The final typeset version of this paper is available from Nature at the following address:
http://dx.doi.org/10.1038/s41586-018-0386-6

When citing this paper, use the following citation:

Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H, Huerta-Cepas J, Medema MH, Maltz MR, Mundra S, Olsson PA, Pent M, Põlme S, Sunagawa S, Ryberg M, Tedersoo L, Bork P: **Structure and function of the global topsoil microbiome**. *Nature*, 560, 7717, 233–237 (2018). doi: 10.1038/s41586-018-0386-6

Structure and function of the global topsoil microbiome

Mohammad Bahram[1,2,3*]†, Falk Hildebrand[4*], Sofia K. Forslund[4,5,6], Jennifer L. Anderson[2], Nadejda A. Soudzilovskaia[7], Peter M. Bodegom[7], Johan Bengtsson-Palme[8,9], Sten Anslan[1,10], Luis Pedro Coelho[4], Helery Harend[1], Jaime Huerta-Cepas[4,11], Marnix H. Medema[12], Mia R. Maltz[13], Sunil Mundra[14], Pål Axel Olsson[15], Mari Pent[1], Sergei Põlme[1], Shinichi Sunagawa[4,16], Martin Ryberg[2], Leho Tedersoo[17]†, Peer Bork[4, 18,19]†

[1]Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St., 51005 Tartu, Estonia. [2]Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden. [3]Department of Ecology, Swedish University of Agricultural Sciences, Ulls väg 16, 756 51 Uppsala, Sweden. [4]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [5]Current address: Experimental and Clinical Research Center, a cooperation of Charité-Universitätsmedizin Berlin and Max Delbruck Center for Molecular Medicine, 13125 Berlin, Germany. [6]Current address: Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany. [7]Conservation Biology Department, Institute of Environmental Sciences, CML, Leiden University, Einsteinweg 2, 2333 CC Leiden, The Netherlands. [8]Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, SE-413 46 Göteborg, Sweden. [9]Centre for Antibiotic Resistance research (CARe) at University of Gothenburg, Göteborg, Sweden. [10]Braunschweig University of Technology, Zoological Institute, Mendelssohnstr. 4, 38106 Braunschweig, Germany. [11]Current address: Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Campus de Montegancedo-UPM, 28223-Pozuelo de Alarcón (Madrid) Spain. [12]Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, Wageningen, The Netherlands. [13]Center for Conservation Biology, University of California Riverside, USA Germany. [14]Section for Genetics and Evolutionary Biology (Evogene), Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, 0316 Oslo, Norway. [15]Biodiversity Unit, Department of Biology, Ecology building, Lund University, SE-223 62 Lund, Sweden. [16]Current address: Department of Biology, Institute of Microbiology, ETH Zurich, 8092 Zurich, Switzerland. [17]Natural History Museum, University of Tartu, 14A Ravila, 50411 Tartu, Estonia. [18]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany. [19]Department of Bioinformatics, University of Würzburg, 97074 Würzburg, Germany.

*: These authors contributed equally.
†: Corresponding author.

Summary

Soils harbour some of Earth's most diverse microbiomes and are essential for both nutrient cycling and carbon storage. To understand soil functioning, it is necessary to model the global distribution patterns, biotic and environmental associations of the diversity and structure of both bacterial and fungal communities, and their functional gene repertoires[1-

[4]. By leveraging metagenomics and metabarcoding of global topsoil samples (189 sites, 7560 subsamples), we show that bacterial, but not fungal, genetic diversity is highest in temperate habitats and that microbial gene composition varies more strongly with environmental variables than geographic distance. We demonstrate that fungi and bacteria show global niche differentiation associated with contrasting diversity responses to precipitation and soil pH. Furthermore, we provide evidence for strong bacterial-fungal antagonism, inferred from antibiotics resistance genes, in topsoil and ocean habitats, indicating a substantial role of biotic interactions in shaping microbial communities. Our results suggest that both competition and environmental filtering affect bacterial and fungal abundance, composition and their encoded gene functions, implying spatially different relative contributions of these microbes to global nutrient cycling.

Bacteria and fungi dominate terrestrial soil habitats in terms of biodiversity, biomass, and their influence over essential soil processes[5]. Specific roles of microbial communities in biogeochemical processes are reflected by their taxonomic composition, biotic interactions and gene- functional potential[1-4]. While microbial biogeography studies have focused largely on single taxonomic groups, and on how their diversity and composition respond to local abiotic soil factors (e.g. pH[6,7]), both 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 compounds[8], which select for antibiotic resistant (AR) bacteria and effectively increase relative antibiotic resistance gene (ARG) abundance. Here we employed metagenomics and DNA metabarcoding (16S, 18S, 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 world[9] (Extended Data Figure 1a; Supplementary Table 1). Altogether 58,000 topsoil subsamples were collected from 0.25-ha plots from 1450 sites (40 subsamples per site), harbouring homogeneous vegetation that were minimally affected by humans. We minimized biases and shortcomings in sampling[10] as well as technical variation including batch effects[11] 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 gut[12] and ocean[13] 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 (OGs) using the eggNOG database[14]. 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 (miTags[15]), complemented by operational taxonomic units (OTUs) based on clustering 18S rRNA and internal transcribed spacer (ITS)[16] genes for soil fungi and 16S rRNA genes for soil bacteria at 97% similarity threshold (see Methods). 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 determined for each sampling site (see Methods). Archaea were poorly represented in our metabarcoding (<1% of OTUs) and metagenomics data (<1% 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 plants[17], applies to microbial global distribution patterns[10]. We found that contrary to the typical LDG, 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 ocean[13], although the pattern was relatively weak for taxonomic diversity herein (Figure 1a, c; Extended Data Figure 1b,2). The deviation of several bacterial phyla (5 of 20) from the general trends may be explained by responses to edaphic and climate factors weakly related to latitude (Extended Data Figure 1b) or contrasting effects at lower taxonomic levels (Supplementary discussion). In contrast, the LDG does apply to

overall fungal taxonomic diversity, and to 3 of 5 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 (Figure 1b,d; Extended Data Figure 2c). The LDG was negligible for oceanic fungi ($p>0.05$)[13], possibly due to their lower dispersal limitation and paucity of plant associations. While fungal taxonomic diversity decreased poleward, the total fungal biomass (inferred from PLFA markers) and the fungi-to-bacteria biomass ratio increased poleward, partly due to decline of bacterial biomass decreased with latitude (Extended Data Figure 3a-c).

We tested the extent to which deterministic processes (such as competition and environmental filtering; i.e. the niche theory) *versus* neutral processes (dispersal and drift; the neutral theory) explain distributions of fungal and bacterial taxa and functions[18]. In bacteria, environmental variation correlated strongly with taxonomic composition (partial Mantel test accounting for geographic distance between samples: $r_{Env|Geo}=0.729$, $p=0.001$) and moderately with gene functional composition ($r_{Env|Geo}=0.100$, $p=0.001$), 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 suggested by Baas Becking[19] and observed for oceanic prokaryotes[13].

For fungi, both geographic distance and environmental parameters were correlated with taxonomic composition (ITS data: $r_{Geo|Env}=0.307$, $p=0.001$; $r_{Env|Geo}=0.208$, $p=0.001$; 18S data: $r_{Geo|Env}=0.193$, $p=0.001$; $r_{Env|Geo}=0.333$, $p=0.001$). Environmental distance (but not geographic distance) correlated with composition of fungal functional genes ($r_{Env|Geo}=0.197$, $p=0.001$), as also observed for bacteria. The relatively weaker correlation of fungi with environmental variation is consistent with results from local scales[7]. Thus, at both global and local scales, different processes appear to underlie community assembly of fungi and bacteria.

To more specifically investigate the association of environmental parameters with the distribution of taxa and gene functions on a global scale, we used multiple regression modelling (see Methods). We found that bacterial taxonomic diversity, composition, richness and biomass

as well as relative abundance of major bacterial phyla can be explained by soil pH and nutrient concentration, and to a lesser extent by climatic variables (Extended Data Figures 4,5; Supplementary Table 4). Bacterial community composition responded most strongly to soil pH, followed by climatic variables, particularly mean annual precipitation (MAP; Extended Data Figures 4,5). This predominant role of pH agrees with studies from local to continental scales[6], and may be ascribed to the direct effect of pH or confounded variables such as concentration of calcium and other cations[6]. The relative abundance of genes encoding several metabolic and transport pathways were strongly increased with pH (Extended Data Figure 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 Figure 2d), indicating a deeper evolutionary niche specialization in bacteria[20]. Together with global biomass patterns (Extended Data Figure 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. In contrast, phylogenetic overdispersion in temperate bacterial communities, may result from greater competitive pressure[20] or nutrient availability as predicted by the niche theory[21].

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 Figure 5a-h). The steeper LDG in gene functions than in taxa (Figure 1a,c) may thus relate to the stronger association of specific metabolic functions to climate than to local soil conditions. While soil and climate variables exhibited comparable correlations with fungal taxa, soil carbon-to-nitrogen (C/N) ratio was the major predictor for fungal biomass and relative abundance and composition of gene functions (Extended Data Figures 3g,4b,d; Supplementary Table 4). We hypothesize that compared to bacteria, global distribution of fungi is more limited by resource availability due 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 factors[7] and their direct competition. Gene functional diversity of both bacteria and fungi responded to MAP and soil pH, albeit in opposite directions (Extended Data Figure 5c,d,g,h; Supplementary Table 3). This may partly explain the observed inverse pattern of gene functional diversity across the latitudinal gradient, i.e. niche differentiation, between bacteria and fungi (Figure 1; Extended Data Figure 2). While increasing precipitation seems to favour higher fungal diversity, it is associated with higher B/F biomass and abundance ratios (Extended Data Figure 3d,g; Extended Data Figure 5f,h). The increasing proportion of fungi towards higher latitudes may be explained by competitive advantages perhaps due 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 previously[22]. As competition for resources affect the biomass of fungi and bacteria[22,23], we hypothesized that B/F biomass ratio is related to the prevalence of fungi and bacterial AR capacity because of broader activities of fungi than bacteria in utilizing complex carbon substrates[24] as well as increased antibiotic production of fungi in high C/N environments[25]. Consistent with this hypothesis, we found that both fungal biomass and the B/F biomass ratio correlated with ARG relative abundance (Extended Data Figure 6) and that most fungal OG subcategories, particularly those involved in biosynthesis of antibiotic and reactive oxygen species, increased with soil C/N ratio (Supplementary Table 4; Supplementary results). We also found that ARG relative abundance in topsoil is more strongly related to fungal relative abundance ($r=0.435$, $p<10^{-9}$) and B/F abundance ratio ($r=-0.445$, $p<10^{-12}$; Figure 2b) than to bacterial relative abundance ($r=0.232$, $p=0.002$, based on miTags), which is supported by our external validation dataset (fungal relative abundance $r=0.637$, $p<10^{-15}$; B/F abundance ratio $r=-0.621$, $p<10^{-15}$; bacterial relative abundance $r=0.174$, $p=0.036$). Also, topsoil ARG relative abundance was significantly negatively correlated with bacterial phylogenetic diversity and OTU richness based on 16S rRNA gene (Extended Data Figures 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 soil C/N ratio and moisture, rather than pH – the predominant driver of bacterial diversity (Extended Data Figure 3g, Supplementary results) – affect B/F abundance ratio that in turn affects ARG relative abundance at the global scale (Figure 2c). In line increased antibiotics production in high competition environments, 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 CAZyme genes involved in degrading fungal carbohydrates (r=0.501, $p<10^{-12}$). ARG relative abundance was also strongly correlated with C/N ratio in the external validation dataset (r=0.505, $p<10^{-10}$).

While the concomitant increase in AR potential and relative abundance of bacteria (as potential ARG carriers) was expected, the strong correlation of fungal relative abundance with ARG relative abundance 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 (based on miTags), taxa with relatively low genomic ARG content (Supplementary Table 6) were most strongly negatively correlated with ARG relative abundance (Figure 3a). In contrast, ARGs were strongly positively correlated with the relative abundance of Proteobacteria, which have the greatest average number of ARGs per genome[26] among bacteria (Supplementary Table 6), and the fungal phyla Ascomycota and Zygomycota *s.lat*. (including Zoopagomycota and Mucoromycota) in both the global soil and the external validation sets (Figure 3a,b; Extended Data Figure 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 Results). Among bacterial ARGs, the relative abundance of efflux pumps and beta-lactamases, which act specifically on fungal-derived antibiotics, were significantly correlated to the relative abundance of Ascomycota (Extended Data Figure 10a; Supplementary Table **7**). Actinobacteria, encompassing antibiotics-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 likely 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 ARG distribution in 139 water samples from the global Tara Oceans project[13] (see Methods; Supplementary Table 1; Extended Data Figure 8a): the fungi-like stramenopile class Oomycetes (water moulds) and the fungal phylum Chytridiomycota constituted the groups most strongly associated with bacterial ARG relative abundance (Figure 3a,c, Extended Data Figures 9b,d,10b,d). Although there is little direct evidence that oomycetes produce antibiotics, their high antagonistic activity can trigger bacteria[29] and other organisms including fungi[30] to produce antibiotics (Supplementary Discussion). As in topsoil, bacterial phylogenetic diversity was significantly negatively correlated with ARG relative abundance in ocean samples (Extended Data Figure 7b,c). In addition, the ARG relative abundance declined with increasing distance from the nearest coast in ocean samples (Extended Data Figure 8b), which may reflect the effect of a decreasing nutrient gradient along distance from the coast on the pattern of bacteria and fungi abundance and in turn ARG abundance. 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). In particular, global distribution of soil bacteria and fungi was 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 of B/F abundance ratio and ARG abundance, our data suggest that environmental variables that impact B/F abundance ratio may have consequences for microbial interactions and favouring fungi- or bacteria-driven soil nutrient cycling. This unprecedented view of global patterns of microbial distributions implies that global climate change may differentially affect bacterial and fungal 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.

Acknowledgments The authors thank Ingrid Liiv for technical and laboratory assistance; Sebastian Waszak for constructive comments on the manuscript; Yan P. Yuan and Anna Glazek for bioinformatics support; and Alexander Holm Viborg for help in retrieving the CAZY database. We also thank Vladimir Benes, Rajna 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 Axson 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).

Author Contributions Author Contributions M.B., L.T. and P.B. conceived the project. L.T. supervised DNA extraction and sequencing. M.B., F.H., S.F., J.L.A., M.R. and P.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.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.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 significant 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.

Author Information The authors declare no competing financial interests. Correspondence should be addressed to P.B. (bork@embl.de), L.T. (leho.tedersoo@ut.ee) or M.B. (bahram@ut.ee). Requests for materials should be addressed to M.B. (bahram@ut.ee) and F.H. (falk.hildebrand@embl.de).

References

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 (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 (2015).

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**, 1078-+ (2014).

17 Willig, M. R., Kaufman, D. & Stevens, R. Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis. *Annu*. *Rev*. *Ecol*. *Evol*. *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*. (WP Van Stockum & Zoon NV, 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. & J Tranvik, 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. *Cryptogamie: 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*., D529-D534 (2017).
27 Bérdy, J. Thoughts and facts about antibiotics: where we are now and where we are heading. *J*. *Antibiot*. **65**, 385-395 (2012).
28 Andersen, N. R. & Rasmussen, P. The constitution of clerocidin 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-saprolegniasis compound isolated from the deep-sea fungus, Penicillium coralligerum YK-247. *J*. *Antibiot*. **70**, 911 (2017).

Figure Legends

Figure 1 | Fungal and bacterial diversity exhibit contrasting patterns across the latitudinal gradient. Latitudinal distributions of bacterial (left columns) and fungal (right columns) taxonomic (**a** and **b;** n=188 biologically independent samples) and gene functional (**c** and **d**; n=189 biologically independent samples) diversity in the global soil samples. The order of polynomial fit was chosen based on the corrected Akaike Information Criterion (AICc; see Methods) of 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= $p<10^{-4}$). 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 choice of index, see Extended Data Figure 2b, c). The latitudinal distribution of the high-level biome (tropical, temperate and boreal-arctic) is given at the top of **a**) and **b**).

Figure 2 | Global relative abundance of antibiotic resistance genes (ARGs) can be explained by a combination of biotic and abiotic factors. **a**, Pairwise Spearman correlation matrix of main biotic and abiotic determinants of ARG relative abundance. **b,** B/F abundance ratio significantly correlated with ARG relative abundance on a global scale. **c,** Structural equation modelling (SEM) of ARG relative abundance of soil (green) and ocean (blue) datasets (explaining 44% and 51% of variation, respectively; Supplementary Table 5). The goodness of fit was acceptable (Soil: RMSEA=0.00, PCLOSE=0.989, n=189 biologically independent samples; Ocean: RMSEA=0.059, PCLOSE=0.302, n=139 biologically independent samples). Abbreviations: C/N, carbon to nitrogen ratio; N, nitrates; Bacteria/Fungi (B/F), the ratio of bacterial to fungal abundance/biomass; Bacterial richness, bacterial OTU (>97% similarity) richness based on metabarcoding dataset; Abundance, relative abundance of miTags determined as fungi or bacteria; Biomass (nmol/g), absolute biomass based on PLFA analysis; MAP: mean annual precipitation; MAT: mean annual temperature; n.a.: not applicable; n.s.: not significant ($p>0.05$, $q>0.1$).

Figure 3 | Fungi are the main determinants of antibiotic resistance gene (ARG) relative abundance in soils and oceans. a, The association between ARG relative abundance and major bacterial and fungal (incl. fungal-like protist) phyla in metagenomic samples from soil and ocean. Outer circle colour corresponds to the Pearson correlation coefficient. Circle fill colour

corresponds to significance after adjustment for multiple testing (q-value), as indicated in the legend. **b-c**, Relationships (non-parametric correlations) between the relative abundances of 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 analysis 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.

METHODS

Soil sample preparation

Composite soil samples from 1450 sites worldwide were collected using highly standardized protocols[16]. The sampling was conducted broadly across the most influential known environmental gradient – 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 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 sufficient to minimize spatial autocorrelation and to cover most areas of the globe. Total DNA was extracted from 2.0 g of soil from each sample using the PowerMax Soil DNA Isolation kit (MoBio, Carlsbad, CA, USA). 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 Inc., San Diego, CA, USA) 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 806RB[32] (although this pair may discriminate against certain groups of Archaea and Bacteria such as Crenarchaeota/Thaumarchaeota (and SAR11, see ref. [33]) and eukaryote primers 1389f and 1510r[34] were used. While the resolution of 16s rRNA sequencing is limited to genus (and higher) -level assignments, 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 barcode[16]. DNA samples were amplified using the following PCR conditions: 95 ˚C for 15 min, followed by 30 cycles of 95 ˚C for 30 s,

50 ˚C 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 H_2O, 5 µl 5× HOT FIREPol Blend MasterMix (Solis Biodyne, Tartu, Estonia), 0.5 µl each primer (200nM) 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 and 25 cycles, respectively. The amplicons were purified (FavorPrep™ Gel/PCR Purification Kit; Favorgen), 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 based on 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}, K, Ca and Mg; the content of ^{12}C, ^{13}C, ^{14}N and ^{15}N were determined using an elemental analyzer (Eurovector, Milan, Italy) coupled with an isotope ratio mass spectrometer[55].

To calculate the absolute abundance of bacteria and fungi using an independent approach, bacterial and fungal biomass were estimated from Phospholipid Fatty Acids (PLFAs)[35] in nmol/g as follows. Lipids were extracted from 2 g freeze dried soil in a one-phase solution of chloroform, methanol and citrate buffer[36]. 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 column[36]. 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 derivatised to fatty acid methyl esters (FAMEs). The FAMEs from neutral (NLFAs) and phospholipids (PLFAs) were dried, using speed vac, and then dissolved in hexane before analysis on a gas-chromatograph as described by ref.[37]. Fungal biomass was estimated as the concentration of PLFA 18:2ω6,9 and bacterial biomass from the sum of nine PLFAs (i15:0, i16:0, i17:0, a15:0, a17:0, cy17:0, cy19:0, 10Me17:0 and 10Me18:0)[36]. The nomenclature of fatty acids follows Frostegård et al.[37].

Acquisition of metadata from public databases

Climate data including monthly temperature and precipitation were obtained from the WorldClim database (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 (www.sage.wisc.edu/atlas/maps.php). Samples were categorized into 11 biomes[9], 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, savannas 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 based on inquiry from local authorities or collaborators and evidence from the field.

Metagenome analysis

Most soil microbes 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 pathways[38], but we can safely assume a strong correspondence between gene functional potential and the resulting ecosystem functioning[39] or enzyme activities[40].

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 ≥0.01[41], 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 software[42]. 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 FLASH[43]. The merged and unmerged reads were mapped against functional reference sequence databases (see below) using DIAMOND 0.8.10 in blastx mode44 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 percent identity to the subject sequence. Finally, aligned reads were filtered to those, having an alignment %identity >50% and matching with an e-value <1e-9 (see below for parameter choice).

The functional databases to which metagenomic reads were mapped included gene categories related to ROS sources (peroxidases genes databases[45,46], KEGG[47] (Kyoto Encyclopedia of Genes and Genomes) and CAZyme genes (www.cazy.org, accessed 22.11.2015)[48]. To facilitate interpretation of the results, the relative abundance of CAZyme genes were summed based on the substrates for each gene family. Substrate utilization information for CAZyme families was obtained from ref. [49,50] as well as CAZypedia (http://www.cazypedia.org/index.php?title=Carbohydrate-binding_modules&oldid=9411). Based on the KEGG Ortholog (KO) abundance matrices we calculated SEED functional module abundances. For functional annotations of metagenomic reads, we used *in silico* annotation based on a curated database of the orthologous gene family resource eggNOG 4.5[14].

For all databases that included taxonomic information (eggNOG, KEGG, CAZy), reads were mapped competitively against all kingdoms and assigned into prokaryotic and eukaryotic groups, based on the best bit score in the alignment and the taxonomic annotation provided with the

database at kingdom level. All functional abundance matrices were normalized by the total number of reads used for mapping in the statistical analysis, unless mentioned otherwise (e.g. rarefied in the case of diversity analysis, see below). 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 in using relative abundance of genes, our analysis shows, which potential functions are relatively more important. Without any normalisation, such analyses cannot be performed. It is currently difficult to test the absolute numbers, due to limitations to reliably quantify soil DNA resulting from differences in extraction efficiency and level of degradation.

To identify ARGs in our metagenome samples, the merged and unmerged reads were mapped to a homology expansion (see ref. [51]) of the Antibiotic Resistance gene Data Base (ARDB). Only hits surpassing the minimum sequence identity values as listed in the ARDB for each family were taken further into account. While there exist newer ARG databases, only the ARDB presently have 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 also for single reads, not merely for full-length genes. Thus, the cut-offs curated for e.g. ResFams[52] or CARD[53] are inappropriate, since they are defined in the length-dependent bit score space. The ARDB cut-offs, however, are defined as sequence identities, thus in principle applicable also to shorter than full-length sequences. Because of these technical limitations, we used a soil gene catalogue to determine CARD based ARG abundance matrices (see further on).

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

miTag taxa abundance estimation

We used a miTag approach[15] to determine bacterial and fungal community composition from metagenome sequence data. First, SortMeRNA[54] was used to extract and blast search rRNA genes against the SILVA LSU/SSU database. Reads approximately matching these databases with e-values $<10^{-1}$ were further filtered with custom Perl and C++ scripts, using FLASH to attempt merging all matched read pairs. In case read pairs could not be merged, as happens if the overlap between them 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 Silva LSU/SSU databases, lambda[55] was used. Using the lowest common ancestor (LCA) algorithm adapted from LotuS (version 1.462)[42], we determined the identity of filtered reads based on lambda hits. This included a filtering step, where 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 by 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/classes. The abundance of

582 miTag sequences matching bacteria and fungi was used to determine B/F abundance ratio. While
583 LSU/SSU assessments refer to number of fungal cells rather than number of discrete
584 multicellular fungi, since this can apply to all samples equally, it is not systematically biased for
585 comparing the trends of bacterial to fungal abundance across samples.
586
587 **External metagenomic datasets**
588 We validate and compare the global trends with those on a smaller scale, we used a regional
589 scale dataset of 145 topsoil generated and processed using the same protocol as our global
590 dataset (Supplementary Table 1).
591
592 In addition, to compare patterns of ARG diversity in soils and oceans on a global scale, we re-
593 analysed the metagenomics datasets of the Tara Oceans[13], including all size fractions
594 (Supplementary Table 1). After quality filtering, 41,790,928,650 out of 43,076,016,494 reads
595 were retained from the Tara Oceans dataset.
596
597 The quality-filtered reads from all datasets were mapped to the corresponding databases using
598 Diamond, with the exception that no merging of read pairs was attempted, because the chances
599 of finding overlapping reads were too low (with a read length of 100 bp and insert size of 300 bp
600 (Tara Oceans). Sequences for SSU/LSU miTags were extracted from these metagenomics
601 datasets as described above. ARG abundance matrices were also obtained from the Tara Oceans
602 project based on the published gene catalogues annotated using a similar approach as in the
603 current study.
604
605 **Gene catalogue construction**
606 To create a gene catalogue, we first searched for complete reference genes that matched to read
607 pairs in our collection using bowtie2[56] with the options "--no-unal --end-to-end". The resulting
608 bam files were sorted and indexed using samtools 1.3.1[57] and the
609 jgi_summarize_bam_contig_depths provided with MetaBat[58] was used to create a depth profile
610 of genes from the reference databases that were covered with ≥95% nucleotide identity. This cut-
611 off is commonly used in constructing gene catalogues[13,59] and chosen to delineate genes
612 belonging to the same species. Using the coverage information, we extracted all genes that had at
613 least 200bp with ≥1× coverage by reads from our topsoil metagenomes. The reference databases
614 included an ocean microbial gene catalogue[13], a gut microbial gene catalogue[12], as well as all
615 genes extracted from 25,038 published bacterial genomes[26]. Altogether 273,723 and 2,376 and
616 8,642 genes from proGenomes, IGC and Tara database, respectively, could be matched to soil
617 reads and were used in the gene catalogue.
618
619 The majority of genes in our catalogue were assembled from the topsoil samples presented here.
620 To reduce the likelihood of chimeric reads, each sample was assembled separately using Spades
621 3.7-0 (development version obtained from the authors)[60] in metagenomic mode with the
622 parameters "--only-assembler -m 500 --meta -k 21,33,67,111,127". Only sdm[42] filtered paired
623 reads were used in the assembly, with the same read filtering parameters as described above.
624 Resulting assemblies had an average N50 of 469 bases (total of all assemblies 21,538 MBp). The
625 low N50 reflects difficulties in the assembly of soil metagenomes, most likely reflecting the vast
626 microbial genetic diversity of these ecosystems. We further *de novo* assembled reads from two
627 other deep sequencing soil[61] and sediment studies[62], 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). 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/samples that were assembled as described above. Additionally, we integrated gene predictions from soil metagenomes downloaded from MG-RAST[63] (Supplementary Table 1). Assembly was not attempted for these samples due to the absence of paired end reads, and relatively low read depth; rather, only long reads or assemblies directly uploaded to MG-RAST with ≥400bp 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 2.6.1[64] 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 DIAMOND[44]. 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 which are 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 on the gene catalogue with bwa[65], requiring >45 nt overlap and >95% identity. The average mapping rate was 26.2 ± 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, due to the low overall mapping rate. Further, using minimap2[66] to find genes at 95% similarity threshold, we compared the soil gene catalogue with the Tara Oceans gene catalogue[13], human gut gene catalogue[12] and the proGenomes prokaryotic database[26]. The gene catalogue nucleotide and amino acid sequences and abundance matrix estimates from rtk[67] have been deposited at http://vm-lux.embl.de/~hildebra/Soil_gene_cat/.

CARD ARG abundance estimation

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. Based on the 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±7.4% of all metagenomic reads could be mapped to the gene catalogue).

Processing of metabarcoding sequence data

673 The LotuS pipeline[42] was used for bacterial 16S rRNA amplicon sequence processing. Reads
674 were demultiplexed with modified quality-filtering settings for MiSeq reads, increasing strictness
675 to avoid false positive OTUs. These modified options were the requirement of correctly detected
676 forward 16S primer, trimming of reads after an accumulated error of 1 and rejecting reads below
677 28 average quality or, exceeding an estimated accumulated error >2.5 with a probability of
678 ≥0.01[41]. Further, we required each unique read (reads preclustered at 100% identity) to be
679 present 8 or more times in at least one sample, 4 or more times in at least two samples, or three
680 or more times in at least three samples. In total 27,883,607 read pairs were quality-filtered and
681 clustered with uparse[68] at 97% identity. Chimeric OTUs were detected and removed based on
682 both reference-based and *de novo* chimera checking algorithms, using the RDP reference
683 database (http://drive5.com/uchime/rdp_gold.fa) in uchime[68], resulting in 13,070,436 high-
684 quality read pairs to generate and estimate the abundance of bacterial OTUs The seed sequence
685 for each OTU cluster was selected from all read pairs assigned to that OTU, selecting the read
686 pair with the highest overall quality and closest to the OTU centroid. Selected OTU seed read
687 pairs were merged with FLASH[43] and a taxonomic identity was assigned to each OTU by
688 aligning full-length sequences with lambda[55] to the SILVA v123 database[69] and the LotuS least
689 common ancestor (LCA) algorithm. This was performed using the following LotuS command
690 line options: "-p miSeq -derepMin 8:1,4:2,3:3 –simBasedTaxo 2 –refDB SLV -thr 8". OTU
691 abundances per sample were summed to class and phylum level per sample, according to their
692 taxonomic classification, to obtain taxa abundance matrices. However, the choice of clustering
693 method (e.g. Swarm) and identity threshold had little effect on retrieved OTU richness
694 (comparison with 99% threshold: $r=0.977$, $p<10^{-15}$; comparison with Swarm clustering: $r=0.979$
695 $p<10^{-15}$).
696
697 For eukaryotic 18S rRNA genes, we used the same options in LotuS, except that reads were
698 rejected if they did not occur at least six times each in a minimum of two samples or at least four
699 times each in a minimum of three samples. This was done to account for lower sequencing depth
700 in 18S rRNA compared to 16S rRNA dataset. Further, the database to annotate fungal taxonomy
701 was extended to include general annotations of SILVA and information from unicellular
702 eukaryotes (PR2 database[70]). Of 7,462,813 reads, 2,890,093 passed quality filtering. The fungal
703 ITS metabarcoding dataset[16] was downloaded and used in addition to 18S data in specific
704 analyses, such as finding associated fungal OTUs with ARG relative abundance. The resulting
705 taxon abundance matrix was further filtered to remove sequences of chloroplast origin for all
706 three metabarcoding experiments.
707
708 Full-length sequences representing OTUs were aligned using the SILVA reference alignment as
709 a template in mothur[71]. A phylogenetic tree was constructed using FastTree2[72] with the
710 maximum-likelihood method using default settings. This program uses the Jukes-Cantor models
711 to correct for multiple substitutions.
712
713 **Parametrization and validation of metagenomics approach**
714 Although we used state-of-art molecular approaches, there are several potential limitations
715 regarding our analyses related to the used technologies. All metagenomics and amplicon-based
716 analysis are affected by taxonomic biases in sequence databases, while (PCR-free) miTag as well
717 as amplicon sequencing are biased due to differential ribosomal gene copy number across
718 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 MetaPhlAn2[73] using standard parameters: while we observed that most fungal phyla are present abundantly in our samples, MetaPhlAn2 detected Ascomycota only in 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%). Since these results are clearly deviating from our miTag, 16S, 18S and ITS based analysis, 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 where 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 iMessi[74]. 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 based on the reference genomes at e-value < $1e^{-9}$ ($r=0.95\pm0.01$, $p<0.001$). Further, 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 e^{-9}, alignment length ≥20 amino acids, and similarity≥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 based on mixed fungal and bacterial genomes ($r=0.94\pm0.05$, $p<0.001$).

Estimating fungal antibiotics production

We also specifically screened for fungal gene clusters directly associated with antibiotic activity, based on 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, antibiotics producing, MiBIG protein domains, we downloaded all published non-redundant fungal genomes deposited in JGI (Supplementary Table 13) as well as all non-redundant fungal genes deposited in NCBI. The set of MiBIG, and fungal derived genes was screened with custom HMMs for domains from secondary metabolite production (specifically these were dmat, AMP-binding, Condensation, PKS_KS and Terpene synthesis domains). All identified domains were aligned together with the MiBIG domains using Clustal Omega[75] 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) where 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 before-mentioned sequence identity cut-offs and rejecting domains of reads that were mapping to bacterial NOGs.

Statistical analyses

Data normalization and diversity estimates

All statistical analyses were performed using specific packages in R (version 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 (version 2.2.1)[76]. ARG matrices were normalized by the total number of merged and singleton reads. 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, OG and CAZyme gene richness (function *specnumber*) and diversity (function *diversity*, based on the Inverse Simpson index). The latter measure accounts for both richness and evenness, and it gives more weight to abundant groups compared to Shannon Index. Our results were robust to choice of index, and the various diversity indices highly correlated in the present dataset (e.g. 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 Figure 2b,c). Since evenness and richness were highly correlated in all datasets, we report the results based on diversity index that represent both richness and evenness. The rarefaction process was repeated for calculating taxonomic and gene functional diversity and richness based on the average of 100 rarefied datasets.

Phylogenetic diversity was calculated based on Faith's Phylogenetic Diversity (PD) metric in Picante package of R[77]. 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/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 based on 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 ($\partial^{15}N$) and carbon ($\partial^{13}C$) 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 impact on fungal or bacterial diversity, such as altitude, age of vegetation, plant diversity and community (the first two PCA 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 OG- and OTU-based community composition, variable selection was performed in the *Forward.sel* function of Packfor (version 0.0-8/r109)[78] according to the coefficient of determination (threshold, r^2=0.01). All functional and taxonomic compositional matrices were transformed using Hellinger transformation prior to statistical analysis. Further, Mantel tests and partial Mantel tests were used to test the effects of geographical *vs*. environmental distances on OTU and OG compositional similarity as implemented in vegan. Mantel tests allow testing the correlation of two distance matrices, whereas 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 OG 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 based on 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 (OG) 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 (RF) analysis[79] using the R-package randomForest (version 4.6-10).

To examine latitudinal gradients of diversity at phylum level (Figure 2), the diversity of OTUs assigned to each phylum was calculated based on 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 (version 1.3-2)[80] with 1000 iterations. To model latitudinal gradients and environmental associations of diversity and biomass (Figure 1, Extended Data Figure 3), we compared the goodness of fit estimates between first and second order polynomial models based on 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 employed depending on the importance of the spatial component. The model variance structure (Gaussian, exponential, spherical and linear) was evaluated based on AICc. Following selection of variance structure, variables were combined in a set of models with specified variance structure (i.e. 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 5 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 based on the lowest AIC values. Because of non-linear 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 10-fold 'leave-one-out' cross-validation. For this, we used 1000 randomly sampled 90%-data subsets for model training and predicting the withheld data. To minimize biases due to the partitioning of the data and potential overfitting, the average of 1000 resulting determination coefficients are reported as cross-validated r^2 (r^2cv) 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/classes) and the ITS OTUs most strongly associated with ARG relative abundance, as implemented in the mixOmics (version 5.0-4)[81] package. ARG composition and taxonomic community matrices (miTags classes/phyla and ITS OTUs) were normalized by 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 based on 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 based on 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 (version 2.0-2)[82]. 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, Chicago, IL) by including predictors of the best GLS model. In *a priori* models, all indirect and direct links between variables were established based on 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 based on Chi-square test to evaluate the difference between observed and estimated by model covariance matrices (non-significant value indicates that the model fits the observed data). We also used Root Mean Square Error of Approximation (RMSEA) and *PCLOSE* (p-value for test of close fit) to assess the discrepancy between the observed data and model per degree of freedom, which is less sensitive to sample size compared to chi-square test (RMSEA < 0.08 and PCLOSE > 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 enable to address 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 Winkler2 projection, using the maptools (version 0.8-36) package[83].

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 soil[84], seasonal variation in soils is expected to have a minor impact[85] (in contrast to ocean). In addition, the vast majority of our samples were collected during 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[6,7,86], which indicates the robustness of our results. Nonetheless, validation of the proposed models needs to be performed by other researchers with extended data or an independent dataset, particularly by including samples from under-sampled regions (Extended Data Figure 1a) and from different seasons (to account for seasonality). For example, there were some under-sampled regions in our dataset (e.g. North Asia) lowering precision of our models for those regions. Unfortunately, there are no published global datasets with comparable sampling protocols used that could be directly compared and used for model validation, and we encourage future studies that will make this possible.

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

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.

31 Tedersoo, L. *et al*. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys* **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 surveys. *mSystems* **1**, e00009-00015 (2016).

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 Frostegård, Å., Tunlid, A. & Bååth, E. Use and misuse of PLFA measurements in soils. *Soil Biol*. *Biochem*. **43**, 1621-1625 (2011).

36 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).

37 Frostegård, Å., Tunlid, A. & Bååth, 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).

38 Prosser, J. I. Dispersing misconceptions and identifying opportunities for the use of'omics' in soil microbial ecology. *Nat*. *Rev*. *Microbiol*. **13**, 439-446 (2015).

39 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** (2012).

40 Trivedi, P. *et al*. Microbial regulation of the soil carbon cycle: evidence from gene–enzyme relationships. *ISME J*. **10**, 2593-2604 (2016).

41 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-e40 (2016).

42 Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P. & Raes, J. LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* **2**, 30 (2014).

43 Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).

44 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat*. *Methods* **12**, 59-60 (2015).

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 (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 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

66 Li, H. Minimap2: fast pairwise alignment for long DNA sequences. *arXiv:1708.01492* (2017).

67 Saary, P., Forslund, K., Bork, P. & Hildebrand, F. RTK: efficient rarefaction analysis of large datasets. *Bioinformatics* **33**, 2594-2595 (2017).

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

69 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** (2007).
70 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*., gks1160 (2012).
71 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).
72 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5** (2010).
73 Truong, D. T. *et al*. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat*. *Methods* **12**, 902 (2015).
74 Mende, D. R. *et al*. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* **7**, e31386 (2012).
75 Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol*. *Syst*. *Biol*. **7** (2011).
76 Oksanen, J. *et al*. Vegan: community ecology package. *R Package* (2015).
77 Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol*. *Conserv*. **61** (1992).
78 Dray, S., Blanchet, F. & Legendre, P. packfor: Forward selection with permutation. *R Package* (2013).
79 Breiman, L. Random forests. *Mach*. *learn*. **45**, 5-32 (2001).
80 Suzuki, R. & Shimodaira, H. pvclust: hierarchical clustering with p-values via multiscale bootstrap resampling. *R Package* (2009).
81 Lê Cao, K. *et al*. mixOmics: Omics Data Integration Project. *R Package* (2017).
82 Friedman, J. & Hastie, T. glmnet: lasso and elastic-net regularized generalized linear models. *R Package* (2015).
83 Bivand, R. *et al*. Maptools: Tools for reading and handling spatial objects. *R package* (2015).
84 Carini, P. *et al*. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat*. *Microbiol*. **2**, 16242 (2016).
85 Žifčáková, 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).
86 Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **103**, 626-631 (2006).

Extended Data legends

Extended Data Figure 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 indicated 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 our results. **b**, Scatterplots showing the 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 based on Inverse Simpson Index. Latitude: absolute latitude; MAP: mean annual precipitation; MAT: mean annual temperature; C/N: carbon to nitrogen ratio.

Extended Data Figure 2 | Contrasting microbial structure and function in major terrestrial biomes. a-d, The average total biomass (n=152 biologically independent samples) as well as richness, diversity and relative abundance (n=188 biologically independent samples) of fungi and bacteria across samples categorized into major terrestrial biomes, including tropical (moist and dry tropical forests and savannas), temperate (coniferous and deciduous forests, grasslands and shrublands, and Mediterranean biomes) and boreal-arctic ecosystems: total biomass (**a**); richness (**b**); diversity (**c**); phylogenetic structure including Nearest relative index (NRI) and Nearest taxon index (NTI) (see Methods) (**d**). **e-i**, Relative abundance of major phyla (n=188 biologically independent samples) and functional categories (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 based on phospholipid-derived fatty acids (PLFA) analysis (see Methods). Different letters denote significant differences between groups (shown in the legend) at the 0.05 probability level based on Kruskal–Wallis test corrected for multiple testing. Additional details for these comparisons are presented in Supplementary Table 14. Taxonomic and gene functional diversity indices were calculated based on Inverse Simpson Index. The centre values and error bars represent mean and SD, respectively.

Extended Data Figure 3 | Significant decline of bacterial to fungal biomass ratio with increasing latitude due to the joint effect of climate and soil fertility. a, The second order polynomial relationship of absolute latitude and the total biomass of bacteria (n=152 biologically independent samples). **b**, The relationship of absolute latitude and the total biomass of fungi. **c**, The relationship of absolute latitude and the ratio of bacterial to fungal (B/F) biomass. **d-f**, The relationship of B/F biomass ratio and mean annual precipitation (MAP), mean annual temperature (MAT) and carbon to nitrogen ratio (C/N), as the main correlated environmental variables with B/F biomass ratio. Linear regression analysis (Pearson correlation) was used in **b-f** (n=152 biologically independent samples). **g,** Pairwise Spearman 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 (see Methods): bacteria, Chi square=15.37, df=11, P=0.166; RMSEA=0.041, PCLOSE=0.573, n=189; fungi, Chi square=7.74, df=12, P=0.805; RMSEA=0.00, PCLOSE=0.970, n=189). Biomass (nmol/g) was measured based on phospholipid-derived fatty acids (PLFA) analysis. pH, soil pH representing soil pH and its quadratic term; Ca, calcium; Mg, magnesium; P, phosphorous; K, potassium; C, carbon; N, nitrogen; $d^{15}N$, nitrogen stable isotope signature; $d^{13}C$, carbon stable isotope signature; PET, potential of evapotranspiration; Fire, time from the last fire disturbance; NPP, net primary productivity.

Extended Data Figure 4 | Environment has stronger effect on bacterial taxa and functions than those of fungi. Correlation and best random forest model for major taxonomic (**a** and **b**; n=188 biologically independent samples) and functional (**c** and **d**; n=189 biologically independent samples) categories of bacteria (left column) and fungi (right column) 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 genes (OG) 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, i.e. decrease in the prediction accuracy (estimated with out-of-bag cross-validation) as a result of permutation of a given variable. Colours represent Spearman correlations. pH, soil pH; Ca, calcium; Mg, magnesium; P, phosphorous; K, potassium; C, carbon; N, nitrogen; $d^{15}N$, nitrogen stable isotope signature; $d^{13}C$, carbon stable isotope signature; C/N, carbon to nitrogen ratio; Latitude, absolute latitude; MAP, mean annual precipitation; MAT, mean annual temperature; PET, potential of evapotranspiration; Fire, time from the last fire disturbance.

Extended Data Figure 5 | Niche differentiation between bacteria and fungi is likely related to precipitation and soil pH. Contrasting effect of pH and mean annual precipitation (MAP) on bacterial (16S; left columns) and fungal (18S; right columns) taxonomic (n=188 biologically independent samples) and gene functional (n=189 biologically independent samples) diversity in the global soil samples: **a, b**, Relationship of soil pH and taxonomic diversity of bacteria (**a**) and fungi (**b**); c**, d**, Relationship of soil pH and gene functional diversity of bacteria (**c**) and fungi (**d**); **e, f,** Relationship of MAP and taxonomic diversity of bacteria (**e**) and fungi (**f**); **g, h,** Relationship of MAP and gene functional diversity of bacteria (**g**) and fungi (**h**). Lines represent regression lines of best fit. The choice of degree of polynomial was determined by a goodness of fit (see Methods). Colours denote biomes as indicated in the legend. MAP: mean annual precipitation. Taxonomic and gene functional diversity indices were calculated based on Inverse Simpson Index. **i-l,** Non-metric multidimensional scaling (NMDS) plots of trends in taxonomic (16S and 18S-based datasets) and gene functional composition (OGs from metagenomes) of bacteria (left column) and fungi (right column) based on Bray-Curtis dissimilarity**.** 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 Figure 6 | Fungal biomass is significantly related to the relative abundance of antibiotic resistance genes (ARG). a, Increase in fungal biomass is related to ARG relative abundance. **b**, Bacterial biomass is unrelated to ARG relative abundance. **c**, ARG relative abundance is inversely correlated with Bacteria-to-Fungi biomass ratio. Biomass (nmol/g) was measured based on Phospholipid Fatty Acids (PLFA) analysis (see Methods). Spearman correlation was used (n=152 biologically independent samples).

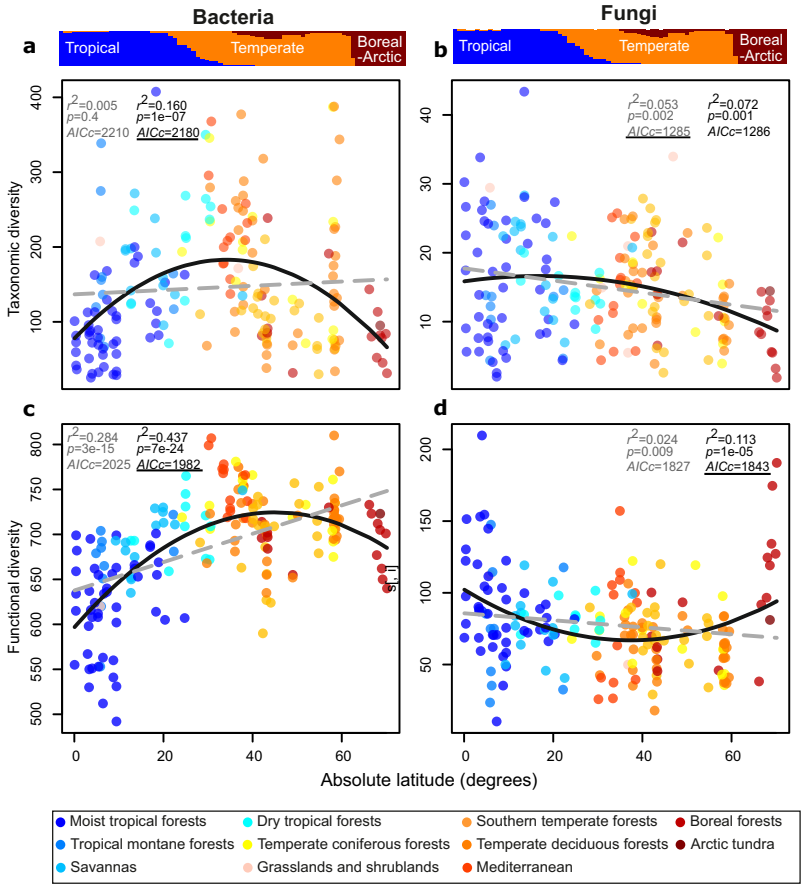
Extended Data Figure 7 | Topsoil and ocean bacterial phylogenetic diversity is negatively correlated with the abundance of antibiotic resistance genes. a, b, Spearman correlations between ARG relative abundance and bacterial phylogenetic diversity (Faith's index; see Methods) in soil (n=188 biologically independent samples). (**a**) and ocean (n=139 biologically independent samples). (**b**) at the global scale. Similar trends were observed for richness (r=-

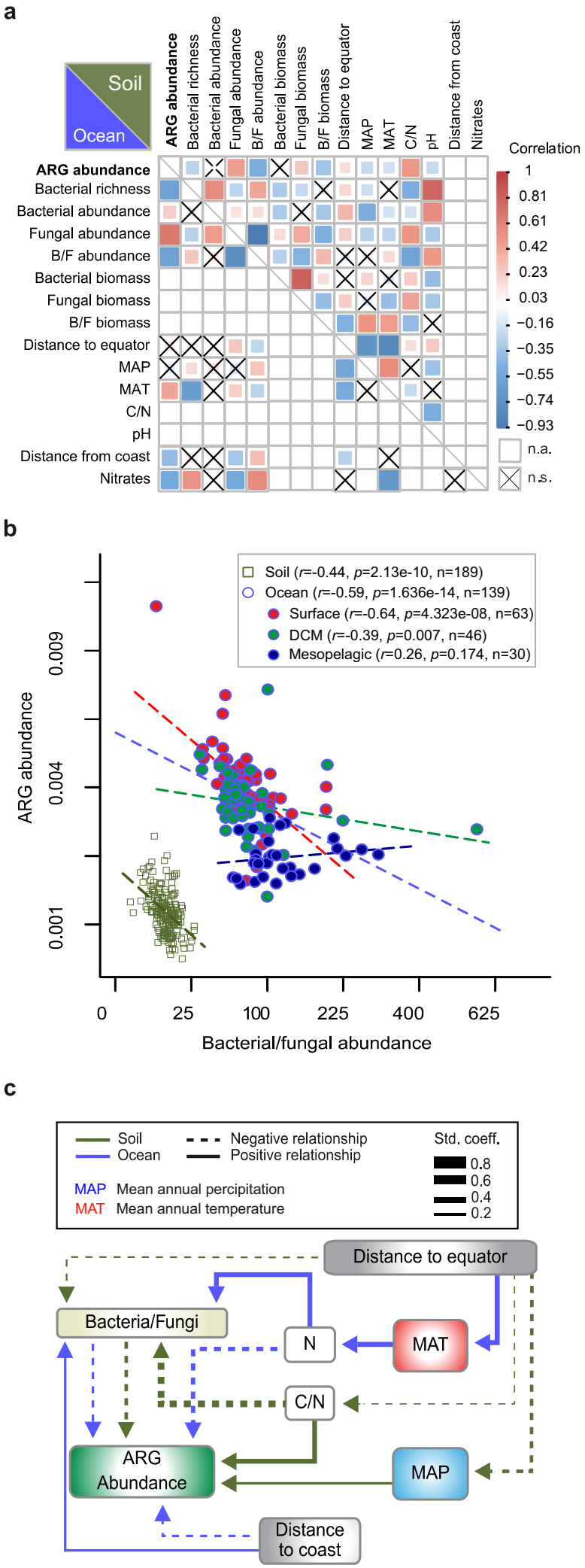
0.219, p=0.007 and r=-0.659, $p<10^{-15}$) in soil and ocean, respectively). **c,** Global map of observed bacterial phylogenetic diversity (Faith's index; see Methods) at the sampled sites. Note that hotspots of bacterial diversity do not correspond to hotspots of ARG relative abundance (See Extended Data Figure 8).

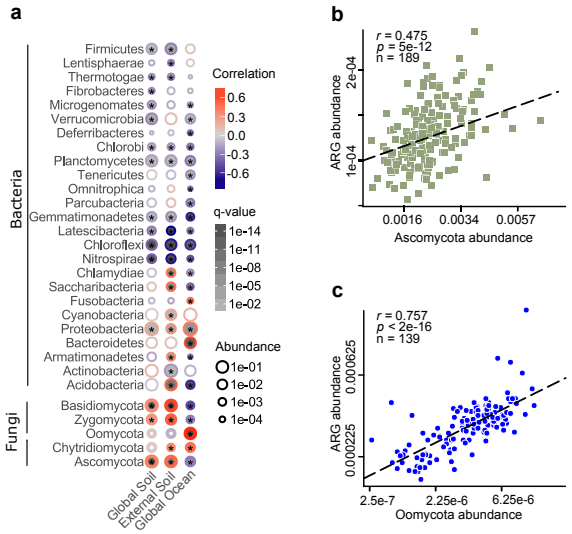
Extended Data Figure 8 | Antibiotic resistance gene (ARG) relative abundance within and between terrestrial and oceanic ecosystems. a, Heat map of observed antibiotic resistance gene (ARG) relative abundance at the global scale. Squares and circles correspond to soil and to ocean samples, respectively. ARG abundance is given on three relative scales for these three datasets. **b,** ARG relative abundance in ocean samples (across depths) declines with distance from land (n=139 biologically independent samples), a pattern which was significant at two water depths, including surface (red) and deep chlorophyll maximum (DCM; green), but not at mesopelagic (blue). Spearman correlation statistics for specified comparisons are given in the legends. Dotted lines display Spearman correlations across the whole dataset and within the three depth categories, respectively. n: number of biologically independent samples.

Extended Data Figure 9 | Antibiotic resistance gene (ARG) relative abundance in both ocean and topsoil samples can be modelled by the relative abundance of fungi and fungi-like protists. a, **b**, Correlation circle indicating the relationships among fungal classes and ARG relative abundance as well as the first two partial least squares regression (PLS) components. 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 based on the relative abundance of fungal taxa in soil (**c**) and ocean (**d**). The two principal axes were chosen based on leave-one-out cross-validation (LOOCV) and explained 42% (LOOCV: R^2=0.401) and 71% (LOOCV: r^2=0.684) of the variation of ARG relative abundance in soil and ocean, respectively. Only taxa significantly associated with ARG relative abundance 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 Figure 10 | Fungal classes are among the main taxa associated with antibiotic resistance gene (ARG) relative abundance, diversity and richness in different habitats. a, b, Heat map derived from sPLS analysis showing correlation of total ARG relative abundance, richness and diversity to that of the main taxonomic classes in soil (**a**) and ocean (**b**) metagenomes (see also the supplementary results 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 ARG relative abundance to that of the main eukaryotic and prokaryotic taxa in soil (**c**) and ocean (**d**) based on sparse partial least square (sPLS) regression analysis. All matrices were normalized by library size and Hellinger transformation. Fungal and fungal-like classes are shown in bold text. See Supplementary Table 15 for ARG gene letter abbreviations.


Bacteria
Fungi
a
b
c
d
Tropical
Temperate
Boreal
-Arctic
Taxonomic diversity
Functional diversity
Absolute latitude (degrees)
r2=0.005
p=0.4
AICc=2210
r2=0.160
p=1e-07
AICc=2180
r2=0.053
p=0.002
AICc=1285
r2=0.072
p=0.001
AICc=1286
r2=0.284
p=3e-15
AICc=2025
r2=0.437
p=7e-24
AICc=1982
r2=0.024
p=0.009
AICc=1827
r2=0.113
p=1e-05
AICc=1843
Moist tropical forests
Dry tropical forests
Southern temperate forests
Boreal forests
Tropical montane forests
Temperate coniferous forests
Temperate deciduous forests
Arctic tundra
Savannas
Grasslands and shrublands
Mediterranean


a
Soil
Ocean
ARG abundance
Bacterial richness
Bacterial abundance
Fungal abundance
B/F abundance
Bacterial biomass
Fungal biomass
B/F biomass
Distance to equator
MAP
MAT
C/N
pH
Distance from coast
Nitrates
Correlation
1
0.81
0.61
0.42
0.23
0.03
−0.16
−0.35
−0.55
−0.74
−0.93
n.a.
n.s.
b
Soil (r=-0.44, p=2.13e-10, n=189)
Ocean (r=-0.59, p=1.636e-14, n=139)
Surface (r=-0.64, p=4.323e-08, n=63)
DCM (r=-0.39, p=0.007, n=46)
Mesopelagic (r=0.26, p=0.174, n=30)
ARG abundance
0.009
0.004
0.001
0
25
100
225
400
625
Bacterial/fungal abundance
c
Soil
Ocean
Negative relationship
Positive relationship
Std. coeff.
0.8
0.6
0.4
0.2
MAP Mean annual percipitation
MAT Mean annual temperature
Distance to equator
Bacteria/Fungi
N
MAT
C/N
ARG Abundance
MAP
Distance to coast


a
Bacteria
Fungi
Firmicutes
Lentisphaerae
Thermotogae
Fibrobacteres
Microgenomates
Verrucomicrobia
Deferribacteres
Chlorobi
Planctomycetes
Tenericutes
Omnitrophica
Parcubacteria
Gemmatimonadetes
Latescibacteria
Chloroflexi
Nitrospirae
Chlamydiae
Saccharibacteria
Fusobacteria
Cyanobacteria
Proteobacteria
Bacteroidetes
Armatimonadetes
Actinobacteria
Acidobacteria
Basidiomycota
Zygomycota
Oomycota
Chytridiomycota
Ascomycota
Global Soil
External Soil
Global Ocean
Correlation
0.6
0.3
0.0
-0.3
-0.6
q-value
1e-14
1e-11
1e-08
1e-05
1e-02
Abundance
1e-01
1e-02
1e-03
1e-04
b
r = 0.475
p = 5e-12
n = 189
ARG abundance
1e-04
2e-04
0.0016
0.0034
0.0057
Ascomycota abundance
c
r = 0.757
p < 2e-16
n = 139
ARG abundance
0.000225
0.000625
2.5e-7
2.25e-6
6.25e-6
Oomycota abundance