nature portfolio

Corresponding author(s): Brett Baker

Last updated by author(s): 12/06/2025

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 *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

Sample acquisition: Sediment samples were collected and processed during three expeditions: BH (Bohai Sea, August 2018), GE2 (Guaymas Basin, Gulf of California, December 2008 and 2009), and GE3 (Guaymas Basin, November 2018). Bohai Sea (BH, 17-30.5 mbsl) and Guaymas Basin cores (GB, ~2000 mbsl) were collected on the R/V Chuangxin Yi and HOV Alvin, respectively. Sampling through de novo binning from the BH and GE2 included in this study have been described previously (Gong et al., 2022 and Langwig et al., 2021). Briefly, BH sediment samples were collected using a box-sampler and after removing the top water samples were taken with an 11 cm in diameter polyvinyl chloride (PVC) tube. Samples from 2 cm depth horizons were collected, using pre-drilled side-holes. Guaymas Basin sediment samples (GE2 and GE3) were sampled with polycarbonate cores approximately 45-60 cm in length with a diameter 6.25 cm. Depending on the core length, the cores were split into depth horizons listed for GE3 in Supplementary Table 1. Sediment samples were frozen at -80°C until extractions could take place onshore.

Guaymas Basin Expedition 3 described first in this study.
Extraction and Sequencing: We split five sediment cores from 12 depth horizons from the GE3 expedition for DNA extraction (Supplementary Fig. 1). DNA was extracted using the DNeasy PowerSoil Kit (Qiagen, Germantown, Maryland, USA), quantified using a QUBIT 2.0 fluorometer (Thermo-Fisher, Singapore), and concentrated with ethanol precipitation. Extracted DNA was sent to North Carolina State Genomics Sciences Laboratory for quality control, library preparation, and sequencing. Samples ran separately on an Illumina NovaSeq S4 150x2 PE flow cell (~10 billion reads/flow cell). Three samples below the threshold for NovaSeq sequencing were sequenced with Illumina NextSeq at the University of Delaware.

Assembly and Binning: We utilized FastQC v0.11.5 and fastp v0.21.0 to assess the initial quality of the GE3 sequencing reads, removing poly-G tails (an artifact of NovaSeq sequencing), Illumina adapter sequences, reads matching human alignments, and short and low-quality reads (https://www. bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed reads were interleaved with BBTools Reformat v38.18 (http://

sourceforge.net/projects/bbmap). The assembler, MEGAHIT v1.2.9, used succinct de Bruijn graphs to assemble the short-read fragments into longer contiguous reads (contigs) using a variety of kmer sizes. Assemblies for each sample greater than 2.5kb were used for de novo tetranucleotide, coverage, and abundance-based binning with MaxBin v2.2.7, CONCOCT, and MetaBAT via MetaWRAP v1.3.2. Das Tool v1.1.2-2 extracted an optimized set of non-redundant bins for each assembly from all three binning programs and MAGs were cleaned with mmgenome.

Obtaining 404 Asgard MAGs required the massive sequencing effort (BH, GE2, and GE3) of 44 samples, >13,839 MAGs, and 34.5 billion reads (15 TB).

Data analysis

DNA extraction and sequencing of Guaymas Basin Expedition 3:
BBTools Reformat v38.18
CheckM v1.2.1
CONCOCT
Das Tool v1.1.2-2
fastp v0.21.0
FastQC v0.11.5
MaxBIN v2.2.7
MEGAHIT v1.2.9
MetaBAT via MetaWRAP v1.3.2

Curation of Asgard archaea database:
CheckM v1.2.1
CompareM v0.1.2
CoverM v0.7.0 and v0.6.1
get_gc_content.pl (https://github.com/mrmckain/)
GTDB-Tk v2.1.1
MEBS v2
miComplete v1.1.1
R package (cooccur v1.3)
SeqKit v2.3.0
SingleM/Sandpiper 0.3.0

Functional annotations:
Bacterial_dating_aerobic_predictor (https://github.com/wwood/bacterial_dating_aerobic_predictor)
dbCAN2 (CAZy)
EggNOGv6
GenomeSPOT
HADEG (Hydrocarbon Aerobic Degradation Enzymes and Genes)
HydDB: A web tool for hydrogenase classification and analysis diamond v0.9.26.127
Interproscan v5.61-93.0-64
KofamKOALA v1.3.0 (KEGG)
MEBS/Pfam v34
MEROPS v12.1
PeroxiBase/RedoxiBase
Prodigal v2.6.3

Phylogenetic analyses of the expanded Asgard genomic catalog:
BioKIT v0.0.9
dRep v2.2.1
FAMSA v2.0.1
GTDB release 214
GTR+G4+C60 model adapted to SR4 (https://github.com/xgrau/recoded-mixture-models)
IQ-TREE v2.1.3
Prokka v1.14.6
PSI-BLAST v2.10.0+
trimAl v1.4.rev22

Metabolic guilds definition:
MEBS v2/Pfam v34

Comparative genomic analysis by lineage, guild, and environmental source:
EggNOGv6
get_homologues (https://github.com/eead-csic-compbio/get_homologues)
KofamKOALA v1.3.0 (KEGG)
MEBS v2/Pfam v34
Pangenome.R (https://figshare.com/s/f139faeb05653d1adf6b)

Key eukaryotic-like and metabolic genes and proteins:
IQ-TREE v2.0.7
MAFFT auto v7.490
trimAl v1.4.rev15

Phylogenetic and structural analyses of the ETC subunits:
AlphaFold2 v2.3.2
AlphaFold2-multimer v3
AlphaFold3

BMGE v1.12
Clustal Omega v1.2.4
CombFold
CUDA v12.0
FoldSeek and Foldseek-multimer v10.941cd33
Geneious Prime v2023.2.1
hmmsearch v3.3.2
HydDB: A web tool for hydrogenase classification and analysis diamond v0.9.26.127
hmmsearchTable (hmmer v3.3)
hydrogenase_motif_finder.py (https://github.com/keappler/Scripts/blob/main/hydrogenase_motif_finder.py)
IQ-TREE v.2.1.4-beta, v2.0.7
LocalColabFold v1.5.2 and v1.5.5
MAFFT v7.490
MMseq2 v17.6904f
mTM-align version 20220104
NCBI conserved domain search (CDD)
UCSF ChimeraX v1.8

Visualization:
Affinity Designer v1.10.6
Biorender
ESPript 3.0
iTOL v6
Jalview v2.11.4.0
PAE Viewer
UCSF ChimeraX v1.8

Custom scripts: Are listed above and provided in figshare (https://figshare.com/s/f139faeb05653d1adf6b).

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

The genomic sequences associated with the study are deposited by the sequencing project in NCBI under BioProjects PRJNA743900 (BH), PRJNA692327 (GE2), and PRJNA1112871 (GE3). Supplementary Table 1 includes the BioProject identifiers, BioSample identifiers, and genome accession numbers. Supplementary Table 9 and 11 includes lists of the visualized Protein Data Bank structures. All raw data underlying phylogenomic (alignments and resulting phylogenetic trees) and structural analyses have been deposited into Figshare https://figshare.com/s/f139faeb05653d1adf6b.

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

Ecological, evolutionary & environmental sciences study design

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

Study description	To expand the genomic catalog of Asgard archaea, we chose two ocean floor ecosystems rich in Asgard diversity, the shallow (17-30.5 m) coastal Bohai Sea (BH, Eastern China) and the deep sea hydrothermal system in (2,000 m) Guaymas Basin (GB, Gulf of California). We combined 10 Tb of new sequencing (GE3; the largest known deep-sea genomic database) with previously reported efforts (BH & GE2), and extensive manual curation to obtain over 13,000 MAGs. From these, a total of 404 new MAGs were classified as Asgardarchaeota, expanding the known diversity of this group by 46.5%. The Guaymas Basin in the Gulf of California is a known hotspot for microbial biodiversity. Several previous expeditions have described the microbial community dynamics and geochemistry. The GE3 sequencing effort first described here resulted in 5,606 archaeal and bacterial MAGs (>50% complete, <10% multiple gene copies based on CheckM v1.2.1) from >10 Tb of sequencing data at 1 billion reads per sample across 13 samples, representing the largest known DNA database from the deep-sea. Metagenomic sequencing enables a large-scale view of the novel lineages and microbial community interactions within this unique ecosystem.
Research sample	BH: First described in Gong et al., 2022. GE2: First described in Langwig et al., 2021. GE3: First described here. This sequencing expedition includes three Alvin dives from which we obtained five cores split into 12 depth horizons. The names for the corresponding MAGs are Dive_Core(s)_Bin#. The first two portions of the name refer to the Alvin dive and cores to assist in tracking the sediment samples across different studies. The horizon numbers increase from shallow to deeper layers of sediment (i.e, H1 shallowest, H2 middle layer, H3 deepest). Additional information on the sampling sites (i.e., location, depth, and temperature) of GE3 are included in Supplementary Table 1.
Sampling strategy	The Bohai Sea sediment samples were taken at previously describe sites (M3, M8, and BHB10). Similarly, the Guaymas Basin sediment was collected near known hydrothermal vent sites (Southern mat mounds, Cathedral Hills, and previously identified Marker 14 Aceto Balsamico, etc.) in the Guaymas Basin. Bohai Sea sediment samples were collected using a box-sampler. After removing the top water, an 11 cm in diameter polyvinyl chloride (PVC) tube was used core samples. Core was sampled t 2 cm depth horizons, using pre-drilled side-holes. Guaymas Basin sediment samples (GE2 and GE3) were sampled with polycarbonate cores approximately 45-60 cm in length with a diameter 6.25 cm. Depending on the core length, the cores were split into depth horizons listed in Supplementary Table 1.
Data collection	In addition to sediment collection, core temperature and sediment depth was collected for GE3 and provided in Supplementary Table 1. Approximate elevation is provided for all MAGs across sampling expedition. Addtional chemical and temperature measurements are described in previous published descriptions of GE2 and BH sediment samples.
Timing and spatial scale	Bohai Sea (BH) sediment was collected August 2018 at three sites M3 (26.5 meters below sea level, 38.668056 N 119.5475 E), M8 (30.5 meters below sea level, 39.69 N 120.65 E), and BHB10 (17 meters below sea level, 38.751389 N 118.156389 E). Guaymas Basin Expedition 2 (GE2) sediment was collected from the Gulf of California, Mexico (27°N 0.388, 111°W 24.560) between December 2008-2009 during two HOV Alvin dives: D4486 and D4573. Samples were taken from approximately 2000 meters below sea level. More information on the Alvin dives is currently available in the Alvin dive logs under cruises AT15-40 and AT15-56 (https://sealog.whoi.edu/sealog-alvin/). Guaymas Basin Expedition 3 (GE3) sediment was collected from the Gulf of California, Mexico (27°N 0.388, 111°W 24.560) in November 2018 during three HOV Alvin dives: D4993, D4994, and D4998. Samples were taken from approximately 2000 meters below sea level. More information on the Alvin dives is currently available in the Alvin dive logs under cruise AT42-05 (https://sealog.whoi.edu/sealog-alvin/).
Data exclusions	Public Asgardarchaeota MAGs were excluded if they did not pass completeness and contamination control checks described in the manuscript. Additional MAGs were excluded from analysis if the genomes were binned from the same public assemblies to avoid potential of duplicated scaffolds/genomes, which would bias results.
Reproducibility	Detailed methods descriptions are provided for reproducibility, including specific flags and versions for public software. Custom scripts are included in figshare and/or github as described in previous section and data availability statement. Changes in public databases may alter metabolic annotations, but dates and versions are listed to mitigate the impact of updates.
Randomization	Randomization is not necessary for the description of microbial communities with metagenomes, especially low abundance microorganisms. These analyses instead rely on evolution relationships and ecological distribution instead of treatments or control groups. Randomization is not necessary for this type of research.
Blinding	Blinding is not necessary to this study as the methods are based on comparisons of ecological characteristics and evolutionary relationships not assessments impacted by treatment effects. This technique is not required or expected for this type of research.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	BH sea sediments were collected in coastal Bohai Sea off the coast of China. Guaymas Basin sediments were collected in the Gulf of California. More detailed description of temperature, elevation, etc. is provided in Supplementary Table 1 for GE3, first described by this study.
Location	Bohai Sea (BH) sediment was collected August 2018 at three sites M3 (26.5 meters below sea level, 38.668056 N 119.5475 E), M8 (30.5 meters below sea level, 39.69 N 120.65 E), and BHB10 (17 meters below sea level, 38.751389 N 118.156389 E). Guaymas Basin Expedition 2 (GE2) sediment was collected from the Gulf of California, Mexico (27°N 0.388, 111°W 24.560) between December 2008-2009 during two HOV Alvin dives: D4486 and D4573. Samples were taken from approximately 2000 meters below sea level. Guaymas Basin Expedition 3 (GE3) sediment was collected from the Gulf of California, Mexico (27°N 0.388, 111°W 24.560) in November 2018 during three HOV Alvin dives: D4993, D4994, and D4998. Samples were taken from approximately 2008-2012.3 meters below sea level.
Access & import/export	NA
Disturbance	NA

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