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

—

Lewis A. Jones1, William Gearty2, Bethany J. Allen3,4, Kilian Eichenseer5, Christopher D. Dean6, Sofía Galván1, Miranta Kouvari6,7, Emma M. Dunne8, Pedro L. Godoy9,10, Cecily Nicholl6, Lucas Buffan11, Erin M. Dillon12,13, Joseph T. Flannery-Sutherland14, and Alfio Alessandro Chiarenza1

—

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

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

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

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

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

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

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

*8Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), GeoZentrum Nordbayern, 91054 Erlangen, Germany.*

*9Laboratório de Paleontologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, 14040-901 Brazil.*

*10Department of Anatomical Sciences, Stony Brook University, Stony Brook, NY, 11794 USA.*

*11Département de Biologie, École Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 69342 Lyon Cedex 07, France.*

*12Smithsonian Tropical Research Institute, Balboa, Republic of Panama.*

*13Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, CA 93106, USA.*

*14School of Earth Sciences, University of Bristol, BS8 1RL, Bristol, UK*

—

**Corresponding author:** LewisAlan.Jones@uvigo.es

Abstract

1. The open-source programming language ‘R’ has become a standard tool in the palaeobiologists’ toolkit. Its popularity amongst the community continues to grow with published articles often citing the usage of R, and R packages. While several packages exist to support palaeobiological analysis, there is a lack of agreed standards for data preparation and of frameworks to support this. Consequently, data preparation steps are often unclear and not reproducible, even when code is provided. Moreover, due to a lack of code accessibility and documentation, palaeobiologists are often forced to ‘reinvent the wheel’ to find solutions to issues already solved by other members of the community.
2. Here, we introduce palaeoverse, a community-driven R package for quantitative palaeobiological research. The package is freely available and has three core principles: (1) streamline data preparation and analyses; (2) enhance code readability; and (3) improve reproducibility of results. During development of the package, we combined our experience with that of the wider palaeobiological community to establish what tools are needed. The latter was collated by conducting an online survey distributed via social media and email.
3. We first report the findings of the survey which shaped the development of the package. Subsequently, we describe and demonstrate the functionality available in palaeoverse and provide usage examples. Finally, we discuss the resources we have made available for the community, and the future plans for the palaeoverse project.
4. palaeoverse is the first community-driven R package in palaeobiology, developed with the intention of bringing palaeobiologists together to establish agreed standards for high-quality quantitative research. The package provides a user-friendly platform for preparing data for analysis with well-documented code to enhance transparency. The functionality available in palaeoverse improves code reproducibility and accessibility, which is beneficial for both the review process and future research.

# Keywords

Analytical Palaeobiology, Computational Palaeobiology, R programming, Readable, Reusable, Reproducible

# Introduction

Since the development of large palaeontological datasets from the 1970s, palaeontologists have increasingly adopted computational approaches to quantitatively address questions about the history of life on Earth (Sepkoski, 1978). Today, the majority of sub-disciplines within palaeontology regularly use large datasets to perform experiments *in silico*. This has initiated a ‘Golden Age’ of palaeontology (Sepkoski and Ruse, 2009), where extensive datasets of various formats are used to test macroevolutionary and macroecological hypotheses (Quental and Marshall, 2013; Mannion et al., 2014; Zaffos, Finnegan and Peters, 2017; Close et al., 2020a). The growth and increasing availability of such datasets has meant that coding is now an integral part of palaeobiology, and is widely used to clean (Zizka et al., 2019; Flannery-Sutherland et al., 2022), analyse (Guillerme, 2018; Kocsis et al., 2019), and visualise data (Bell and Lloyd, 2015), as well as build models (Silvestro, Salamin and Schnitzler, 2014; Starrfelt and Liow, 2016) and implement simulations (Fraser, 2017; Barido-Sottani et al., 2019; Furness et al., 2021; Jones et al., 2021). Whilst software has been developed in languages such as C++ (Garwood, Spencer and Sutton, 2019) and Python (Silvestro et al., 2014), the programming language R is currently the most popular in palaeobiology. This is due to the wide range of tools–in the form of R packages–available to help users analyse their data. Many of these tools are often borrowed or repurposed from ecology (Chao et al., 2014; Oksanen et al., 2020), while others have been developed to specifically handle fossil data (Lloyd, 2016; Kocsis et al., 2019).

Despite this, few packages explicitly focus on preparing data for analyses, forcing users to construct custom scripts. This can result in distinct differences in code style and practices amongst the community, including code legibility and documentation. Consequently, custom scripts can be inaccessible to other users (Filazzola and Lortie, 2022). Although increasingly requested by journals, code is not always provided as supplementary material, or made available in online repositories (e.g. GitHub, Zenodo, Dryad). A lack of accessible code can lead to research results being unreproducible, preventing future studies from extending the work. Even when code is available, it might be poorly documented or written in a way that is specific to the dataset being analysed. Such code might require extensive reworking before it can be applied to other data. Consequently, researchers often have to ‘reinvent the wheel’, putting time and effort into writing code that already exists, but is unavailable, inaccessible, and/or difficult to repurpose (Filazzola and Lortie, 2022). Such issues are exacerbated by the absence of community standards in how data should be prepared for analyses; differing approaches utilised by different researchers result in a lack of consistency between studies, making comparison between results challenging. There is therefore a well-established need for both protocols and tools for preparing palaeontological data for further analysis.

Here, we introduce the R package palaeoverse, a community-driven toolkit for streamlining palaeobiological analyses and improving code accessibility and reproducibility. Our approach differs from other palaeontological R packages in that it aims to bring the palaeobiological community together to establish consensus on the steps taken in data preparation for analysis, and how these steps should be implemented. The package contains functions developed in line with current researcher demands to cleanse, prepare and explore occurrence datasets for further analysis. These needs were established via a survey conducted by members of a new working group. The functionality of palaeoverse is purposefully flexible, and can be applied to a wide variety of occurrence datasets. In this paper, we report results from our survey, describe and detail the functionality of palaeoverse, and illustrate its features with usage examples.

# Community survey

We conducted a public survey to collect the opinions of the palaeobiological community, and determine which tools were most needed. The survey was distributed via social media (i.e. Twitter) and email, and included questions related to researchers’ previous experience, pre-existing code (to identify potential contributions), and what functionality they would considered useful in a new palaeobiological toolkit. The type of data participants (*n* = 35) typically work with, the tasks commonly carried out when working with this data, and the tools they would like to have access to are summarised in [Figure 1](#fig-survey). Our participants work with a wide range of data ([Figure 1](#fig-survey)), and described the checking and transformation of data as the most commonly employed task. A wide variety of functions were requested by survey participants, with data plotting, time binning, and data access commonly suggested ([Figure 1](#fig-survey)). Over 40% of participants also indicated that they were willing to contribute code to palaeoverse, highlighting the potential for a community-driven project. Specific details regarding the survey and responses can be found in Supplementary Material 1.

|  |
| --- |
| Figure 1: Summary of responses to the palaeoverse survey. (A) The types of palaeontological data that survey participants typically work with. (B) Tasks that respondents routinely carry out in their analyses, and the functions they would find useful in the palaeoverse package. |

# Package description

palaeoverse is a R package that provides auxiliary functions to support data preparation for palaeobiological analysis. A summary of the functions currently available in palaeoverse is provided in Table 1, with further description provided in the Features section. To demonstrate the functionality and versatility of the package, we also provide usage examples.

## Installation

The palaeoverse package can be installed from CRAN using the install.packages function in R (R Core Team, 2022):

install.packages("palaeoverse")

If the user prefers, the development version of palaeoverse can be installed from GitHub via the remotes R package (Csárdi et al., 2021):

remotes::install\_github("palaeoverse-community/palaeoverse")

Following either of these steps, palaeoverse can be loaded via the library function in R:

library("palaeoverse")

## Data

Functionality in palaeoverse was not developed to provide solutions for any one data source (e.g. the Paleobiology Database (<https://paleobiodb.org/#/>), the Geobiodiversity Database (<http://www.geobiodiversity.com>), Neptune (<https://nsb.mfn-berlin.de/>), etc.). Instead, functions are designed around the assumption that most user’s data would be stored in an occurrence dataframe. As such, this is the expected input data format for most functions in palaeoverse, and we avoid unnecessary transformation of data by the user wherever possible. Functions are therefore flexible and can be applied to various data sources with ease. In most instances, the returned object from a function is also a dataframe, which we consider the easiest data structure for most users to understand and work with. This may be undesirable for some advanced users of R, but transforming data structure should be straightforward for these users.

## Functions

A summary of the functions available in palaeoverse is provided in [Table 1](#tbl-summary). Detailed descriptions of the functions are provided herein.

Table 1: A summary table of the functions available in the palaeoverse R package

| **Function** | **Description** |
| --- | --- |
| axis\_geo | Add a geological time scale axis to a plot |
| bin\_lat | Bin fossil occurrences into latitudinal bins |
| bin\_spatial | Bin fossil occurrences into spatial bins |
| bin\_time | Bin occurrences into time bins (choice of approaches) |
| data | Datasets: ‘tetrapods’, ‘reefs’, ‘interval\_key’, ‘GTS2012’, and ‘GTS2020’ |
| group\_apply | Apply a function over user-defined groups |
| lat\_bins | Generate latitudinal bins |
| look\_up | Link user-specified interval names to the International Geological Time Scale |
| palaeorotate | Reconstruct the palaeogeographic coordinates of fossil occurrences |
| phylo\_check | Check taxon names against tips in a phylogeny and/or remove tips from the tree |
| tax\_check | Check for spelling mistakes in taxon names and flag potential issues |
| tax\_range\_geo | Calculate the geographic range of taxa (choice of approaches) |
| tax\_range\_time | Calculate and plot the temporal range of taxa |
| tax\_expand\_lat | Convert taxon latitudinal ranges to bin-level pseudo-occurrences |
| tax\_expand\_time | Convert taxon temporal ranges to interval-level pseudo-occurrences |
| tax\_unique | Calculate the number of unique taxa in a dataset of occurrences |
| time\_bins | Generate stratigraphic time bins or near-equal length time bins |

### Example datasets

Two occurrence datasets (tetrapods and reefs) are provided in palaeoverse to enable reproducible examples within function documentation. The tetrapods dataset is a compilation of Carboniferous–Early Triassic tetrapod occurrences (*n* = 5,270) from the Paleobiology Database. The dataset includes variables relevant to common palaeobiological analyses, covering the taxonomic identification of fossils and their geological, geographical and environmental context. The reefs dataset is a compilation of Phanerozoic reef occurrences (*n* = 4,363) from the PaleoReefs Database (Kiessling and Krause, 2022). This dataset includes information on the biological, geological, and geographical context of each reef. With the exception of the removal of superfluous columns and the renaming of some columns to improve clarity, both datasets are unaltered from source. Additional information on both datasets can be accessed via ?tetrapods or ?reefs, once the package is loaded.

# Load occurrence data  
data("tetrapods")  
data("reefs")

### Time bins

We have developed time\_bins to enable access to two popular Geological Time Scales (GTS): GTS2012 and GTS2020 (Gradstein et al., 2012, 2020). Both GTS2012 and GTS2020 are included in the package as reference datasets. The time\_bins function allows users to generate temporal bins at different temporal ranks (i.e. stage, epoch, period, era, or eon) using these datasets for a specified interval input:

# Get stage-level time bins  
time\_bins(interval = "Phanerozoic", rank = "stage", plot = TRUE)

|  |
| --- |
| Figure 2: Phanerozoic stage-level time bins. Plot depicts the uneveness in duration of stratigraphic time bins. Bar colour filling follows the established colour scheme of the International Commision on Stratigraphy (<https://stratigraphy.org/>). |

As is evident from [Figure 2](#fig-stages), GTS temporal bins are highly uneven in duration. Several previous studies have attempted to circumvent this issue by generating near-equal-length time bins by grouping stages towards a target bin length (Mannion et al., 2015; Close et al., 2020a). time\_bins enables users to generate near-equal-length time bins following this approach ([Figure 3](#fig-equal)) to a specified target size:

# Generate near-equal length time bins  
time\_bins(interval = "Phanerozoic", rank = "stage", size = 15, plot = TRUE)

|  |
| --- |
| Figure 3: Phanerozoic near-equal-length time bins. Plot depicts composite stratigraphic bins (grouping stage-level bins) for the Phanerozoic of a target bin size of 15 million years. **Note:** time bins are still uneven, but less so than stage-level bins. |

The appropriate set of time bins to use will depend upon the nature of subsequent analyses. Near-equal-length bins might be more desirable for calculating evolutionary rates through time, while GTS bins are defined on observed phenomena in the geological record, reflecting prior knowledge of cohesive biological units separated by some form of transition. Additional functionality in time\_bins allows the user to assign occurrences to the generated bins if absolute ages are known (e.g. from radiometric dating). However, the bespoke bin\_time function (discussed below) is likely to be the preferred option for most fossil occurrence data which often have an age range.

### Occurrence binning

Fossil occurrences are frequently ‘binned’ into distinct time intervals to enable quantification of changes (e.g. biodiversity or disparity) through geological time. The function bin\_time allows users to assign occurrences into time bins generated by the function time\_bins, or those defined by the user:

# Generate temporal bins  
bins <- time\_bins()  
# Assign occurrences to bins  
bin\_time(occdf = tetrapods, bins = bins, method = "mid")

Whilst binning occurrences with tightly defined temporal limits is straightforward, those with poorly constrained maximum and minimum ages can span several intervals, and therefore cannot be easily assigned to a single bin. Palaeontologists have identified numerous solutions to try and address this problem (Lloyd et al., 2012; Silvestro et al., 2016; Davies et al., 2017; Dean, Chiarenza and Maidment, 2020; Franeck and Liow, 2020), but there is currently no consensus on the best methodological approach and subsequent implementation. The bin\_time function provides five approaches defined by the ‘method’ argument: ‘mid’ (assigned based on the midpoint of the temporal range of the occurrence), ‘majority’ (assigned to the bin which covers the majority of the temporal range of the occurrence), ‘all’ (assigned to all bins within the temporal range of the occurrence), ‘random’, (assigned randomly to bins (with equal probability) within the temporal range of the occurrence, repeated up to assigned ‘reps’), and ‘point’ (assigned randomly from a uniform distribution over the temporal range of the occurrence, repeated up to assigned ‘reps’). We hope that formally including these options within the bin\_time function will encourage palaeontologists to routinely explore and compare the outcomes of various binning approaches with ease.

In recent years, palaeobiology has developed a heightened interest in the spatial structure of the fossil record, with studies focused on understanding the spatial distribution of biodiversity, and the processes that drive them (Vilhena and Smith, 2013; Antell et al., 2020; Close et al., 2020b; Chiarenza et al., 2022; Flannery-Sutherland, Silvestro and Benton, 2022; Jones et al., 2022). In order to support such analyses, bin\_spatial has been developed for palaeoverse. The function allows the user to assign occurrence data into equal-area grid cells using discrete hexagonal grids via the h3jsr package (O’Brien, 2022). Additional functionality allows simultaneous assignation of occurrence data to cells of a finer-scale (i.e. a ‘sub-grid’) within the primary grid. This might be desirable for users to evaluate differences in the amount of area occupied by occurrences within their primary grid cells. This functionality also allows the user to rarefy across sub-grid cells within primary cells to further standardise spatial sampling.

# Assign data to equal-area spatial bins  
bin\_spatial(occdf = reefs, spacing = 250)  
bin\_spatial(occdf = reefs, spacing = 250, sub\_grid = 50)

Understanding the latitudinal distribution of biodiversity in deep time has also gained research interest in recent years (Powell, 2009; Mannion et al., 2012, 2014; Allen et al., 2020; Song et al., 2020; Jones et al., 2021). To ease implementation of such analyses, we have developed two functions lat\_bins and bin\_lat which can be used to generate latitudinal bins of a given size, and assign occurrence data to those respective bins.

# Generate latitudinal bins  
bins <- lat\_bins(size = 15)  
# Assign occurrences to bins  
bin\_lat(occdf = tetrapods, bins = bins)

### Palaeogeographic reconstruction

Using the present-day coordinates of fossil occurrences, plate rotation models can be used to reconstruct their location at time of deposition. Existing fossil databases provide reconstructed coordinates for occurrences from only one or two of the many plate rotation models available (if any), and it is not always transparent which model (or version of the model) has been used. This lack of transparency is reflected in some published articles that only cite the use of the GPlates (Boyden et al., 2011; Müller et al., 2018) to reconstruct palaeocoordinates, yet lack specifics on which plate rotation model was used with the GPlates web service or desktop software. Furthermore, the uncertainty in palaeogeographic reconstructions is underappreciated; reconstructed coordinates are often treated as being well-established, rather than model-based estimates. Finally, both published and unpublished data (e.g. museum specimens) exists outside of these databases which researchers might require palaeocoordinates for.

We have developed the function palaeorotate to address these shortcomings. The function allows palaeocoordinates to be reconstructed within R using two different approaches (‘point’ and ‘grid’). The first makes use of the GPlates Web Service, and allows point data to be rotated to specific ages using the available models (see <https://gwsdoc.gplates.org>). The second approach uses reconstruction files of pre-generated palaeocoordinates to spatiotemporally link occurrences’ modern coordinates and age estimates with their respective palaeocoordinates. These rotation files were generated using a 1º x 1º spatial grid, and allows palaeocoordinates to be generated efficiently for large datasets. Additional functionality when using the reconstruction files allows the user to calculate the palaeolatitudinal range between reconstructed coordinates, as well as the great circle distance between the two most distant points (i.e. palaeogeographic uncertainty).

# Add midpoint age for rotation  
tetrapods$age <- (tetrapods$max\_ma + tetrapods$min\_ma) / 2  
# Palaeorotate occurrences and return uncertainty  
palaeorotate(occdf = tetrapods, method = "grid", uncertainty = TRUE)

### Taxon-related features

When working with large occurrence datasets, errors can easily creep into data. One frequently encountered issue is spelling variations of the same taxon name. This can have undesirable consequences when calculating metrics such as taxonomic richness or abundance. The tax\_check function computes character string distances between taxonomic names via the heuristic Jaro distance metric (Jaro, 1989). This metric provides a measure of dissimilarity between character strings of 0 (exact match) to 1 (completely dissimilar). During function call, the user defines a threshold for string dissimilarity to identify potential synonyms. In tax\_check, Jaro distances are calculated via the stringdistmatrix function from the stringdist package (van der Loo, 2014). While this function helps researchers to perform a spell check on their dataset, this function is no replacement for taxonomic vetting.

# Check for taxonomic errors  
tax\_check(occdf = tetrapods, name = "genus")

The function tax\_unique is provided to improve the accuracy of richness estimates from fossil occurrence data. Palaeobiologists routinely discard occurrences not identified to their desired taxonomic resolution. For example, if analysis is conducted at species-level, occurrences identified to genus level (or above) are discarded from the dataset. However, these occurrences can represent unique species, and their removal can impact richness estimation. The tax\_unique function reduces the amount of unique taxa being discarded by retaining fossils which are identified to a coarser taxonomic resolution than the desired level, but must represent a clade not already in the filtered dataset. For instance, with three fossil occurrences identified as *Tyrannosaurus rex*, *Spinosaurus aegyptiacus*, and Diplodocidae indet., the latter would be discarded under species-level analysis (i.e. a species richness of two). However, this occurrence clearly represents a different species to the two already present in the dataset. Using tax\_unique, Diplodocidae is treated as an additional species (i.e. a species richness of three). However, the implementation is also conservative: if multiple coarsely identified occurrences exist in the dataset, these are collapsed to the minimum number of possible species (i.e. two occurrences of Diplodocidae indet. would be treated as only one species). This method is similar to the ‘cryptic’ diversity measure introduced by Mannion et al. (2011).

# Evaluate unique taxa  
tax\_unique(occdf = tetrapods)

Two functions exist in palaeoverse for computing taxon ranges. The first, tax\_range\_time, can be called to calculate and plot the temporal range of taxa. The function identifies all unique taxa provided in the occurrence dataframe and finds their first and last appearance dates. The second, tax\_range\_geo, can be called to calculate the geographic range of taxa. This function allows the user to specify one of four different approaches (Darroch et al., 2020): (1) the area of a convex hull; (2) the (palaeo-)latitudinal range; (3) the maximum great-circle distance; and (4) the number of occupied equal-area grid cells. Similar to tax\_range\_time, the function will identify all unique taxa provided, and calculate these metrics based on the available occurrences of each taxon.

# Remove NA data  
tetrapods <- subset(tetrapods, !is.na(order))  
# Compute temporal range of orders  
tax\_range\_time(occdf = tetrapods, name = "order", plot = TRUE)

|  |
| --- |
| Figure 4: Temporal range of tetrapod orders in the palaeoverse example dataset. |

# Compute latitudinal range of orders  
tax\_range\_geo(occdf = tetrapods, name = "order", method = "lat")

The provided tax\_expand\_time and tax\_expand\_lat functions are complementary to the taxonomic range functions. They convert temporal or latitudinal range data to bin-level pseudo-occurrences. These pseudo-occurrences serve to fill in ghost ranges, in which a taxon is presumed to be present, but no record exists. While these pseudo-occurrences should not be treated as equivalent to actual occurrence data, such data can be useful for performing statistical analyses where bin-level data is required.

### Phylogeny wrangling

The function phylo\_check compares a list of taxonomic names to the list of tip names in a user-provided phylogeny. This information can be provided as a table describing the presence or absence of each taxon in the list and/or tips, or as counts of taxa present only in the list, only in the phylogeny, or in both. The function can also be used to trim the phylogeny to only include branches whose tip names are included within the list of taxonomic names.

### Additional features

Datasets are frequently explored within groups in palaeobiology, whether it be time bins, collections or otherwise. The group\_apply function has been included to allow users to run functions over a single, or multiple grouping variables with ease.

# Compute the number of occurrences per collection  
group\_apply(occdf = tetrapods, group = "collection\_no", fun = nrow)

A common feature request from our survey was the ability to add the ‘Geological Time Scale’ to time-series plots in base R, with similar behaviour to the deeptime R package (Gearty, 2022) for ggplot2 (Wickham, 2016). To address this request, the axis\_geo function has been developed for the palaeoverse package ([Figure 5](#fig-axis)).

plot(x = 540:0, xlab = "Time (Ma)", ylab = "User-variable",  
 xlim = c(540, 0), xaxt = "n", type = "l", lwd = 2)  
axis\_geo(side = 1, intervals = "periods")

|  |
| --- |
| Figure 5: Example Phanerozoic plot. Plot depicts the usage of the axis\_geo function for adding the Geological Time Scale to a base R plot. |

A common difficulty faced by palaeontologists is that the temporal information associated with fossil occurrence data is often asynchronous, and not directly comparable. Temporal data may be provided as either interval names or numeric ages, and might conform to different time scales (e.g. international geological stages, or North American land mammal ages). While interval names tend to be relatively stable over time, numerical age estimates are frequently updated with improved dating techniques, or the collection of new data. Consequently, where possible, interval names should be used to correlate occurrences from different stratigraphic time scales. The look\_up function is provided to help assign a common time scale–typically international stages–to occurrence data. This is achieved with a user-defined table that links arbitrary interval names to corresponding stages on a common time scale. Numerical ages for the assigned stages can be provided by the user (see example dataset interval\_key), or looked up in GTS2012 or GTS2020.

look\_up(occdf = tetrapods)

# Resources

To support the aims and use of palaeoverse, we have made several resources available to the palaeobiological community. Firstly, we have built a package website (<http://www.palaeoverse.org>) which provides information on how to contribute to palaeoverse, how to report issues and bugs, and a general community code of conduct. Secondly, a poster and ‘cheatsheet’ is provided via the website which serves as a visual guide on functionality. Finally, a vignette is provided to demonstrate usage examples with real fossil occurrence data.

# Future perspectives

While the initial development of the palaeoverse R package was led by the authors of this manuscript, it was also informed by the opinions of 35 additional researchers (survey participants). Our hope is that palaeoverse will evolve into a community-driven package, with contributions from the wider palaeontological community broadening available functionality. To support this aim, we have provided guidance on how the community can contribute to palaeoverse on the package website (<http://www.palaeoverse.org>). Our working group also has the wider aim of establishing community standards and consensus in computational palaeobiological research, and enhancing the potential for comparison between studies through providing a space for discussion of these issues. We hope to assist in making code more familiar and readable to fellow researchers, prevent researchers from ‘reinventing the wheel’ for common procedures, and improve the overall reproducibility of research through the use of computational tools which have been vetted and accepted by the broader community.

The development of the palaeoverse R package marks an initial effort to both streamline palaeontological analysis pipelines and unite the computational palaeobiology community. Future efforts will see the expansion of the palaeoverse ‘universe’, with the development Shiny applications to support non-R users and teaching exercises, tutorials to offer guidance for new researchers, and workshops to provide practical experience. We hope these efforts foster collaboration and the sharing of resources amongst the community. Finally, we warmly welcome the community to join these efforts and have established a community space accordingly to help facilitate the process (<https://groups.google.com/g/palaeoverse>).

# Acknowledgements

The authors are extremely grateful to all survey respondents who helped to shape the development of palaeoverse. The contributions of LAJ, SG, and AAC were supported by the European Research Council under the European Union’s Horizon 2020 research and innovation program (grant agreement 947921; MAPAS project). AAC was also supported by a Juan de la Cierva-formación 2020 fellowship funded by FJC2020-044836-I / MCIN /AEI / 10.13039 /501100011033 by the European Union “NextGenerationEU”/ PRTR. The contributions of BJA were supported by an ETH+ grant (BECCY). The contributions of CDD were supported by a Royal Society grant (RF\_ERE\_210013). This is Paleobiology Database publication no XXX. **Please add and update where appropriate :-)**.

# Authors’ contributions

LAJ conceived the project. All authors contributed to developing the project. LAJ, BJA, WG, KE, CD, and JF contributed the code. All authors contributed to testing and reviewing the code. SG processed the survey results and produced the figures. All authors contributed to writing the manuscript. **Please add and update where appropriate, you have all contributed in some way :-)**.

# Data accessibility

The palaeoverse R package is hosted on CRAN (XXX) and is available on GitHub (<https://github.com/palaeoverse-community/palaeoverse>). All example datasets are bundled with the R package.

# References

Allen, B.J., Wignall, P.B., Hill, D.J., Saupe, E.E. and Dunhill, A.M. (2020) The latitudinal diversity gradient of tetrapods across the permo-triassic mass extinction and recovery interval. *Proceedings of the Royal Society B*, **287**, 20201125.

Antell, G.S., Kiessling, W., Aberhan, M. and Saupe, E.E. (2020) [Marine biodiversity and geographic distributions are independent on large scales](https://doi.org/10.1016/j.cub.2019.10.065). *Current Biology*, **30**, 115–121.e5.

Barido-Sottani, J., Pett, W., O’Reilly, J.E. and Warnock, R.C. (2019) FossilSim: An r package for simulating fossil occurrence data under mechanistic models of preservation and recovery. *Methods in Ecology and Evolution*, **10**, 835–840.

Bell, M.A. and Lloyd, G.T. (2015) *Strap: An r Package for Plotting Phylogenies Against Stratigraphy and Assessing Their Stratigraphic Congruence*. Wiley Online Library.

Boyden, J.A., Müller, R.D., Gurnis, M., Torsvik, T.H., Clark, J.A., Turner, M., et al. (2011) Next-generation plate-tectonic reconstructions using GPlates.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sande, E.L., Ma, K.H., Colwell, R.K., et al. (2014) Rarefaction and extrapolation with hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs*, **84**, 45–67.

Chiarenza, A.A., Mannion, P.D., Farnsworth, A., Carrano, M.T. and Varela, S. (2022) Climatic constraints on the biogeographic history of mesozoic dinosaurs. *Current Biology*, **32**, 570–585.

Close, R.A., Benson, R.B.J., Alroy, J., Carrano, M.T., Cleary, T.J., Dunne, E.M., et al. (2020a) [The apparent exponential radiation of phanerozoic land vertebrates is an artefact of spatial sampling biases](https://doi.org/10.1098/rspb.2020.0372). *Proceedings of the Royal Society B: Biological Sciences*, **287**, 20200372.

Close, R., Benson, R.B., Saupe, E., Clapham, M. and Butler, R. (2020b) The spatial structure of phanerozoic marine animal diversity. *Science*, **368**, 420–424.

Csárdi, G., Hester, J., Wickham, H., Chang, W., Morgan, M. and Tenenbaum, D. (2021) [*Remotes: R Package Installation from Remote Repositories, Including ’GitHub’*](https://CRAN.R-project.org/package=remotes).

Darroch, S.A., Casey, M.M., Antell, G.S., Sweeney, A. and Saupe, E.E. (2020) High preservation potential of paleogeographic range size distributions in deep time. *The American Naturalist*, **196**, 454–471.

Davies, T.W., Bell, M.A., Goswami, A. and Halliday, T.J. (2017) Completeness of the eutherian mammal fossil record and implications for reconstructing mammal evolution through the cretaceous/paleogene mass extinction. *Paleobiology*, **43**, 521–536.

Dean, C.D., Chiarenza, A.A. and Maidment, S.C. (2020) Formation binning: A new method for increased temporal resolution in regional studies, applied to the late cretaceous dinosaur fossil record of north america. *Palaeontology*, **63**, 881–901.

Filazzola, A. and Lortie, C. (2022) A call for clean code to effectively communicate science. *Methods in Ecology and Evolution*.

Flannery-Sutherland, J.T., Raja, N.B., Kocsis, Á.T. and Kiessling, W. (2022) [Fossilbrush: An r package for automated detection and resolution of anomalies in palaeontological occurrence data](https://doi.org/10.1111/2041-210X.13966). *Methods in Ecology and Evolution*, **n/a**.

Flannery-Sutherland, J.T., Silvestro, D. and Benton, M.J. (2022) Global diversity dynamics in the fossil record are regionally heterogeneous. *Nature Communications*, **13**, 1–17.

Franeck, F. and Liow, L.H. (2020) Did hard substrate taxa diversify prior to the great ordovician biodiversification event? *Palaeontology*, **63**, 675–687.

Fraser, D. (2017) Can latitudinal richness gradients be measured in the terrestrial fossil record? *Paleobiology*, **43**, 479–494.

Furness, E.N., Garwood, R.J., Mannion, P.D. and Sutton, M.D. (2021) Evolutionary simulations clarify and reconcile biodiversity-disturbance models. *Proceedings of the Royal Society B*, **288**, 20210240.

Garwood, R.J., Spencer, A.R. and Sutton, M.D. (2019) RE voSim: Organism-level simulation of macro and microevolution. *Palaeontology*, **62**, 339–355.

Gearty, W. (2022) [*Deeptime: Plotting Tools for Anyone Working in Deep Time*](https://CRAN.R-project.org/package=deeptime).

Gradstein, F.M., Ogg, J.G., Schmitz, M. and Ogg, G. (2012) *The Geologic Time Scale 2012*. elsevier.

Gradstein, F.M., Ogg, J.G., Schmitz, M.D. and Ogg, G.M. (2020) *Geologic Time Scale 2020*. Elsevier.

Guillerme, T. (2018) dispRity: A modular r package for measuring disparity. *Methods in Ecology and Evolution*, **9**, 1755–1763.

Jaro, M.A. (1989) Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, **84**, 414–420.

Jones, L.A., Dean, C.D., Mannion, P.D., Farnsworth, A. and Allison, P.A. (2021) Spatial sampling heterogeneity limits the detectability of deep time latitudinal biodiversity gradients. *Proceedings of the Royal Society B*, **288**, 20202762.

Jones, L.A., Mannion, P.D., Farnsworth, A., Bragg, F. and Lunt, D.J. (2022) Climatic and tectonic drivers shaped the tropical distribution of coral reefs. *Nature communications*, **13**, 1–10.

Kiessling, W. and Krause, C. (2022) [PaleoReefs database (PARED)](https://doi.org/10.5281/zenodo.6037852).

Kocsis, Á.T., Reddin, C.J., Alroy, J. and Kiessling, W. (2019) The r package divDyn for quantifying diversity dynamics using fossil sampling data. *Methods in Ecology and Evolution*, **10**, 735–743.

Lloyd, G.T. (2016) Estimating morphological diversity and tempo with discrete character-taxon matrices: Implementation, challenges, progress, and future directions. Biological journal of the linnean society. *Biological Journal of the Linnean Society*, **118**, 131–151.

Lloyd, G.T., Pearson, P.N., Young, J.R. and Smith, A.B. (2012) Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea. *Paleobiology*, **38**, 569–584.

Mannion, P.D., Benson, R.B., Carrano, M.T., Tennant, J.P., Judd, J. and Butler, R.J. (2015) Climate constrains the evolutionary history and biodiversity of crocodylians. *Nature communications*, **6**, 1–9.

Mannion, P.D., Benson, R.B., Upchurch, P., Butler, R.J., Carrano, M.T. and Barrett, P.M. (2012) A temperate palaeodiversity peak in mesozoic dinosaurs and evidence for late cretaceous geographical partitioning. *Global Ecology and Biogeography*, **21**, 898–908.

Mannion, P.D., Upchurch, P., Benson, R.B. and Goswami, A. (2014) The latitudinal biodiversity gradient through deep time. *Trends in ecology & evolution*, **29**, 42–50.

Mannion, P.D., Upchurch, P., Carrano, M.T. and Barrett, P.M. (2011) [Testing the effect of the rock record on diversity: A multidisciplinary approach to elucidating the generic richness of sauropodomorph dinosaurs through time](https://doi.org/10.1111/j.1469-185X.2010.00139.x). *Biological Reviews*, **86**, 157–181.

Müller, R.D., Cannon, J., Qin, X., Watson, R.J., Gurnis, M., Williams, S., et al. (2018) [GPlates: Building a virtual earth through deep time](https://doi.org/10.1029/2018GC007584). *Geochemistry, Geophysics, Geosystems*, **19**, 2243–2261.

O’Brien, L. (2022) [*H3jsr: Access Uber’s H3 Library*](https://CRAN.R-project.org/package=h3jsr).

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2020) [*Vegan: Community Ecology Package*](https://CRAN.R-project.org/package=vegan).

Powell, M.G. (2009) The latitudinal diversity gradient of brachiopods over the past 530 million years. *The Journal of Geology*, **117**, 585–594.

Quental, T.B. and Marshall, C.R. (2013) How the red queen drives terrestrial mammals to extinction. *Science*, **341**, 290–292.

R Core Team. (2022) [*R: A Language and Environment for Statistical Computing*](https://www.R-project.org/). R Foundation for Statistical Computing, Vienna, Austria.

Sepkoski, J.J. (1978) A kinetic model of phanerozoic taxonomic diversity i. Analysis of marine orders. *Paleobiology*, **4**, 223–251.

Sepkoski, D. and Ruse, M. (2009) *The Paleobiological Revolution: Essays on the Growth of Modern Paleontology*. University of Chicago Press.

Silvestro, D., Salamin, N. and Schnitzler, J. (2014) PyRate: A new program to estimate speciation and extinction rates from incomplete fossil data. *Methods in Ecology and Evolution*, **5**, 1126–1131.

Silvestro, D., Zizka, A., Bacon, C.D., Cascales-Minana, B., Salamin, N. and Antonelli, A. (2016) Fossil biogeography: A new model to infer dispersal, extinction and sampling from palaeontological data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **371**, 20150225.

Song, H., Huang, S., Jia, E., Dai, X., Wignall, P.B. and Dunhill, A.M. (2020) Flat latitudinal diversity gradient caused by the permian–triassic mass extinction. *Proceedings of the National Academy of Sciences*, **117**, 17578–17583.

Starrfelt, J. and Liow, L.H. (2016) How many dinosaur species were there? Fossil bias and true richness estimated using a poisson sampling model. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **371**, 20150219.

van der Loo, M.P.J. (2014) [The stringdist package for approximate string matching](https://CRAN.R-project.org/package=stringdist). *The R Journal*, **6**, 111–122.

Vilhena, D.A. and Smith, A.B. (2013) Spatial bias in the marine fossil record. *PLoS One*, **8**, e74470.

Wickham, H. (2016) [*Ggplot2: Elegant Graphics for Data Analysis*](https://ggplot2.tidyverse.org). Springer-Verlag New York.

Zaffos, A., Finnegan, S. and Peters, S.E. (2017) Plate tectonic regulation of global marine animal diversity. *Proceedings of the National Academy of Sciences*, **114**, 5653–5658.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., et al. (2019) CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, **10**, 744–751.