worldwide. The sampling was conducted broadly across the most influential known environmental gradient (latitude), taking advantage of a global “natural laboratory” to study the impact of climate on diversity across vegetation, biome and soil types and to enable testing the effects of environmental parameters, spatial distance, and biotic interactions in structuring microbial communities. We carefully selected 189 representative sites for different vegetation types separated by spatial distances sufficient to minimize spatial autocorrelation and to cover most areas of the globe. No formal power analysis was undertaken but based on experiences from previous studies of soil and ocean metagenomes, the present sample size was deemed sufficient.

Data collection

Soil samples were metagenomically sequenced at the Estonian Genomics Center (Tartu, Estonia) (18S, metagenomics) and EMBL GeneCore facility (Heidelberg, Germany) (16S). Soil physical parameters were measured at the Estonian University of Life Sciences (Estonia). Soil PFLAs were measured at Lund University (Sweden) by Pål Axel Olsson and Nadejda A. Soudzilovskaia. Soil metagenomic and metataxonomic reads were demultiplexed (if required) and quality controlled with sdm (part of the Lotus and MATAFILER pipeline). For detailed description, please see Methods.

Timing and spatial scale

Samples were collected between 2010 - 2014 (main global soil samples) and 2011-2016 (independent soil samples). No specific month was preselected and collection time depended on the availability of collaborators to collect samples. For detailed description, please see Methods.

Data exclusions

Desert and mangrove samples were only used in metabarcoding analysis and were excluded in comparative analysis of functional and taxonomic patterns. Samples from desert (n=8: G4010,G4034,S357,S359,S411,S414,S418 and S421) and mangrove (n=1: G4023) biomes yielded sufficient DNA for metabarcoding, but not metagenomic sequencing; thus these samples were used for examining global trends of taxonomic diversity but excluded from all comparisons between functional and taxonomic diversity. One sample (S017) contained no 16S sequences; thus, altogether 197 and 189 samples were used for metabarcoding and metagenomics analyses, respectively.

Reproducibility

We used cross-validation where appropriate throughout the manuscript and included additional soil sites after initial analysis to support the main theories derived from the initial dataset. We could reproduce the main findings on independent datasets (i.e. using the independent or the public soil dataset to reproduce main ARG trends).

Randomization

The statistics being mostly correlative, no randomization was necessary. For these involving group of soil sites, these were tested with the appropriate statistical tests.

Blinding

Since this is an exploratory study without specified groups being compared, but rather a correlative analysis, no blinding was possible in data collection. We used cross-validation where appropriate throughout the manuscript and included additional sites after initial analysis to support the main theories derived from the initial dataset.

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions

Field work was carried out in sites that were minimally affected by human disturbance, across various biomes and regions of the world. Temperature and rainfall were obtained for the soil samples from the climate database and did not influence sampling time nor location. For detailed description, please see Methods.

Location	The global soil samples were collected from around the world. The Tara Oceans samples were collected from all major oceans. The Forest samples were collected mostly in Estonia. The exact GPS coordinates for these samples are available Online.
Access and import/export	All samples were collected in full accordance with local and international law, with negligible impact on the study sites, the required permissions have been documented in our previous study (Tedersoo, L. et al. 2014. "Global Diversity and Geography of Soil Fungi." <i>Science</i> 346(6213)).
Disturbance	Given the small size of soil cores (5 cm diam. to 5 cm depth), the sampling procedure (incl. trampling) caused minimal disturbance to the environment.

Reporting for specific materials, systems and methods

Materials & experimental systems

- | | |
|-------------------------------------|-----------------------------|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | Unique biological materials |
| <input checked="" type="checkbox"/> | Antibodies |
| <input checked="" type="checkbox"/> | Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | Palaeontology |
| <input checked="" type="checkbox"/> | Animals and other organisms |
| <input checked="" type="checkbox"/> | Human research participants |

Methods

- | | |
|-------------------------------------|------------------------|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | ChIP-seq |
| <input checked="" type="checkbox"/> | Flow cytometry |
| <input checked="" type="checkbox"/> | MRI-based neuroimaging |