

Corresponding author(s): Jamie Thompson, jbt38@bath.ac.ukLast updated by author(s): Jun 29, 2023

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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒	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 <i>Give P values as exact values whenever suitable.</i>
☐	☒	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	OneTwoTree was used to collect nucleotide data (https://doi.org/10.1111/1755-0998.12927). Spatial data were sourced from the Global Biodiversity Information Facility (https://www.gbif.org/). Spatial variables were sourced from various commonly-used datasets, including https://doi.org/10.1038/sdata.2018.40 , https://github.com/ISRICWorldSoil/SoilGrids250m and https://figshare.com/articles/dataset/Global_Aridity_Index_and_Potential_Evapotranspiration_ET0_Climate_Database_v2/7504448/3 . Details are available in the manuscript, with links present in the code. Further variables collected manually were from books and a previous publication described in the manuscript (e.g. Anderson, 2000 and Hernandez-Hernandez et al., 2014).
Data analysis	OneTwoTree was used to assemble the first alignments, which were cleaned with AMAS (https://github.com/marekborowiec/AMAS), and RAXML V8 was used to reconstruct the phylogeny (https://cme.h-its.org/exelixis/web/software/raxml/). TreePL estimated divergence timings (https://github.com/blackrim/treePL) and BAMM estimated diversification rates (http://bamm-project.org/). R version 4.1.3 was used to analyse resulting data, with packages too numerous to name here but including XGBoost (https://cran.r-project.org/web/packages/xgboost/index.html) and diversitree (https://cran.r-project.org/web/packages/diversitree/index.html). Details on R packages can be found in the manuscript and in the code.

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](#)

All data, including accession numbers used to reconstruct the phylogeny, and variable data, is available at <https://github.com/jamie-thompson/cactaceae>. Further sources of data are cited within the manuscript text.

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

We used Machine Learning and phylogenetic models to explain the complexity underlying the diversification of Cacti (family Cactaceae). We reconstructed a large phylogeny with Maximum Likelihood, and time-calibrated this with robust estimates of lineage age. BAMM-estimated tip speciation rates were analysed with Machine Learning (XGBoost), with a dataset of 39 variables of potential relevance. Following this, phylogenetic models (BiSSE, QuaSSE and HiSSE) confirmed the role of each powerful variable and visualised the relationship with diversification rate.

Research sample

A dataset of ~1,000 cactus species were analysed in a phylogenetic framework, to reveal and rank the most powerful drivers of diversification. The phylogeny was created from publicly available nucleotide data and publicly available data were used to analyse their relationships with diversification rate.

Sampling strategy

All available data was used at every stage. All publicly available nucleotide data were used to reconstruct the phylogeny. All publicly available distribution data were used for estimating median latitude and 33 other variables. The remaining variables, for which data

were collected largely manually, maximised species within the tree.

Data collection Data collection utilised publicly available data (e.g. Genbank sequences, GBIF coordinates, published size estimates). All data were collected by JBT, GBIF data were cleaned by THH and range size estimated by THH.

Timing and spatial scale Data were collected between January 2021 and September 2022. The spatial extent was the Americas, where cacti are naturally present.

Data exclusions Some nucleotide sequences were excluded by JBT in the alignment stage, if they belonged to a locus present in very few species, or if they were represented by a better sequence for that locus in the same species. Some GBIF coordinates were excluded in the curation stage by THH, if they fell outside natural ranges as described by experts.

Reproducibility Each stage was run several times over the course of analysis. The Machine Learning code was replicated by JBT and GMK. The QuaSSE code was replicated in a few cases to ensure stability of results. BISSE, HiSSE and BAMM used standard procedures, with frameworks used multiple times within our lab and beyond. Furthermore, the XGBoost code framework we used has been used several times, including in a Nat Comms paper (<https://doi.org/10.1038/s41467-020-16498-w>). Code for extracting variables and running XGBoost models are available at <https://github.com/jamie-thompson/cactaceae>.

Randomization This was not relevant for our study, which investigated relationships between variables and diversification rates.

Blinding Blinding is not relevant, since we analysed relationships between variables and diversification rates.

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