nature portfolio

Corresponding author(s): *Double-anonymous peer review submissions: write DAPR and your manuscript number here instead of author names.*

Last updated by author(s): *YYYY-MM-DD*

Reporting Summary

Nature Portfolio 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 Portfolio 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	The samples taken as part of the MFD sampling campaign were registered and associated with the appropriate metadata using codeREADr [https://www.codereadr.com] .
Data analysis	R v4.2.3 (lubridate v1.9.2 and terra v1.7.55, dm v1.1_18, tidymodels v1.1.1, ranger v0.16.0, vip v0.4.1, yardstick v1.3.0, rstatix v0.7.2 packages, ggspatial v1.1.9, rnaturalearth v1.0.1), R v4.3.2 (tidyverse v2.0.0 and vegan v2.6-6.1 packages), R v4.4.1 (UpSetR v1.4.0 and ComplexUpset v1.3.3, tidyverse v2.0.0, vegan v2.6-6.1, parallelDist v0.2.6 packages), fastp v0.23.2, GNU-parallel v20220722, pigz v2.4, Guppy v. 5.0.11, https://github.com/SorenKarst/longread_umi v0.3.2, Lima v2.6.0, cutadapt v3.4, custom script , usearch v11, SINA v.1.6.0, the custom scripts remove_incomplete_seqs_from_sina_aln.py, Remove_gaps_in_fasta.py and trim_euk_RNA_operons_3ndf-1510R.sh. Custom scripts available in github: https://github.com/cmc-aau/mfd_wiki/wiki. HMMER v3.3.2, R v4.4.1 (tidyverse v2.0.0, egan v2.6-6-1, iNEXT v3.0.1, ampvis2 v2.8.9, parallelDist v0.2.6, ape v5.8, codep v1.2-3, stats v4.4.1, bootstrap v0.1 packages). . MegaHit v1.2.9, Minimap2 v2.24, SAMtools v1.16.1, MetaBAT2 v2.12.1, CheckM2 v1.0.2, GTDB-Tk v2.3.0, Bakta v1.8.1, tRNAscan-SE v2.0.9, barrnap v0.9, dRep v2.6.2, CoverM v0.6.1, SingleM v0.18.0, SingleM v1.0.0beta7, Bin Chicken v0.9.6, fastANI v1.32, Prodigal v2.63, DIAMOND v2.1.8, ARB v7.0, iTOL v6, GraftM v0.14. CheckM2 v1.0.2, GTDB-Tk v2.3.2, GTDB-Tk v2.4.0, fxtract v2.3, ITOL v6, MAFFT v7.490, TrimAl v1.4.1, IQ-TREE v2.2.0.3. IQ-TREE v2.4.0. IQ-TREE v2.1.2, Additional R packages: readxl v1.4.3, data.table v1.15.4, wesanderson v0.3.7, viridisLite v0.4.2, ggpubr v0.6.0, patchwork v1.2.0, ggpmisc v0.5.6, hexbin v1.28.3. Inkscape v1.4.2, Adobe Illustrator 2024, DRAM v1.4.6, dbCAN HMMdb v13, ARB v7.0, Rfam v14.7, HMMER v3.3.2, sylph v0.6.1, anvi'o v8, GraftM v0.14, Muscle5, RStudio 2024.04.2 and 2023.12 , DADA2 v1.26.0, bcl2fastq2 v2.20.0, Bakta v1.8.1, tRNAscan-SE v2.0.0, barrnap v0.9, myloasm v0.1.0, Flye v2.9.3, mmseq2 v14.7e284.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data in NCBI are accessible through the BioProject PRJNA1071982. This includes the sequencing data: bacterial UMI 16S rRNA gene reads (PromethION), bacterial rRNA operon data (PACBIO), eukaryotic rRNA operon (PACBIO), and metagenomes (Illumina), which are available in NCBI SRA. Additionally, the high-quality MAGs, based on completeness and contamination (>90 % completeness <5 % contamination), are available in NCBI GenBank under BioProject PRJNA1071982. The long-read MAG data is in ENA under BioProject ID: PRJEB586.

Data in Zenodo are accessible through the link https://doi.org/10.5281/zenodo.15535748. This includes the medium quality MAGs, metagenome assemblies and MFG 16S rRNA gene reference database (clustered and unclustered, incorporating sequences from SILVA v138.1 SSURef NR99, EMP500, AGP70, MiDAS 4 and 5, and Overgaard et al. 2022). We furthermore uploaded the amplicon data (UMI 16S rRNA gene, 16S from operons and 18S from operons) to GBIF (Global Biodiversity Information Facility) as separate datasets which can be accessed through the links https://doi.org/10.15468/33qqsm, https://doi.org/10.15468/yr5rmw and https://doi.org/10.15468/ea7jvq, respectively. Data (metadata and data files) in GitHub are accessible through https://github.com/cmc-aau/mfd_wiki/wiki.

Public data used in this study include: GTDB r220 and r214 https://data.gtdb.ecogenomic.org/releases/, SingleM GTDB R220 reference metapackage v4.3.0 (https://zenodo.org/records/11323477), SILVA v138.1 https://www.arb-silva.de/documentation/release-1381/, PR2 database v5.0.0 https://github.com/pr2database/pr2database/releases, MiDAS 4 and 5 databases: https://www.midasfieldguide.org/guide/downloads, AGP70 database ENA PRJEB32674: https://www.ebi.ac.uk/ena/browser/view/PRJEB32674, Overgaard et al., 2022. NCBI SRA PRJNA787301 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA787301/, MIDORI2 GB257 database: https://www.reference-midori.info/download.php, Greengenes2 https://ftp.microbio.me/greengenes_release/2022.10/ , GEM database: https://genome.jgi.doe.gov/portal/GEMs/GEMs.home.html, GlobDB release 220: https://globdb.org/downloads, KEGG database release 109: https://www.kegg.jp/kegg/docs/relnote.html, dbCAN HMMdb and dbCAN3 server: https://bcb.unl.edu/dbCAN2/, MEROPS Peptidase Database: https://www.ebi.ac.uk/merops/, UniRef100 database: https://www.uniprot.org/uniref.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This information is not relevant to the study.
Reporting on race, ethnicity, or other socially relevant groupings	This information is not relevant to the study.
Population characteristics	This information is not relevant to the study.
Recruitment	This information is not relevant to the study.
Ethics oversight	This information is not relevant to the study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	The study is "Microflora Danica", an atlas of Danish environmental microbiomes, encompassing 10,683 shotgun metagenomes and 450 full-length 16S and 18S rRNA datasets linked to a detailed 5 level habitat classification scheme. We determine that while human-disturbed habitats have high alpha diversity, the same species reoccur, revealing hidden homogeneity and underlining the importance of natural systems for total species (gamma) diversity. The study further delves into the distribution of nitrifiers, a functional group closely linked to climate change, challenging existing perceptions regarding habitat preference and discovering several novel lineages as more abundant than canonical ones.

Research sample	The samples all represent environmental microbiomes described by their metagenomic sequencing content.
Sampling strategy	Samples were collected as part of the Microflora Danica campaign or contributed by collaborators. The number of samples reflects our aim to cover the various habitats of Denmark. Our sample size was determined by feasibility and reach. We have described the sampling approach in detail in the methods of the manuscript and Supplementary Data 5, detailing procurement processes of samples contributed by Microflora Danica consortium members, and the Microflora Danica sampling campaign.
Data collection	Data collection and metadata curation: the metadata associated with the samples were recorded by our collaborators with the codeREADr at the sampling sites. Metadata include information such as fieldsample_barcode (the unique sample identifier), project_id (unique identifier of the subproject), longitude and latitude, sitename (common name of the sampling site), sampling_date, five levels of the MFD habitat ontology (mfd_sampletype, mfd_areatype, mfd_hab1, mfd_hab2, mfd_hab3) and custom comments. Subsampling was performed following the procedure described in Jensen et al. 2024 with the modification that the samples were divided into a total of three Matrix™ 1.2 mL 2D barcoded tubes [Thermo Scientific™].
Timing and spatial scale	The samples were collected between 2008-04-15 and 2023-02-13; however, 9292 samples (87% of the total) were taken between 2020 and 2022. The sampling covered the country of Denmark.
Data exclusions	No samples were excluded a priori. In several analysis a threshold (e.g., minimum metagenomics reads) was applied to subset the data to a reliable set. In all those cases, the criterion is reported in the Methods section.
Reproducibility	For the vast majority of the samples an aliquot has been stored. All the data and code are publicly available.
Randomization	The study does not require randomisation.
Blinding	The study does not require blinding.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Palaeontology and Archaeology

Specimen provenance	15 soil samples in the study were collected as top soil at archaelogical excavation sites in Ribe and Aalborg.
Specimen deposition	All data has been deposited together with the remaining samples in the study.
Dating methods	n/a

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	No ethical approval or guidance was required, as the samples were not archaeological specimens.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	No plants were used in this study
Novel plant genotypes	No plants were used in this study
Authentication	No plants were used in this study