nature portfolio

Corresponding author(s): Wuqin Xu, Pan Li, Jinfang Zheng, Guangyong Chen.

Last updated by author(s): May 28, 2025

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	All experiments were performed based on three datasets: 1.Simulated datasets, which were generated using Seq-Gen v1.3.5 (https://github.com/rambaut/Seq-Gen). 2.Plant dataset, whose raw sequences were manually retrieved from the NCBI Nucleotide Database (https://www.ncbi.nlm.nih.gov/nucleo).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. 3.Bordetella dataset, whose raw sequences were manually downloaded from https://bigsdb.pasteur.fr/bordetella/.
Data analysis	The alignment and tree inference tools used MAFFT v.7 (https://mafft.cbrc.jp) and RAxML v.8.2 (https://github.com/stamatak/standard-RAxML), respectively. The Robinson-Foulds (RF) distance are calculated using ETE3(http://etetoolkit.org). 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), 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 and pretrained model weights are deposited in Github: https://github.com/danruod/PhyloTune.

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

Simulated datasets are available at https://github.com/danruod/PhyloTune/blob/main/data/simulated.zip.
Plant dataset is available at https://github.com/danruod/PhyloTune/blob/main/data/Plant.zip.
Bordetell dataset is available at https://github.com/danruod/PhyloTune/blob/main/data/Bordetella.zip.
The experimental data supporting the findings of this study can be obtained through Source Date files.

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	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 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	1. Simulated dataset: datasets of different sizes are generated according to the experimental settings, and the largest dataset contains 100 sequences. 2. Plant dataset: contains 157,742 nucleotide sequences from 19,887 species. 3. Bordetella dataset: contains 3,011 nucleotide sequences from 21 species.
Data exclusions	For novelty detection and taxonomic classification evaluations of PhyloTune, we analyzed all sequences in the test set of the Simulated/Plant/ Bordetella datasets without excluding any data. For the example tree in Figure 2d of the manuscript, we selected species with at least five (out of nine) molecular marker records from the training set of the Plant dataset to ensure sufficient number of sequences to build a tree. For the resolution of phylogenetic trees constructed with high-attention regions for Plant dataset, we selected orders containing at least two genera based on the sequences used in Figure 2d of the manuscript. For attention heatmap analysis of PhyloTune for Plant dataset, we randomly selected 1,000 sequences for each molecular marker from the test set to eliminate the effect of imbalanced number of sequences for molecular markers.
Replication	We provide the model parameters of PhyloTune 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/PhyloTune.
Randomization	1. Simulated dataset: For each randomly generated ground-truth tree, we simulate sequence data ten independent times to minimize the influence of random variation and ensure robustness of the evaluation. 2. Plant dataset: Training, validation, and test sets are randomly sampled according to predefined proportions, stratified by taxonomic rank to maintain representative distributions. 3. Bordetella dataset: Training, validation, and test sets are randomly sampled based on proportional stratification, ensuring consistent representation across taxonomic ranks.
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

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	Not applicable; The sequences for the Plant dataset were manually retrieved from the NCBI nucleotide database.
Novel plant genotypes	Not applicable; The sequences for the Plant dataset were manually retrieved from the NCBI nucleotide database.
Authentication	Not applicable; The sequences for the Plant dataset were manually retrieved from the NCBI nucleotide database.