nature portfolio

| Corresponding author(s): | Chen. Chen. |
|----------------------------|--------------|
| Last updated by author(s): | Jul 10, 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.

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| \sim . | | | | |
|----------|-----|------|----|--------|
| St | าลา | TIC. | tı | \sim |

| n/2 | Cor | nfirmed |
|-----|-----|--|
| n/a | COI | mmed |
| | X | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| | X | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| | X | 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. |
| X | | A description of all covariates tested |
| x | | 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> |
| X | | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| | x | 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 experiments were performed based on a dataset we created called PlantSeqs, whose original sequences were manually retrieved from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide).

Data analysis

The alignment and tree inference tools used MAFFT v.7 (https://mafft.cbrc.jp/alignment/server/index.html) and RAxML v.8.2 (https://github.com/stamatak/standard-RAxML), respectively.

Other data analyses including novelty detection and taxonomic classification used Python v3.11.9 (https://www.python.org/), PyTorch v2.0.1 (https://pytorch.org/), transformers v4.42.3 (https://huggingface.co/docs/transformers/installation), NumPy v1.24.3 (https://www.numpy.org/), pandas v2.0.2 (https://pandas.pydata.org/), scikit-learn v1.4.2 (https://scikit-learn.org/stable/), SciPy v1.10.1 (https://scipy.org/), seaborn v0.13.2 (https://seaborn.pydata.org/) and matplotlib v3.7.1 (https://matplotlib.org/).

The code developed in this manuscript is deposited in Github (https://github.com/danruod/BERTPhylo).

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

PlantSeqs dataset is available at: https://drive.google.com/drive/folders/1wAQVjLYqIRA_0Xk9I3XvOsh_A-sdb9zZ.

The experimental data supporting the findings of this study can be obtained through the submitted source date file.

Research involving human participants, their data, or biological material

Policy information about studies with <u>human participants or human data</u>. See also policy information about <u>sex, gender (identity/presentation)</u>, <u>and sexual orientation</u> and <u>race</u>, <u>ethnicity and racism</u>.

| Reporting on sex and gender | Not applicable, the research does not involve human participants. | |
|--|---|--|
| Reporting on race, ethnicity, or other socially relevant groupings | Not applicable, the research does not involve human participants. | |
| Population characteristics | Not applicable, the research does not involve human participants. | |
| Recruitment | Not applicable, the research does not involve human participants. | |
| Ethics oversight | Not applicable, the research does not involve human participants. | |

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 resear | ch. 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

The original sequences for PlantSeqs were manually retrieved from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/). We collected sequences under the Embryophyta (land plants) from nine most commonly used molecular markers: atpB, matK, rbcL, rpl32-trnL, trnL-trnF, trnH-psbA, 5S rRNA, 28S rRNA, and ITS, covering the coding genes (CDS) and intergenic spacer (IGS) regions of chloroplast DNA, and nuclear DNA. To ensure that each taxon has a sufficient number of sequences for training and evaluation, we selected the 150 genera with the largest number of sequences to generate the PlantSeqs dataset, which ultimately contained 157,742 nucleotide sequences from 19,887 species.

Data exclusions

For novelty detection and taxonomic classification evaluations of BERTPhylo, we analyzed all sequences in the test set of PlantSeqs without excluding any data.

For the example tree in Figure 2 of the manuscript, we selected species with at least five (out of nine) molecular marker records from the training set of PlantSeqs to ensure sufficient number of sequences to build a tree.

For the resolution of phylogenetic trees constructed with high-attention regions, we selected orders containing at least two genera based on the sequences used in Figure 2 of the manuscript.

For attention heatmap analysis of BERTPhylo, we randomly selected 1,000 sequences for each molecular marker from the test set of PlantSeqs to eliminate the effect of imbalanced number of sequences for molecular markers.

Replication

We provide the model parameters of BERTPhylo and the experimental results can be reproduced based on the model and evaluation code. The model parameters and code are available through the Github repository: https://github.com/danruod/BERTPhylo.

Randomization

To ensure that each taxon has a sufficient number of sequences for training and evaluation, we selected the 150 genera with the largest number of sequences to generate the PlantSeqs dataset. For the top 100 most abundant genera, 50 and 100 samples were randomly selected to form the validation and test sets, respectively, with the remaining samples forming the training set. For the other 50 genera, 100 samples were downsampled to form an additional part of the test set.

Blinding

Not applicable; we are not making a comparison between different groups.

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 | Methods |
|--|--|
| n/a Involved in the study | n/a Involved in the study |
| X Antibodies | ChIP-seq |
| x Eukaryotic cell lines | Flow cytometry |
| Palaeontology and archaeology | MRI-based neuroimaging |
| X Animals and other organisms | |
| Clinical data | |
| Dual use research of concern | |
| ☐ X Plants | |
| Plants | |
| Seed stocks Not applicable; The sequence | tes for PlantSeqs were manually retrieved from the NCBI nucleotide database. |
| Novel plant genotypes Not applicable; The sequence | ces for PlantSeqs were manually retrieved from the NCBI nucleotide database. |
| Authentication Not applicable; The sequence | ces for PlantSeqs were manually retrieved from the NCBI nucleotide database. |