Received: 2 November 2022 | Accepted: 10 March 2023

DOI: 10.1111/2041-210X.14099

Methods in Ecology and Evolution

APPLICATION

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 S. C. Nicholl[6] | Lucas Buffan[10] | Erin M. Dillon[11,12] | Joseph T. Flannery-Sutherland[13] | Alfio Alessandro Chiarenza[1]

[1]Grupo de Ecoloxía Animal, Departamento de Ecoloxía e Bioloxía Animal, Universidade de Vigo, Vigo, Spain; [2]Division of Paleontology, American Museum of Natural History, New York, New York, USA; [3]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland; [4]Computational Evolution Group, Swiss Institute of Bioinformatics, Lausanne, Switzerland; [5]Department of Earth Sciences, Durham University, Durham, UK; [6]Department of Earth Sciences, University College London, London, UK; [7]Life Sciences Department, Natural History Museum, London, UK; [8]Laboratório de Paleontologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brazil; [9]Department of Anatomical Sciences, Stony Brook University, Stony Brook, New York, USA; [10]Département de Biologie, École Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 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, California, USA and [13]School of Earth Sciences, University of Bristol, Bristol, UK

Correspondence
Lewis A. Jones
Email: lewisalan.jones@uvigo.es

Funding information
ETH+ grant (BECCY); Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Number: 2022/05697-9; H2020 European Research Council, Grant/Award Number: 947921; Juan de la Cierva-formación 2020 fellowship (European Union "NextGenerationEU"/PRTR), Grant/Award Number: FJC2020-044836-I/MCIN/AEI/10.13039/501100011033; Juan de la Cierva-formación 2021 fellowship (European Union "NextGenerationEU"/PRTR), Grant/Award Number: FJC2021-046695-I/MCIN/AEI/10.13039/501100011033; Lerner-Gray Postdoctoral Research Fellowship from the Richard Gilder Graduate School (American Museum of Natural History); Population Biology Program of Excellence Postdoctoral Fellowship (University of Nebraska-Lincoln School of Biological Sciences); Royal Society, Grant/Award Number: RF_ERE_210013, RGF_EA_180318 and RGF_R1_180020

Handling Editor: Daniele Silvestro

Abstract

1. The open-source programming language 'R' has become a standard tool in the palaeobiologist's toolkit. Its popularity within the palaeobiological community continues to grow, with published articles increasingly citing the usage of R and R packages. However, there are currently a lack of agreed standards for data preparation and available frameworks to support the implementation of such standards. Consequently, data preparation workflows 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 to aid data preparation and exploration 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. To develop these aims, we assessed the analytical needs of the broader palaeobiological community using an online survey, in addition to incorporating our own experiences.
3. In this work, 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 our future plans for the broader Palaeoverse project.
4. palaeoverse is a community-driven R package for 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 open-source 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, reproducible, reusable

1 | INTRODUCTION

Since the development of large palaeontological datasets from the 1970s onwards, palaeontologists have increasingly adopted computational approaches to address questions about the history of life on Earth (Benton, 1999; Sepkoski, 1978). Today, most sub-disciplines within palaeontology regularly use large datasets to perform experiments in silico. This has initiated a 'Golden Age' of palaeontology (Sepkoski & Ruse, 2009), where extensive datasets of various formats are used to test macroevolutionary and macroecological hypotheses (e.g. Close, Benson, Alroy, et al., 2020; Mannion et al., 2014; Quental & Marshall, 2013; Zaffos et al., 2017). The growth and increasing availability of such datasets has made coding an integral part of palaeobiological research. Today, palaeobiologists commonly use code to clean (e.g. Flannery-Sutherland, Raja, et al., 2022; Zizka et al., 2019), analyse (e.g. Guillerme, 2018; Kocsis et al., 2019), and visualise data (e.g. Bell & Lloyd, 2015), as well as to build models (e.g. Silvestro et al., 2014; Starrfelt & Liow, 2016) and implement simulations (e.g. Barido-Sottani et al., 2019; Fraser, 2017; Furness et al., 2021; Jones et al., 2021). Although software has been developed in languages such as C++ (e.g. Garwood et al., 2019) and Python (e.g. 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 work with their data. Many of these tools are often borrowed or repurposed from ecology (e.g. Chao et al., 2014; Oksanen et al., 2020), while others have been developed to specifically handle fossil data (e.g. Kocsis et al., 2019; Lloyd, 2016).

In spite of the growth of analytical tools, 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 in code legibility and documentation. Accordingly, custom scripts can be inaccessible to other users (Filazzola & Lortie, 2022). Although increasingly requested by journals, code is also not always provided as supplementary material nor made available in online repositories (e.g. GitHub, Zenodo, Dryad). A lack of available 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, and as such it might require extensive reworking before it can be applied to other data. Consequently, researchers are often forced to 'reinvent the wheel', putting time and effort into writing code that already exists but is unavailable, inaccessible, and/or difficult to repurpose (Filazzola & Lortie, 2022). Such issues are exacerbated by the absence of community standards for 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. Thus, there is 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 that align with current researcher needs to clean, 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 the survey, describe and detail the functionality of palaeoverse, and illustrate its features with usage examples.

2 | COMMUNITY SURVEY

To assess the needs of the palaeobiological community, we conducted an online survey. The survey was distributed via social media (Twitter) and email, and it included questions related to

researchers' previous experience, their pre-existing code (to identify potential contributions), and what functionality they would consider to be useful in a new palaeobiological toolkit. We summarise the types of data that survey participants typically work with, the tasks commonly carried out when working with these data, and the tools they would like to have access to, in Figure 1. We found that survey participants ($n = 35$) work with a wide range of data (Figure 1) and that the checking and transformation of data is the most. A wide variety of functions were requested by survey participants, with data plotting, time binning, and data access commonly suggested (Figure 1). 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 the Supporting Information.

3 | PACKAGE DESCRIPTION

After conducting the community survey, we combined participant input with our own experience to develop a toolkit for palaeobiologists: the palaeoverse R package. The package provides auxiliary functions to support data preparation and exploration for palaeobiological analysis. A summary of the functions currently available in palaeoverse is provided in Table 1, with further description provided in the Functions section. To demonstrate the functionality and versatility of the package, we also provide usage examples.

3.1 | Installation

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

```
install.packages("palaeoverse")
```

If preferred, 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 installation, palaeoverse can be loaded via the library function in R:

```
library("palaeoverse")
```

3.2 | Data

Functionality in palaeoverse was designed to be compatible with occurrence dataframes, such as those downloaded from the Paleobiology Database (https://paleobiodb.org/#/), the Geobiodiversity Database (http://www.geobiodiversity.com), or the Neptune Sandbox Berlin database (https://nsb.mfn-berlin.de/). Functionality is purposely flexible in palaeoverse 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

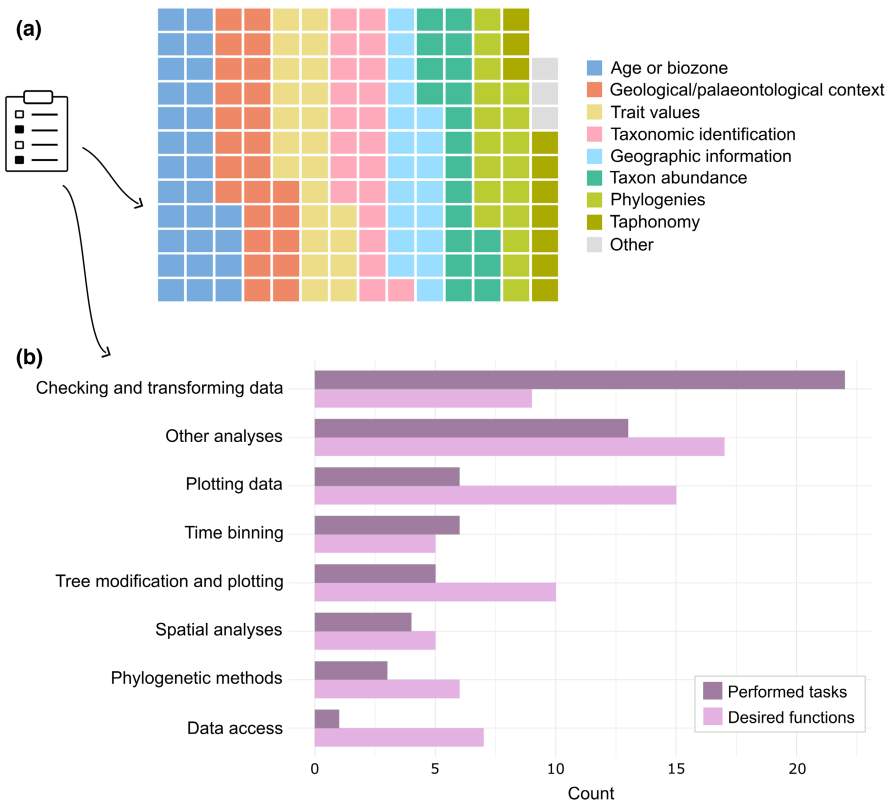

FIGURE 1 Summary of responses to the palaeoverse survey. (a) The types of palaeontological data that survey participants typically work with. Each box represents an individual check within a check-box list, in which participants could check multiple boxes. (b) Tasks that survey participants routinely carry out in their own analyses (dark pink), and the functions they would find useful in the palaeoverse package (light pink).

TABLE 1 A summary table of the functions currently available in the palaeoverse R package and respective dependencies. Base R dependencies are highlighted with an asterisk.

Function	Description	Dependency
axis_geo	Add a geological time scale axis to a plot	deeptime (Gearty, 2023)
bin_lat	Bin fossil occurrences into latitudinal bins	—
bin_space	Bin fossil occurrences into spatial bins	h3jsr (O'Brien, 2023), sf (Pebesma, 2018)
bin_time	Bin fossil occurrences into time bins (choice of approaches)	stats* (R Core Team, 2022)
data	Datasets: 'tetrapods', 'reefs', 'interval_key', 'GTS2012', and 'GTS2020'	—
group_apply	Apply a function over user-defined groups	stats* (R Core Team, 2022)
lat_bins	Generate latitudinal bins	graphics* (R Core Team, 2022)
look_up	Link user-specified interval names to the International Geological Time Scale	—
palaeorotate	Reconstruct the palaeogeographical coordinates of fossil occurrences	curl (Ooms, 2023), geosphere (Hijmans, 2022), httr (Wickham, 2022), h3jsr (O'Brien, 2023), pbapply (Solymos & Zawadzki, 2023), sf (Pebesma, 2018), stats* (R Core Team, 2022), utils* (R Core Team, 2022)
phylo_check	Check taxon names against tips in a phylogeny and/or remove tips from the tree	ape (Paradis & Schliep, 2019)
tax_check	Check for spelling mistakes in taxon names and flag potential issues	stats* (R Core Team, 2022), stringdist (van der Loo, 2014)
tax_range_space	Calculate the geographic range of taxa (choice of approaches)	geosphere (Hijmans, 2022), grDevices (R Core Team, 2022), h3jsr (O'Brien, 2023)
tax_range_time	Calculate and plot the temporal range of taxa	graphics* (R Core Team, 2022)
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	stats* (R Core Team, 2022)
time_bins	Generate stratigraphic time bins or near-equal length time bins	—

users to understand and work with. Although this might be undesirable for some advanced R users, transforming data structures should be straightforward for these users.

3.3 | Functions

A summary of the functions available in palaeoverse along with their respective dependencies is provided in Table 1. All functions available in palaeoverse are novel in either their functionality or implementation, and they collectively provide a flexible and versatile toolkit for palaeobiological research. Detailed descriptions of the functions are provided herein.

3.3.1 | Example datasets

Two occurrence datasets (tetrapods and reefs) are provided in palaeoverse to support reproducible examples within the function documentation. The tetrapods dataset is a compilation of Carboniferous–Early Triassic tetrapod occurrences (n = 5270) 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 = 4363) from the PaleoReefs Database (Kiessling & Krause, 2022). This dataset includes information on the biological, geological, and geographical context of each reef. Except for the removal of superfluous columns and the renaming of some columns to improve clarity, both datasets are unaltered from their sources. Additional information about both datasets can be accessed via ?tetrapods or ?reefs once the package is loaded.

3.3.2 | Time bins

We 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 extract

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)
```

As is evident from Figure 2, GTS temporal bins are highly uneven in duration. Previous studies have attempted to circumvent this issue by generating near-equal-length time bins by grouping stages towards a target bin length (e.g. Close, Benson, Alroy, et al., 2020; Mannion et al., 2015). time_bins enables users to generate near-equal-length time bins following this approach (Figure 3) to a specified target size:

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

Nevertheless, 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.

3.3.3 | Occurrence binning

Fossil occurrences are frequently 'binned' into distinct time intervals to quantify changes (e.g. biodiversity or disparity) through geological time, as described by the survey participants (Figure 1). 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")
```

Although binning occurrences with tightly defined temporal limits is straightforward and has been implemented in other R packages (e.g. Lloyd, 2016), 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 tackle this problem (e.g. Davies et al., 2017; Dean et al., 2020; Franeck & Liow, 2020; Lloyd et al., 2012; Silvestro et al., 2016), but there is currently no consensus on the best methodological approach or 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 using a user-defined probability 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, palaeobiologists have developed a heightened interest in the spatial structure of the fossil record, with many studies focused on understanding the spatial distribution of biodiversity and the processes that drive them (Figure 1; Antell et al., 2020; Chiarenza et al., 2022; Close, Benson, Saupe, et al., 2020; Flannery-Sutherland, Silvestro, & Benton, 2022; Jones et al.,

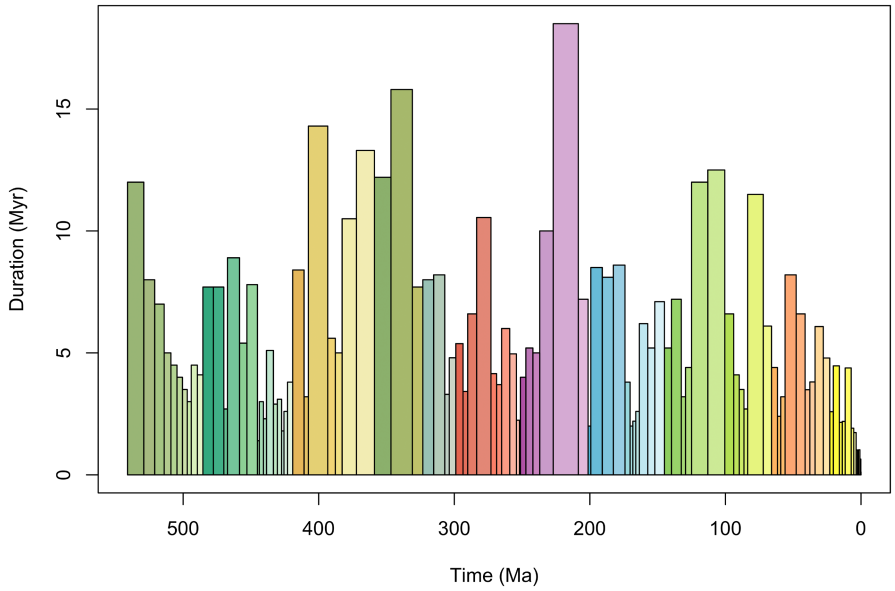

FIGURE 2 Phanerozoic stage-level time bins. Plot depicts the uneveness in duration of stratigraphic time bins. Bar colour filling follows the established colour scheme of the International Commission on Stratigraphy (https://stratigraphy.org/).

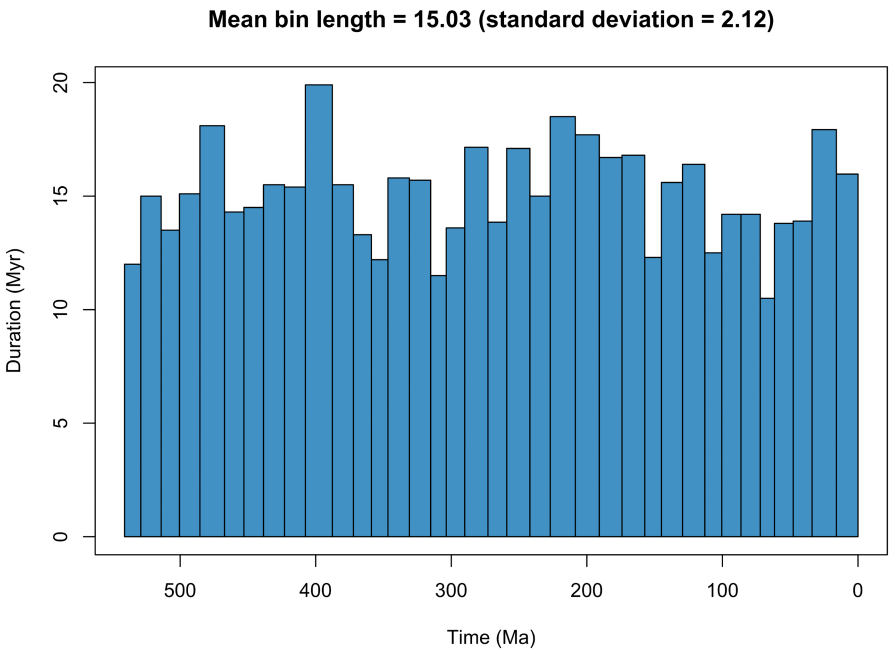

FIGURE 3 Phanerozoic near-equal-length time bins. Plot depicts composite stratigraphic bins (grouping stage-level bins) for the Phanerozoic of a target bin size of 15 million years. Note: time bins are still uneven but less so than stage-level bins.

2022; Vilhena & Smith, 2013). In order to support such analyses, bin_space 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, 2023). 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.

```
# Assign data to equal-area spatial bins
bin_space(occdf = reefs, spacing = 250)
bin_space(occdf = reefs, spacing = 250, sub_grid = 50)
```

Understanding the latitudinal distribution of biodiversity in deep time has also gained research interest in recent years (Allen et al., 2020; Jones et al., 2021; Mannion et al., 2012, 2014; Powell, 2009; Song et al., 2020). 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)
```

3.3.4 | Palaeogeographical reconstruction

Using the present-day coordinates of fossil occurrences, plate rotation models can be used to reconstruct their location at the 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 clear 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 GPlates to reconstruct palaeocoordinates, yet lack specifics on which plate rotation model was used with the GPlates Web Service or desktop application (Müller et al., 2018). Furthermore, the uncertainty in palaeogeographical reconstructions is often underappreciated; reconstructed coordinates are treated as being well-established, rather than model-based estimates. Finally, online databases do not provide palaeocoordinates for all known samples. Both published and unpublished data (e.g. museum specimens) exist outside of online databases for which researchers might require palaeocoordinates.

We have developed the function palaeorotate to address this challenge. The function allows palaeocoordinates to be reconstructed within R using two different approaches: 'point' and 'grid'. The first approach uses 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 reconstruction files were generated using an equal-area hexagonal grid (~100 km spacings) via the h3jsr package (O'Brien, 2023) and allow palaeocoordinates to be generated efficiently for large datasets. Furthermore, these reconstruction files allow the user to calculate the palaeolatitudinal range between reconstructed coordinates from different models, as well as the great circle distance between the two most distant points (i.e. the palaeogeographical uncertainty). Finally, to encourage transparency in palaeobiological research, the function also reports additional information such as the plate rotation model(s) used.

```
# 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)
```

3.3.5 | 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 the dissimilarity between character strings as a value between 0 (exact match) and 1 (completely dissimilar). When calling the function, 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). This function is provided to help researchers perform a spell check on their dataset, with additional functionality available in the fossilbrush package (Flannery-Sutherland, Raja, et al., 2022). However, it should be made clear that this is no replacement for thorough taxonomic vetting.

```
# Check for taxonomic errors
tax_check(taxdf = 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 an analysis is conducted at species level, occurrences identified to the 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 number of unique taxa being discarded by retaining fossils which are identified to a coarser taxonomic resolution than the desired level yet 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) because this occurrence represents a different species than the two already present in the dataset. Yet, the implementation is also conservative: if multiple coarsely identified occurrences exist in the dataset, they 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, genus = "genus", family = "family",
  order = "order", class = "class", resolution = "genus")
```

Two functions exist in palaeoverse for computing taxon ranges. The first, tax_range_time, can be used 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 (Figure 4). The second, tax_range_space, 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 and proportion 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)
```

```
# Compute latitudinal range of orders
tax_range_space(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. Although 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 are required.

3.3.6 | Phylogeny wrangling

The function phylo_check compares a list of taxonomic names to the list of tip names in a user-provided phylogeny using the ape package (Paradis & Schliep, 2019). This comparison 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.

3.3.7 | Additional features

Datasets are frequently explored within groups in palaeobiology, such as time bins, collections, or regions. The group_apply function

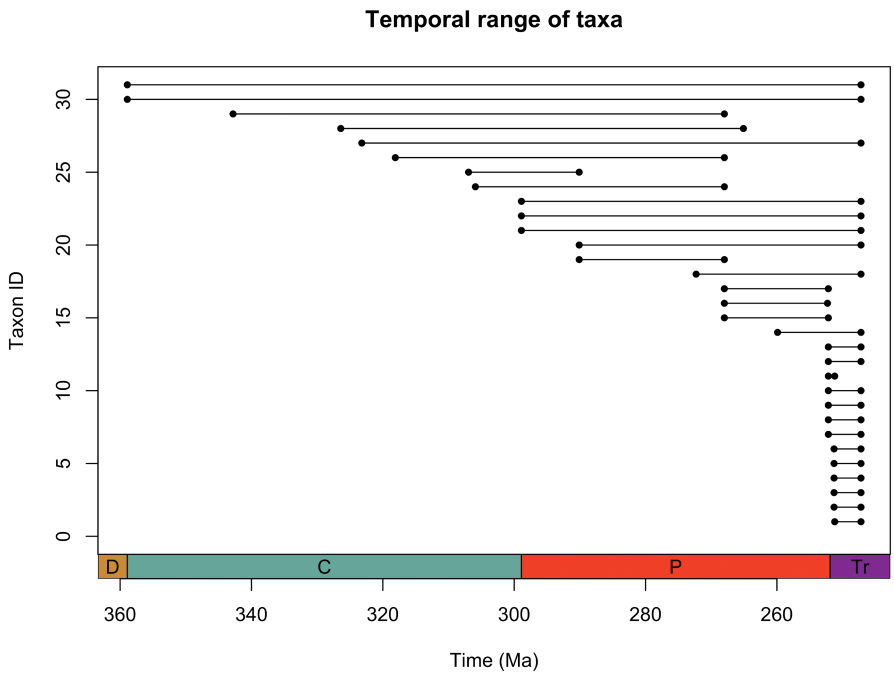

FIGURE 4 Temporal range of tetrapod orders in the palaeoverse example dataset.

has been included to allow users to run functions over 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 difficulty faced by palaeontologists is that the temporal information associated with fossil occurrence data is often asynchronous and not directly comparable. Temporal data might be provided as either character-based interval names or numeric ages and might use different time scales (e.g. international geological stages or North American land mammal ages). Although interval names tend to be relatively stable over time, numerical age estimates are frequently updated with improved dating techniques or with the collection of new data. Consequently, where possible, interval names should be used to correlate occurrences from different stratigraphic time scales. We provide the look_up function to help assign a common time scale–typically international stages–to occurrence data. This is achieved with a user-defined table that links chosen interval names to corresponding stages on a common time scale (see example dataset interval_key). Numerical ages for the assigned stages can be provided by the user or looked up in GTS2012 or GTS2020 (the default). This functionality therefore enables numerical ages to be assigned to datasets only containing character-based interval names (e.g. "Maastrichtian").

```
reefs <- look_up(occdf = reefs,
                 early_interval = "interval",
                 late_interval = "interval",
                 int_key = interval_key)
```

Finally, a common feature request from our survey (Figure 1) 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, 2023) for ggplot2 (Wickham, 2016). To fulfil this request, the axis_geo function has been developed for the palaeoverse package (Figure 5).

```
# Palaeorotate reef dataset
reefs <- palaeorotate(occdf = reefs, age = "interval_mid_ma")

# Plot palaeolatitudinal distribution through time
plot(x = reefs$interval_mid_ma, y = reefs$p_lat,
     xlab = "Time (Ma)", ylab = "Palaeolatitude (\u00B0)",
     xlim = c(541, 0), xaxt = "n", type = "p", pch = 20)
# Add Geological Time Scale
axis_geo(side = 1, intervals = "periods")
```

4 | RESOURCES

To support the aims and use of palaeoverse, we have made several resources available to the palaeobiological community. First, we have built a package website (http://palaeoverse.palaeoverse.org), which provides information on how to contribute to palaeoverse, how to report issues and bugs, and a general community code of conduct. Second, we have established a Google Group to foster collaboration and discussion around the issues faced by the community, such as establishing standards for data preparation (https://groups.google.com/g/palaeoverse).

5 | FUTURE PERSPECTIVES

Palaeoverse is envisioned as a community project. While the initial development of the palaeoverse R package was led by the authors of this manuscript, it was also informed by the perspectives of 35

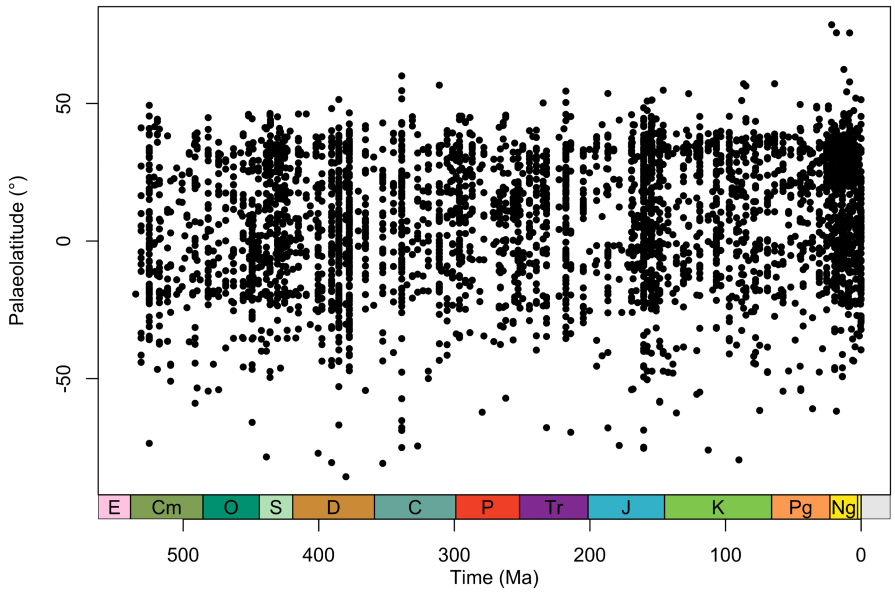

FIGURE 5 Example Phanerozoic plot of the palaeolatitudinal distribution of reefs through time. The plot demonstrates the usage of the axis_geo function for adding the Geological Time Scale to a base R plot.

additional researchers (survey participants). Our hope is that palaeoverse will evolve into a community-driven package by welcoming contributions from the wider palaeontological community to broaden the available functionality. To support this aim, we provide guidance on how community members can contribute to palaeoverse on the package website (http://palaeoverse.palaeoverse.org). Our working group also has the wider aim of establishing community standards and consensus in computational palaeobiological research and facilitating comparison across studies. Through the palaeoverse R package, 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 that have been vetted and accepted by the broader community.

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

AUTHOR CONTRIBUTIONS

Lewis A. Jones conceived the project. All authors contributed to developing the project. Lewis A. Jones, Bethany J. Allen, William Gearty, Kilian Eichenseer, Christopher D. Dean and Joseph T. Flannery-Sutherland contributed the code. All authors contributed to testing and reviewing the code. Sofía Galván processed the survey results and produced the survey figures. All authors contributed to writing the manuscript.

ACKNOWLEDGEMENTS

The authors are extremely grateful to all survey participants who helped to shape the development of palaeoverse. Special thanks are given to Emma M. Dunne, who participated in numerous discussions, and shared her experience with the development team. Thanks are also given to two anonymous reviewers that helped improve this manuscript. The contributions of L.A.J., S.G., and A.A.C. were supported by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement 947921; MAPAS project). L.A.J. was also supported by a Juan de la Cierva-formación 2021 fellowship (FJC2021-046695-I/MCIN/AEI/10.13039/501100011033) from the European Union "NextGenerationEU"/PRTR. A.A.C. was also supported by a Juan de la Cierva-formación 2020 fellowship (FJC2020-044836-I/MCIN/AEI/10.13039/501100011033) from the European Union "NextGenerationEU"/PRTR. The contributions of W.G. were supported by the Population Biology Program of Excellence Postdoctoral Fellowship from the University of Nebraska-Lincoln School of Biological Sciences and the Lerner-Gray Postdoctoral Research Fellowship from the Richard Gilder Graduate School at the American Museum of Natural History. The contributions of B.J.A. were supported by an ETH+ grant (BECCY). The contributions of C.D.D. (RF_ERE_210013), M.K. (RGF_EA_180318) and C.S.C.N. (RGF_R1_180020) were supported by Royal Society grants. The contributions of P.L.G. were supported by a FAPESP postdoctoral grant (2022/05697-9). This is Paleobiology Database publication no. 450.

CONFLICT OF INTEREST STATEMENT

We declare we have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14099.

DATA AVAILABILITY STATEMENT

The palaeoverse R package is hosted on CRAN (https://cran.r-project.org/web/packages/palaeoverse/) and is available on GitHub (https://github.com/palaeoverse-community/palaeoverse). The code is also archived in Zenodo through continuous integration (Jones et al., 2023). All example datasets are bundled with the R package. All code is released under a GPL (>=3) licence.

ORCID

Lewis A. Jones https://orcid.org/0000-0003-3902-8986
William Gearty https://orcid.org/0000-0003-0076-3262
Bethany J. Allen https://orcid.org/0000-0003-0282-6407
Kilian Eichenseer https://orcid.org/0000-0002-0477-8878
Christopher D. Dean https://orcid.org/0000-0001-6471-6903
Sofía Galván https://orcid.org/0000-0002-3092-4314
Miranta Kouvari https://orcid.org/0000-0002-5442-6221
Pedro L. Godoy https://orcid.org/0000-0003-4519-5094
Cecily S. C. Nicholl https://orcid.org/0000-0003-2860-2604
Lucas Buffan https://orcid.org/0000-0002-2353-1432
Erin M. Dillon https://orcid.org/0000-0003-0249-027X
Joseph T. Flannery-Sutherland https://orcid.org/0000-0001-8232-6773
Alfio Alessandro Chiarenza https://orcid.org/0000-0001-5525-6730

REFERENCES

Allen, B. J., Wignall, P. B., Hill, D. J., Saupe, E. E., & Dunhill, A. M. (2020). The latitudinal diversity gradient of tetrapods across the permo-triassic mass extinction and recovery interval. *Proceedings of the Royal Society BB: Biological Sciences*, *287*(1929), 20201125.

Antell, G. S., Kiessling, W., Aberhan, M., & Saupe, E. E. (2020). Marine biodiversity and geographic distributions are independent on large scales. *Current Biology*, *30*(1), 115–121.e5. https://doi.org/10.1016/j.cub.2019.10.065

Barido-Sottani, J., Pett, W., O'Reilly, J. E., & 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*(6), 835–840.

Bell, M. A., & Lloyd, G. T. (2015). Strap: An r package for plotting phylogenies against stratigraphy and assessing their stratigraphic congruence. In *Palaeontology* (Vol. *58*, pp. 379–389). Wiley Online Library.

Benton, M. J. (1999). The history of life: Large data bases in palaeontology. In D. Harper (Ed.), *Numerical palaeobiology* (pp. 249–283). Wiley.

Chao, A., Gotelli, N. J., Hsieh, T. C., Sande, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (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., & Varela, S. (2022). Climatic constraints on the biogeographic history of mesozoic dinosaurs. *Current Biology*, *32*(3), 570–585.

Close, R., Benson, R. B., Saupe, E., Clapham, M., & Butler, R. (2020). The spatial structure of phanerozoic marine animal diversity. *Science*, *368*(6489), 420–424.

Close, R. A., Benson, R. B. J., Alroy, J., Carrano, M. T., Cleary, T. J., Dunne, E. M., Mannion, P. D., Uhen, M. D., & Butler, R. J. (2020). The apparent exponential radiation of phanerozoic land vertebrates is an artefact of spatial sampling biases. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1924), 20200372. https://doi.org/10.1098/rspb.2020.0372

Csárdi, G., Hester, J., Wickham, H., Chang, W., Morgan, M., & 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., & Saupe, E. E. (2020). High preservation potential of paleogeographic range size distributions in deep time. *The American Naturalist*, *196*(4), 454–471.

Davies, T. W., Bell, M. A., Goswami, A., & 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*(4), 521–536.

Dean, C. D., Chiarenza, A. A., & 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*(6), 881–901.

Filazzola, A., & Lortie, C. (2022). A call for clean code to effectively communicate science. *Methods in Ecology and Evolution*, *13*(10), 2119–2128. https://doi.org/10.1111/2041-210X.13961

Flannery-Sutherland, J. T., Raja, N. B., Kocsis, Á. T., & Kiessling, W. (2022). Fossilbrush: An r package for automated detection and resolution of anomalies in palaeontological occurrence data. *Methods in Ecology and Evolution*, *13*(11), 2404–2418. https://doi.org/10.1111/2041-210X.13966

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

Franeck, F., & Liow, L. H. (2020). Did hard substrate taxa diversify prior to the great ordovician biodiversification event? *Palaeontology*, *63*(4), 675–687.

Fraser, D. (2017). Can latitudinal richness gradients be measured in the terrestrial fossil record? *Paleobiology*, *43*(3), 479–494.

Furness, E. N., Garwood, R. J., Mannion, P. D., & Sutton, M. D. (2021). Evolutionary simulations clarify and reconcile biodiversity-disturbance models. *Proceedings of the Royal Society B*, *288*(1949), 20210240.

Garwood, R. J., Spencer, A. R., & Sutton, M. D. (2019). REvoSim: Organism-level simulation of macro and microevolution. *Palaeontology*, *62*(3), 339–355.

Gearty, W. (2023). Deeptime: Plotting tools for anyone working in deep time. https://CRAN.R-project.org/package=deeptime

Gradstein, F. M., Ogg, J. G., Schmitz, M., & Ogg, G. (2012). *The geologic time scale 2012*. Elsevier.

Gradstein, F. M., Ogg, J. G., Schmitz, M. D., & 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*(7), 1755–1763.

Hijmans, R. J. (2022). Geosphere: Spherical trigonometry. https://CRAN.R-project.org/package=geosphere

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*(406), 414–420.

Jones, L. A., Dean, C. D., Mannion, P. D., Farnsworth, A., & Allison, P. A. (2021). Spatial sampling heterogeneity limits the detectability of deep time latitudinal biodiversity gradients. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1945), 20202762.

Jones, L. A., Gearty, W., Eichenseer, K., Dean, C., Allen, B., & Flannery-Sutherland, J. (2023). Palaeoverse-community/palaeoverse: v.1.1.1 (Version v.1.1.1) [Computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.7728639

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

Kiessling, W., & Krause, C. (2022). *PaleoReefs database* (*PARED*) (Version 1.0) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.6037852

Kocsis, Á. T., Reddin, C. J., Alroy, J., & Kiessling, W. (2019). The r package divDyn for quantifying diversity dynamics using fossil sampling data. *Methods in Ecology and Evolution*, *10*(5), 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., & Smith, A. B. (2012). Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea. *Paleobiology*, *38*(4), 569–584.

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

Mannion, P. D., Benson, R. B., Upchurch, P., Butler, R. J., Carrano, M. T., & Barrett, P. M. (2012). A temperate palaeodiversity peak in mesozoic dinosaurs and evidence for late cretaceous geographical partitioning. *Global Ecology and Biogeography*, *21*(9), 898–908.

Mannion, P. D., Upchurch, P., Benson, R. B., & Goswami, A. (2014). The latitudinal biodiversity gradient through deep time. *Trends in Ecology & Evolution*, *29*(1), 42–50.

Mannion, P. D., Upchurch, P., Carrano, M. T., & 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. *Biological Reviews*, *86*(1), 157–181. https://doi.org/10.1111/j.1469-185X.2010.00139.x

Müller, R. D., Cannon, J., Qin, X., Watson, R. J., Gurnis, M., Williams, S., Pfaffelmoser, T., Seton, M., Russell, S. H. J., & Zahirovic, S. (2018). GPlates: Building a virtual earth through deep time. *Geochemistry, Geophysics, Geosystems*, *19*(7), 2243–2261. https://doi.org/10.1029/2018GC007584

O'Brien, L. (2023). 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., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *Vegan: Community ecology package*. https://CRAN.R-project.org/package=vegan

Ooms, J. (2023). Curl: A modern and flexible web client for r. https://CRAN.R-project.org/package=curl

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528.

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Powell, M. G. (2009). The latitudinal diversity gradient of brachiopods over the past 530 million years. *The Journal of Geology*, *117*(6), 585–594.

Quental, T. B., & Marshall, C. R. (2013). How the red queen drives terrestrial mammals to extinction. *Science*, *341*(6143), 290–292.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sepkoski, D., & Ruse, M. (2009). *The paleobiological revolution: Essays on the growth of modern paleontology*. University of Chicago Press.

Sepkoski, J. J. (1978). A kinetic model of phanerozoic taxonomic diversity i. *Analysis of Marine Orders. Paleobiology*, *4*(3), 223–251.

Silvestro, D., Salamin, N., & Schnitzler, J. (2014). PyRate: A new program to estimate speciation and extinction rates from incomplete fossil data. *Methods in Ecology and Evolution*, *5*(10), 1126–1131.

Silvestro, D., Zizka, A., Bacon, C. D., Cascales-Minana, B., Salamin, N., & 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*(1691), 20150225.

Solymos, P., & Zawadzki, Z. (2023). *Pbapply: Adding progress bar to '*apply' functions*. https://CRAN.R-project.org/package=pbapply

Song, H., Huang, S., Jia, E., Dai, X., Wignall, P. B., & Dunhill, A. M. (2020). Flat latitudinal diversity gradient caused by the permian–triassic mass extinction. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(30), 17578–17583.

Starrfelt, J., & 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*(1691), 20150219.

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

Vilhena, D. A., & Smith, A. B. (2013). Spatial bias in the marine fossil record. *PLoS ONE*, *8*(10), e74470.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wickham, H. (2022). *Httr: Tools for working with URLs and HTTP*. https://CRAN.R-project.org/package=httr

Zaffos, A., Finnegan, S., & Peters, S. E. (2017). Plate tectonic regulation of global marine animal diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(22), 5653–5658.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., & Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744–751.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Supporting Information S1: 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.

How to cite this article: Jones, L. A., Gearty, W., Allen, B. J., Eichenseer, K., Dean, C. D., Galván, S., Kouvari, M., Godoy, P. L., Nicholl, C. S. C., Buffan, L., Dillon, E. M., Flannery-Sutherland, J. T., & Chiarenza, A. A. (2023). palaeoverse: A community-driven R package to support palaeobiological analysis. *Methods in Ecology and Evolution*, *14*, 2205–2215. https://doi.org/10.1111/2041-210X.14099