nature portfolio

Corresponding author(s): Sharvari Sunil Gadegaonkar

Last updated by author(s): Nov 13, 2025

# 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<br>*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<br>*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**

Seawater samples were collected during the SWINGS cruise (January–March 2021) across 13 stations spanning the Subtropical, Subantarctic, Polar Frontal, and Antarctic Zones. Samples were classified into four surface water types and eight water masses based on temperature, salinity, oxygen, depth, and location. For each sample, 6L of seawater was collected using Niskin bottles mounted on a CTD rosette. Particle-attached (PA) prokaryotes were captured on 0.8µm filters, while free-living (FL) prokaryotes were concentrated on 0.22µm Sterivex units. A total of 23 FL and 19 PA samples were used for metagenomic analyses. Filtered samples were stored at −80°C until DNA extraction. DNA was extracted using the DNeasy PowerWater Kit with modifications for Sterivex filters, including lysozyme and Proteinase K treatments, and quantified with a QuantiFluor® dsDNA fluorometer. Metagenomic sequencing was performed on an Illumina NovaSeq 6000 platform using 2 × 150bp paired-end chemistry at Fasteris SA (Switzerland). The analysis consisted of 42 metagenomes generated from this collection, which have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB75506.

**Data analysis**

Raw metagenomic reads were quality-checked using FastQC v0.11.9 and preprocessed with Trimmomatic v0.32 to remove low-quality bases and adapter sequences. High-quality reads were assembled per sample using MEGAHIT v1.2.9 (--min-contiglen 1000, --presets meta-large), and ORFs were predicted with Prodigal v2.6.3 in metagenomic mode. A non-redundant protein set was generated across all samples using CD-HIT v4.8.1 (-c 1, -aS 1, -g 1), yielding 18,497,675 unique proteins. Reads were mapped to this set with Salmon v1.4.0 (--meta, --seqBias, --gcBias) and normalized as genes per kilobase million (GPM). Nitrogen-transforming genes were annotated against NCycDB using Diamond-based NcycProfiler, with additional querying of nifB via GhostKOALA and KofamKOALA using the KEGG database tools.

Metagenome-assembled genomes (MAGs) were generated from contigs ≥2.5kb using MetaWRAP v1.3, integrating MaxBin2 and MetaBAT2, refined, and evaluated for completeness and contamination with CheckM v1.2.2. High-quality MAGs were dereplicated with dRep (≥95% ANI), resulting in 556 non-redundant MAGs. MAG coverage was quantified via Bowtie2 alignments, processed with samtools, and visualized using Anvi'o v7.1. MAG functional annotation employed Prodigal-predicted ORFs, NCycDB, KEGG, GhostKOALA, KofamKOALA, and METABOLIC v4.0.

Spearman's rank correlation (R v4.4.2, corrplot) assessed relationships between environmental parameters and N-transforming gene abundances (pairwise.complete.obs). GPM values were Hellinger-transformed to reduce dominance effects and manage zero values. Physical (depth, temperature, salinity, density) and chemical ($O_2$, nutrients, trace metals) parameters were included for both free-living and particle-attached communities, following standard analytical protocols (Zhang et al., 2024).

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

Raw metagenomic reads for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB75506.

## 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 | NA |
| Reporting on race, ethnicity, or other socially relevant groupings | NA |
| Population characteristics | NA |
| Recruitment | NA |
| Ethics oversight | NA |

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 based on the metagenomic analysis of 42 metagenomes collected during the South West INdian GEOTRACES GS02 Section - (https://swings.geotraces.org/en/homepage-english/) |
| Research sample | Across 13 stations spanning the Subtropical, Subantarctic, Polar Frontal, and Antarctic Zones. Samples were classified into four surface water types and eight water masses. Total of 42 metagenomes were extracted comprising of distant surface waters and water masses with the depth profile consisting of surface to 4000m. |
| Sampling strategy | 6L of seawater was collected using Niskin bottles mounted on a CTD rosette. Particle-attached (PA) prokaryotes were captured on 0.8µm filters, while free-living (FL) prokaryotes were concentrated on 0.22µm Sterivex units. A total of 23 FL and 19 PA samples were used for metagenomic analyses. |
| Data collection | A total of 42 metagenomes were collected during the SWINGS GEOTRACERS cruise, comprising of distant surface waters and water masses. Filtered samples were stored at −80°C until DNA extraction. DNA was extracted using the DNeasy PowerWater Kit with modifications for Sterivex filters, including lysozyme and Proteinase K treatments, and quantified with a QuantiFluor® dsDNA fluorometer. Metagenomic sequencing was performed on an Illumina NovaSeq 6000 platform using 2 × 150bp paired-end chemistry at Fasteris SA (Switzerland). |
| Timing and spatial scale | All metadata information of the samples collected in presented in Table S4. |

| | |
|---|---|
| Data exclusions | No data was excluded |
| Reproducibility | NA |
| Randomization | Particle-attached (PA) prokaryotes were captured on 0.8µm filters, while free-living (FL) prokaryotes were concentrated on 0.22µm Sterivex units. A total of 23 FL and 19 PA samples were used for metagenomic analyses. |
| Blinding | NA |

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 |

## Plants

| | |
|---|---|
| Seed stocks | NA |
| Novel plant genotypes | NA |
| Authentication | NA |