nature portfolio

Corresponding author(s): Song-Can Chen and Yong-Guan Zhu

Last updated by author(s): Jun 1, 2026

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 genomic data was downloaded from GTDB r214 and IMG/VR v7.0 via FTP protocol. No specific software was used to collect the data.
Data analysis	HMMER v3.2.1, Diamond v0.9.14.115, R v3.6.2, cd-hit v4.8.1, MUSCLE v5, TrimAl v1.4.rev15, IQ-Tree v2.2.0, Python ETE v3 package, PhyloBayes v4.1c, R package treeio v1.18.1, WebPlotDigitizer (https://automeries.io/), HYPHY v2.5, R ape v5.7 package, R geosphere v1.5-1.8 package, R philentropy v0.7.0, GenomeSPOT (https://github.com/cultivarium/GenomeSPOT), R randomForest v4.7-1.1, R caret v6.90, R rfPermute v2.5.1, R RFA v0.2, R caper v1.0.3 package, R stats v3.6.2, RAxML v8.2.12, MAFFT v7.407, VIBRANT v1.2.0, R castor v1.8.3, ecceTERA v1.2.4, Ranger-DTL v2.0.

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 core dataset of all retrieved sox gene clusters in bacteria and archaea, alongside with their metadata information, and the gene context of sox, rDsr, and sHdr

system, and the predicted oxygen requirement, optimal pH, temperature and salinity in sox-hosting microorganisms are provided in Supplementary Information. The fasta file containing all bacterial, archaeal, and viral protein sequences was deposited in https://github.com/RHarperR/fastafile. Multiple sequence alignments and phylogenetic trees of SoxBYZ, SoxAX, SoxCD, rDsr and sHdr, as well as those of the species tree are deposited in https://github.com/RHarperR/treefile.

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	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

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 explored the evolutionary history of genetic systems involved in dissimilatory sulfur oxidation process.
Research sample	The study relied on the genomic datasets of prokaryotes and viruses collected from GTDB r214 database and IMG/VR v7.0.
Sampling strategy	No sample size calculation was conducted, as the study utilized the most comprehensive available genome datasets at the time of analysis to maximize phylogenetic and genomic representation.
Data collection	The genomic datasets were downloaded from GTDB r214 and IMG/VR database via FTP protocols.
Timing and spatial scale	The genomic datasets from GTDB r214 and JGI/IMG comprise globally distributed species, with samples collected across diverse geographic regions from the advent of next-generation sequencing (mid-2000s) to the present.
Data exclusions	All genomic data was utilized for screening of genes involved in sulfur oxidation, which were further used for phylogenetic reconstruction.
Reproducibility	All analyses including the phylogenetic reconstruction, molecular dating, and machine learning are reproducible.
Randomization	In the random forest analysis, the allocation of species was not random, but based on the presence of truncated Sox system, SoxCD, and/or rDsr system.
Blinding	Blinding is not a standard or necessary practice in molecular evolution studies because genomic analyses are objective, computational, and reproducible. The field relies on statistical rigor and independent validation rather than subjective interpretation, making blinding irrelevant.

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

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.