nature portfolio

Corresponding author(s): Claude LEMIEUX

Last updated by author(s): Jan 17, 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	No software was used for data collection
Data analysis	All software tools used for data analysis in this study are either open source or custom scripts developed in-house. These scripts are freely available under open-source licenses and can be accessed via the following link of the Pombert Lab GitHub repositories: https://github.com/PombertLab/Publication_scripts/tree/master/2025_Pmarina_MS_scripts. Open-Source Software Quality Control and Filtering: FastQC v0.12.1: Quality assessment of short-read sequencing data. AfterQC v0.9.6 and Fastp v0.23.4: Filtering and trimming of Illumina reads. Filtlong v0.2.1: Filtering of long-read data (parameters: min_length=4000, keep_percent=90). Genome Assembly and Polishing: CANU v2.2: Long-read genome assembly. HAPO-G v1.2: Short-read polishing of assemblies. Purge Haplotigs v1.1.3: Removal of redundant haplotigs (parameter: align_cov=70). Gene Prediction and Annotation: MAKER v2.31.10, BRAKER v2.1.4, Augustus v3.3.3: Gene model prediction. RNAmmer v1.2, tRNAscan-SE v2.0: Annotation of rRNAs and tRNAs. Trinity v2.12.0: RNA-seq normalization. HiSAT2 v2.1.0: RNA-seq read mapping. Web-Apollo v2.5.0: Manual curation of gene models.

InterProScan v5: Protein domain annotation.
DIAMOND: Fast protein alignment against SwissProt/TrEMBL databases.
Assembly and Annotation Metrics:
MultiQC v1.28: Compilation of sequencing and assembly statistics.
QUAST v5.2: Assembly quality assessment.
BUSCO v5.8.2: Genome completeness evaluation using the chlorophyta_odb10 dataset.
Comparative Genomics:
OrthoFinder v2.3.11: Orthogroup inference.
UpSetR v1.4.0: Visualization of shared gene families.
SYNY v1.2: Chromosomal synteny analysis.
VarScan2 v2.4.6: Variant calling.
RepeatModeler v2.0.6, RepeatMasker v4.1.5: Repeat annotation.
EDTA v2.2.2: LTR retrotransposon annotation.
Circos v0.69-9: Integrated genome visualization.
fitdistplot (R): Allele frequency distribution plotting.
Jellyfish v2.3.1: Tool for memory-efficient counting of k-mers in DNA.
GenomeScope v2.07: Reference-free profiling of polyploid genomes.
Transcriptomics and Differential Expression:
DESeq2: Differential gene expression analysis in R.
Functional Genomics:
KEGG Mapper: Pathway analysis.
BLASTp: Protein homology searches.
DeepLoc v2.1: Subcellular localization prediction.
HMMER: CAZyme identification via dbCAN3.
Phylogenetics:
MUSCLE v3.7, TrimAl v1.3, Phyutility v2.2.6: Sequence alignment and trimming.
IQ-TREE v2.1.3: Phylogenetic inference with ultrafast bootstrap.

Custom Scripts
A2BG pipeline: Functional annotation of predicted proteins.
SSRG pipeline: Variant calling and ploidy inference.
keep_longest_reads.pl v0.9b and read_len_plot.py v0.5g: Visualization of long-read metrics.
check_for_telomeres.pl v0.4b: Chromosome completeness assessment.
gene_expressed.pl v0.5: Gene expression quantification.
kmer_ploidy_counts.sh: shell script used to assess genome ploidy by counting k-mers.

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 Pseudoscourfieldia marina genomes were deposited in NCBI under the following accession numbers:
Pseudoscourfieldia marina CCMP1203, principal haplotype (BioProject: PRJNA849025; BioSample: SAMN29039285; Accession number: GCA_049487715.1)
Pseudoscourfieldia marina CCMP1203, alternate haplotype (BioProject: PRJNA1090107; BioSample: SAMN29039285; Accession number: GCA_049487705.1)
Pseudoscourfieldia marina UIO007 (BioProject: PRJNA1014779; BioSample: SAMN37341342; Accession number: GCA_049488355.1)

Sequencing data were deposited in the NCBI Sequence Read Archive under the following accessions:

Pseudoscourfieldia marina CCMP1203
SRR28475442, RNAseq Illumina NovaSeq sequencing
SRR28475443, Random Oxford Nanopore sequencing (MinION_2)
SRR28475444, Random Oxford Nanopore sequencing (MinION_1)
SRR28475445, Random Illumina NovaSeq sequencing
SRR33318468, Replica S1 of RNAseq AVITI sequencing
SRR33318467, Replica S2 of RNAseq AVITI sequencing
SRR33318466, Replica S3 of RNAseq AVITI sequencing

Pseudoscourfieldia marina UIO007
SRR28476104, RNAseq Illumina NovaSeq sequencing
SRR28476105, Random Oxford Nanopore sequencing
SRR28476106, Random Illumina NovaSeq sequencing
SRR33319090, Replica S1 of RNAseq AVITI sequencing
SRR33319089, Replica S2 of RNAseq AVITI sequencing
SRR33319088, Replica S3 of RNAseq AVITI sequencing

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	This study involved no human participant.
Reporting on race, ethnicity, or other socially relevant groupings	This study involved no human participant.
Population characteristics	This study involved no human participant.
Recruitment	This study involved no human participant.
Ethics oversight	This study involved no human participant.

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

Life sciences study design

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

Sample size	Pseudoscourfieldia marina is a unicellular green alga with a limited number of strains currently maintained in public culture collections. For this study, we selected two genetically distinct strains, CCMP1203 (from the National Center for Marine Algae and Microbiota, USA) and UIO007 (from the Norwegian Culture Collection of Algae, Norway), to capture representative genomic and transcriptomic diversity within the species. While additional strains exist, these two were chosen based on availability, prior characterization, and suitability for comparative genomic analysis. The inclusion of both strains provides sufficient biological context to address the study's objectives, particularly regarding genome architecture, gene expression, and evolutionary insights.
Data exclusions	No data were excluded from the genomic or transcriptomic analyses. All sequencing reads passing quality control thresholds were retained and incorporated into downstream analyses. For long-read sequencing, fragments shorter than 4 kb were filtered out using Filtlong to improve assembly quality, as per standard practice. In transcriptomic analyses, genes with low read counts across all replicates were excluded during DESeq2 normalization, following recommended guidelines to ensure statistical robustness.
Replication	All experimental procedures were performed with appropriate replication. For transcriptomic analyses, three biological replicates were used for each Pseudoscourfieldia marina strain (CCMP1203 and UIO007), ensuring reproducibility and statistical robustness in differential gene expression analyses. Genomic sequencing was conducted independently for both strains, and assembly and annotation pipelines were applied consistently. Custom scripts and bioinformatic tools were run using standardized parameters across datasets to ensure replicable results.
Randomization	Randomization was not applicable to this study. The research involved comparative genomic and transcriptomic analyses of two defined Pseudoscourfieldia marina strains obtained from culture collections. Experimental conditions were standardized across replicates and strains to minimize variability, but random assignment was not relevant given the fixed identity and biological nature of the samples.
Blinding	Blinding was not applicable to this study. The research involved genomic and transcriptomic analyses of two distinct Pseudoscourfieldia marina strains obtained from culture collections. Sample identity was inherently known throughout experimental procedures due to the nature of strain-specific comparisons and the requirement for accurate genome assembly and annotation. Data processing and analysis pipelines were standardized and automated to minimize bias.

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