Nucleic Acids Research, 2023, 1–7
https://doi.org/10.1093/nar/gkad943
Database issue

SPIRE: a Searchable, Planetary-scale mIcrobiome REsource

Thomas S.B. Schmidt [1,*,†], Anthony Fullam [1,†], Pamela Ferretti [1], Askarbek Orakov [1], Oleksandr M. Maistrenko [1], Hans-Joachim Ruscheweyh [2], Ivica Letunic [3], Yiqian Duan [4], Thea Van Rossum [1], Shinichi Sunagawa [2], Daniel R. Mende [5], Robert D. Finn [6], Michael Kuhn [1], Luis Pedro Coelho [4,7] and Peer Bork [1,8,9,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
[2]Institute of Microbiology, Department of Biology and Swiss Institute of Bioinformatics, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland
[3]Biobyte solutions GmbH, Bothestr. 142, 69117 Heidelberg, Germany.
[4]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China
[5]Department of Medical Microbiology, Amsterdam University Medical Centers, Amsterdam, The Netherlands
[6]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, United Kingdom
[7]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, Woolloongabba, Queensland, Australia
[8]Department of Bioinformatics, Biozentrum, University of Würzburg, 97074 Würzburg, Germany
[9]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany

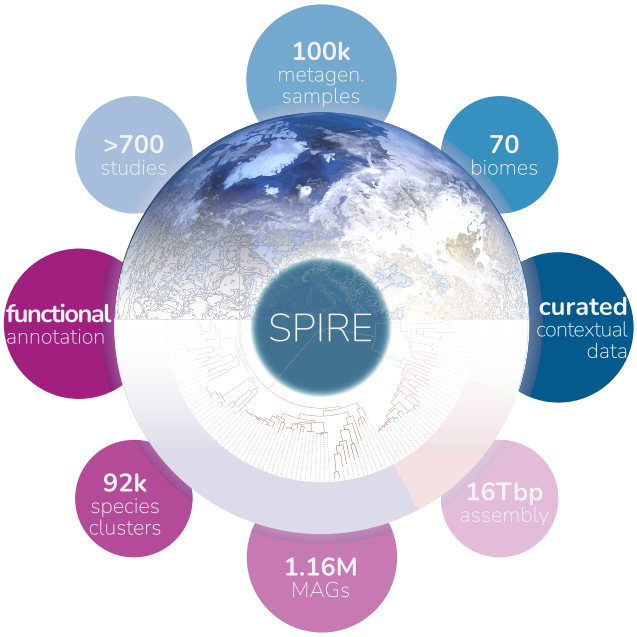
[*]To whom correspondence should be addressed. Tel: +49 6221 387 8526; Email: peer.bork@embl.org
Correspondence may also be addressed to Thomas S.B. Schmidt. Email: sebastian.schmidt@embl.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Present addresses:
Pamela Ferretti, Section of Genetic Medicine, Division of Biological Sciences, University of Chicago, Chicago, USA.
Oleksandr M. Maistrenko, Royal Netherlands Institute for Sea Research (NIOZ), Department of Marine Microbiology & Biogeochemistry, 1797 SZ, 't Horntje (Texel), The Netherlands.

Abstract

Meta'omic data on microbial diversity and function accrue exponentially in public repositories, but derived information is often siloed according to data type, study or sampled microbial environment. Here we present SPIRE, a Searchable Planetary-scale mIcrobiome REsource that integrates various consistently processed metagenome-derived microbial data modalities across habitats, geography and phylogeny. SPIRE encompasses 99 146 metagenomic samples from 739 studies covering a wide array of microbial environments and augmented with manually-curated contextual data. Across a total metagenomic assembly of 16 Tbp, SPIRE comprises 35 billion predicted protein sequences and 1.16 million newly constructed metagenome-assembled genomes (MAGs) of medium or high quality. Beyond mapping to the high-quality genome reference provided by proGenomes3 (http://progenomes.embl.de), these novel MAGs form 92 134 novel species-level clusters, the majority of which are unclassified at species level using current tools. SPIRE enables taxonomic profiling of these species clusters via an updated, custom mOTUs database (https://motu-tool.org/) and includes several layers of functional annotation, as well as crosslinks to several (micro-)biological databases. The resource is accessible, searchable and browsable via http://spire.embl.de.

Received: August 18, 2023. **Revised:** October 1, 2023. **Editorial Decision:** October 10, 2023. **Accepted:** October 11, 2023

Graphical abstract



Introduction

Life on Earth is dominated by microbes: bacteria, archaea and small eukaryotes shape our world by driving biogeochemical cycles across ecosystems (1), they enable macroscopic life as plant and animal symbionts (2), and they represent by far the greatest biodiversity among known life (3). Yet most of this diversity remains biological 'dark matter' (4): although meta'omic techniques enable their study directly from sequencing data, the vast majority of microbes eludes laboratory cultivation and only a small fraction of the functional space encoded by microbial genes has been characterized (5,6). While sampling efforts have increased exponentially and generated petabytes of data in recent years (7), most major microbial habitats remain understudied to the extent that almost every newly sequenced metagenome adds 'novel' species (as inferred from metagenome-assembled genomes, MAGs) and thousands of 'novel' genes of unknown function to the census (8).

The bulk of metagenomic data is generated in individual studies to address specific research questions. Heterogeneity in sample preparation (9), sequencing protocols and bioinformatic processing workflows (10,11) complicate comparisons of findings across studies. Several initiatives have sought to integrate and consolidate datasets by re-processing them using consistent pipelines. For example, QIITA (12), MGnify (7) or the Microbe Atlas Project (https://microbeatlas.org/) host millions of amplicon samples, whereas other projects, such as curatedMetagenomicData (13), GMrepo (14) and the OceanMicrobiomicsDatabase (15), focus on taxonomic and functional profiles of human-associated or ocean metagenomes. Large MAG catalogs for multiple biomes are hosted online as part of the DOE's IMG/M (16) and EBI's MGnify (7) resources. Moreover, the Genome Taxonomy Database (GTDB, 17) has advanced the field by consistently organizing both isolate genomes and quality-filtered MAGs into a common prokaryotic reference tree that guides standardized, phylogeny-informed taxonomies (18–20). The GTDB encompasses 85 205 species-level genome clusters across 181 phyla (as of release r214, April 2023), two thirds of which are represented only by MAGs, while also providing widely used tools for genome quality control (21) and taxonomic classification (22). Overall, existing resources focus on either providing large gene or genome catalogs, on functional and taxonomic profiling, or on harmonizing contextual data given heterogeneous data submission and annotation practices, and are often restricted to individual microbial habitats or cordon data on different habitats off into distinct subsets.

Here we introduce SPIRE, a Searchable, Planetary-scale, Integrated mIcrobiome REsource to study microbial diversity and function at global habitat, geographical and phylogenetic scales. As detailed below, SPIRE version1 encompasses 99 146 consistently processed whole-genome shotgun metagenomic samples from 739 distinct studies, integrated across environments and amended with manually curated contextual data, based on a newly developed lightweight 'microntology' of 92 terms describing microbial habitats and lifestyles. SPIRE combines 1.16 million newly constructed MAGs of medium or high quality (23) with the 907k high-quality reference genomes in proGenomes3 (24), clustered into 133 402 species-level genome clusters, 78 804 of which are unclassifiable at species level using current tools (22). Species clusters are profilable using mOTUs (25) via an updated custom database and pre-computed taxonomic profiles across all 99k metagenomic samples will be released as part of the resource. SPIRE further comprises 35 billion metagenomically called open reading frames (ORFs) with various layers of functional annotation, linked to clusters in the Global Microbial Gene Catalogue (GMGC, 8). SPIRE provides consistent integration of these heterogeneous data modalities and is designed to interoperate with other (micro-)biological resources, such as proGenomes (24, https://progenomes.embl.de), the GMGC (8, https://gmgc.embl.de), eggNOG (26, http://eggnog6.embl.de) and metaMap (https://metamap.biobyte.de/), among others. The resource can be accessed, browsed, and searched via https://spire.embl.de.

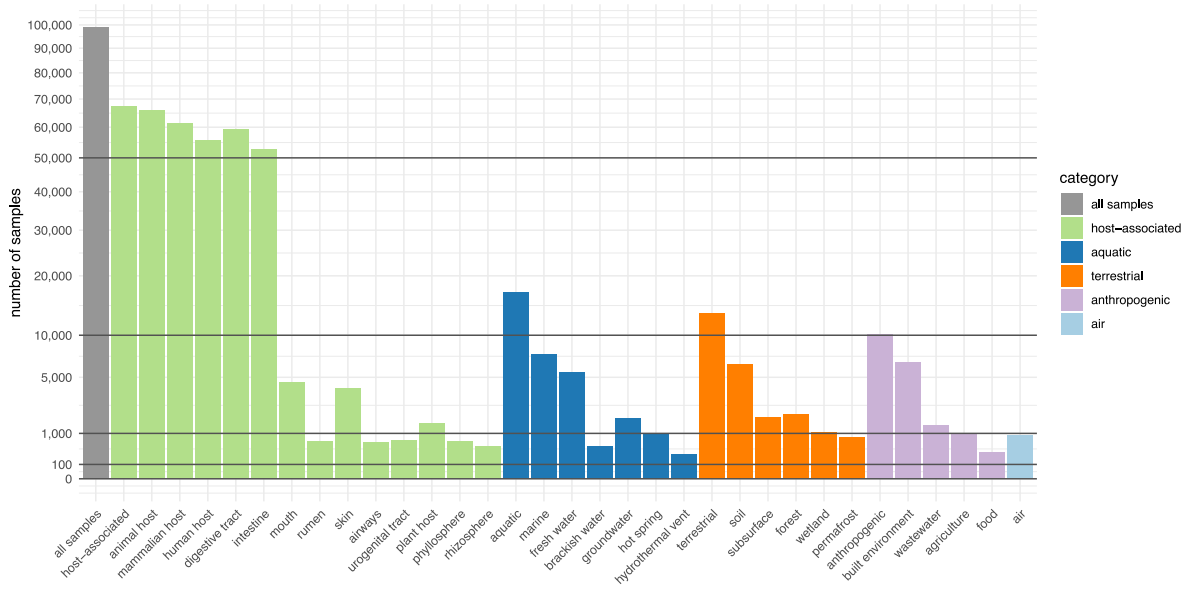


Figure 1. Overview of sampled habitats in SPIRE, as a subset of annotated 'microntology' terms (see table S1). Microntology terms are assigned using a 'multi-tag' system, meaning that individual samples can be annotated with multiple terms of varying granularity and redundantly within a flat hierarchy (e.g. a human fecal metagenome will be annotated as 'host-associated, animal host, mammalian host, human host, digestive tract, intestine', whereas a mangrove-associated sample carries tags from both the 'aquatic' and 'terrestrial' term space, while moreover possibly being annotated as 'host-associated, plant host'). Shown above is the total number of samples annotated to a subset of microntology terms under this system.

Database construction and characteristics

Metagenome collection and dataset curation

The core dataset underlying SPIRE was defined using a semi-automatic process, combining three data sources: (i) samples in the European Nucleotide Archive (ENA) meeting the criteria 'library_source = METAGENOMIC AND library_strategy = WGS AND instrument_platform = ILLUMINA AND base_count >= 10^9 AND average read length >= 100' were selected from all projects where >=20 samples satisfied the above criteria as of Sep 30th 2022; (ii) metagenomic samples available via the JGI's IMG/M resource (27) on Sep 30th 2019 (to comply with JGI data policies and embargo periods); (iii) manually selected 'allowlisted' studies of particular interest (e.g. providing data on exotic environments). For the resulting list, ENA project accessions were manually matched to publications where possible; in case of data submitted by the JGI, where each metagenomic sample is associated with a distinct project accession, 'studies' were defined based on matched publications and as consistent groups based on sample metadata provided via IMG/M.

The metagenomic sample set was further filtered and curated by (i) removing amplicon and isolate genome sequencing datasets erroneously annotated as shotgun metagenomes; (ii) identifying and removing erroneously submitted datasets (e.g. where both mates in 'paired end' data were identical); (iii) identifying and removing duplicates (submitted under distinct project or sample accessions); (iv) removing samples from controlled experimental setups (e.g. laboratory mice, pathogen challenges or defined *in vitro* communities); (v) flagging special cases such as microcosms, paleobiological samples or pre-enriched samples; (vi) resolving misfits with the European Nucleotide Archive (ENA) and Sequence Read Archive (SRA) data model, e.g. if distinct biological samples were erroneously submitted under the same biosample accession, but distinct experiment or run accessions; (vii) identifying and combining technical replicates (distinct experiment accessions) for the same biological sample. For the resulting list, raw sequencing data was downloaded from the ENA.

Following these steps, the final dataset in SPIRE comprises 99 146 metagenomic samples across 739 distinct studies.

Curation of contextual data and overview of sampled environments

Contextual data for each metagenomic sample was sourced (i) via annotation fields in ENA, (ii) via IMG/M metadata tables where applicable and (iii) directly from matched publications. Information was consolidated into common fields (e.g. latitude and longitude data were manually harmonized across different submitted formats). All samples were manually annotated against a newly developed '*microntology*' (see Table S1), a shallow and lightweight ontology of 92 terms to describe microbial habitats and lifestyles, crosslinked to terms in established resources such as the EnvO (28) or UBERON (29) ontologies. SPIRE sample annotation uses a 'multiple tag' system, meaning that each sample is described using a combination of concurrent tags, rather than one specific term in a (deep) hierarchy, allowing an annotation with increased flexibility, yet compatibility to established ontologies. As a result, for example, 68% of the ~100k samples in SPIRE are annotated as 'host-associated' (66.5% as animal-associated, 56% as human-associated, 1.5% as plant-associated); 17% are aquatic samples (including 7.6% marine and 5.5% fresh water); 13.5% are terrestrial (including 6.4% soil samples); 10.3% are from anthropogenic or human-impacted environments (including 6.6% from built environments); see Figure 1 for details. Moreover, data included in SPIRE cover pole-to-pole latitudes, with samples from ~200 countries and territories.

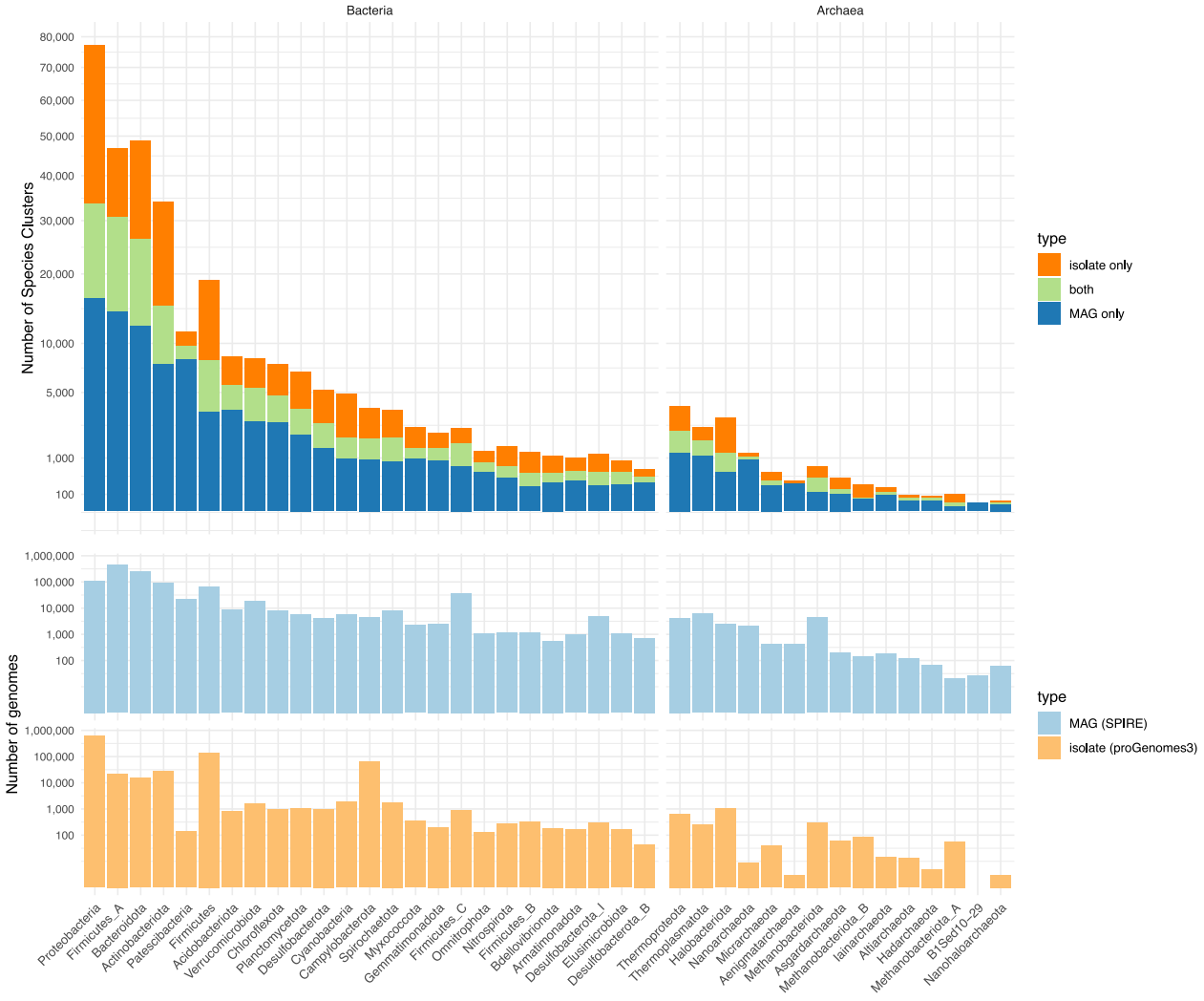


Figure 2. Representation of taxonomic groups covered in SPIRE. Shown are the total number of species clusters (top) and total number of genomes (bottom) for the largest 25 bacterial and largest 15 archaeal phyla represented in SPIRE. Orange hues indicate clusters and genomes of isolates, as downloaded from proGenomes3 (progenomes.embl.de; 'isolate only'). Blue hues indicate clusters and genomes introduced in SPIRE ('MAGs only'). Green indicates species clusters that contain both isolate genomes and MAGs. See Supplementary Table S2 for taxonomic classifications of all species clusters included in SPIRE.

Metagenomic sequence processing

Data processing was implemented in a Nextflow pipeline (30) to enable robustness and reproducibility. Downloaded sequence data was quality-trimmed and filtered using NGless (31) as described previously (8). Reads were assembled into contigs using megahit v1.2.9 (32) with default parameters, separately for each sample, resulting in a total of 210M contigs and a total assembly length of 16 Tbp. From these, a total of 35 billion open reading frames (ORFs) were called using prodigal v2.6.3 (using -p meta) (33) and further processed as described below. Sequences of tRNA genes were called using tRNAscan v2.0 (34) using the 'general' model; rRNA genes were called using Barrnap v0.9 (https://github.com/tseemann/barrnap); putative CRISPR sequences were identified using MinCED (https://github.com/ctSkennerton/minced).

Metagenome-assembled genomes (MAGs)

MAGs were binned from pre-filtered contig sets per sample (length >= 1000 bp) using metaBAT2 (v2.12.1) (35), resulting in a total of 3 023 270 genome bins which were further filtered to 1.16 million bins of 'medium or high quality' (CheckM2-inferred completeness >= 50% and contamination <= 10% (21); passing default clade consistency score filters and <= 5% estimated contamination in GUNC (36); CheckM2 v0.1.3, GUNC v1.0.1). Of these, 73.3% were mapped to the 41 171 species-level clusters of high quality reference genomes in proGenomes3 (24) based on a two-step procedure: (i) extraction of 40 'specI' marker genes using fetchMG (37), followed by MAPseq-mapping (38) of each marker gene with parameters calibrated for high specificity and a consensus call across hits per query MAG; (ii) fastANI-derived (39) Average Nucleotide Identity (ANI) of >= 95% to species representative genomes. The remaining 309 020 unmapped MAGs were clustered into 92 134 species-level groups in a two-step procedure: (i) single linkage preclustering at 90% whole genome ANI using mash (40); (ii) resolution of mash pre-clusters into 95% ANI average linkage clusters using fastANI (39) and fastcluster (41). A full list of species clusters included in SPIRE with consensus taxonomic classifications is included as Supplementary Table S2. In an independent approach, all 1.16 million medium and high qual-

ity bins were mapped against the mOTUs 3.1 database (25) and unmapped MAGs were clustered into 84 287 novel mOTUs based on 10 marker genes, resulting in a novel, extended mOTU-profilable database. The ANI- and mOTU marker gene-based partitions of the data were highly concordant at an Adjusted Mutual Information (42,43) of 0.98.

All SPIRE MAGs were taxonomically classified using gtdb-tk v2.11 against release r207 (22) and consensus taxonomy for species clusters at each taxonomic level was assigned based on a majority vote, with manual resolution of a few remaining conflicting labels. The 92 134 MAG-based species clusters were classified into 178 different phyla (926 clusters representing 1 185 MAGs remained unclassified at phylum level) and 11 082 named species (79 782 clusters representing 198 384 MAGs remained unclassified at species level; Figure 2). This large proportion of 'novel' species relative to the GTDB may in part be due to a conservative parametrization of the gtdb-tk classifier (favoring specificity over sensitivity), but it indicates that SPIRE covers a vast diversity of previously uncharacterized and undescribed microbial diversity. Notably, 28 856 SPIRE clusters unclassified at species level contain more than a single genome.

Functional annotation

Detection of orthologs and inference of putative function for metagenomically-called ORFs (see above) were performed using eggNOG-mapper v2 (44,45). ORFs were further annotated for putative roles in antibiotics resistance using DeepARG (46) and abricate v1.0.1 (https://github.com/tseemann/abricate) against the MEGARes (version 2020-04-19, 47) and VFDB (version 2020–04-19, 48) reference databases.

Database design

SPIRE relies on a mongoDB database as its foundation. Within this system, a repository of samples/MAGs and their attributes is stored. This data can be conveniently accessed through the web-based interface. Structured data such as annotation of genes and genomes is stored in a relational database management system to allow complex and time efficient queries.

Website

SPIRE is accessible, browsable, searchable and downloadable via spire.embl.de. The main access modes are *by habitat/sample* (searching based on accessions or metadata tags), *by taxon* (based on clade names and species-level clusters) and *by genome* (individual genomes within clusters). These modes are inter-accessible (e.g. browsing from a sample to a specific taxon observed therein, for which then multiple genomes can be accessed) and at each level, link-outs to relevant independent or third party databases are provided. We invite user contributions, suggestions for improvements and bug reports under spire.embl.de/contribute.

Outlook

Given the exponential growth of publicly available metagenomic data, we anticipate biennial updates of the underlying data for SPIRE. We will continue to develop and update the processing pipeline to address rising computational demands and integrate novel or improved tools. Moreover, we will seek to extend the range of available functional annotations at gene and genome level, within the limits of computational scalability. Finally, and most importantly, we will continue to further integrate SPIRE with other resources such as proGenomes (24), eggNOG (26), the GMGC (8) and other ongoing efforts.

Discussion

SPIRE provides the largest sets of consistently processed metagenomes, newly generated MAGs and profilable microbial species clusters to date. Combined with a high degree of curation and integration of various data modalities (MAGs, contigs, genes, profiles, etc.), SPIRE is the most comprehensive resource available to study microbial diversity and function. Covering a broad range of habitats and geography, SPIRE enables true 'planetary-scale' analyses of microbiomes across various environments, including so far understudied ones. At the same time, SPIRE encompasses large amounts of 'novel', previously undescribed microbial diversity both at the gene and genome level. We are confident that SPIRE will enable and simplify a wide range of analyses for end users, ranging from the characterization of individual taxa or gene clusters of interest against a global data canvas, to truly 'planetary-scale' studies of microbial life across habitats and phylogeny.

Data availability

All raw data underlying SPIRE v1 is publicly available via the European Nucleotide Archive and/or the Sequence Read Archive. No new sequencing data was generated for this study. The derived and curated data described above is freely accessible and downloadable via spire.embl.de. SPIRE is released under a Creative Commons Attribution-ShareAlike 4.0 International License.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The authors would like to thank members of the Bork and Coelho teams for helpful discussions on designing the resource. We also acknowledge the EMBL IT Services for providing support and access to the EMBL high-performance computing infrastructure. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Funding

European Molecular Biology Laboratory; German Research Foundation (DFG) ['NFDI4Microbiota' to P.B.]; German Federal Ministry of Education and Research [LAMarCK, 031L0181A to P.B.]; The Science and Technology Commission of Shanghai Municipality [22JC1410900 to LPC]; NCCR Microbiomes [51NF40_180575 to S.S.].

Conflict of interest statement

None declared.

References

1. Falkowski,P.G., Fenchel,T. and Delong,E.F. (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science*, **320**, 1034–1039.
2. Gilbert,J.A. and Neufeld,J.D. (2014) Life in a world without microbes. *PLoS Biol.*, **12**, e1002020.
3. Hug,L.A., Baker,B.J., Anantharaman,K., Brown,C.T., Probst,A.J., Castelle,C.J., Butterfield,C.N., Hernsdorf,A.W., Amano,Y., Ise,K., *et al.* (2016) A new view of the tree of life. *Nat. Microbiol.*, **1**, 16048.
4. Rinke,C., Schwientek,P., Sczyrba,A., Ivanova,N.N., Anderson,I.J., Cheng,J.-F., Darling,A., Malfatti,S., Swan,B.K., Gies,E.A., *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.
5. Vanni,C., Schechter,M.S., Acinas,S.G., Barberán,A., Buttigieg,P.L., Casamayor,E.O., Delmont,T.O., Duarte,C.M., Eren,A.M., Finn,R.D., *et al.* (2022) Unifying the known and unknown microbial coding sequence space. *Elife*, **11**, e67667.
6. Río,Á.R., del,del Río, Á.R., Giner-Lamia,J., Cantalapiedra,C.P., Botas,J., Deng,Z., Hernández-Plaza,A., Paoli,L., Schmidt,T.S.B., Sunagawa,S., *et al.* (2022) Functional and evolutionary significance of unknown genes from uncultivated taxa. bioRxiv doi: https://doi.org/10.1101/2022.01.26.477801, 27 January 2022, preprint: not peer reviewed.
7. Richardson,L., Allen,B., Baldi,G., Beracochea,M., Bileschi,M.L., Burdett,T., Burgin,J., Caballero-Pérez,J., Cochrane,G., Colwell,L.J., *et al.* (2022) MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.*, **51**, D753–D759.
8. Coelho,L.P., Alves,R., Del Río,Á.R., Myers,P.N., Cantalapiedra,C.P., Giner-Lamia,J., Schmidt,T.S.B., Mende,D.R., Orakov,A., Letunic,I., *et al.* (2022) Towards the biogeography of prokaryotic genes. *Nature*, **601**, 252–256.
9. Costea,P.I., Zeller,G., Sunagawa,S., Pelletier,E., Alberti,A., Levenez,F., Tramontano,M., Driessen,M., Hercog,R., Jung,F.-E., *et al.* (2017) Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.*, **35**, 1069.
10. Sczyrba,A., Hofmann,P., Belmann,P., Koslicki,D., Janssen,S., Dröge,J., Gregor,I., Majda,S., Fiedler,J., Dahms,E., *et al.* (2017) Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.
11. Meyer,F., Fritz,A., Deng,Z.-L., Koslicki,D., Lesker,T.R., Gurevich,A., Robertson,G., Alser,M., Antipov,D., Beghini,F., *et al.* (2022) Critical assessment of metagenome interpretation: the second round of challenges. *Nat. Methods*, **19**, 429–440.
12. Gonzalez,A., Navas-Molina,J.A., Kosciolek,T., McDonald,D., Vázquez-Baeza,Y., Ackermann,G., DeReus,J., Janssen,S., Swafford,A.D., Orchanian,S.B., *et al.* (2018) Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods*, **15**, 796–798.
13. Pasolli,E., Schiffer,L., Manghi,P., Renson,A., Obenchain,V., Truong,D.T., Beghini,F., Malik,F., Ramos,M., Dowd,J.B., *et al.* (2017) Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods*, **14**, 1023–1024.
14. Dai,D., Zhu,J., Sun,C., Li,M., Liu,J., Wu,S., Ning,K., He,L.-J., Zhao,X.-M. and Chen,W.-H. (2022) GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Res.*, **50**, D777–D784.
15. Paoli,L., Ruscheweyh,H.-J., Forneris,C.C., Hubrich,F., Kautsar,S., Bhushan,A., Lotti,A., Clayssen,Q., Salazar,G., Milanese,A., *et al.* (2022) Biosynthetic potential of the global ocean microbiome. *Nature*, **607**, 111–118.
16. Nayfach,S., Roux,S., Seshadri,R., Udwary,D., Varghese,N., Schulz,F., Wu,D., Paez-Espino,D., Chen,I.-M., Huntemann,M., *et al.* (2021) A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.*, **39**, 499–509.
17. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumeil,P.-A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.
18. Parks,D.H., Chuvochina,M., Waite,D.W., Rinke,C., Skarshewski,A., Chaumeil,P.-A. and Hugenholtz,P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004.
19. Rinke,C., Chuvochina,M., Mussig,A.J., Chaumeil,P.-A., Davín,A.A., Waite,D.W., Whitman,W.B., Parks,D.H. and Hugenholtz,P. (2021) A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol*, **6**, 946–959.
20. Parks,D.H., Chuvochina,M., Chaumeil,P.-A., Rinke,C., Mussig,A.J. and Hugenholtz,P. (2020) A complete domain-to-species taxonomy for bacteria and archaea. *Nat. Biotechnol.*, **38**, 1098.
21. Chklovski,A., Parks,D.H., Woodcroft,B.J. and Tyson,G.W. (2023) CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods*, **20**, 1203–1212.
22. Chaumeil,P.-A., Mussig,A.J., Hugenholtz,P. and Parks,D.H. (2022) GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, **38**, 5315–5316.
23. Bowers,R.M., Kyrpides,N.C., Stepanauskas,R., Harmon-Smith,M., Doud,D., Reddy,T.B.K., Schulz,F., Jarett,J., Rivers,A.R., Eloe-Fadrosh,E.A., *et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, **35**, 725–731.
24. Fullam,A., Letunic,I., Schmidt,T.S.B., Ducarmon,Q.R., Karcher,N., Khedkar,S., Kuhn,M., Larralde,M., Maistrenko,O.M., Malfertheiner,L., *et al.* (2022) proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Res.*, **51**, D760–D766.
25. Ruscheweyh,H.-J., Milanese,A., Paoli,L., Karcher,N., Clayssen,Q., Keller,M.I., Wirbel,J., Bork,P., Mende,D.R., Zeller,G., *et al.* (2022) Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome*, **10**, 212.
26. Hernández-Plaza,A., Szklarczyk,D., Botas,J., Cantalapiedra,C.P., Giner-Lamia,J., Mende,D.R., Kirsch,R., Rattei,T., Letunic,I., Jensen,L.J., *et al.* (2023) eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.*, **51**, D389–D394.
27. Chen,I.-M.A., Chu,K., Palaniappan,K., Ratner,A., Huang,J., Huntemann,M., Hajek,P., Ritter,S., Varghese,N., Seshadri,R., *et al.* (2020) The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Res.*, **49**, D751–D763.
28. Buttigieg,P.L., Pafilis,E., Lewis,S.E., Schildhauer,M.P., Walls,R.L. and Mungall,C.J. (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semantics*, **7**, 57.
29. Mungall,C.J., Torniai,C., Gkoutos,G.V., Lewis,S.E. and Haendel,M.A. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
30. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
31. Coelho,L.P., Alves,R., Monteiro,P., Huerta-Cepas,J., Freitas,A.T. and Bork,P. (2019) NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. *Microbiome*, **7**, 84.
32. Li,D., Liu,C.-M., Luo,R., Sadakane,K. and Lam,T.-W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674–1676.
33. Hyatt,D., Chen,G.-L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.*, **11**, 119.

34. Chan,P.P. and Lowe,T.M. (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.*, **1962**, 1–14.
35. Kang,D.D., Li,F., Kirton,E., Thomas,A., Egan,R., An,H. and Wang,Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, **7**, e7359.
36. Orakov,A., Fullam,A., Coelho,L.P., Khedkar,S., Szklarczyk,D., Mende,D.R., Schmidt,T.S.B. and Bork,P. (2021) GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.*, **22**, 178.
37. Mende,D.R., Sunagawa,S., Zeller,G. and Bork,P. (2013) Accurate and universal delineation of prokaryotic species. *Nat. Methods*, **10**, 881–884.
38. Rodrigues,J.F.M., Schmidt,S.B.T., Tackmann,J. and von Mering,C. (2017) MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics*, **33**, 3808–3810.
39. Jain,C., Rodriguez-R,L.M., Phillippy,A.M., Konstantinidis,K.T. and Aluru,S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 5114.
40. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
41. Müllner,D. (2013) fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. *J. Stat. Softw.*, **53**, 1–18.
42. Vinh,N.X., Epps,J. and Bailey,J. (2009) Information theoretic measures for clusterings comparison: is a correction for chance necessary? *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*.
43. Schmidt,T.S.B., Matias Rodrigues,J.F. and von Mering,C. (2015) Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ. Microbiol.*, **17**, 1689–1706.
44. Huerta-Cepas,J., Forslund,K., Coelho,L.P., Szklarczyk,D., Jensen,L.J., von Mering,C. and Bork,P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.*, **34**, 2115–2122.
45. Cantalapiedra,C.P., Hernández-Plaza,A., Letunic,I., Bork,P. and Huerta-Cepas,J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, **38**, 5825–5829.
46. Arango-Argoty,G., Garner,E., Pruden,A., Heath,L.S., Vikesland,P. and Zhang,L. (2018) DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, **6**, 23.
47. Bonin,N., Doster,E., Worley,H., Pinnell,L.J., Bravo,J.E., Ferm,P., Marini,S., Prosperi,M., Noyes,N., Morley,P.S., *et al.* (2023) MEGARes and AMR++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic Acids Res.*, **51**, D744–D752.
48. Liu,B., Zheng,D., Zhou,S., Chen,L. and Yang,J. (2022) VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.*, **50**, D912–D917.

Received: August 18, 2023. **Revised:** October 1, 2023. **Editorial Decision:** October 10, 2023. **Accepted:** October 11, 2023