nature portfolio

Corresponding author(s): Stephan Köstlbacher,Thijs Ettema

Last updated by author(s): Jul 2, 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	ColabFold v1.3.0, ColabFold v1.5.2, ESMfold v2.0.0
Data analysis	Custom code is available at GitHub (https://github.com/stephkoest/structural_genomics). PSI-BLAST v2.10.0+, FoldSeek v6.29e2557, MMseqs2 v14.7e284, HHpred v3.3.0, HMMer web server, IQ-TREE V2.0.3 and v2.1.3, MAFFT v7.505, web server of PROMALS3D, CheckM v1.2.1, GTDB-Tk v2.3.2, Prokka v1.14.6, EggNOG v5.0,FAMSA v2.2.2, InterProScan v5.57-90.0, ChimeraX v1.6.1,R v4.2.1, r3dmol v0.1.2

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

Asgard archaeal genome data was obtained from NCBI GenBank (https://www.ncbi.nlm.nih.gov/nucleotide) and identifiers can be found in Data S1. Archaeal outgroup proteomes were downloaded from Genome Taxonomy Database (GTDB) release 214 (https://data.gtdb.ecogenomic.org/releases/release214/214.0/

genomic_files_reps/gtdb_proteins_aa_reps_r214.tar.gz) and proteins for gene phylogenies from release 207 (https://data.gtdb.ecogenomic.org/releases/release207/207.0/genomic_files_reps/gtdb_proteins_aa_reps_r207.tar.gz). All predicted structures, original multiple sequence alignments, and IQ-TREE outputs are available at figshare (https://figshare.com/s/adfec4c8d101ef774355). The uncollapsed phylogenies can be found on the iTOL 105 website: https://itol.embl.de/tree/621451922103993416998883333 (CINP, Figure 4A), https://itol.embl.de/tree/137224252121998116998682285 (COMMD, Figure 4C), https://itol.embl.de/tree/62145192210319901699902102 (Ufm1, Figure 4D).

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	Does not apply. We mostly worked with prokaryotic genomes. Where eukaryotic data was used, sex and gender were of no relevance to the questions and therefore not subject of this manuscript.
Reporting on race, ethnicity, or other socially relevant groupings	This study does not involve human subjects, social groupings, or research related to race, ethnicity, or other socially relevant groupings.
Population characteristics	No population level data was analysed.
Recruitment	No participants were recruited.
Ethics oversight	None.

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	This study investigates the evolution of eukaryotic signature proteins (ESPs) in Asgard archaea using structural modeling, sequence-based annotation, and phylogenetic analysis. The study analyzes 936 genomes to understand the contributions of archaeal proteins to the origins of eukaryotic cellular complexity. The experimental units are computational models of protein structures.
Research sample	The research sample comprises 936 publicly available Asgard archaeal genomes, including metagenome-assembled genomes (MAGs). These genomes were selected to maximize phylogenetic diversity and representation across known Asgard clades. All data were sourced from NCBI or from collaborators (Baker et al.) and will be made publicly available on NCBI.
Sampling strategy	Genomes were selected based on quality thresholds: completeness >50% and contamination <10%, assessed using CheckM. Sampling aimed to represent phylogenetic diversity within Asgard archaea. No predetermined sample size calculation was performed due to the exploratory nature of the study, but the dataset was deemed sufficient for robust evolutionary and structural analyses
Data collection	Genomic data were downloaded from NCBI. Structural data were generated using ESMfold and ColabFold for protein structurepredictions.
Timing and spatial scale	Data was obtained with a cutoff date of October 6, 2022.
Data exclusions	Genomes with completeness <=50% or contamination >=10% were excluded to ensure robust downstream analyses. Protein clusters with low structural prediction confidence.
Reproducibility	All computational analyses were performed using open-source software with defined parameters. Custom scripts and workflows are publicly available on GitHub (https://github.com/stephkoest/structural_genomics).
Randomization	Randomization was not applicable, as this study systematically analyzed publicly available genomes and structural data.
Blinding	Blinding was not applicable, as the study involved automated computational workflows without subjective assessments.

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	n/a
Novel plant genotypes	n/a
Authentication	n/a