

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

3 —
4 Lewis A. Jones¹, William Gearty², Bethany J. Allen^{3,4}, Kilian Eichenseer⁵, Christopher D. Dean⁶, Sofia
5 Galván¹, Miranta Kouvari^{6,7}, Pedro L. Godoy^{8,9}, Cecily Nicholl⁶, Lucas Buffan¹⁰, Erin M. Dillon^{11,12},
6 Joseph T. Flannery-Sutherland¹³, and Alfio Alessandro Chiarenza¹

7 —
8 ¹*Grupo de Ecoloxía Animal, Departamento de Ecoloxía e Bioloxía Animal, Universidade de Vigo, 36310*
9 *Vigo, Spain.*

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

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

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

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

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

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

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

20 ⁹*Department of Anatomical Sciences, Stony Brook University, Stony Brook, NY, 11794 USA.*

21 ¹⁰*Département de Biologie, École Normale Supérieure de Lyon, Université Claude Bernard Lyon 1,*
22 *69342 Lyon Cedex 07, France.*

23 ¹¹*Smithsonian Tropical Research Institute, Balboa, Republic of Panama.*

24 ¹²*Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, CA*
25 *93106, USA.*

26 ¹³*School of Earth Sciences, University of Bristol, BS8 1RL, Bristol, UK*

27 —
28 **Corresponding author:** LewisAlan.Jones@uvigo.es



31

Abstract

32

1. The open-source programming language ‘R’ has become a standard tool in the palaeobiologist’s toolkit. Its popularity within the palaeobiology 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 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 the future plans for the broader **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 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.

56

Keywords

57

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

59

60 **Introduction**

61 Since the development of large palaeontological datasets from the 1970s onwards, palaeontologists have
62 increasingly adopted computational approaches to address questions about the history of life on Earth
63 (Sepkoski, 1978; Benton and Harper, 1999). Today, most sub-disciplines within palaeontology regularly
64 use large datasets to perform experiments *in silico*. This has initiated a ‘Golden Age’ of palaeontology
65 (Sepkoski and Ruse, 2009), where extensive datasets of various formats are used to test macroevolutionary
66 and macroecological hypotheses (e.g. Quental and Marshall, 2013; Mannion et al., 2014; Zaffos, Finnegan
67 and Peters, 2017; Close et al., 2020a). The growth and increasing availability of such datasets has made
68 coding an integral part of palaeobiological research. Today, palaeobiologists commonly use code to clean
69 (e.g. Zizka et al., 2019; Flannery-Sutherland et al., 2022a), analyse (e.g. Guillerme, 2018; Kocsis et al.,
70 2019), and visualise data (e.g. Bell and Lloyd, 2015), as well as build models (e.g. Silvestro, Salamin and
71 Schnitzler, 2014; Starrfelt and Liow, 2016) and implement simulations (e.g. Fraser, 2017; Barido-Sottani
72 et al., 2019; Furness et al., 2021; Jones et al., 2021). Whilst software has been developed in languages such
73 as C++ (e.g. Garwood, Spencer and Sutton, 2019) and Python (e.g. Silvestro et al., 2014), the programming
74 language R is currently the most popular in palaeobiology. This is due to the wide range of tools—in the
75 form of R packages—available to help users work with their data. Many of these tools are often borrowed
76 or repurposed from ecology (e.g. Chao et al., 2014; Oksanen et al., 2020), while others have been developed
77 to specifically handle fossil data (e.g. Lloyd, 2016; Kocsis et al., 2019).

78 In spite of the growth of analytical tools, few packages explicitly focus on preparing data for analyses,
79 forcing users to construct custom scripts. This can result in distinct differences in code style and practices
80 amongst the community, including code legibility and documentation. Accordingly, custom scripts can be
81 inaccessible to other users (Filazzola and Lortie, 2022). Although increasingly requested by journals, code
82 is also not always provided as supplementary material nor made available in online repositories
83 (e.g. GitHub, Zenodo, Dryad). A lack of available code can lead to research results being unreproducible,
84 preventing future studies from extending the work. Even when code is available, it might be poorly
85 documented or written in a way that is specific to the dataset being analysed, and as such it may require
86 extensive reworking before it can be applied to other data. Consequently, researchers are often forced to
87 ‘reinvent the wheel’, putting time and effort into writing code that already exists, but is unavailable,
88 inaccessible, and/or difficult to repurpose (Filazzola and Lortie, 2022). Such issues are exacerbated by the
89 absence of community standards for how data should be prepared for analyses; differing approaches utilised
90 by different researchers result in a lack of consistency between studies, making comparison between results
91 challenging. Thus, there is a well-established need for both protocols and tools for preparing
92 palaeontological data for further analysis.

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

102 **Community survey**

103 To assess the needs of the palaeobiological community, we conducted an online survey. The survey was
104 distributed via social media (Twitter) and email, and included questions related to researchers' previous
105 experience, pre-existing code (to identify potential contributions), and what functionality they consider to
106 be useful in a new palaeobiological toolkit. We summarise the types of data participants typically work
107 with, the tasks commonly carried out when working with this data, and the tools they would like to have
108 access to in Figure 1. We found that survey participants ($n = 35$) work with a wide range of data (Figure 1)
109 and the checking and transformation of data is the most commonly employed task. A wide variety of
110 functions were requested by survey participants, with data plotting, time binning, and data access commonly
111 suggested (Figure 1). Over 40% of participants also indicated that they were willing to contribute code to
112 `palaeoverse`, highlighting the potential for a community-driven project. Specific details regarding the
113 survey and responses can be found in the Supplementary Material.

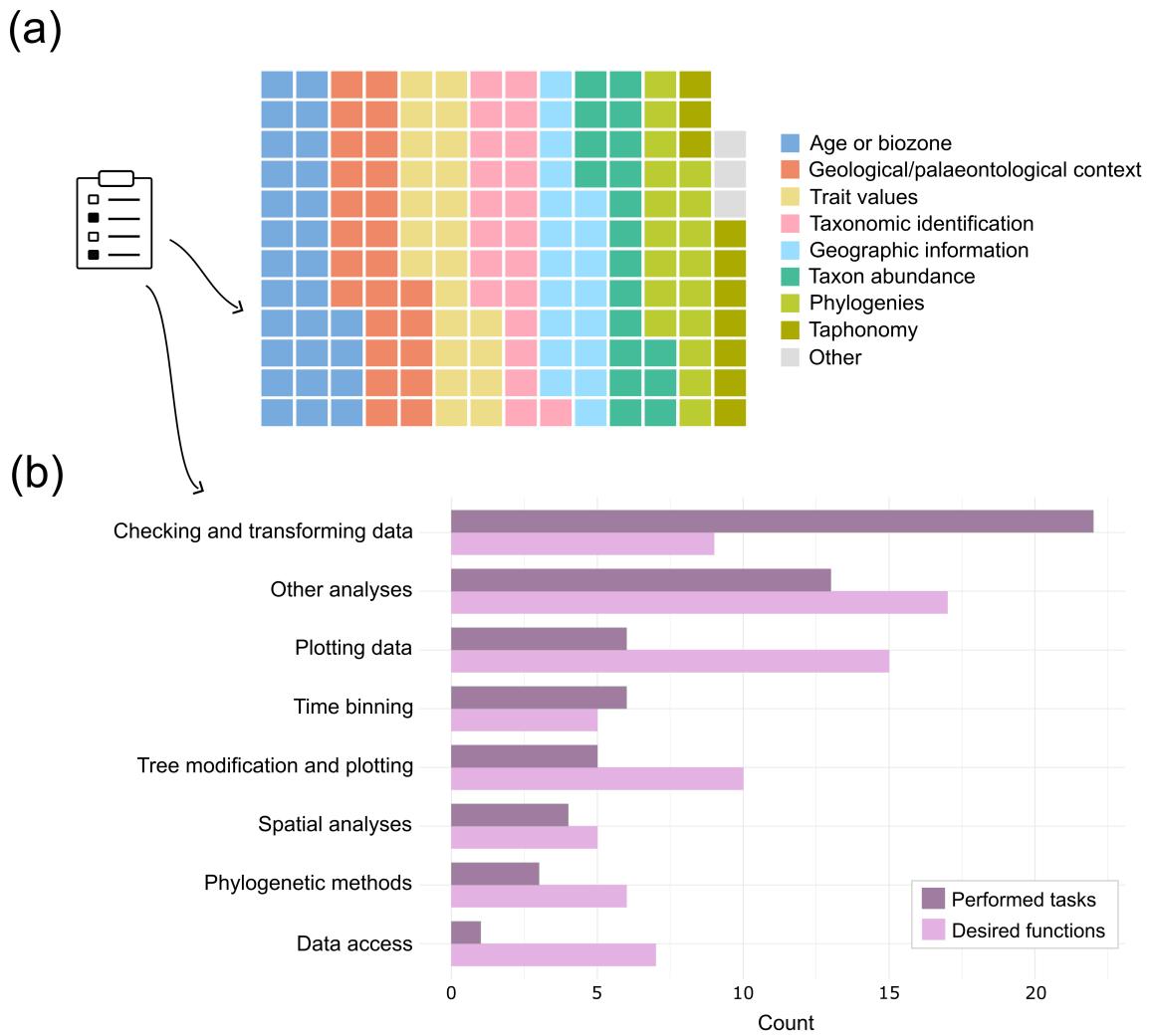


Figure 1: Summary of responses to the `palaeoverse` survey. (a) The types of palaeontological data that survey participants typically work with. Each box represents an individual check within a checkbox list, in which participants could check multiple boxes. (b) Tasks that respondents routinely carry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

114 **Package description**

115 After conducting the community survey, we combined participant input with our own experience to develop
 116 a toolkit for palaeobiologists, the `palaeoverse` R package. The package provides auxiliary functions to
 117 support data preparation and exploration for palaeobiological analysis. A summary of the functions
 118 currently available in `palaeoverse` is provided in Table 1, with further description provided in the

119 Features section. To demonstrate the functionality and versatility of the package, we also provide usage
120 examples.

121 **Installation**

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

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

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

```
127 remotes::install_github("palaeoverse-community/palaeoverse")
```

128 Following installation, `palaeoverse` can be loaded via the `library` function in R:

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

130 **Data**

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

134 Functionality is purposely flexible in `palaeoverse` and can be applied to various data sources with ease.

135 In most instances, the returned object from a function is also a dataframe, which we consider the easiest
136 data structure for most users to understand and work with. Although this might be undesirable for some
137 advanced R users, transforming data structures should be straightforward for these users.

138 **Functions**

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

142 Detailed descriptions of the functions are provided herein.

143 Table 1: A summary table of the functions currently available in the `palaeoverse` R package and respective dependencies. Base R dependencies
 144 are highlighted with an asterisk.

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

146 **Example datasets**

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

157 **Time bins**

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

```
162 # Get stage-level time bins
163 time_bins(interval = "Phanerozoic", rank = "stage", plot = TRUE)
```

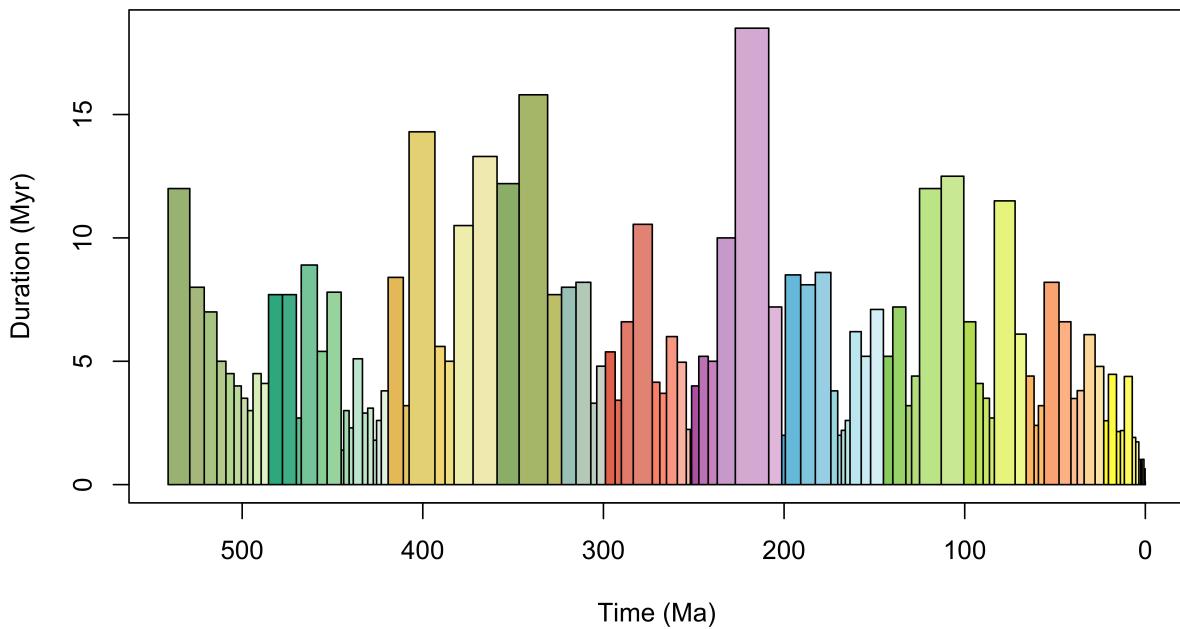


Figure 2: Phanerozoic stage-level time bins. Plot depicts the unevenness in duration of stratigraphic time bins. Bar colour filling follows the established colour scheme of the International Commission on Stratigraphy (<https://stratigraphy.org/>).

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

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

Mean bin length = 15.03 (standard deviation = 2.12)

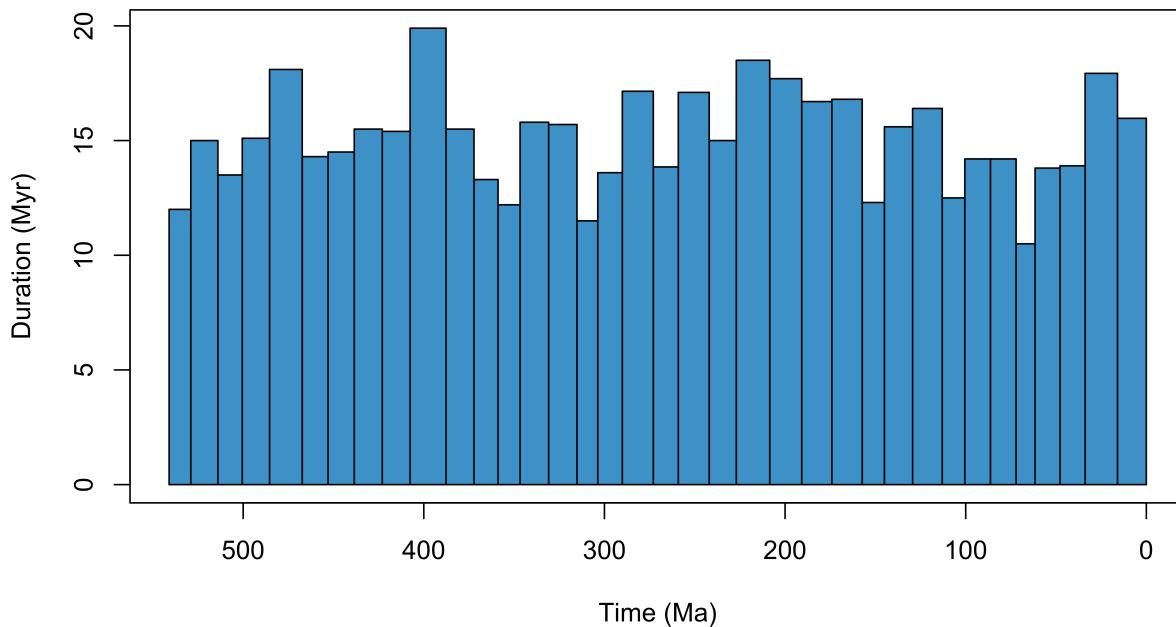


Figure 3: Phanerozoic near-equal-length time bins. Plot depicts composite stratigraphic bins (grouping stage-level bins) for the Phanerozoic of a target bin size of 15 million years. **Note:** time bins are still uneven but less so than stage-level bins.

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

177 Occurrence binning

178 Fossil occurrences are frequently ‘binned’ into distinct time intervals to enable quantification of changes
179 (e.g. biodiversity or disparity) through geological time, as described by the respondents to our survey
180 (Figure 1). The function `bin_time` allows users to assign occurrences into time bins generated by the
181 function `time_bins`, or those defined by the user:

182 `# Generate temporal bins`
183 `bins <- time_bins()`

```
184 # Assign occurrences to bins
185 bin_time(occdf = tetrapods, bins = bins, method = "mid")
186 Whilst binning occurrences with tightly defined temporal limits is straightforward and has been
187 implemented in other R packages (e.g. Lloyd, 2016), those with poorly constrained maximum and minimum
188 ages can span several intervals, and therefore cannot be easily assigned to a single bin. Palaeontologists
189 have identified numerous solutions to tackle this problem (e.g. Lloyd et al., 2012; Silvestro et al., 2016;
190 Davies et al., 2017; Dean, Chiarenza and Maidment, 2020; Franeck and Liow, 2020), but there is currently
191 no consensus on the best methodological approach or subsequent implementation. The bin_time function
192 provides five approaches defined by the ‘method’ argument: ‘mid’ (assigned based on the midpoint of the
193 temporal range of the occurrence), ‘majority’ (assigned to the bin which covers the majority of the temporal
194 range of the occurrence), ‘all’ (assigned to all bins within the temporal range of the occurrence), ‘random’
195 (assigned randomly to bins with equal probability within the temporal range of the occurrence, repeated up
196 to assigned ‘reps’), and ‘point’ (assigned randomly using a user-defined probability distribution over the
197 temporal range of the occurrence, repeated up to assigned ‘reps’). We hope that formally including these
198 options within the bin_time function will encourage palaeontologists to routinely explore and compare
199 the outcomes of various binning approaches with ease.
```

200 In recent years, palaeobiologists have developed a heightened interest in the spatial structure of the fossil
201 record, with studies focused on understanding the spatial distribution of biodiversity and the processes that
202 drive them (Figure 1) (Vilhena and Smith, 2013; Antell et al., 2020; Close et al., 2020b; Chiarenza et al.,
203 2022; Flannery-Sutherland, Silvestro and Benton, 2022b; Jones et al., 2022). In order to support such
204 analyses, `bin_space` has been developed for `palaeoverse`. The function allows the user to assign
205 occurrence data into equal-area grid cells using discrete hexagonal grids via the `h3jsr` package (O’Brien,
206 2023). Additional functionality allows simultaneous assignation of occurrence data to cells of a finer-scale
207 (i.e. a ‘sub-grid’) within the primary grid. This might be desirable for users to evaluate differences in the
208 amount of area occupied by occurrences within their primary grid cells.

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

212 Understanding the latitudinal distribution of biodiversity in deep time has also gained research interest in
213 recent years (Powell, 2009; Mannion et al., 2012, 2014; Allen et al., 2020; Song et al., 2020; Jones et al.,
214 2021). To ease implementation of such analyses, we have developed two functions, `lat_bins` and
215 `bin_lat`, which can be used to generate latitudinal bins of a given size and assign occurrence data to those
216 respective bins.

```
217 # Generate latitudinal bins  
218 bins <- lat_bins(size = 15)  
219 # Assign occurrences to bins  
220 bin_lat(occdf = tetrapods, bins = bins)
```

221 Palaeogeographic reconstruction

222 Using the present-day coordinates of fossil occurrences, plate rotation models can be used to reconstruct
223 their location at the time of deposition. Existing fossil databases provide reconstructed coordinates for
224 occurrences from only one or two of the many plate rotation models available (if any), and it is not always
225 clear which model (or version of the model) has been used. This lack of transparency is reflected in some
226 published articles that only cite the use of GPlates to reconstruct palaeocoordinates, yet lack specifics on
227 which plate rotation model was used with the GPlates Web Service or desktop application (Müller et al.,
228 2018). Furthermore, the uncertainty in palaeogeographic reconstructions is often underappreciated;
229 reconstructed coordinates are treated as being well-established, rather than model-based estimates. Finally,
230 online databases do not provide palaeocoordinates for all known samples. Both published and unpublished
231 data (e.g. museum specimens) exists outside of online databases for which researchers might require
232 palaeocoordinates.

233 We have developed the function `palaeorotate` to address these shortcomings. The function allows
234 palaeocoordinates to be reconstructed within R using two different approaches: ‘point’ and ‘grid’. The first
235 approach makes use of the GPlates Web Service and allows point data to be rotated to specific ages using
236 the available models (see <https://gwsdoc.gplates.org>). The second approach uses reconstruction files of pre-
237 generated palaeocoordinates to spatiotemporally link occurrences’ modern coordinates and age estimates
238 with their respective palaeocoordinates. These reconstruction files were generated using an equal-area
239 hexagonal grid (~100 km spacings) via the `h3jsr` package (O’Brien, 2023), and allow palaeocoordinates
240 to be generated efficiently for large datasets. Furthermore, these reconstruction files allows the user to
241 calculate the palaeolatitudinal range between reconstructed coordinates, as well as the great circle distance
242 between the two most distant points (i.e. the palaeogeographic uncertainty). Finally, to encourage
243 transparency in palaeobiological research, the function also reports additional information such as the plate
244 rotation model used.

```
245 # Add midpoint age for rotation  
246 tetrapods$age <- (tetrapods$max_ma + tetrapods$min_ma) / 2  
247 # Palaeorotate occurrences and return uncertainty  
248 palaeorotate(occdf = tetrapods, method = "grid", uncertainty = TRUE)
```

249 **Taxon-related features**

250 When working with large occurrence datasets, errors can easily creep into data. One frequently encountered
251 issue is spelling variations of the same taxon name. This can have undesirable consequences when
252 calculating metrics such as taxonomic richness or abundance. The `tax_check` function computes character
253 string distances between taxonomic names via the heuristic Jaro distance metric (Jaro, 1989). This metric
254 provides a measure of dissimilarity between character strings of 0 (exact match) to 1 (completely
255 dissimilar). During function call, the user defines a threshold for string dissimilarity to identify potential
256 synonyms. In `tax_check`, Jaro distances are calculated via the `stringdistmatrix` function from the
257 `stringdist` package (van der Loo, 2014). This function is provided to help researchers perform a spell
258 check on their dataset, with further similar functionality available in the `fossilbrush` package (Flannery-
259 Sutherland et al., 2022a). However, it should be made clear that this is no replacement for thorough
260 taxonomic vetting.

```
261 # Check for taxonomic errors
262 tax_check(taxdf = tetrapods, name = "genus")
```

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

```
278 # Evaluate unique taxa
279 tax_unique(occdf = tetrapods, genus = "genus", family = "family",
280             order = "order", class = "class", resolution = "genus")
```

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

```

289 # Remove NA data
290 tetrapods <- subset(tetrapods, !is.na(order))
291 # Compute temporal range of orders
292 tax_range_time(occdf = tetrapods, name = "order", plot = TRUE)
  
```

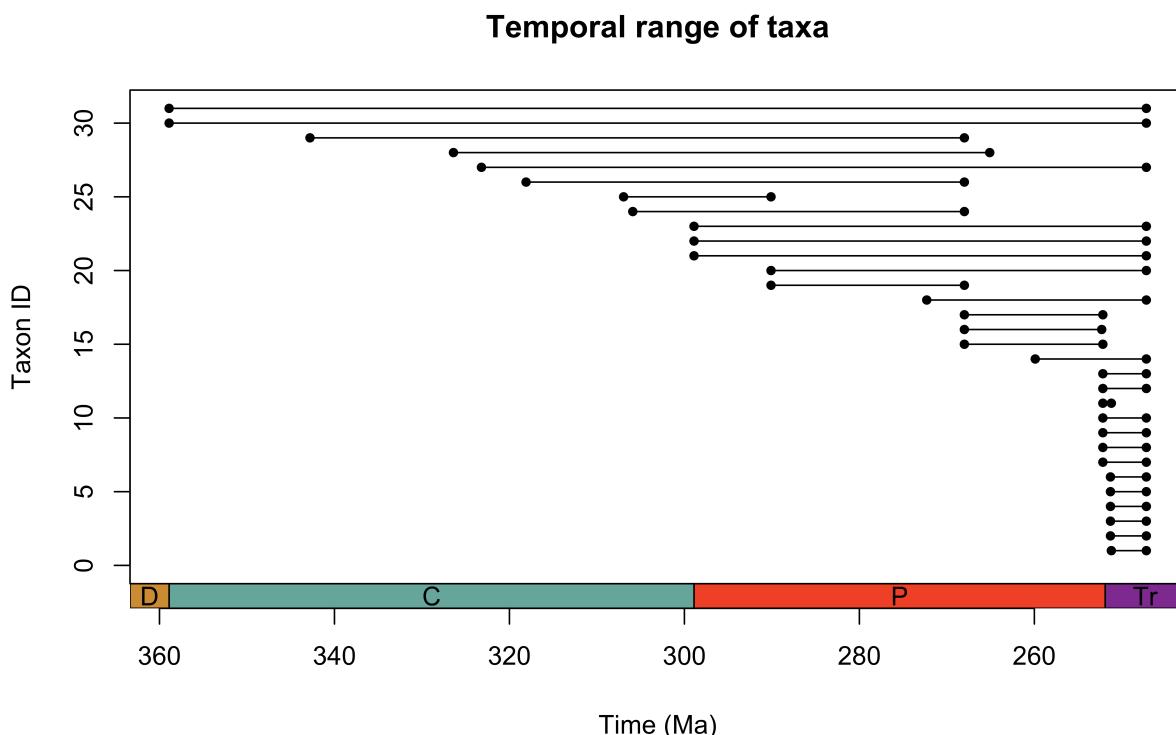


Figure 4: Temporal range of tetrapod orders in the `palaeoverse` example dataset.

```

293 # Compute Latitudinal range of orders
294 tax_range_space(occdf = tetrapods, name = "order", method = "lat")
  
```

295 The provided `tax_expand_time` and `tax_expand_lat` functions are complementary to the taxonomic
 296 range functions. They convert temporal or latitudinal range data to bin-level pseudo-occurrences. These
 297 pseudo-occurrences serve to fill in ghost ranges, in which a taxon is presumed to be present, but no record

298 exists. While these pseudo-occurrences should not be treated as equivalent to actual occurrence data, such
299 data can be useful for performing statistical analyses where bin-level data is required.

300 **Phylogeny wrangling**

301 The function `phylo_check` compares a list of taxonomic names to the list of tip names in a user-provided
302 phylogeny using the `ape` package (Paradis and Schliep, 2019). This comparison can be provided as a table
303 describing the presence or absence of each taxon in the list and/or tips, or as counts of taxa present only in
304 the list, only in the phylogeny, or in both. The function can also be used to trim the phylogeny to only
305 include branches whose tip names are included within the list of taxonomic names.

306 **Additional features**

307 Datasets are frequently explored within groups in palaeobiology, such as time bins, collections or regions.
308 The `group_apply` function has been included to allow users to run functions over a single, or multiple
309 grouping variables, with ease.

```
310 # Compute the number of occurrences per collection
311 group_apply(occdf = tetrapods, group = "collection_no", fun = nrow)
```

312 A common difficulty faced by palaeontologists is that the temporal information associated with fossil
313 occurrence data is often asynchronous, and not directly comparable. Temporal data may be provided as
314 either character-based interval names or numeric ages, and might conform to different time scales
315 (e.g. international geological stages, or North American land mammal ages). Although interval names tend
316 to be relatively stable over time, numerical age estimates are frequently updated with improved dating
317 techniques, or the collection of new data. Consequently, where possible, interval names should be used to
318 correlate occurrences from different stratigraphic time scales. The `look_up` function is provided to help
319 assign a common time scale—typically international stages—to occurrence data. This is achieved with a user-
320 defined table that links chosen interval names to corresponding stages on a common time scale (see example
321 dataset `interval_key`). Numerical ages for the assigned stages can be provided by the user, or looked up
322 in `GTS2012` or `GTS2020` (the default). This functionality therefore enables numerical ages to be assigned
323 to datasets only containing character-based interval names (e.g. “Maastrichtian”).

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

328 Finally, a common feature request from our survey (Figure 1) was the ability to add the ‘Geological Time
329 Scale’ to time-series plots in base R, with similar behaviour to the `deeptime` R package (Gearty, 2023) for

330 ggplot2 (Wickham, 2016). To address this request, the `axis_geo` function has been developed for the
331 `palaeoverse` package (Figure 5).

```
332 # Palaeorotate reef dataset
333 reefs <- palaeorotate(occdf = reefs, age = "interval_mid_ma")
334 # Plot palaeolatitudinal distribution through time
335 plot(x = reefs$interval_mid_ma, y = reefs$p_lat,
336       xlab = "Time (Ma)", ylab = "Palaeolatitude (\u00b0)",
337       xlim = c(541, 0), xaxt = "n", type = "p", pch = 20)
338 # Add Geological Time Scale
339 axis_geo(side = 1, intervals = "periods")
```

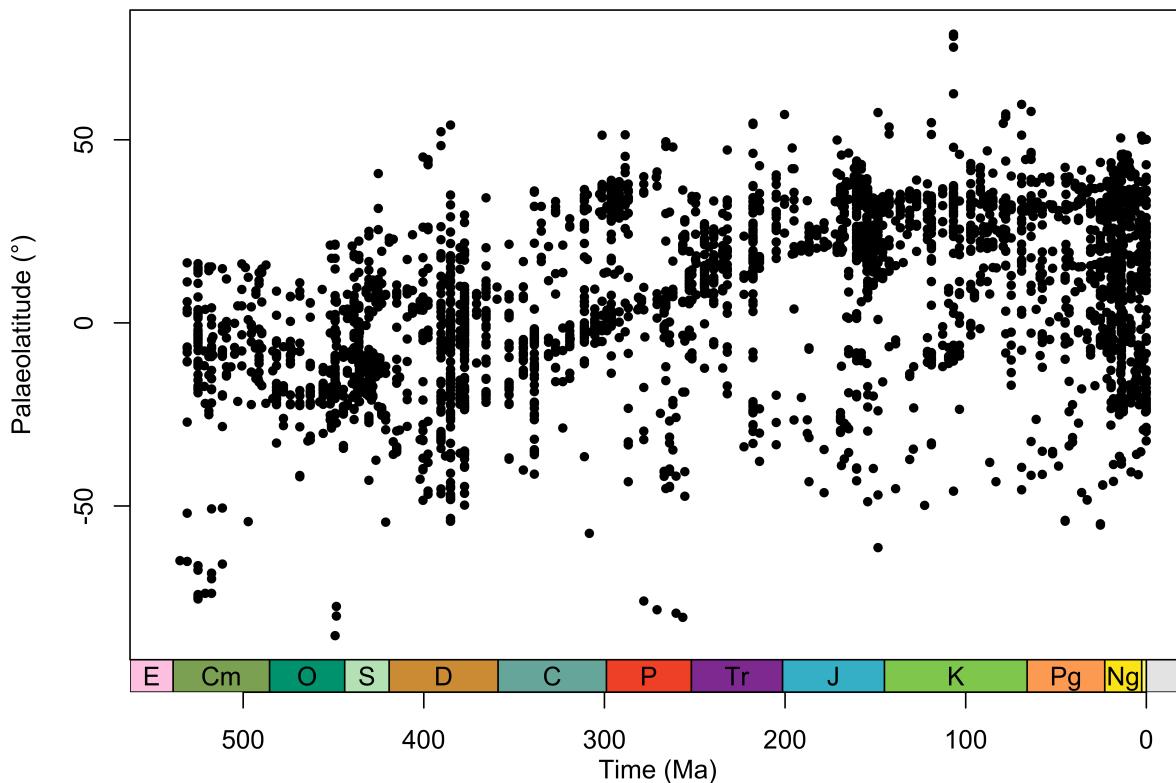


Figure 5: Example Phanerozoic plot of the palaeolatitudinal distribution of reefs through time. The plot demonstrates the usage of the `axis_geo` function for adding the Geological Time Scale to a base R plot.

340 **Resources**

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

347 **Future perspectives**

348 Palaeoverse is envisioned as a community project. While the initial development of the **palaeoverse** R
349 package was led by the authors of this manuscript, it was also informed by the perspectives of 35 additional
350 researchers (survey participants). Our hope is that **palaeoverse** will evolve into a community-driven
351 package by welcoming contributions from the wider palaeontological community to broaden available
352 functionality. To support this aim, we provide guidance on how the community can contribute to
353 **palaeoverse** on the package website (<http://palaeoverse.palaeoverse.org>). Our working group also has
354 the wider aim of establishing community standards and consensus in computational palaeobiological
355 research and facilitating comparisons across studies. Through the **palaeoverse** R package, we hope to
356 assist in making code more familiar and readable to fellow researchers, prevent researchers from
357 ‘reinventing the wheel’ for common procedures, and improve the overall reproducibility of research
358 through the use of computational tools which have been vetted and accepted by the broader community.

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

367 **Acknowledgements**

368 The authors are extremely grateful to all survey respondents who helped to shape the development of
369 `palaeoverse`. Special thanks are given to Emma M. Dunne whom participated in numerous discussions,
370 and shared her experience with the development team. Thanks are also given to two anonymous reviewers
371 that helped improve this manuscript. The contributions of LAJ, SG, and AAC were supported by the
372 European Research Council under the European Union’s Horizon 2020 research and innovation program
373 (grant agreement 947921; MAPAS project). AAC was also supported by a Juan de la Cierva-formación
374 2020 fellowship funded by FJC2020-044836-I / MCIN /AEI / 10.13039 /501100011033 from the European
375 Union “NextGenerationEU”/PRTR. The contributions of WG were supported by the Population Biology
376 Program of Excellence Postdoctoral Fellowship from the University of Nebraska-Lincoln School of
377 Biological Sciences and the Lerner-Gray Postdoctoral Research Fellowship from the Richard Gilder
378 Graduate School at the American Museum of Natural History. The contributions of BJA were supported
379 by an ETH+ grant (BECCY). The contributions of CDD (RF_ERE_210013), MK (RGF_EA_180318) and
380 CN (RGF_R1_180020) were supported by Royal Society grants. The contributions of PLG were supported
381 by a FAPESP postdoctoral grant (2022/05697-9). This is Paleobiology Database publication no XXX.

382 **Authors’ contributions**

383 LAJ conceived the project. All authors contributed to developing the project. LAJ, BJA, WG, KE, CD, and
384 JFS contributed the code. All authors contributed to testing and reviewing the code. SG processed the
385 survey results and produced the survey figures. All authors contributed to writing the manuscript.

386 **Data accessibility**

387 The `palaeoverse` R package is hosted on CRAN (<https://cran.r-project.org/web/packages/palaeoverse/>)
388 and is available on GitHub (<https://github.com/palaeoverse-community/palaeoverse>). All example datasets
389 are bundled with the R package. All code is released under a GPL (>= 3) license.

390 **References**

- 391 Allen, B.J., Wignall, P.B., Hill, D.J., Saupe, E.E. and Dunhill, A.M. (2020) The latitudinal diversity
392 gradient of tetrapods across the permo-triassic mass extinction and recovery interval. *Proceedings of the*
393 *Royal Society B*, **287**, 20201125.
- 394 Antell, G.S., Kiessling, W., Aberhan, M. and Saupe, E.E. (2020) Marine biodiversity and geographic
395 distributions are independent on large scales. *Current Biology*, **30**, 115–121.e5.

- 396 Barido-Sottani, J., Pett, W., O'Reilly, J.E. and Warnock, R.C. (2019) FossilSim: An r package for
397 simulating fossil occurrence data under mechanistic models of preservation and recovery. *Methods in*
398 *Ecology and Evolution*, **10**, 835–840.
- 399 Bell, M.A. and Lloyd, G.T. (2015) *Strap: An r Package for Plotting Phylogenies Against Stratigraphy*
400 and Assessing Their Stratigraphic Congruence. Wiley Online Library.
- 401 Benton, M.J. and Harper, D. (1999) The history of life: Large databases in palaeontology. *Numerical*
402 *palaeobiology*, 249–283.
- 403 Chao, A., Gotelli, N.J., Hsieh, T.C., Sande, E.L., Ma, K.H., Colwell, R.K., et al. (2014) Rarefaction and
404 extrapolation with hill numbers: A framework for sampling and estimation in species diversity studies.
405 *Ecological Monographs*, **84**, 45–67.
- 406 Chiarenza, A.A., Mannion, P.D., Farnsworth, A., Carrano, M.T. and Varela, S. (2022) Climatic
407 constraints on the biogeographic history of mesozoic dinosaurs. *Current Biology*, **32**, 570–585.
- 408 Close, R.A., Benson, R.B.J., Alroy, J., Carrano, M.T., Cleary, T.J., Dunne, E.M., et al. (2020a) The
409 apparent exponential radiation of phanerozoic land vertebrates is an artefact of spatial sampling biases.
410 *Proceedings of the Royal Society B: Biological Sciences*, **287**, 20200372.
- 411 Close, R., Benson, R.B., Saupe, E., Clapham, M. and Butler, R. (2020b) The spatial structure of
412 phanerozoic marine animal diversity. *Science*, **368**, 420–424.
- 413 Csárdi, G., Hester, J., Wickham, H., Chang, W., Morgan, M. and Tenenbaum, D. (2021) *Remotes: R*
414 *Package Installation from Remote Repositories, Including 'GitHub'*.
- 415 Darroch, S.A., Casey, M.M., Antell, G.S., Sweeney, A. and Saupe, E.E. (2020) High preservation
416 potential of paleogeographic range size distributions in deep time. *The American Naturalist*, **196**, 454–
417 471.
- 418 Davies, T.W., Bell, M.A., Goswami, A. and Halliday, T.J. (2017) Completeness of the eutherian mammal
419 fossil record and implications for reconstructing mammal evolution through the cretaceous/paleogene
420 mass extinction. *Paleobiology*, **43**, 521–536.
- 421 Dean, C.D., Chiarenza, A.A. and Maidment, S.C. (2020) Formation binning: A new method for increased
422 temporal resolution in regional studies, applied to the late cretaceous dinosaur fossil record of north
423 america. *Palaeontology*, **63**, 881–901.
- 424 Filazzola, A. and Lortie, C. (2022) A call for clean code to effectively communicate science. *Methods in*
425 *Ecology and Evolution*, **13**, 2119–2128.
- 426 Flannery-Sutherland, J.T., Raja, N.B., Kocsis, Á.T. and Kiessling, W. (2022a) Fossilbrush: An r package
427 for automated detection and resolution of anomalies in palaeontological occurrence data. *Methods in*
428 *Ecology and Evolution*, **13**, 2404–2418.
- 429 Flannery-Sutherland, J.T., Silvestro, D. and Benton, M.J. (2022b) Global diversity dynamics in the fossil
430 record are regionally heterogeneous. *Nature Communications*, **13**, 1–17.
- 431 Franeck, F. and Liow, L.H. (2020) Did hard substrate taxa diversify prior to the great ordovician
432 biodiversification event? *Palaeontology*, **63**, 675–687.
- 433 Fraser, D. (2017) Can latitudinal richness gradients be measured in the terrestrial fossil record?
434 *Paleobiology*, **43**, 479–494.

- 435 Furness, E.N., Garwood, R.J., Mannion, P.D. and Sutton, M.D. (2021) Evolutionary simulations clarify
436 and reconcile biodiversity-disturbance models. *Proceedings of the Royal Society B*, **288**, 20210240.
- 437 Garwood, R.J., Spencer, A.R. and Sutton, M.D. (2019) REvoSim: Organism-level simulation of macro
438 and microevolution. *Palaeontology*, **62**, 339–355.
- 439 Gearty, W. (2023) *Deeptime: Plotting Tools for Anyone Working in Deep Time*.
- 440 Gradstein, F.M., Ogg, J.G., Schmitz, M. and Ogg, G. (2012) *The Geologic Time Scale 2012*. Elsevier.
- 441 Gradstein, F.M., Ogg, J.G., Schmitz, M.D. and Ogg, G.M. (2020) *Geologic Time Scale 2020*. Elsevier.
- 442 Guillerme, T. (2018) dispRity: A modular r package for measuring disparity. *Methods in Ecology and
443 Evolution*, **9**, 1755–1763.
- 444 Hijmans, R.J. (2022) *Geosphere: Spherical Trigonometry*.
- 445 Jaro, M.A. (1989) Advances in record-linkage methodology as applied to matching the 1985 census of
446 tampa, florida. *Journal of the American Statistical Association*, **84**, 414–420.
- 447 Jones, L.A., Dean, C.D., Mannion, P.D., Farnsworth, A. and Allison, P.A. (2021) Spatial sampling
448 heterogeneity limits the detectability of deep time latitudinal biodiversity gradients. *Proceedings of the
449 Royal Society B*, **288**, 20202762.
- 450 Jones, L.A., Mannion, P.D., Farnsworth, A., Bragg, F. and Lunt, D.J. (2022) Climatic and tectonic drivers
451 shaped the tropical distribution of coral reefs. *Nature communications*, **13**, 1–10.
- 452 Kiessling, W. and Krause, C. (2022) PaleoReefs database (PARED).
- 453 Kocsis, Á.T., Reddin, C.J., Alroy, J. and Kiessling, W. (2019) The r package divDyn for quantifying
454 diversity dynamics using fossil sampling data. *Methods in Ecology and Evolution*, **10**, 735–743.
- 455 Lloyd, G.T. (2016) Estimating morphological diversity and tempo with discrete character-taxon matrices:
456 Implementation, challenges, progress, and future directions. *Biological journal of the Linnean society*.
457 *Biological Journal of the Linnean Society*, **118**, 131–151.
- 458 Lloyd, G.T., Pearson, P.N., Young, J.R. and Smith, A.B. (2012) Sampling bias and the fossil record of
459 planktonic foraminifera on land and in the deep sea. *Paleobiology*, **38**, 569–584.
- 460 Mannion, P.D., Benson, R.B., Carrano, M.T., Tennant, J.P., Judd, J. and Butler, R.J. (2015) Climate
461 constrains the evolutionary history and biodiversity of crocodylians. *Nature communications*, **6**, 1–9.
- 462 Mannion, P.D., Benson, R.B., Upchurch, P., Butler, R.J., Carrano, M.T. and Barrett, P.M. (2012) A
463 temperate palaeodiversity peak in mesozoic dinosaurs and evidence for late cretaceous geographical
464 partitioning. *Global Ecology and Biogeography*, **21**, 898–908.
- 465 Mannion, P.D., Upchurch, P., Benson, R.B. and Goswami, A. (2014) The latitudinal biodiversity gradient
466 through deep time. *Trends in Ecology & Evolution*, **29**, 42–50.
- 467 Mannion, P.D., Upchurch, P., Carrano, M.T. and Barrett, P.M. (2011) Testing the effect of the rock
468 record on diversity: A multidisciplinary approach to elucidating the generic richness of sauropodomorph
469 dinosaurs through time. *Biological Reviews*, **86**, 157–181.

- 470 Müller, R.D., Cannon, J., Qin, X., Watson, R.J., Gurnis, M., Williams, S., et al. (2018) GPlates: Building
471 a virtual earth through deep time. *Geochemistry, Geophysics, Geosystems*, **19**, 2243–2261.
- 472 O'Brien, L. (2023) *H3jsr: Access Uber's H3 Library*.
- 473 Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2020) *Vegan: Community Ecology Package*.
- 475 Ooms, J. (2023) *Curl: A Modern and Flexible Web Client for r*.
- 476 Paradis, E. and Schliep, K. (2019) Ape 5.0: An environment for modern phylogenetics and evolutionary
477 analyses in R. *Bioinformatics*, **35**, 526–528.
- 478 Pebesma, E. (2018) Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*,
479 **10**, 439–446.
- 480 Powell, M.G. (2009) The latitudinal diversity gradient of brachiopods over the past 530 million years. *The
481 Journal of Geology*, **117**, 585–594.
- 482 Quental, T.B. and Marshall, C.R. (2013) How the red queen drives terrestrial mammals to extinction.
483 *Science*, **341**, 290–292.
- 484 R Core Team. (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for
485 Statistical Computing, Vienna, Austria.
- 486 Sepkoski, J.J. (1978) A kinetic model of phanerozoic taxonomic diversity i. Analysis of marine orders.
487 *Paleobiology*, **4**, 223–251.
- 488 Sepkoski, D. and Ruse, M. (2009) *The Paleobiological Revolution: Essays on the Growth of Modern
489 Paleontology*. University of Chicago Press.
- 490 Silvestro, D., Salamin, N. and Schnitzler, J. (2014) PyRate: A new program to estimate speciation and
491 extinction rates from incomplete fossil data. *Methods in Ecology and Evolution*, **5**, 1126–1131.
- 492 Silvestro, D., Zizka, A., Bacon, C.D., Cascales-Minana, B., Salamin, N. and Antonelli, A. (2016) Fossil
493 biogeography: A new model to infer dispersal, extinction and sampling from palaeontological data.
494 *Philosophical Transactions of the Royal Society B: Biological Sciences*, **371**, 20150225.
- 495 Solymos, P. and Zawadzki, Z. (2023) *Pbapply: Adding Progress Bar to 'Apply' Functions*.
- 496 Song, H., Huang, S., Jia, E., Dai, X., Wignall, P.B. and Dunhill, A.M. (2020) Flat latitudinal diversity
497 gradient caused by the permian–triassic mass extinction. *Proceedings of the National Academy of
498 Sciences*, **117**, 17578–17583.
- 499 Starrfelt, J. and Liow, L.H. (2016) How many dinosaur species were there? Fossil bias and true richness
500 estimated using a poisson sampling model. *Philosophical Transactions of the Royal Society B: Biological
501 Sciences*, **371**, 20150219.
- 502 van der Loo, M.P.J. (2014) The stringdist package for approximate string matching. *The R Journal*, **6**,
503 111–122.
- 504 Vilhena, D.A. and Smith, A.B. (2013) Spatial bias in the marine fossil record. *PLoS One*, **8**, e74470.
- 505 Wickham, H. (2016) *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

- 506 Wickham, H. (2022) *Htr: Tools for Working with URLs and HTTP*.
- 507 Zaffos, A., Finnegan, S. and Peters, S.E. (2017) Plate tectonic regulation of global marine animal
508 diversity. *Proceedings of the National Academy of Sciences*, **114**, 5653–5658.
- 509 Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., et al. (2019)
510 CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases.
511 *Methods in Ecology and Evolution*, **10**, 744–751.