

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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	All the data collected for the study are included in the following databases: Integrated Microbial Genomes and Microbiomes (IMG/M) database, NCBI non-redundant (nr) database, NCBI organelle database, NCBI RefSeq plastomes and mitochondrial genomes, GTDB database, SILVA ribosomal RNA gene database, and Protist Ribosomal Reference database (PR2).
Data analysis	The workflow (NSGTree v0.5.1) used to infer phylogenies in this manuscript is available on GitHub (https://github.com/NeLLi-team/nsgtree). Published softwares used in our study includes, MetaBAT 2, Prodigal v2.6.3, CheckM v1.1.3, seqkit v2.5.1, skani v0.1.4, IQ-TREE v2.0, hmmsearch v3.1/b2, MAFFT v7, trimAl v1.4, PhyloDM v..2.0, MCL v22-282, Infernal v1.1, blastn v2.13.0, DIAMOND v2.1.8, GeSeq v2.03, tRNAscan-SE v2.07, Geneious Prime v2025.0.2, progressiveMauve v1.1.3, and OrthoFinder v2.5.5.

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 plastid MAGs generated in our study are publicly available through the European Nucleotide Archive under project accession PRJEB106049 and individual accessions are listed in Supplementary Data 3. Additional data generated in our study are available as supplementary dataset via FigShare [<https://figshare.com/s/5b6a0c38129c2d3dc572>]. The supplementary dataset includes concatenated protein alignments used for phylogenetic inferences, phylogenetic trees, annotations of plastid MAGs in GenBank format, and plastid MAGs in FASTA format.

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	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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](https://www.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	We used bioinformatics approach to identify and analyze plastid sequences from global metagenomic data, including phylogenetic reconstruction of plastid origin and distribution in eukaryotes. The study employed metagenomic binning for cyanobacterial data and screening for plastid 16S rRNA sequences, followed by quality control and phylogenetic inferences.
Research sample	Our research sample consists of publicly accessible (unrestricted) datasets largely from IMG/M database up to April 2023. Specifically, 31,152 MAGs for cyanobacterial dataset, 26,247 MAGs for plastid 16S rRNA sequences, 115,969 reference genomes from IMG isolates, 85,205 GTDB bacterial/archaeal datasets. The MAGs represent diverse environmental samples collected globally and deposited in public repositories.
Sampling strategy	We employed a comprehensive inclusion of publicly available data representing a census approach rather than statistical sampling.
Data collection	All data were obtained from the IMG/M (Integrated Microbial Genomes & Microbiomes) database, a publicly accessible repository maintained by the Joint Genome Institute. All MAGs publicly available and non-redundant up to April 2023 were downloaded and processed. Reference genomes were obtained from IMG isolates repository and GTDB (Genome Taxonomy Database).
Timing and spatial scale	Our data collection included all publicly available MAGs deposited in IMG/M, which were collected across global diverse environments by various research groups. Individual data dates and location vary based on original depositor metadata.
Data exclusions	Data publicly available but restricted on use were excluded. Additionally, any data that failed to meet quality control described in the Methods section, such as, contigs shorter than 5 kb, sequences failed to meet completeness thresholds or failed during the contamination assessment were excluded.
Reproducibility	All analyses were performed utilizing publicly accessible data from IMG/M and GTDB databases, and we have provided versions and parameters for computational tools that allows the reproducibility of our study. All quality thresholds and filtering criteria are explicitly stated in the methods.

Randomization

Not applicable

Blinding

Not applicable

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