Corresponding author(s): Mohammad Bahram, Leho Tedersoo, Peer Bork

Reporting Summary

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

Statistical parameters

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

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	For DNA sequencing, we used RTA Version 1.18.54 and bcl2fastq v1.8.4 from illumina.
Data analysis	For statistical analysis we used R version 3.3.2. SEM models were built using AMOS software (SPSS, Chicago, IL). The bioinformatic pipelines used to process metagenomes are available at https://github.com/hildebra/MATAFILER and https://github.com/hildebra/Rarefaction. The pipeline to process amplicon sequences is available at http://psbweb05.psb.ugent.be/lotus/.

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

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

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

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

Study description	This study uses high-throughput sequencing methods, soil chemistry and biomass analysis to characterize soil microbiomes across diverse locations and biomes in relation to environmental variables on a global scale.
Research sample	We used shotgun sequencing and metabarcoding of global soil samples to investigate microorganisms present in samples based on their genetic signature, focusing on dominant soil microbial groups such as Bacteria and Fungi. All soil samples used for the main text were collected by us. All ocean samples are from the Tara Oceans expedition and available at EBI under the project identifiers PRJEB402 and PRJEB7988. All public soil samples (only mentioned in Supplement) are publicly available on MG-RAST or EBI, as specified in the Methods.
Sampling strategy	We collected composite soil samples from 1450 sites worldwide. The sampling was conducted broadly across the most influential known environmental gradient (latitude), taking advantage of a global "natural laboratory" to study the impact of climate on diversity across vegetation, biome and soil types and to enable testing the effects of environmental parameters, spatial distance, and biotic interactions in structuring microbial communities. We carefully selected 189 representative sites for different vegetation types separated by spatial distances sufficient to minimize spatial autocorrelation and to cover most areas of the globe. No formal power analysis was undertaken but based on experiences from previous studies of soil and ocean metagenomes, the present sample size was deemed sufficient.
Data collection	Soil samples were metagenomically sequenced at the Estonian Genomics Center (Tartu, Estonia) (18S, metagenomics) and EMBL GeneCore facility (Heidelberg, Germany) (16S). Soil physical parameters were measured at the Estonian University of Life Sciences (Estonia). Soil PFLAs were measured at Lund University (Sweden) by Pål Axel Olsson and Nadejda A. Soudzilovskaia. Soil metagenomic and metataxonomic reads were demultiplexed (if required) and quality controlled with sdm (part of the LotuS and MATAFILER pipeline). For detailed description, please see Methods.
Timing and spatial scale	Samples were collected between 2010 - 2014 (main global soil samples) and 2011-2016 (independent soil samples). No specific month was preselected and collection time depended on the availability of collaborators to collect samples. For detailed description, please see Methods.
Data exclusions	Desert and mangrove samples were only used in metabarcoding analysis and were excluded in comparative analysis of functional and taxonomic patterns. Samples from desert (n=8: G4010,G4034,S357,S359,S411,S414,S418 and S421) and mangrove (n=1: G4023) biomes yielded sufficient DNA for metabarcoding, but not metagenomic sequencing; thus these samples were used for examining global trends of taxonomic diversity but excluded from all comparisons between functional and taxonomic diversity. One sample (S017) contained no 16S sequences; thus, altogether 197 and 189 samples were used for metabarcoding and metagenomics analyses, respectively.
Reproducibility	We used cross-validation where appropriate throughout the manuscript and included additional soil sites after initial analysis to support the main theories derived from the initial dataset. We could reproduce the main findings on independent datasets (i.e. using the independent or the public soil dataset to reproduce main ARG trends).
Randomization	The statistics being mostly correlative, no randomization was necessary. For these involving group of soil sites, these were tested with the appropriate statistical tests.
Blinding	Since this is an exploratory study without specified groups being compared, but rather a correlative analysis, no blinding was possible in data collection. We used cross-validation where appropriate throughout the manuscript and included additional sites after initial analysis to support the main theories derived from the initial dataset.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

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

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

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging