PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology

Søren Faurby[1,2], Matt Davis[3,4], Rasmus Østergaard Pedersen[3,4], Simon D. Schowanek[3,4], Alexandre Antonelli[1,2,5,6], Jens-Christian Svenning[3,4]

[1]University of Gothenburg, Department of Biological and Environmental Sciences, Gothenburg 405 30, Sweden
[2]Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Göteborg, Sweden
[3]Section for Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Ny Munkegade 114, 8000 Aarhus C, Denmark
[4]Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Aarhus University, Ny Munkegade 114, 8000 Aarhus C, Denmark
[5]Gothenburg Botanical Garden, SE-413 19 Göteborg, Sweden
[6] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02139, USA.

Abstract

Data needed for macroecological analyses are difficult to compile and often hidden away in supplementary material under non-standardized formats. Phylogenies, range data, and trait data often use conflicting taxonomies and require ad hoc decisions to synonymize species or fill in large amounts of missing data. Furthermore, most available data sets ignore the large impact that humans have had on species ranges and diversity. Ignoring these impacts can lead to drastic differences in diversity patterns and estimates of the strength of biological rules. To help overcome these issues, we assembled PHYLACINE, The Phylogenetic Atlas of Mammal Macroecology. This taxonomically integrated platform contains phylogenies, range maps, trait data, and threat status for all 5,831 known mammal species that lived since the last interglacial (~130,000 years ago until present). PHYLACINE is ready to use directly, as all taxonomy and metadata are consistent across the different types of data, and files are provided in easy-to-use formats. The atlas includes both maps of current species ranges and present natural ranges, which represent estimates of where species would live without anthropogenic pressures. Trait data include body mass and coarse measures of life habit and diet. Data gaps have been minimized through extensive literature searches and clearly labelled imputation of missing values. The PHYLACINE database will be archived here as well as hosted online so that users may easily contribute updates and corrections to continually improve the data. This database will be useful to any researcher who wishes to investigate large scale ecological patterns. Previous versions of the database has already provided valuable information and have for instance shown that

megafauna extinctions caused substantial changes in vegetation structure and nutrient transfer patterns across the globe.

Keywords
Body size, diet, distributions, IUCN, mammal, mass, phylogeny, present natural, range maps

Metadata

CLASS I. DATA SET DESCRIPTORS

A. Data set identity: PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology

B. Data set identification code:

(1) Trait_data.csv
(2) Synonymy_table_valid_species_only.csv
(3) Synonymy_table_with_unaccepted_species.csv
(4) Complete_phylogeny.nex
(5) Small_phylogeny.nex
(6) Current ranges
(7) Present_natural ranges
(8) Spatial_metadata.csv

C. Data set description:

Originators:

Søren Faurby; University of Gothenburg, Department of Biological and Environmental Sciences, Gothenburg 405 30, Sweden; Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Göteborg, Sweden.

Jens-Christian Svenning; Section for Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Ny Munkegade 114, 8000 Aarhus C, Denmark; Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Aarhus University, Ny Munkegade 114, 8000 Aarhus C, Denmark.

D. Key words:

mammal, mass, body size, diet, phylogeny, range maps, present natural, IUCN, distributions

CLASS II. RESEARCH ORIGIN DESCRIPTORS

A. Overall project description

Identity:
MegaPast2Future and HISTFUNC

Originator:
Jens-Christian Svenning; Section for Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Ny Munkegade 114, 8000 Aarhus C, Denmark; Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Aarhus University, Ny Munkegade 114, 8000 Aarhus C, Denmark

Period of Study:
2013-Now

Objectives:
MegaPast2Future aims to provide a deeper, synthetic understanding of megafauna ecosystem ecology and its potential role in developing a sustainable, biodiverse future. HISTFUNC aims to understand historical constraints on functional diversity and ecosystem functioning, with mammals as one focus group.

Abstract:
Data needed for macroecological analyses are difficult to compile and often hidden away in supplementary material under non-standardized formats. Phylogenies, range data, and trait data often use conflicting taxonomies and require ad hoc decisions to synonymize species or fill in large amounts of missing data. Furthermore, most available data sets ignore the large impact that humans have had on species ranges and diversity. Ignoring these impacts can lead to drastic differences in diversity patterns and estimates of the strength of biological rules. To help overcome these issues, we assembled PHYLACINE, The Phylogenetic Atlas of Mammal Macroecology. This taxonomically integrated platform contains phylogenies, range maps, trait data, and threat status for all 5,831 known mammal species that lived since the last interglacial (~130,000 years ago until present). PHYLACINE is ready to use directly, as all taxonomy and metadata are consistent across the different types of

data, and files are provided in easy-to-use formats. The atlas includes both maps of current species ranges and present natural ranges, which represent estimates of where species would live without anthropogenic pressures. Trait data include body mass and coarse measures of life habit and diet. Data gaps have been minimized through extensive literature searches and clearly labelled imputation of missing values. The PHYLACINE database will be archived here as well as hosted online so that users may easily contribute updates and corrections to continually improve the data. This database will be useful to any researcher who wishes to investigate large scale ecological patterns. Previous versions of the database has already provided valuable information and have for instance shown that megafauna extinctions caused substantial changes in vegetation structure and nutrient transfer patterns across the globe. Any use of the data requires citation of this publication in *Ecology*, plus this database and the relevant underlying large datasets.

Source of funding:
Support was provided by Carlsberg Foundation Semper Ardens project MegaPast2Future (grant CF16-0005), European Research Council (ERC-2012-StG-310886-HISTFUNC), and VILLUM Investigator project (VILLUM FONDEN, grant 16549).

B. Specific subproject description

Data set overview:
We developed the PHYLACINE database to help address several major problems with macroecological analyses: data sets often have limited/nonrandom taxonomic or spatial extents, lack large amounts of data (Penone et al. 2014, Davis and Pineda-Munoz 2016), do not account for phylogenetic non-independence of traits (Freckleton et al. 2002), examine habitats and patterns already greatly impacted by humans, and are hard to compare with other data sets because of differing taxonomies. Our database is global in scale and includes all 5,831 mammal species that lived during the last ~130,000 years. We strove to eliminate missing data by extensive literature searches, robust phylogenetic imputations, and a novel heuristic-hierarchical Bayesian tree building approach that incorporates multiple data sources. Both current ranges and present natural ranges of species can be compared to estimate macroecological relationships with and without anthropogenic pressures. Furthermore, our entire database is taxonomically integrated and compatible with IUCN Version 2016-3 (2016) so that users can seamlessly move between trait, range,

and phylogenetic data without mismatches in species names. We have even included standardized synonymy tables so that users can easily match the PHYLACINE database to other popular trait databases like EltonTraits 1.0 (Wilman et al. 2014) and the IUCN (2016).

Taxonomy:

Extant species taxonomy followed IUCN Version 2016-3 (2016) except for a handful of extremely poorly known species we judged unlikely to be valid. A list of excluded species and justifications for deviating from IUCN taxonomy can be found in Synonymy_table_with_unaccepted_species.csv. For convenience, we have also made the same synonymy file trimmed to only species we accept in PHYLACINE 1.2 available as Synonomy_table_1.2_valid_species_only.csv. The taxonomy of fossil species was originally assembled for Sandom et al. (2014) by first consulting review publications and original research articles before independently validating each potential species. This taxonomy has since been slightly modified based on new information. Only fossil species that had records from a directly dated site within the last ~130,000 years were included. Species were generally not accepted if they could only be dated by association with indicator fossils (with the sole exception that species from oceanic islands were accepted if found in clear association with humans). We also consistently tried to lump together closely related fossil species that are part of extant lineages to avoid overinflating extinct species lists. For example, even though many extinct forms of North American bison are morphologically distinct and traditionally treated as separate species (Davis 2017), we did not include them here because genetic analysis shows they are part of the same lineage as extant American bison (Shapiro et al. 2004). This will make any estimate of extinction rates from the database conservative. Family assignment generally follows IUCN Version 2016-3 (2016) for extant species and recent taxonomic treatments for extinct ones but it is slightly modified so that all entities we classify as families are constrained to be monophyletic in the phylogeny.

Trait data:

We assembled data on life habit, body mass, diet, island endemicity and IUCN status. These data can be found in Trait_data.csv. We chose life habit, body mass, and diet because these traits are thought to be fundamentally important for an organism's biology and they are widely used in functional diversity analysis (e.g. Safi et al. 2011). They are also the most widely available traits for modern species and can be reasonably inferred for extinct species. Trait values are not tied to an explicit temporal scale (e.g., mass estimates could come from species collected on one expedition or from museum records collected over a hundred years), but it is

unknown how this would affect downstream analysis. Extant species mostly have values for traits that are recorded and averaged over only a few years, while taphonomic time averaging guarantees that most traits for extinct species will be averaged over thousands of years or more. Extant species might be more likely to display extreme trait values that would normally be averaged out when measuring extinct species (Davis and Pineda-Munoz 2016). However, this problem should be mitigated by the increased measurement errors associated with inferring body size or diets for extinct species from osteological proxies, rather than from direct measurements, which could make extreme values for extinct species more likely. Potentially more troublesome, although unlikely to create large biases, is that most extinct species in our database died out during periods with a colder climate than the present. Given the generality and strength of Bergmann's rule in large mammals (Meiri and Dayan 2003), it is possible that some large, extinct species could have evolved smaller body sizes if they had survived into our current, warmer climate. Since our database also contains a data set of a counterfactual, present natural scenario where these extinct species are still alive, we could have slightly overestimated the size of extinct species relative to extant ones. Island endemicity was included as a biogeographic trait because of the well-known effect that the Island Rule has on body size (Faurby and Svenning 2016) and because the coarse spatial resolution of our range maps would make it impossible to determine island endemicity for many species. We included IUCN statuses (IUCN 2016) to easily separate extant from extinct species in the database; to measure how species ranges, relationships, and functional traits impact their extinction risks; and to investigate how extinction risks could in turn affect functional and phylogenetic diversity. The derivation of each trait is discussed individually below.

Life habit:
Life habit data were originally assembled for (Faurby and Svenning 2015a) and consist of non-exclusive, binary measures of whether a species can be considered terrestrial, marine, or freshwater (bats were consistently scored as exclusively aerial). Codings (except for bats) follow the "Systems" designation listed for each species' description on the IUCN website (www.iucnredlist.org). Extinct species scores were phylogenetically imputed or, in very few cases, inferred from the literature.

Body mass:
Body masses were originally assembled for (Faurby and Svenning 2016) from a large number of literature sources although about 60% of the species were given values from the Mass of Mammals database (V.4.1) (Smith et al. 2003). In total, the database contains weights reported from the Mass of Mammals database or

other sources for 4861species. For another 768 species, masses were estimated by morphological correlates or based on a relative that a species was said to resemble in size. Masses for the remaining 202 species were estimated by phylogenetic imputation and clearly labelled. See Missing Data in Class IV, Section C.

Diet:
We decided to record the percentage of three coarse categories (vertebrate prey, invertebrate prey, and plants) in each species' diet because these rough dietary classifications can be accurately imputed phylogenetically (Gainsbury et al. 2018) and often inferred in fossil organisms (Davis 2017). The majority of diet percentages for species (4193 species) came from the EltonTraits 1.0 database (Wilman et al. 2014), 863 additional species were given diets from the MammalDIET database (Kissling et al. 2014) or the MammalDIET 2 database (Gainsbury et al. 2018), and 282 diets were taken or estimated directly from literature sources. The 493 remaining diets were inferred using phylogenetic imputation. See Missing Data in Class IV, Section C.

Both the MammalDIET (Kissling et al. 2014) and EltonTraits (Wilman et al. 2014) databases use standardized keys to convert qualitative dietary descriptions from literature sources into semi-quantitative estimates of diet for extant mammal species. Another diet database we used, MammalDIET 2 (Gainsbury et al. 2018) follows the exact methodology of MammalDIET (Kissling et al. 2014) so any mention of MammalDIET here represents both the original MammalDIET and its extension, MammalDIET 2. We used dietary estimates from these databases at Diet-Certainty ABC for EltonTraits and FillCode 0 or 1 for MammalDIET. This means that diet was typically determined at the specific level. Less frequently, diet could be determined at the genus level but only when the authors of these original databases expertly judged all species in a genus to have similar diet, e.g., a reference said, "the genus is completely herbivorous". We first converted EltonTraits' ten percentage dietary categories into our three dietary categories by summing EltonTraits' categories together in the following way: Diet.Plant = Diet-Fruit + Diet-Nect + Diet-Seed + Diet-PlantO, Diet.Vertebrate = Diet-Vend + Diet-Vect + Diet-Vfish + Diet-Vunk + Diet-Scav, Diet.Invertebrate = Diet-Inv.

To convert MammalDIET's ordinal diet values into percentages, we selected all species that had the same unique combination of ordinal values in MammalDIET. From this group, we selected those species that also occurred in EltonTraits and took the median of their diet percentages in EltonTraits. We assigned this median value to all species lacking a percentage diet within the original group. However, any values of 0 in MammalDIET (meaning no use of a dietary category) were set

to 0 % in our classification, regardless of what the median EltonTraits percentage might have been for that category. Lastly, all values were rescaled to sum to 100 %. For example, there were 246 species in MammalDiet that had an ordinal diet classification of 0 Vertebrate, 2 Invertebrate, and 1 Plant. 216 of these species were also in EltonTraits and had a median percentage diet of 0 % Diet.Vertebrate, 20 % Diet.Invertebrate, and 80 % Diet.Plant. This percentage diet was given to the 30 species in MammalDiet (out of the original 246) that lacked a percentage diet in EltonTraits.

For 282 species, diet percentages were taken or estimated directly from the scientific literature. If percentage values were reported or could be derived precisely, we used those. Alternatively, if it was impossible to collect precise quantitative estimates from the source, we followed the methodology outlined by (Wilman et al. 2014) to estimate diet percentages from qualitative descriptions of diet. In the case that a source stated that the diet of a species was similar to that of a known species, we used the diet of the known species as a proxy.

We stress that although we report the diet of all species in percentages for mathematical convenience, this does not necessarily imply high precision for the data. Intraspecific, and even within individual, diets can be highly variable over time and space (Davis and Pineda-Munoz 2016). The dietary data in PHYLACINE represent best estimates that allow researchers to compare the average gross diets of many different species. They are meant to be used for large-scale, macroecological studies and small differences in diet should not be over emphasized. Researchers interested in detailed descriptions of diets for particular species should consult the primary literature where more information is available.

Island endemicity:
We scored island endemicity by closely examining species' ranges and historical and fossil occurrence records. Because various definitions of island endemicity exist, we used a nested classification that expresses the different degrees of isolation. The strictest category (Occurs only on isolated islands) considers species as island endemics only if they are restricted to islands that have not been connected to a continent during Pleistocene glaciations (which we estimated as islands for which the deepest water level between the island and a continent is more than 110 m deep e.g., New Guinea, Borneo, or Java).

The second and third categories (Occurs on small land bridge islands, Occurs on large land bridge islands**)** include species that occur on land bridge islands:

islands that are separated from the mainland by water no more than 110 m deep. Therefore, these islands would have been part of the mainland during the last glacial maximum. We made the distinction between small (< 1,000 km^2) and large (≥ 1,000 km^2) land bridge islands because the low population sizes on smaller islands may have enabled rapid speciation since the isolation of the island after the last glacial maximum. Therefore, species from such islands may have been island endemic for their entire history, unlike most species endemic to larger land bridge islands. For example, the island of Escudo de Veraguas has only been separated from mainland Panama for ~8,900 years but its small size (4.3 km2) may have been enough to lead to the speciation of the pygmy three-toed sloth (*Bradypus pygmaeus*) from its mainland ancestors (Voirin 2015).

The last category (Occurs on mainland) includes those species that occur on the mainland or used to occur on the mainland during the Holocene. Species such as the Tasmanian devil (*Sarcophilus harrisii*) or the Tasmanian tiger (*Thylacinus cynocephalus*) were hence not considered island endemics under any of our definitions, since they both occurred on mainland Australia until the mid-Holocene (Johnson and Wroe 2016). A strict definition of island endemicity would include only species that occur on isolated islands. A semi-strict definition would include those species as well as species that occur on small land bridge islands. A classical definition would include all species that occur on isolated islands or land bridge islands, regardless of size. Species that never occur on land because they are strictly marine are designated as "Exclusively marine".

IUCN status:
Threat status rankings were taken from IUCN Version 2016-3 (2016). We created an additional, unofficial status "extinct in prehistory" (EP) for 270 species that went extinct before 1500 CE, the baseline for IUCN consideration. Two additional species, *Alouatta seniculus* and *Cebus capucinus*, are oddly treated only at the subspecies level by the IUCN. They were given the status of their subspecies (LC) in both cases.

Phylogeny:
The phylogeny was originally published by Faurby & Svenning (2015b) but has been updated for this database to incorporate new data and new changes in taxonomy. The phylogeny is built based on a hierarchical Bayesian approach where species are placed with various levels of freedom depending on the amount of genetic data they have. As a consequence, family level monophyly is consistently enforced while genera are considered monophyletic unless there is strong genetic evidence against their monophyly. Full details of the construction methodology can be found in

(Faurby and Svenning 2015b). Importantly, we provide a posterior distribution of 1,000 trees, which is intended to recover uncertainties in topology and branch lengths, but given the uncertainties associated, we recommend against using a single consensus or "best" tree. Imagine a genus with four species, three of which have ample genetic data while the fourth species without genetic data is placed based on taxonomy, floating freely in the genus. This uncertainty can be handled in the posterior distribution by examining multiple trees but there is no single tree that meaningfully places the fourth species without genetic data. We provide both a tree that includes all species in the dataset, Complete_phylogeny.nex; and a smaller tree, Small_phylogeny.nex, which only includes species with genetic data and species for which topological placement was unambiguous based on taxonomy. Details of genetic and topological data used to construct the trees can be found in Synonomy_table_1.2.csv. Branch lengths in both phylogenies are recorded in millions of years (Ma).

Current ranges:

Ranges for extant species were originally created for (Faurby and Svenning 2015a) but updated in PHYLACINE 1.2 to reflect newer data from IUCN Version 2016-3 (IUCN 2016). Polygon range maps for extant species were first downloaded from IUCN and projected to Behrmann cylindrical equal area rasters with a cell size of 96.5 km by 96.5 km at 30° North and 30° South. Extinct species have no current ranges so they were given empty rasters. A cell was considered as occupied if any part of the cell was overlapping with the range polygon. For all species, we only used what IUCN considered their current and natural range and thus excluded all parts coded as introduced, extinct, or probably extinct, but we kept ranges coded as reintroduced. For these ranges, we followed IUCN's decisions for all species irrespective of whether or not we agree in the assignment of natural versus introduced. For instance, we disagree with the treatment of red deer (*Cervus elaphus*) on Sardinia and Corsica, which we consider introduced but since IUCN considers them natural, we have kept them in the current ranges. We did deviate from IUCN in a few instances though, where IUCN ranges were lacking or internally inconsistent. *Homo sapiens* are listed but not mapped by IUCN so we assigned any cell in the present natural rasters containing one or more non marine mammal species as part of humans' current range given that people now occupy almost every habitable place on Earth. There are multiple instances where species are listed as extant by IUCN but have only extinct range polygons or vice versa. We deleted the current ranges of two species, *Melomys rubicola* and *Sus bucculentus*, that IUCN ranks as "EX". IUCN lists 43 extant species that lack extant ranges. We generated present natural ranges for 28 of these species but 15 species were not mapped by IUCN and showed no evidence for anthropogenic modification so we did not generate current or present natural

ranges for them. Any inconsistencies we found in IUCN's data were reported to them. Current ranges are stored as geoTIFF files in the Current file folder.

Present natural ranges:
For all species, we estimated their present natural ranges (Peterken 1977), which is potential current natural ranges if species had never experienced strong anthropogenic pressures. This implicitly assumes that all extinct species lacking a current range in our database were driven extinct, at least partially, by humans. We acknowledge that this is unlikely to be true for every species. It is also important to clarify that for extinct species, present natural ranges are not prehistoric ranges, but rather where species would live today if they had not gone extinct. For extant species, present natural ranges are also not always larger than current ranges. Some extant species, like the already mentioned red deer that have had their ranges expanded intentionally or accidentally by humans, had their current ranges reduced to estimate their present natural ranges. Present natural ranges are stored as geoTIFF files in the Present_natural file folder.

The range modifications can be grouped into eight categories of decreasing certainty: 1) range reductions for species with recent, human-induced range expansions (21 species); 2) range expansions based on other sources for species with known range declines (162 species); 3) merger of likely human caused disjunct ranges by filling intervening suitable habitats (194 species); 4) expansion of ranges to entire islands (184 species); 5) expansion of ranges for species with known, or at least highly suspected, range declines to cover suitable areas contiguous with the current range (223 species); 6) estimation of ranges based on the natural ranges of the extant species that the target species co-occurred with at fossil sites (192 species); 7) estimation of ranges based on the natural ranges of the extinct species that the target species co-occurred with at fossil sites (3 species); and 8) unique species-specific modifications in special cases (13 species). For many species, the range modifications involved several of the above processes and we listed them under the type with the perceived lowest certainty. 4824 species were not thought to have their ranges significantly impacted by humans so their present natural range is identical to their current IUCN range; they were coded as: 0) Present Natural range identical to IUCN range. We note that for species with the latter coding, the coding only meant that no other source than IUCN was used and for 80 of these species the present natural range was larger than the current range because IUCN contains part of the polygons coded as extinct or possibly extinct, which we do not include in current range but do include for the present-natural range. Detailed descriptions of how present natural ranges were estimated can be found in the supplemental material of Faurby and Svenning (2015a), where these ranges were first published.

Method 6), estimation of ranges based on the natural ranges of the extant species that the target species co-occurred with at fossil sites, may be the most important method concerning downstream macroecological analysis because it was generally employed for formerly widespread continental species like ground sloths and mammoths that may drive overall species diversity patterns. The method relies on the present natural ranges of extant species that used to co-occur with the extinct species in question. First, literature sources and paleo occurrence databases were used to identify all the currently extant, non-generalist species that were found in the same fossil sites/analysis layers as the focal extinct species. Then, any cell in the present natural ranges where 50% of more of these extant species co-occur was considered a cell where the focal fossil species could also occur in its present natural range. The logic behind this is that if all the species at a fossil site are limited by similar ecological parameters, the current distribution of extant species from that site should be informative for the climatic suitability of the extinct species whose distribution we are trying to infer. This does not assume that species are found in the same area today as they did in the past but rather that the focal species would change its distribution in a similar way to the other species it used to co-occur with if it had not gone extinct. This does implicitly assume a constant age of all fossils found together or at least that they were all deposited under identical climatic conditions. This method generally provides larger distributions for species with wider prehistoric ranges but we stress that it does not necessarily produce larger ranges for species that readily fossilize, a desirable property. For example, the American mastodon, *Mammut americanum*, has a similarly sized present natural range as the apparently equally widespread but much more rarely fossilized skunk species *Brachyprotoma obtusata*.

The reliability of the method was assessed by applying the method to all extant carnivores and ungulates from the USA and Canada. It was found that the method produced diversity estimates that were highly correlated with the present natural diversities when these were instead based on historical range information (ρ=0.856), and this correlation was in fact higher than the correlation between current diversity and present natural diversity (ρ=0.762) (Faurby and Svenning 2015a). We stress that this does not guarantee that the method produces accurate distributions for all species but only that the method appears to be unbiased on average. More detailed descriptions of this and the other methods can be found in the supplemental material of Faurby and Svenning (2015a). We have also provided an updated description of the modifications for each species here in Spatial_metadata.csv. Present natural ranges were mapped using the same resolution and projection as current ranges and again a cell was coded as within the range if any part of the estimated range polygon was within the cell.

Project personnel:
In addition to the authors of the database, we wish to highlight the work by Christopher Sandom who shared first authorship on a paper defining the taxonomy for all large continental extinct mammals.

CLASS III. DATA SET STATUS AND ACCESSIBILITY

A. Status

Latest update:
May 2018. Data compilation is ongoing and we plan to update the database regularly at MegaPast2Future.github.io (http://doi.org/10.5281/zenodo.1250504).

Latest archive date:
n/a

Metadata status:
Metadata updated with dataset in May 2018.

Data verification:
The latest development version of the PHYLACINE database will be hosted at MegaPast2Future.github.io so that any user may easily submit an error report through a simple online form and request that erroneous values be changed. After a number of these error reports are submitted, we will rebuild the database and release a new stable version to the Dryad Digital Data Repository that will supplant the current version of the database already hosted there. This ensures that the database will be updated to reflect new research and findings and that over time, errors can be found and fixed.

B. Accessibility

Storage location and medium:
The database is available in the Dryad Digital Data Repository (https://doi.org/10.5061/dryad.bp26v20). To find news and the latest development version of the database, users should visit MegaPast2Future.github.io (http://doi.org/10.5281/zenodo.1250504).

Contact person:
Søren Faurby, University of Gothenberg, (soren.faurby@bioenv.gu.se).

Copyright restrictions:

Proprietary restrictions:
None.

Costs:
None.

CLASS IV. DATA STRUCTURAL DESCRIPTORS

A. Data set file

Identity:
(1) Trait_data.csv
(2) Synonymy_table_valid_species_only.csv
(3) Synonymy_table_with_unaccepted_species.csv
(4) Complete_phylogeny.nex
(5) Small_phylogeny.nex
(6) Current ranges
(7) Present_natural ranges
(8) Spatial_metadata.csv

Size:
(1) 139,968 records (including header) and 24 fields. Total file size is 2.7 MB.

(2) 134,136 records (including header) and 23 fields. Total file size is 1.4 MB.
(3) 139,104 records (including header) and 23 fields. Total file size is 1.5 MB.
(4) Total file size is 132 MB.
(5) Total file size is 96 MB.
(6) 5,831 individual geoTIFF files. Total folder size is 61.4 MB.
(7) 5,831 individual geoTIFF files. Total folder size is 61.2 MB.
(8) 69,984 records (including header) and 12 fields. Total file size is 978 kB.

Format and storage mode:

(1) UTF-8 text, comma delimited, not compressed. Please note that opening in programs assuming ASCII encoding may render parts of the text unreadable and e.g. in R all users should specify encoding by using fileEncoding="UTF-8" in their read.csv command.
(2) UTF-8 text, comma delimited, not compressed. Please note that opening in programs assuming ASCII encoding may render parts of the text unreadable and e.g. in R all users should specify encoding by using fileEncoding="UTF-8" in their read.csv command.
(3) UTF-8 text, comma delimited, not compressed. Please note that opening in programs assuming ASCII encoding may render parts of the text unreadable and e.g. in R all users should specify encoding by using fileEncoding="UTF-8" in their read.csv command.
(4) Nexus file.
(5) Nexus file.
(6) Individual geoTIFF files for each species in a Behrman cylindrical equal area projection with a resolution of 96.5 km by 96.5 km at 30° North and 30° South. (full projection description following R-notation is "+proj=cea +lon_0=0 +lat_ts=30 +x_0=0 +y_0=0 +datum=WGS84 +units=m +no_defs +ellps=WGS84 +towgs84=0,0,0"). Only the area between 90°N and 60°S (which includes the full range of all terrestrial mammals) is plotted.
(7) Individual geoTIFF files for each species in a Behrman cylindrical equal area projection with a resolution of 96.5 km by 96.5 km at 30° North and 30° South. (full projection description following R-notation is "+proj=cea +lon_0=0 +lat_ts=30 +x_0=0 +y_0=0 +datum=WGS84 +units=m +no_defs +ellps=WGS84 +towgs84=0,0,0"). Only the area between 90°N and 60°S (which includes the full range of all terrestrial mammals) is plotted.
(8) UTF-8 text, comma delimited, not compressed. Please note that opening in programs assuming ASCII encoding may render parts of the text unreadable and e.g. in R all users should specify encoding by using fileEncoding="UTF-8" in their read.csv command.

Header information:
The first rows of 1, 2, 3, and 8 contain variable names (see below).

Row information:
Each row represents data for a single species in 1, 2, 3, and 8.

Alphanumeric attributes:
Mixed.

Special character fields:
"NA" is used to denote missing data. Fields starting with "000" represent comments. The variable "Diet.Method" of Trait_data.csv also uses "000" to represent unknown values in a multipart variable; see details below.

Authentication procedures:
Checksums for the data:
(1) MD5: 3ac7f24102fd64515354149168ae8fb1
(2) MD5: 9b70958959d284759ee748918622ddae
(3) MD5: 7e0eb30a09f7b083ef4446ab8a258d0e
(4) MD5: e4bfdb1832c97ba0415dcbc487eab705
(5) MD5: efd13f8eac4b979fd93221fa27b8220e
(6) None.
(7) None.
(8) MD5: bc1d39b13cb836c26bcc03bad58fed27

B. Variable information

(1) Trait_data.csv

Binomial.1.2

Definition: Binomial name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 5831 binomial names

Order.1.2

Definition: Order name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 29 order names

Family.1.2

Definition: Family name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 169 family names

Genus.1.2

Definition: Genus name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 1400 genus names

Species.1.2

Definition: Specific name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 4114 specific epithets

Terrestrial

Definition: Whether the species spends a significant amount of time on land.
Data Type: Binary
Values: **0** (no), **1** (yes)

Marine

Definition: Whether the species spends a significant amount time in oceans and/or seas.
Data Type: Binary
Values: **0** (no), **1** (yes)

Freshwater

Definition: Whether the species spends a significant amount of time in fresh water.
Data Type: Binary
Values: **0** (no), **1** (yes)

Aerial

Definition: Whether the species is capable of powered flight and thus spends a significant amount of time flying in the air.
Data Type: Binary
Values: **0** (no), **1** (yes)

Life.Habit.Method

Definition: The method used to estimate the life habits (terrestrial, marine, freshwater, aerial).
Data Type: Factor
Values: Three methods

- **Reported** means the life habit was taken from a scientific source.
- **Imputed** means the life habit was estimated using phylogenetic imputation.
- **Taxonomic** means the life habit was estimated using taxonomy. Note that this applies only to bats, which were all listed as aerial.

Life.Habit.Source

Definition: The source from which we collected the life habit value
Data Type: Character
Values: Apart from those values that were inferred via taxonomy or phylogenetic imputation, all life habit values were taken from IUCN 2016-3 (2016).

Mass.g

Definition: Body mass in grams
Data Type: Ratio scale
Values: Body mass estimates ranging from 1.6 g to $1.9*10^8$ g

Mass.Method

Definition: The method used to estimate Mass.g.
Data Type: Character
Values: Five possible methods:

- **Reported** means values were taken from a scientific source.
- **Assumed isometric based on …** means the value was estimated using an isometric relationship based on selected linear measurement of the species and a close relative from a scientific source.
- **As relative of suggested similar size** means another species was used as a proxy to estimate the mass.
- **Estimated based on equation from …** means the value was calculated using a general equation rather than isometric relationship based on close relatives.
- **Imputed** means the mass was estimated using phylogenetic imputation.

Mass.Source

Definition: The literature source from which we collected the body mass value.
Data Type: Character
Values: 699 sources

Mass.Comparison

Definition: The species used as a proxy or comparison for the mass of the target species.
Data Type: Character
Values: 490 species

Mass.Comparison.Source

Definition: The source of the scaling equation used to predict mass or the source from which we collected the linear measurement we used to estimate the mass of the target species (only listed if different from the publication discussing the target species).
Data Type: Character
Values: 272 sources

Island.Endemicity

Definition: The degree to which a species can be considered an island endemic. The categories are nested such that a species labelled as one category might also be found in the environments in the categories listed beneath it. A species labelled as **Occurs on mainland** could also occur on large land bridge islands, small land bridge islands, and isolated islands; a species labelled as **Occurs on large land bridge islands** may also be found on small land bridge islands and isolated islands, and a species labelled as **Occurs on small land bridge islands** may also be found on isolated islands. Strictly marine species are coded as **Exclusively marine** as they never occur on islands or the mainland.

Data Type: Ordinal

Values: Four levels:

- **Exclusively marine**
- **Occurs on mainland**
- **Occurs on large land bridge islands** means that a species occurs on islands greater than 1,000 km^2 that are separated from the mainland by water no more than 110 m deep. Thus, the islands would have been part of the mainland during the last glacial maximum.
- **Occurs on small land bridge islands** means that a species occurs on islands smaller than 1,000 km^2 that are separated from the mainland by water no more than 110 m deep. Thus, the islands would have been part of the mainland during the last glacial maximum.
- **Occurs only on isolated islands** means that the species occurs on islands separated from the mainland by water deeper than 110 m.

IUCN.Status.1.2

Definition: The IUCN Red List status used in the current PHYLACINE Version 1.2. We added a new status (**EP**) to denote species that went extinct in prehistory (prior to 1500 AD).

Data Type: Ordinal

Values: Nine levels:

- **EP** (extinct in prehistory, before 1500 CE)
- **EX** (extinct, after 1500 CE)
- **EW** (extinct in the wild)
- **CR** (critically endangered)
- **EN** (endangered)
- **VU** (vulnerable)
- **NT** (near threatened)
- **LC** (least concern)
- **DD** (data deficient)

Added.IUCN.Status.1.2

Definition: Whether or not the IUCN status was added during the construction of PHYLACINE Version 1.2. Note that added statuses are not official IUCN statuses.

Data Type: Binary

Values: **Yes**, **No**

Diet.Plant

Definition: The percentage of plants and/or fungi in the diet.

Data Type: Ratio scale

Values: Percentage values ranging from 0% to 100%

Diet.Vertebrate

Definition: The percentage of vertebrate prey in the diet.

Data Type: Ratio scale

Values: Percentage values ranging from 0% to 100%

Diet.Invertebrate

Definition: The percentage of invertebrate prey in the diet.
Data Type: Ratio scale
Values: Percentage values ranging from 0% to 100%

Diet.Method

Definition: The method used to estimate the diet.
Data Type: Character
Values: Each entry consists of three words:
The first word denotes how the diet values were generated:

- **Reported** means the diet values have been copied directly from another source or could be inferred precisely from another source (e.g., Eltontraits 1.0).
- **Transformed** means the diet values are based on numerical values from another source but have been mathematically transformed to conform to the PHYLACINE format (e.g., MammalDIET)
- **Estimated** means the diet values have been manually estimated by the authors of this paper, based on the listed evidence.
- **Imputed** means the diet values have been estimated using phylogenetic imputation.

The second word denotes the type of evidence the diet was based upon.

- **Observed** means the estimate is based on a direct measure: visual observation, gut content, scat content, etc.
- **Isotopes** means the estimate is based on stable isotope analysis.
- **Craniodental** means the diet is based on the morphology and/or wear of the teeth, skull, or skull musculature.
- **Expert** means the estimate is based on classifications by experts that lack sufficient empirical justification. This can include studies that either report no data or data that does not explicitly distinguish between a carnivorous, herbivorous, or insectivorous diet. For example, sources that use only C13 isotopes to determine whether a species is a browser or a grazer are considered **Expert** because they don't explicitly measure a species' consumption of meat. Herbivory is implicitly assumed.

The third word denotes whether the diet was reported at the family, genus, or species level.

- **Family** means the source reported diet at a family level. This could reflect uncertainty or similarity between species within the same family.
- **Genus** means the source reported diet at a genus level. This could reflect uncertainty, the prevalent use of genera as a taxonomic unit in fossil studies (e.g., *Equus sp.*), or genuine dietary similarity between congeners.
- **Species** means the source reported diet at a species level.

If no data was available regarding the method or the taxonomic resolution, these were denoted with **000**. For example, **Reported 000 Species** represents data that was reported at the species level but the source did not specify exactly how their numerical dietary estimates were generated.

Diet.Source

Definition: The literature source from which we collected or estimated the percentage values.
Data Type: Character
Values: 95 different literature sources

(2) Synonymy_table_valid_species_only.csv

This table holds synonymy data for only those species used in PHYLACINE Version 1.2. In this table, if a species was not accepted by a certain dataset for whatever reason, its name was replaced with **000 Species not accepted**. The "000" tags these comments, making them easy to programmatically separate from real species names. For example, the dire wolf is a fossil species so the name *Canis dirus* used in PHYLACINE 1.2 was listed as **000 Species not accepted** in the IUCN.2016.3.Genus and IUCN.2016.3.Species columns. If a species in one dataset had a synonym in another dataset, both names were listed. For example, *Aonyx cinereus* in Genus.1.2 and Species.1.2 is listed as *Aonyx cinerea* in EltonTraits1.0.Genus and EltonTraits.1.0.Species.

Binomial.1.2

Definition: Binomial name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 5831 binomial names

Order.1.2

Definition: Order name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 29 order names

Family.1.2

Definition: Family name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 169 family names

Genus.1.2

Definition: Genus name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 1400 genus names

Species.1.2

Definition: Specific name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 4114 specific epithets

Genus.1.1

Definition: Genus name used in PHYLACINE Version 1.1.
Data Type: Character

Values: 1384 genus names

Species.1.1

Definition: Specific name used in PHYLACINE Version 1.1.
Data Type: Character
Values: 4033 specific epithets

Genus.1.0

Definition: Genus name used in PHYLACINE Version 1.0.
Data Type: Character
Values: 1369 genus names

Species.1.0

Definition: Specific name used in PHYLACINE Version 1.0.
Data Type: Character
Values: 4009 specific epithets

EltonTraits.1.0.Genus

Definition: Genus name used in EltonTraits 1.0.
Data Type: Character
Values: 1220 genus names

EltonTraits.1.0.Species

Definition: Specific name used in EltonTraits 1.0.
Data Type: Character
Values: 3672 specific epithets

IUCN.2016.3.Genus

Definition: Genus name used in IUCN 2016-3.
Data Type: Character
Values: 1264 genus names

IUCN.2016.3.Species

Definition: Whether the PHYLACINE Version 1.2 taxonomy differs from the IUCN 2016.3 taxonomy.
Data Type: Character
Values: 3919 specific epithets

Intrafamily.Ages.Generated.From.Genetic.Data

Definition: Whether intrafamily ages were determined from raw sequence data (**Yes**) or simulated based on Birth-Death models (**No**).
Data Type: Binary
Values: **Yes**, **No**

Analyzed.Intrafamily.Phylogeny

Definition: The version of PHYLACINE when intrafamily topology was most recently analyzed.
Data Type: Factor

Values: Four levels:

- **Only one species in family and therefore no intrafamily analysis**
- **Version 1.0**
- **Version 1.1**
- **Version 1.2**

Included.In.Small.Tree

Definition: Whether the species was included in the small phylogenetic tree where placement was determined by genetic data or unambiguous taxonomic rules.

Data Type: Binary

Values: **Yes, No**

Hierarchical.Level

Definition: The highest hierarchical level at which the species was included when building the tree.

Data Type: Ordinal

Values: Four levels (a few species have specific modifications that are described with their level):

- **1** means that the species was included in the analyses of the interfamily relationship of therian or marsupial mammals.
- **2** means that the species was included in the first analysis of the relationship within each family.
- **3** means that the species was only included in the genus level phylogeny.
- **4** means that the species was placed without genetic data.

Markers

Definition: The number of independent markers used to assign the placement of the species. For the few analyses where several mitochondrial markers are used to improve taxonomic coverage they are still only counted as one. For species included at Hierarchical.Level **1**, the number given is the number used for family level or lower analyses. Therefore, species in families with only one or two species and where we considered the topology fully settled were given values of 0 because the interfamily analyses was enough to place them precisely on the tree.

Data Type: Integer

Values: 0 to 57

Bases

Definition: The number of base pairs used to assign the placement of the species. For species included at Hierarchical.Level **1**, the number given is the number used for family level or lower analyses. Therefore, species in families with only one or two species and where we considered the topology fully settled were given values of 0 because the interfamily analyses was enough to place them precisely on the tree.
Data Type: Integer
Values: 0 to 55,504 base pairs

Placement

Definition: Denotes the placement of all species assigned without genetic data. Parentheses are used to help clarify monophyletic groupings in cases where several placements are possible.
Data Type: Character
Values: Individual descriptions of placement

Placement.Source

Definition: The reference or logic behind the placement of all species mentioned in the Placement column.
Data Type: Character
Values: 114 sources

Taxonomy.Source

Definition: The taxonomic reference or argument for adding or deleting the species from the list of species accepted by IUCN 2016-3.
Data Type: Character
Values: 82 sources

Dating.Source

Definition: The reference and/or argument for the occurrence of a species in the late Pleistocene and/or Holocene.
Data Type: Character
Values: 77 sources

(3) Synonymy_table_with_unaccepted_species.csv

This table uses the same variables and format as Synonymy_table_valid_species_only.csv but it includes all species used in EltonTraits 1.0 and IUCN 2016-3, not just those species accepted in PHYLACINE Version 1.2. For descriptions of variables, refer to Synonymy_table_valid_species_only.csv above.

(8) Spatial_metadata.csv

Binomial.1.2

Definition: Binomial name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 5831 binomial names

Order.1.2

Definition: Order name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 29 order names

Family.1.2

Definition: Family name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 169 family names

Genus.1.2

Definition: Genus name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 1400 genus names

Species.1.2

Definition: Specific name used in PHYLACINE Version 1.2.
Data Type: Character
Values: 4114 specific epithets

Certainty.Level

Definition: The general method used to construct a species' present natural range and the relative level of certainty of this method. Lower numbers imply more certainty and fewer assumptions by the authors. Species' present natural ranges could be estimated based on several methods so they are listed at the most uncertain method used.
Data Type: Ordinal

Values: Nine Levels:

- **0) Present Natural range identical to IUCN range**
- **1) Range reductions for species with recent, human-induced range expansions**
- **2) Range expansions based on other sources for species with known range declines**
- **3) Merger of likely human caused disjunct ranges by filling intervening suitable habitats**
- **4) Expansion of ranges to entire islands**
- **5) Expansion of ranges for species with known, or at least highly suspected, range declines to cover suitable areas contiguous with the current range**
- **6) Estimation of ranges based on the natural ranges of the extant species that the target species co-occurred with at fossil sites**
- **7) Estimation of ranges based on the natural ranges of the extinct species that the target species co-occurred with at fossil sites**
- **8) Unique species-specific modifications**

Modification

Definition: The ad hoc modifications made to the initial range map to create the final present natural range for a species. For most species, this initial range map was the IUCN extant range. For many extinct species which have no IUCN extant range, their initial range was modeled based on co-occurrence patterns (method 6 and 7 above) and we instead mention any additional species-specific modifications.

Data Type: Character

Values: Individual descriptions of changes

Motivation

Definition: The motivation and/or references for the changes described in the Modification variable.

Data Type: Character

Values: Individual descriptions of motivation and references

Last.Manual.Present.Natural.Modification

Definition: The last time the present natural range was manually edited.

Data Type: Factor

Values: Four levels (including NA's):

- **NA** marks species where the present natural range was the same as the current range and thus, never manually edited.
- **Version 1.0**
- **Version 1.1**
- **Version 1.2**

Number.Cells.Current.Range

Definition: The number of raster cells in the current range.

Data Type: Integer

Values: 0 to 37,542 cells

Number.Cells.Present.Natural.Range

Definition: The number of raster cells in the present natural range.

Data Type: Integer

Values: 0 to 37,542 cells

Change.In.Cells

Definition: The difference between the number of cells of the present natural and the number of cells of the current range.

Data Type: Integer

Values: -14,047 to 9239 cells gained from the current to the present natural range of species.

C. Data anomalies

Data reliability:

We stress that the data is intended only for large-scale analyses across broad sets of mammal taxa, and if used for narrow species sets or where specific species values are very important, then users should go to the original data sources to check the scoring, and likely even beyond to find the best data for these particular species. We have worked hard to make the best possible database covering all mammals but this means that a number of uncertainties and idiosyncrasies that are important for individual clades could not be incorporated.

Missing data:

The philosophy behind the PHYLACINE database is that it should contain no missing data in any of its constituent data sets so that users can immediately perform analyses on the data without making ad-hoc data filling decisions. We minimized missing data by exhaustively searching the literature for data not covered in our source data sets but some blank cell values were inevitable. For trait data, missing values were estimated using the Phylopars phylogenetic imputation method with the R package Rphylopars v. 0.2.9 (Goolsby et al. 2017). Phylopars was run with all default parameters including a Brownian motion evolutionary model. Values were imputed for the life habits of 376 species, the diets of 493 species, and the masses of 202 species. Imputed values are clearly marked and should be removed before some phylogenetic analyses to avoid circularity. Before imputation, binomial life habits were treated as continuous, masses were log10 transformed and percentage diets were transformed into two new variables (described below). Imputation was run on all

1000 trees in the posterior distribution and the median values of the imputed trait means for each species from each tree were used as the final predicted trait values. Compositional data, like diet percentages that must add to 100%, can be difficult to impute because Phylopars treats each diet category as an independent variable. In order to minimize any potential artifacts, before imputation, we first transformed the three diet percentage categories into two variables. These two new variables were the x and y coordinates locating each species on a ternary diagram of diet with 100% plant, 100% invertebrate, and 100% vertebrate consumption as its three vertices. After imputation, imputed values were moved to the closest point within this triangle that represented an integer percent diet before being back transformed to three percentage diet categories. For life habits, all estimates <0.5 were treated as 0 and >=0.5 treated as 1. Phylopars is technically not meant for non-continuous data like our binary life habit traits but it is important to include these life habits in our imputations as the Phylopars method uses covariation among traits to refine predictions; whether a species is marine or aerial will have a huge influence on its imputed body mass value. Treating these variables as continuous traits still generates reasonable values. Only two species, *Bubalus grovesi* and *Sivacobus sankaliai*, actually had a life habit variable that was imputed to be more than 0.05 away from a binary value of 0 or 1. The freshwater habit is ambiguous for both of these species and their interpolated values (0.21 and 0.38, respectively) were rounded down to a binary value of 0 making them fully terrestrial.

For the phylogeny, two sets of trees are provided: one including all 5831 species in the database where species without genetic information were allowed to move freely based on taxonomy, and one where only the 4253 species that can be reliably placed using genetic data or hardcoded taxonomical constraints were included. Even though not all of these 4253 species had genetic data, some species could still be placed reliably without it. For example, in a genus with only two species, if one species has genetic data, the other species lacking genetic data can be safely positioned as sister to the first species without any ambiguous topology (assuming the genus is monophyletic). Only 15 species lack both current and present natural ranges (described in Class II, Section B above). These species are recognized by the IUCN, but the IUCN has not created range maps for them yet. Extinct species, of course, lack current ranges so they are given empty rasters in this data set.

Comparing trait variability between different groups in the PHYLACINE dataset should always be performed with extreme caution. This is because source data may often be at a coarser scale than imputed data. For example, EltonTraits (Wilman et al. 2014) gives percentage diet rounded to the nearest 10% (e.g., 20%, 30%, 40%, etc.). However, imputed values were not rounded to this level and may appear to be more

finely resolved (e.g., 37%, 56%, etc.). It is possible that a researcher using PHYLACINE could find that one clade is more variable in diet than another but this might just be the result of the first clade having fewer raw, rounded diets from EltonTraits than the second clade. Because imputed diets are not rounded to the nearest 10%, they are also much less likely than reported diets to be 100% in any one category, e.g., an imputed diet may be 99% plant and 1% invertebrate. Users should take this into consideration when selecting species from the database that have a strict diet. For example, if a user wanted to select strict herbivores from the database, she should consider taking all species that have more than 90% plant in their diet rather than only those species whose diet is listed as 100% plant.

Taxonomic oddities:
Family assignment is in flux within Cingulata (Mitchell et al. 2016) and several extinct clades usually classified as separate families or *Incertae sedis* are likely nested within a family of extant species. We therefore classified all species as belonging to a single, family level cluster called "CingulataFam". Also, some species have not received formal binomial names yet and are referred to in the scientific literature only by specimen codes e.g., *Geocapromys sp. nov.* A (Turvey 2009). We consistently named these species as *Genus_spCode* in the database. For example, *Geocapromys_spA* or *Homo_spDenisova*. A few undescribed species lack generic assignments and were instead named based on the level of known taxonomic resolution. For example, one unnamed caprine species was referred to as *CapriniGen_spA* signifying that it is a species that belongs to some (possibly unnamed) genus within the tribe Caprini.

CLASS V. SUPPLEMENTAL DESCRIPTORS

A. **Data acquisition**

Acquisition methods:
See Class II, Section B above.

Data entry verification procedures:
See supplemental material for the main databases that we acquired data from: EltonTraits (Wilman et al. 2014), Mass of Mammals (Smith et al. 2003), MammalDIET (Kissling et al. 2014), and IUCN (IUCN 2016).

B. **Quality assurance/quality control procedures**

Procedures:

Data was continuously and rigorously checked for logical errors throughout the assembly process. Almost all database assembly was performed programmatically through R and the version control system Git. This cuts down on potential errors arising from manually editing and manipulating spreadsheets of data. Any errors we noticed could be programmatically tracked back to their source and corrected. We, however, cannot guarantee that the database is without errors and/or inconsistencies and encourage all users to report any errors at MegaPast2Future.github.io so they can be corrected as soon as possible.

C. **Related materials**

Description:

More detailed descriptions of data and methods can be found in the papers were the constituent parts of this database were first published (Sandom et al. 2014, Faurby and Svenning 2015b, 2015a, 2016).

D. **Computer programs and data-processing algorithms**

Description:

The PHYLACINE database is hosted on GitHub at MegaPast2Future.github.io and was assembled with the statistical language R version 3.4 (R Development Core Team 2015) using Git version 2.14 for version control.

E. **Archiving**

Archival Procedures:

The database is available from the Dryad Digital Data Repository (https://doi.org/10.5061/dryad.bp26v20). The latest development versions of the database can always be found at MegaPast2Future.github.io (http://doi.org/10.5281/zenodo.1250504).

F. **Publications and results**

List of publications using parts and previous versions of the database:

Sandom, C., S. Faurby, B. Sandel, and J. C. Svenning. 2014. Global late Quaternary megafauna extinctions linked to humans, not climate change. Proceedings of the Royal Society of London Series B 281:20133254.

Faurby, S., and J. C. Svenning. 2015. A species-level phylogeny of all extant and late Quaternary extinct mammals using a novel heuristic-hierarchical Bayesian approach. Molecular Phylogenetics and Evolution 84:14–26.

Doughty, C. E., S. Faurby, and J. C. Svenning. 2015. The impact of the megafauna extinctions on savanna woody cover in South America. Ecography 39:213–222.

Faurby, S., and J. C. Svenning. 2015. Historic and prehistoric human-driven extinctions have reshaped global mammal diversity patterns. Diversity and Distributions 21:1155–1166.

Svenning, J. C., P. B. M. Pedersen, C. J. Donlan, R. Ejrnaes, S. Faurby, M. Galetti, D. M. Hansen, B. Sandel, C. J. Sandom, J. W. Terborgh, and F. W. M. Vera. 2016. Science for a wilder Anthropocene: synthesis and future directions for trophic rewilding research. Proceedings of the National Academy of Sciences 113:898–906.

Faurby, S., and J. C. Svenning. 2016. The asymmetry in the Great American Biotic Interchange in mammals is consistent with differential susceptibility to mammalian predation. Global Ecology and Biogeography 25:1443–1453.

Doughty, C. E., J. Roman, S. Faurby, A. Wolf, A. Haque, E. S. Bakker, Y. Malhi, J. B. Dunning Jr, and J. C. Svenning. 2016. Global nutrient transport in a world of giants. Proceedings of the National Academy of Sciences 113:868–873.

Faurby, S., and J. C. Svenning. 2016. Resurrection of the Island Rule: human-driven extinctions have obscured a basic evolutionary pattern. American Naturalist 187:812–820.

Doughty, C. E., S. Faurby, A. Wolf, Y. Malhi, and J. C. Svenning. 2016. Changing NPP consumption patterns in the Holocene: from megafauna-‘liberated’ NPP to ‘ecological bankruptcy’. The Anthropocene Review:1–14.

Faurby, S., and M. B. Araújo. 2017. Anthropogenic impacts weaken Bergmann's rule. Ecography 40:683–684.

Sandom, C. J., S. Faurby, J. C. Svenning, D. Burnham, A. Dickman, A. E. Hinks, E. A. Macdonald, W. J. Ripple, J. Williams, and D. W. Macdonald. 2017. Learning from the past to prepare for the future: felids face continued threat from declining prey. Ecography 41:140-152.

Svenning J. C., and S. Faurby. (2017). Prehistoric and historic baselines for trophic rewilding in the Neotropics. Perspectives in Ecology and Conservation 15: 282-291.

G. History of data set usage

Data request history:
None.

Data set update history:
The PHYLACINE database was originally published as several different data sets. This is the first time that all of the constituent parts have been combined together into one complete database under the name PHYLACINE for publication. For Version 1.2, all the original data sets have been revised and crosschecked against each other to ensure consistency and ease of use as described above. Previous versions of the database are described below.

Version 0.0:
The original taxonomy of species that have gone extinct since the last interglacial was compiled for (Sandom et al. 2014).

Version 1.0:
Based on IUCN 2012-1 taxonomy and the taxonomy of extinct species developed for Sandom et al. (2014), we generated several interconnected data sets describing the functional diversity, phylogenetic diversity, and spatial range of all mammals from the latest Quaternary. We built the first phylogeny of all these species for Faurby and Svenning (2015b). The life habits used in Trait_data.csv and the present natural ranges were originally compiled for Faurby and Svenning (2015a). The body sizes and island endemicity statuses used in Trait_data.csv were originally assembled for Faurby and Svenning (2016).

Version 1.1:
For Version 1.1, major modifications of the phylogeny were performed. The relationships of a number of families totaling 1,903 species were reanalyzed including information published since the original phylogeny was published. For a number of families totaling 1,443 species, we let the relative ages be based directly on sequence data rather than simulated data as before in Version 1.0. We initially used the simulation approach described in (Faurby and Svenning 2015b) because it enabled us

to be more flexible with topological modifications of the tree. The simulation procedure enabled us to place un-sampled species without any constraints on the internal topology of species with data. For the 1,443 species mentioned above, complex modifications however, were not needed either because all species had data or the desired topological placement of the missing species could be proposed without influencing the placement of species missing genetic data. For example, in Faurby and Svenning (2015b), we describe how the method we employed can place a species without genetic data as sister to any species within a group of species without enforcing their monophyly. But if the group in question is monophyletic with a posterior support of 1.0, this is also trivial to do with simple constraints. We also discussed a method where we can place a species without genetic data as sister to a group of species when this group is monophyletic or as sister to a part of it when it is not. When such a group is monophyletic with a posterior support of 1.0, such placement is likewise trivial to handle with topological constraints instead. The 1,443 species all belonged to groups where all clades relevant to the placement of the missing species had a posterior support of 1.0.

Both the current and present natural ranges were reanalyzed to match new knowledge and IUCN taxonomy 2015-4.

Version 1.2:
Version 1.2 is the current version of the database published here.

Review history:
None.

Questions and comments from secondary users:
None.

Acknowledgments

We thank all the researchers who performed the original field work and analyses necessary for the data that we compiled and are making available here. We especially thank the authors of the EltonTraits (Wilman et al. 2014), Mass of Mammals (Smith et al. 2003), MammalDIET (Kissling et al. 2014), and IUCN (IUCN 2016) databases. We are also grateful to the Carlsberg Foundation, European Research Council, and VILLUM FONDEN.

Literature Cited

Davis, M. 2017. What North America's skeleton crew of megafauna tells us about community disassembly. Proceedings of the Royal Socieity B 284:20162116.

Davis, M., and S. Pineda-Munoz. 2016. The temporal scale of diet and dietary proxies. Ecology and Evolution 6:1883–1897.

Faurby, S., and J. C. Svenning. 2015a. Historic and prehistoric human-driven extinctions have reshaped global mammal diversity patterns. Diversity and Distributions 21:1155–1166.

Faurby, S., and J.-C. Svenning. 2015b. A species-level phylogeny of all extant and late Quaternary extinct mammals using a novel heuristic-hierarchical Bayesian approach. Molecular Phylogenetics and Evolution 84:14–26.

Faurby, S., and J.-C. Svenning. 2016. Resurrection of the Island Rule: human-driven extinctions have obscured a basic evolutionary pattern. American Naturalist 187:812–820.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. The American Naturalist 160:712–726.

Gainsbury, A. M., O. J. S. Tallowin, and S. Meiri. 2018. An updated global data set for diet preferences in terrestrial mammals: testing the validity of extrapolation. Mammal Review doi: 10.1111/mam.12119

Goolsby, E. W., J. Bruggeman, and C. Ané. 2017. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. Methods in Ecology and Evolution 8:22–27.

IUCN. 2016. IUCN red list of threatened species. Version 2016-3. http://www.iucnredlist.org.

Johnson, C. N., and S. Wroe. 2016. Causes of extinction of vertebrates during the Holocene of mainland Australia: arrival of the dingo, or human impact? The Holocene 13:941–948.

Kissling, W. D., L. Dalby, C. Fløjgaard, J. Lenoir, B. Sandel, C. Sandom, K. Trøjelsgaard, and J.-C. Svenning. 2014. Establishing macroecological trait datasets: digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. Ecology and Evolution 4:2913–2930.

Meiri, S., and T. Dayan. 2003. On the validity of Bergmann's rule. Journal of Biogeography 30:331–351.

Mitchell, K. J., A. Scanferla, E. Soibelzon, R. Bonini, J. Ochoa, and A. Cooper. 2016. Ancient DNA from the extinct South American giant glyptodont *Doedicurus sp.* (Xenarthra: Glyptodontidae) reveals that glyptodonts evolved from Eocene armadillos. Molecular Ecology 25:3499–3508.

Penone, C., A. D. Davidson, K. T. Shoemaker, M. Di Marco, C. Rondinini, T. M. Brooks, B. E. Young, C. H. Graham, and G. C. Costa. 2014. Imputation of missing data in life-history trait datasets: which approach performs the best? Methods in Ecology and Evolution 5:961–970.

Peterken, G. F. 1977. Habitat conservation priorities in British and European woodlands. Biological Conservation 11:223–236.

R Development Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Safi, K., M. V. Cianciaruso, R. D. Loyola, D. Brito, K. Armour-Marshall, and J. A. F. Diniz-Filho. 2011. Understanding global patterns of mammalian functional and phylogenetic diversity. Philosophical Transactions of the Royal Society B: Biological Sciences 366:2536–2544.

Sandom, C., S. Faurby, B. Sandel, and J.-C. Svenning. 2014. Global late Quaternary megafauna extinctions linked to humans, not climate change. Proceedings of the Royal Society of London Series B 281:20133254.

Shapiro, B., A. J. Drummond, A. Rambaut, M. C. Wilson, P. E. Matheus, A. V. Sher, O. G. Pybus, M. T. P. Gilbert, I. Barnes, J. Binladen, E. Willerslev, A. J. Hansen, G. F. Baryshnikov, J. A. Burns, S. Davydov, J. C. Driver, D. G. Froese, C. R. Harington, G. Keddie, P. Kosintsev, M. L. Kunz, L. D. Martin, R. O. Stephenson, J. Storer, R. Tedford, S. Zimov, and A. Cooper. 2004. Rise and fall of the Beringian steppe bison. Science 306:1561–1565.

Smith, F. A., S. K. Lyons, S. Ernest, K. Jones, D. Kaufman, T. Dayan, P. Marquet, J. Brown, and J. Haskell. 2003. Body mass of late quaternary mammals. Ecology 84:3403.

Turvey, S. T. 2009. Holocene mammal extinctions. Pages 41–62 *in* S. T. Turvey, editor. Holocene Extinctions. Oxford University Press, Oxford.

Voirin, B. 2015. Biology and conservation of the pygmy sloth, *Bradypus pygmaeus*. Journal of Mammalogy 96:703–707.

Wilman, H., J. Belmaker, J. Simpson, C. de la Rosa, M. M. Rivadeneira, and W. Jetz. 2014. EltonTraits 1.0: species-level foraging attributes of the world's birds and mammals. Ecology 95:2027.