nature research

Corresponding author(s): Peer Bork

Last updated by author(s): May 18, 2021

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 our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒	A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Only open source software was used to retrieve the data sets. Custom scripts were provided as Supplemental Material. These are written in Python (3.6.4) using Jug (1.6.6), pandas (0.22.0), and requests (2.14.2).
Data analysis	Only open source software was used for data analysis. Custom algorithms and scripts were provided as Supplemental Material. These are written in Python (3.6.4) using Jug (1.6.6), NumPy (1.12.1), SciPy (0.19.1), and scikit-learn (0.19.0), as well as Haskell (Stackage LTS 10.2). Additional command line tools used were NGLess (0.9.1), eggnog-mapper (2.0.0), and diamond (0.8.36), MetaGeneMark (2.8), RNACode (0.3), mmseqs2 (fd3db05699decf550f428782e1b382a9b7f490e1), ETE3 (3.1.1), FastTree (2.1), ClustalOmega (1.2.4).

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 and reviewers. 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

All data is publicly available. Suppl. Table 1 lists the accession numbers of all the samples.

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

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

Study description	We analysed the distribution of genes and functions by building a global gene catalog, including genes defined at different clustering levels (from species to broad gene families). The presence and abundance of the genes in each metagenomes was quantified by mapping the short reads to the catalog and subsequently, the observed patterns were analysed in the context of existing literature and ecological theory.
Research sample	This study re-analyses publicly available data. In particular, it includes all studies available on the European Nucleotide Archive (ENA) in early 2017 which (1) contained shotgun metagenomic data, (2) with at least 1 million Illumina reads per sample, (3) an average of at least 75bp per read, and (4) at least 100 samples. The initial list of samples was automatically generated by querying ENA and later manually curated to remove mislabeled samples. Additionally, the dataset was manually enriched by including dog gut and soil microbiomes which the authors had access to (even though they were not all publicly available at the time). Metadata was retrieved from ENA or the original publication by manual curation. Genomes were obtained from the ProGenomes database.
Sampling strategy	The sample size was not pre-defined. Rather, all samples which fulfilled the quality criteria listed above were included.
Data collection	The data was retrieved from the European Nucleotide Archive (ENA) using scripts which automatically identified samples which fulfilled the criteria listed above.
Timing and spatial scale	Data was collected without timing or spatial limitations.
Data exclusions	Some datasets are mis-labeled on ENA, thus leading the automated scripts to erroneously include them even though they do not actually fulfill the pre-defined criteria. They were excluded by manual curation. For some analyses, only samples that retained at least 1 million reads after quality control (which may reduce the number of reads) were used as indicated in the methods section.
Reproducibility	Not applicable: the study is a meta-analysis and includes all available data.
Randomization	Not applicable: the study is a meta-analysis and there is no randomized component in the computational methods.
Blinding	In this study, it was not possible to meaningfully blind the researchers during data collection. Note that the data was publicly available and only technical criteria were defined (described above).

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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