Supplementary Information

palaeoverse: a community-driven R package to support palaeobiological analysis

—

Lewis A. Jones[1], William Gearty[2], Bethany J. Allen[3,4], Kilian Eichenseer[5], Christopher D. Dean[6], Sofía Galván[1], Miranta Kouvari[6,7], Pedro L. Godoy[8,9], Cecily Nicholl[6], Lucas Buffan[10], Erin M. Dillon[11,12], Joseph T. Flannery-Sutherland[13], and Alfio Alessandro Chiarenza[1]

—

[1]Centro de Investigación Mariña, Grupo de Ecoloxía Animal, Universidade de Vigo, 36310 Vigo, Spain.

[2]Division of Paleontology, American Museum of Natural History, New York, NY, 10024 USA.

[3]Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland.

[4]Computational Evolution Group, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

[5]Department of Earth Sciences, Durham University, South Road, DH1 3LE, Durham, United Kingdom.

[6]Department of Earth Sciences, University College London, Gower Street, WC1E 6BT, London, United Kingdom.

[7]Life Sciences Department, Natural History Museum, Cromwell Road, SW7 5BD, London, United Kingdom.

[8]Laboratório de Paleontologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, 14040-901 Brazil.

[9]Department of Anatomical Sciences, Stony Brook University, Stony Brook, NY, 11794 USA.

[10]Département de Biologie, École Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 69342 Lyon Cedex 07, France.

[11]Smithsonian Tropical Research Institute, Balboa, Republic of Panama.

[12]Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, CA 93106, USA.

[13]School of Earth Sciences, University of Bristol, BS8 1RL, Bristol, UK

—

Corresponding author: LewisAlan.Jones@uvigo.es

Survey

We conducted an online public survey to collect the opinions of the palaeobiological community, and determine which tools were most needed. The survey was advertised via social media (i.e. Twitter) and a mailing list (not included here for the sake of privacy). The survey was opened on the 26th of May 2022 via the following link (http://www.tinyurl.com/palaeoverse). For the purpose of this manuscript, answers were collected until the 9th of August, 2022. However, the survey remains open for the community to respond to. In total, 35 participants from 14 different countries completed the survey, most of which are affiliated with European and North American institutions. This observation likely reflects a geographical bias in the reach of the survey, and demonstrates the additional effort that will need to be made to reach further afield in the future.

Below, we include the conducted survey for the sake of completeness:

---

palaeoverse: towards a community-driven R package

palaeoverse is an R package being developed by palaeobiologists, for palaeobiologists.

The aim of palaeoverse is to generate a community-driven software package of generic functions for the palaeobiology community. The package does not aim to provide implementations of statistical approaches, rather it provides auxiliary functions to help streamline analyses, and improve code readability and reproducibility.

As part of the initial development of palaeoverse, we would like to hear from you, the palaeobiology community! What generic functions do you wish existed to streamline your work? What takes up more of your precious time than it should? For example, time and spatial binning of data, palaeorotating fossil occurrences, checking species names for errors… this kind of thing! Let us know what functions you think would be useful for you, and for the wider community.

1. Name (short answer text)
2. Email address (short answer text)
3. Affiliation (long answer text)
4. Do you wish to join our mailing list, and receive updates on palaeoverse? (If so, please include your email above) (select one of the following)
    - Yes
    - No

5. What types of palaeontological data do you typically use or are interested in? (Multiple choice)
    - Taxonomic identifications
    - Age or biozone
    - Geographic information (modern or at time of deposition)
    - Geological/palaeoenvironmental context
    - Taxon abundance
    - Taphonomy
    - Trait values, classifications or descriptions
    - Phylogenies
    - Other
6. What R packages (or other tools) have you used previously to clean or explore palaeontological data sets? (long answer text)
7. What kinds of tasks do you typically carry out when cleaning and exploring palaeontological data? (long answer text)
8. Please specify and detail generic functions that you feel would be useful to include in the palaeoverse package. If you have suggestions for more than three functions, please submit a second form.
    - Function 1 (long answer text)
    - Function 2 (long answer text)
    - Function 3 (long answer text)
9. Do you have pre-existing code (for the above functions or other functions) you would like to contribute to palaeoverse? (select one of the following)
    - Yes
    - No
10. If yes, what function is your code performing? (long answer text)

Survey responses

Below we provide a summary of the responses to survey questions not documented in the main text (6–8). Some questions (1–4) are omitted for the sake of privacy. Where appropriate, we grouped survey responses into the following distinct categories to aid summary:

- Data access

- Checking and transforming data
- Plotting data
- Time binning
- Spatial analyses
- Tree modification and plotting
- Phylogenetic analyses
- Other analyses

Responses to questions related to the tasks participants usually carry out

Checking and transforming data

- Transforming data into portable structures
- Organising based on multiple data types
- Grouping things by character state
- Restructuring and filtering
- Cross-referencing PBDB with other datasets for taxonomy errors (synonyms/misspellings)
- Checking for occurrences outside expected taxon duration, or freshwater species in marine sediments
- Checking that column names are identical between imported datasets from the literature
- Reshaping into tidy/long format
- Transforming species abbreviations into full species names
- Formatting and matching data with trees
- Processing age/taxon information for phylogenetic analysis
- Merging files using specific specimen ids, cleaning repeat specimens, fixing age ranges across time intervals, updating species names and connecting to past species names
- Checking taxonomic assignments/spellings, verifying geochronology and environmental setting, checking for missing data, reviewing spatial distribution, reviewing data summaries by taxonomic group
- Refining chronostratigraphy for PBDB collections
- Standardising taxonomy
- Surveying and cleaning biostratigraphic data on PBDB
- Taxonomic harmonisation, filtering non-pollen and converting to percentages

- Aligning multiple ecological abundance, paleoenvironmental and depositional datasets (sometimes from different csv files, collected by different people or from different core intervals) in time/space and with age-depth model outputs
- Calculating derived variables from raw data
- Reformatting dataframes for various analyses/packages which might require data to be formatted in a different way, especially when some variables are measured per sample (relative abundance) and others are measured per individual specimen (taphonomy scores)
- Tidying taxonomic identifications
- Trimming datasets above a threshold

Plotting data

- Exploratory figures
- Plotting census data as a diagram to see faunal trends, nMDS, etc
- Skimming through the dataset and plotting temporal/morphological data
- Data visualisation
- Visualising variables to look for distributions, outliers, and preliminary trends
- Overlaying multiple paleo time series plots with aligned axes

Time binning

- Temporal subsetting/assignment
- Binning data by time period
- Time binning with variable binning rules (midpoint, range-through, random age assignment, exclusion of multi-bin spanning data)
- Binning
- Time binning

Tree modification and plotting

- Stitching two phylogenetic trees together
- Acquiring FADs and LADs, matching taxon names to tree tip labels, reading character data with missing data or polymorphisms, rooting trees
- Tree calibration
- Time-calibrating phylogenies
- Converting TNT tree files into NEXUS format

- Tree plotting, pruning, assigning max ages or nodes, exploration of evolutionary rates and shape changes, phylogeny building, etc

Phylogenetic methods

- Estimating phylogenetic signal, phylogenetic regressions and multimodel inference, predictive modelling for paleobiological inference, ancestral state reconstructions, analysis and mapping of evolutionary rates
- Phylogenetic comparative methods and multivariate statistics, origination analyses, model fitting and ancestral state reconstructions
- Tree plotting, pruning, assigning max ages or nodes, exploration of evolutionary rates and shape changes, phylogeny building, etc

Spatial analyses

- Geographical analyses
- Spatial subsetting/assignment
- Converting extant latitude and longitude to extinct latitude and longitude
- Aggregate collections within a given radius into palaeocommunities/sites

Data Access

- GBIF guide

Other analyses

- Diversity analyses using different metrics
- Resampling and appropriate rarefaction
- Creating tip priors/contrast matrices
- Applying an age model
- Comparing morphological/isotope measurements across space and time
- Paleoenvironmental reconstructions
- Ecological coupling between aquatic and terrestrial systems
- Human-environment links
- Trend analyses
- Multi-proxy comparison/analysis
- Age modelling

- Sensitivity analyses
- Relative abundance

Responses to questions related to the functions that participants consider useful to be included in palaeoverse

Checking and transforming data

- Removing taxonomic equivalents, species with too little data and uninformative characters
- Checking for typos in taxa names, especially between two vectors; checking if taxa names are formatted as binomial entities, i.e. checking if they follow a "Genus_species" format and return an error message if not.
- Cross-matching PBDB and phylogenetic data (bonus: some functionality for visualising this data)
- Cleaning specimen description data into components when single specimen number has multiple entries
- Detection of unusual spatiotemporal occurrences of PBDB records
- Taxonomic harmonization
- Resolving to taxonomic authorities, such as Paleobiology Database or World Register of Marine Species
- Prepping datasets for RevBayes analyses
- Merging duplicate species or lumping specified species

Plotting data

- Plotting data on stratigraphic-geological timeline with appropriate colours
- Range charts
- Visualising geographic data
- Adding geologic timescales to plots in base R graphics rather than ggplot
- Plotting spindle diagrams given a tree and diversity (or abundance) data through time
- Perspective and stacked time slice disparity plots
- coord_geo (similar to deeptime R package) to add geologic timescale to a plot
- core_photograph to automatically look up core images remotely, download them, shrink them, add them to the plot
- Faunal diagram (eg % abundance of different species vs age or depth) from census data
- Heatmaps to overview faunal composition

- Stratigraphic charts
- Summary plots (e.g. richness by geologic unit, time interval, and/or taxonomic groups)
- Adding geological scales to ggplot2
- Plotting
- Overlaying multiple time series plots over an aligned time axis

Time binning

- Time binning
- Age binning
- FADs to LADs ranging with gap info
- Age binning of PBDB data
- Biostrat or geological epoch to quantitative age
- Time binning function with variable binning rules (midpoint age, assign randomised age in FAD-LAD range then bin, range-through binning)
- Translating periods/stages into up-to-date absolute ages
- Converting and organising input data according to age (across different age types - calBP, 14C, etc)
- Updating timescale ages with new chronostratigraphic works
- Tools to facilitate time binning and alignment of coeval data points across time intervals, especially when dealing with age-depth model uncertainty, different temporal resolutions, etc

Tree modification and plotting

- Tree slicing
- Treestich: attach two trees together
- Plotting unambiguous synapomorphies on a tree like MacClade used to do
- Plotting up tree spaces
- Finding centroid trees

Phylogenetic methods

- Using ranges in ancestral state estimation, rather than a single mean value
- Grouping taxa by character trait and work with character matrices easily
- Total-evidence phylogeny plot

- Performing cross-validation for predictive models in a phylogenetic context (e.g. for PGLS)
- A function that simply removes all fossil taxa based on non-ultrametricity and performs leave-n-out cross-validation
- Tree stats, Templeton Test, double decay indices, positional congruence, leaf stability
- Comparisons or tests for correlations between the diversification rates of two different clades

Spatial analyses

- Latitudinal bins assignment from palaeolatitudes
- Palaeo-ocean basin assignment from lat/long and date
- Converting long/lat with fossil age to past long/lat
- Spatial binning function using hexagonal grids
- Palaeocoordinating

Data Access

- Gathering data from the Paleobiology Database
- Processing private data into a format to be easily uploaded to Neotoma
- Read TNT files or a way to quickly and easily convert files
- Bringing TNT into R
- Data downloading functions from various databases (e.g. PBDB)
- Getting data
- Compiling a list of relevant packages that perform useful functions to increase exposure

Other analyses

- Rarefaction
- Tools for data simulation to generate null models
- Generalised differencing
- Fossil frequency
- Oldest record
- Parsing stratigraphic range data for use in phylogenetic packages and PyRate
- Age models
- Converting element descriptions to a standardize form which the user can define, using a few specific shared terms in those descriptions

- Integrating paleobiological and stratigraphic data
- Breakpoint and smooth community data analyses
- Age-depth modelling
- Assessing distributions of sampling rates among taxa
- Easy extraction of some results (e.g. clock rates) from Bayesian analyses
- Age model
- Beta diversity time-series analyses. For instance, a function that computes cumulative (from a reference point) and successive (between time points) assemblage turnover
- With pollen data, deciphering relative abundance of terrestrial vs all species

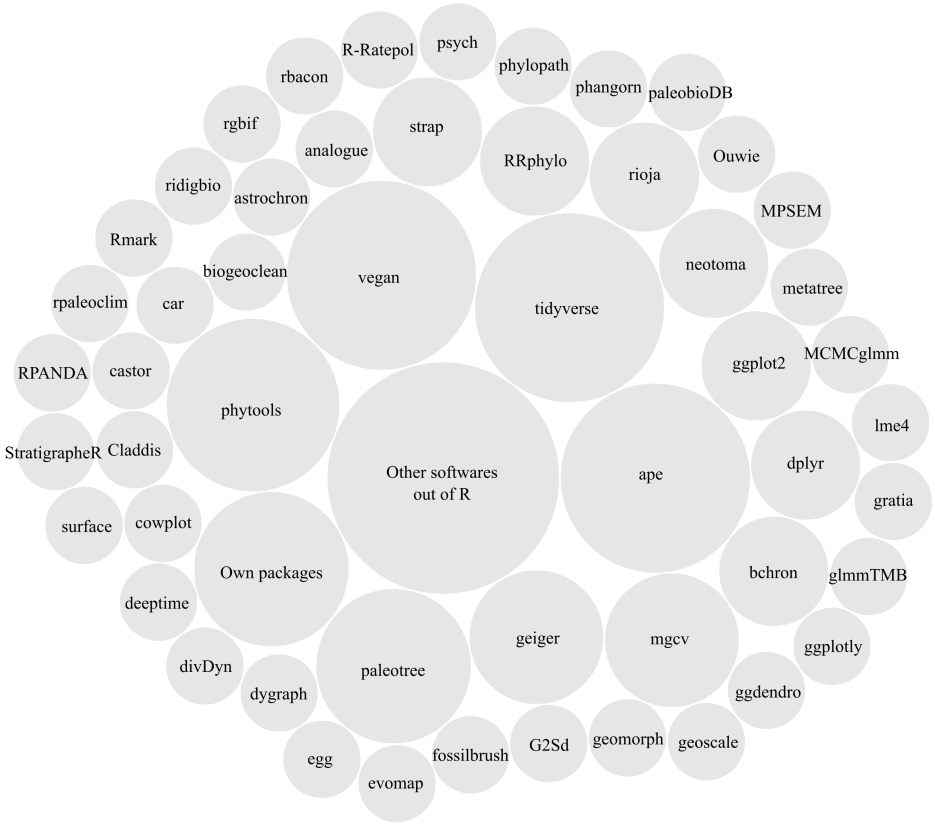

Figure 1: Summary of responses to the palaeoverse survey. Preferred tools for processing and analysing paleontological data. Both resources inside and outside the R environment and R packages are included as categories.