nature research

Corresponding author(s): Philip Hugenholtz

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	Calculation of relative evolutionary divergence values was performed with custom software released under GPL v3.0 and made freely available as PhyloRank v0.0.27 at GitHub.
Data analysis	The following software was used: Prodigal v2.6.3, HMMER v3.1b1, FastTree v2.1.7, NINJA v1.2.2, ExaML v3.0.20, IQ-TREE v1.5.5, RaxML v8.1.11, CheckM v1.0.7, BLASTN v2.2.30+, ssu-align v0.1

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 files for the GTDB taxonomy are available at http://gtdb.ecogenomic.org and include: i) flat file with the GTDB taxonomy defined for 94,759 genomes; ii)

Nature Biotechnology: doi:10.1038/nbt.4229

bootstrapped bac120 tree in Newick format spanning the 21,943 dereplicated genomes and annotated with the GTDB taxonomy; iii) FASTA files for each marker gene and the trimmed concatenated alignment; iv) metadata for all genomes including NCBI, SILVA, and Greengenes taxonomies, completeness and contamination estimates, assembly statistics (e.g., N50), and genomic properties (e.g., GC-content, genomes size); v) FASTA file of 16S rRNA gene sequences identified within the 21,943 dereplicated genomes; and vi) an ARB database containing the bac120 tree. The 3,087 MAGs introduced in this study are available under BioProject PRJNA417962 and on the GTDB website.

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	We propose a standardized bacterial taxonomy based on a concatenated protein phylogeny that conservatively removes polyphyletic groups and normalizes taxonomic ranks based on relative evolutionary divergence.
Research sample	Reference genomes from NCBI along with additional metagenome-assembled genomes (MAGs) recovered as part of this study.
Sampling strategy	Genomes were screen for quality using standard estimates of completeness and contamination, and the dereplicated based on an estimate of average nucleotide identity and described species affiliation.
Data collection	Not applicable.
Timing and spatial scale	Not applicable.
Data exclusions	Poor quality genomes were excluded using a predefined set of criteria. Such genomes were excluded as they may negatively impact the quality of the inferred phylogeny.
Reproducibility	The robustness of the approach used to generate the proposed taxonomy was evaluated by varying tree inference software, evolutionary models, marker sets, and genome datasets.
Randomization	Not applicable.
Blinding	Not applicable.

Did the study involve field work? ☐ Yes ☒ No

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

Nature Biotechnology: doi:10.1038/nbt.4229