nature portfolio

Corresponding author(s): Daan Speth

Last updated by author(s): Oct 15, 2024

# 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 | All sequencing data used in this study was downloaded from NCBI SRA using fasterq-dump (version 2.11.0), as included in the SRA toolkit (version 2.11.0) |
| Data analysis | web:<br>IMNGS https://www.imngs.org/<br>Silva-ACT https://www.arb-silva.de/aligner/<br>iToL https://itol.embl.de/<br>Fasttree (no version provided in Silva-ACT web interface)<br>RaxMLMUSCLE74 (version 3.8.31 (no version provided in Silva-ACT web interface)<br><br>stand-alone:<br>BLASTn (version 2.6.0)<br>coverM (version 0.7.0)<br>Spades (version 3.15)<br>Bandage (version 0.9.0)<br>MEGAHIT (version 1.2.9)<br>anvi'o (version 7.1)<br>anvi'o (development version)<br>prodigal (version 2.6.3) |

DIAMOND (version 2.0.14.152)
UpSetR (version 1.4.0)
fastANI (version 1.33)
ezAAI (version 1.2.2)
pseudofinder (version 1.1.0)
prokka (version 1.14.6)
BLASTp (version 2.12.0+)
BLASTx (version 2.12.0+)
inStrain (version 1.9.0)
cutadapt (version 3.7)
trim-galore (version 0.6.10)
Phyloflash (version 3.4.2)
idba-ud (version 1.1.3)
MUSCLE (version 3.8.31)
IQ-tree (version 2.2.0)
ModelFinder (integrated in IQ-tree, no separate version provided)
UFBoot2 (integrated in IQ-tree, no separate version provided)
hmmer (version 3.3.2)

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

The sequencing data used to construct the symbiont genomes described in this study is publicly available in Genbank under BioProject numbers PRJEB36523 [https://www.ncbi.nlm.nih.gov/bioproject/PRJEB36523/] and PRJNA512237 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA512237/]. The successfully reconstructed genomes are available in Genbank under Bioproject number PRJNA1073475 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1073475/] and accession numbers DAWERM000000000, DAWERN000000000, and DAWERO000000000. The previously reconstructed genome that we identified as Candidatus Azosocius agrarius is available in Genbank under accession number CP066692 [https://www.ncbi.nlm.nih.gov/nuccore/CP066692.1/]. Metatranscriptome sequencing data used was publicly available in Genbank under BioProject accession number PRJEB28738 [https://www.ncbi.nlm.nih.gov/bioproject/PRJEB28738/]. 16S rRNA gene amplicon data was accessed through the IMNGS web server (https://www.imngs.org/). Data files for phylogenetic trees discussed in the manuscript, as well as the anvi'o generated genome annotation files used to generate Supplementary Data 2 are available on figshare at https://figshare.com/projects/Groundwater_Azoamicaceae_phylogenies/205756

## 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.*<br>*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.*<br>*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).*<br>*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.)*<br>*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 aimed to discover respiratory endosymbionts related to Candidatus Azoamicus ciliaticola, to assess the metabolic capabilities and environmental distribution of these organisms. |
| Research sample | Samples used in this study were groundwater samples described in previous studies. The details of each sample are available under the biosample accession numbers associated with the bioproject accession numbers listed in the data availability statement. |
| Sampling strategy | Available sequence data were screened for the presence of a gene encoding a NTT type nucleotide transporter that can be considered diagnostic for this type of symbiosis. If reads matching this NTT type transporter could be detected the dataset was retained for further processing. |
| Data collection | Data used was publicly available, and groundwater datasets were selected based on sample type assignment provided by the researchers originally submitting the data. |
| Timing and spatial scale | Our goal of environmental detection and evolutionary analyses of endosymbiont genomes allowed us to use samples taken globally, at any time. We make no claims that these endosymbionts are never present in the environments or locations we could not detect them with available data. |
| Data exclusions | No data were excluded from the analyses |
| Reproducibility | The procedure to reconstruct the genomes from the publicly available sequence data is described in the methods section. |
| Randomization | Our comparative genomics analysis aimed at the evolutionary history of this group of organisms did not require us to divide the available data into randomized groups |
| Blinding | Our comparative genomics analysis aimed at the evolutionary history of this group of organisms did not require blinding |

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.* |