Version dated: May 18, 2015

RH: Morphological data availability in living mammals

Morphological data availability in living mammals

Thomas Guillerme[1,2*], and Natalie Cooper[1,2]

[1]*School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland.*

[2]*Trinity Centre for Biodiversity Research, Trinity College Dublin, Dublin 2, Ireland.*

***Corresponding author.** *Zoology Building, Trinity College Dublin, Dublin 2, Ireland; E-mail: guillert@tcd.ie; Fax: +353 1 6778094; Tel: +353 1 896 2571.*

Abstract

Studying changes in global biodiversity through time and space is essential. For that we need methods to combine both palaeontological and neontological data. One promising method, the Total Evidence method, allows such thing but needs a lot of data. Especially morphological data from living taxa to allow the fossil taxa to accurately branch in the trees. Despite two centuries of morphological studies on living taxa, scientists using and generating such data mainly focus on palaeontological data. Therefore, even in well known groups such as mammal, there is a huge gap in our knowledge of morphological data for living mammals. In this study, using phylogenetic community structure methods, we quantify the availability of data in each mammalian order. And maybe at the end of the paper we propose some discussion on how to improve all that (go in museums!).

()

Introduction

Studying both living and fossil taxa together in essential to fully understand macrovelutionary patterns and processes and is becoming increasingly common among evolutionary biologists (Jackson and Erwin, 2006; Quental and Marshall, 2010; Dietl and Flessa, 2011; Slater and Harmon, 2013; Fritz et al., 2013; Wood et al., 2013). Combining the global biodiversity by including both living and fossil in such studies allows, for example, to improve accuracy of the timing of diversification events (e.g. Ronquist et al., 2012), to better understand relationships among lineages (e.g. Beck and Lee, 2014) or to infer biogeographical patterns through time (e.g. Meseguer et al., 2015). In order to perform such analysis, one must efficiently combine living and fossil taxa data in macroevolutionary models. One trending method, called the Total Evidence method (Eernisse and Kluge, 1993; Ronquist et al., 2012), allows to combine molecular data from living taxa and morphological data from both living and fossil taxa in a supermatrix (e.g. Pyron, 2011; Ronquist et al., 2012; Schrago et al., 2013; Slater and Harmon, 2013; Beck and Lee, 2014; Meseguer et al., 2015). This method not only allows to use all the available data but also allows to treat fossil taxa as tips rather than nodes *via* integrative phylogenetic inference methods such as tip-dating (Ronquist et al., 2012; Drummond et al., 2012; Wood et al., 2013; Matzke, 2014).

However, the Total Evidence method requires, by definition, a lot of data. One must collect both molecular data for living taxa as well as morphological data for both living and fossil taxa, two types of data that require fairly different technical skills (e.g.

Meredith et al., 2011) *vs.* (O'Leary et al., 2013). Additionally, entire sections of this data can sometimes be difficult or impossible to collect for every taxa present in the analysis. For example, fossil have really rarely molecular data available and morphological characters are rarely collected for living taxa when molecular data is available (e.g. Slater, 2013; Beck and Lee, 2014). This difference between both data can lead to topological errors in phylogenetic inference (Guillerme and Cooper, 2015). In fact, the ability of the Total Evidence method to recover correct topology is expected to decrease when there is a low overlap between morphological data from the living and the fossil taxa (Guillerme and Cooper, 2015). The effect of missing data is most important on topology when too few living taxa have available morphological data (Guillerme and Cooper, 2015). For example, if there is no morphological available for any living taxa within an entire clade, it is impossible to link a fossil taxon to this clade because no morphological data will overlap between the fossil (regardless the amount of data available for the fossil) and the living taxa (that have no morphological data). This property of the Total Evidence method can rapidly lead to important topological incongruities because the fossil will only be able to branch to a clade of living taxa that contains morphological data, even if in reality, the fossil taxon does not belong to that clade (Guillerme and Cooper, 2015) (see figure 1.

It is therefore crucial to understand the distribution of the available morphological data for living taxa in a clade before using a Total Evidence approach because missing living taxa with morphological data can lead to two topological

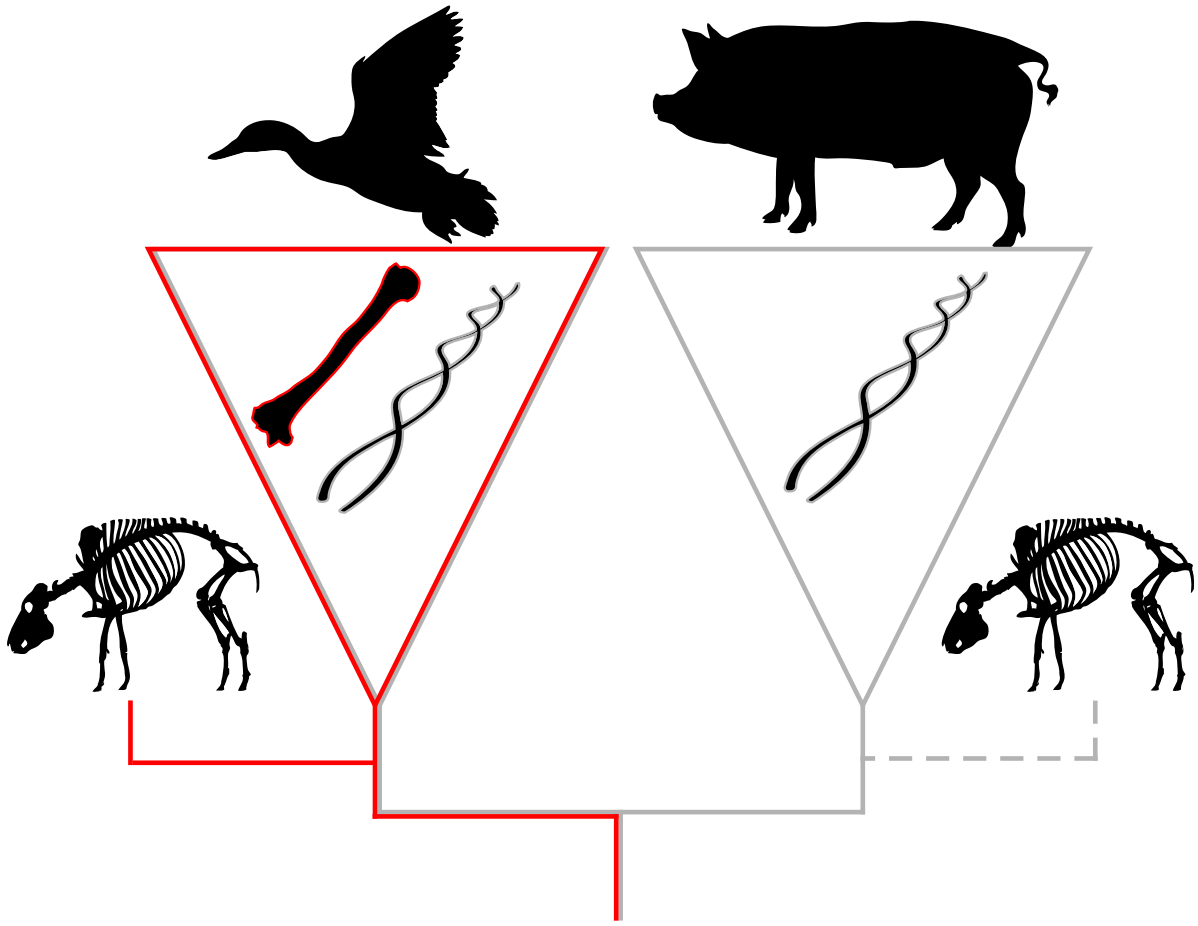

Figure 1: Example of topological errors due to missing morphological data in living taxa. If a phylogeny contains two clades, for example Aves and Mammalia, with molecular data (in grey) for both but only morphological data (in red) for Aves. If an additional Mammalia fossil (with no molecular data) is added to the phylogeny, it will erroneously branch with the Aves clade instead of the Mammalia one because no morphological data will overlap between the fossil Mammalia and the living ones.

artefacts: (1) the impossibility for a fossil to branch in the right clade if there is no morphological data available in this clade (Guillerme and Cooper, 2015); and (2) the higher probability for a fossil to branch within a wrong clade that has more morphological data available for living taxa than the right clade (Guillerme and Cooper, 2015). In this study, we assess the level of available morphological data in living mammals in order to highlight the two caveats described above. We collected living mammals morphological data from @256 phylogenetic matrices available online and measured the proportion of morphological data availability for each mammalian order. Additionally, in the mammalian orders where data was missing, we estimated the structure of the available data to detect if the available data was biased towards certain clades only or if it was randomly distributed using the Net Relatedness Index (Webb et al., 2002).

Material and Methods

Matrices search

To investigate the living mammals with available morphological data, we downloaded cladistic matrices from three major public databases: morphobank (`http://www.morphobank.org/`) (O'Leary and Kaufman, 2011), Graeme Lloyd's website (`http://graemetlloyd.com/`) and Ross Mounce's GitHub repository (`https://github.com/rossmounce`). We downloaded all the matrices containing any

fossil or living mammal taxa from these databases. Additionally we ran a thorough Google Scholar search for matrices that might not have been uploaded on the previously cited data bases. We downloaded the additional morphological characters matrices from the 20 first Google search results matching with our selected keywords and with any of the 35 taxonomic levels (see supplementary materials for detailed description of the procedure). We downloaded 256 matrices containing a total of 9411 operational taxonomic units (OTUs) from the combination of both searches (public repositories and Google Scholar). The list of matrices is available in the supplementary materials.

We then transformed all the matrices to be in the same nexus format. We also standardised the taxonomic nomenclature by fixing invalid binomial inputs to match with the official taxonomic nomenclature rules (i.e. *H. sapiens* was transformed in *Homo sapiens*. We assigned each species as being either living or fossil using a taxonomic matching algorithm. We designated as living OTU all the OTUs that where either present in (Fritz et al., 2009) or (Wilson and Reeder, 2005). We designated as fossil OTUs all the OTUs that where present in the Paleobiology database. For the OTUs neither labelled as living or fossil we tried to decompose the OTUs name (i.e. *Homo_sapiens* became *Homo* and *sapiens* and tried to match the first element with any taxonomic level (Family, Genus, Species, etc.) from (Wilson and Reeder, 2005). The matching OTUs where labelled as living and the ones still not matching where ignored and labelled as non-applicable (NA; see supplementary material for more details on the

taxonomic matching algorithm).

Data availability analysis

Number of characters threshold.— From all the 256 matrices, we selected only the ones that displayed at least 100 morphological characters per OTU. This arbitrary threshold of a minimal number of morphological characters was chosen to be in adequacy with (Guillerme and Cooper, 2015) and (Harrison and Larsson, 2014). Also this threshold avoids biases towards small matrices that could be either not informative (Wagner, 2000) (e.g. too few characters not allowing character overlap among matrices) or made of non-applicable characters (Brazeau, 2011) (e.g. antlers which are sexual dimorphic characters proper to a specific clade).

Data availability.— To assess the data availability per mammal order, we calculated the percentage of OTUs with morphological data for three different taxonomic levels: Family, Genus and Species. We highlighted all the orders containing less than 25% of living taxa with morphological data because their amount of missing data (>75%) was higher than in (Guillerme and Cooper, 2015) and therefore highly probable of having wrong topology due to the missing data. On the opposite, orders with up to 25% of missing data (75% of available data) were shown to have no significant effect on topology (Guillerme and Cooper, 2015). We therefore used this value as a threshold of "good" data sampling.

Available data structure.— For the orders with no morphological data for all its OTUs at

the three different taxonomic levels (Family, Genus and Species), we investigated the structure of the available data to test if it was either (i) randomly distributed, (ii) over-dispersed or (iii) clustered. To measure the structure of the available data we used classic community structure metric from the `picante` R package (Kembel et al., 2010). We compared the structure of the available data for each order to the structure of a potentially fully sampled order (i.e. only the OTUs with available morphological data *versus* all the OTUs). For each orders and taxonomic levels that presented OTUs with no available morphological data, we calculated the Net Relatedness Index (NRI) which quantifies the overall distribution of the data with negative values showing more dispersed data and positive values more clustered data than expected by the null model (random) (Webb et al., 2002). We choose to present only the NRI values because they have been shown to be slightly less sensitive to the structure of the phylogeny (i.e. branch length and topology) (Letcher, 2009; Swenson, 2009) but we also calculated the two other common phylogenetic structure indices: Faith's Phylogenetic Distance (PD) (Faith, 1992) and the Nearest Taxon Index (NTI) (Webb et al., 2002). Both metrics are available in the Supplementary results.

Therefore because I was focusing on the part above and to make it repeatable I didn't did that part yet. However, I still planing on doing it! I just need to find a bit of time to write some code to make it non-GUI (and to be able to use big trees).

All the following procedure is repeatable and available on GitHub.

Results

Data availability

We extracted 1422 living mammal OTUs from the 256 matrices with a minimum of 6 characters and a maximum of 4541 per OTU. After removing all the matrices with less than 100 morphological characters, the number of extracted living mammals OTUs was down to 815. 11/28 orders have less than 25% of taxa with morphological data at a species level and 24/28 orders have less than 75% taxa with available morphological data. At the Genus level however only 3/28 orders have less than 25% of taxa with morphological data and 16/28 have less than 75%. Finally, at the family level no order has less than 25% taxa with available morphological data and only 5/28 have less than 75% (table 1.

Table 1: Proportion of available OTUs with morphological data per order and per taxonomic level. We highlighted in bold the orders that have more than 75% of missing data for each taxonomic level. Note that it is possible that more data is available at a higher taxonomic level (Genus > Species) since if the species name for an OTU was not or miss specified, we still counted the OTU for higher taxonomic level analysis.

Order	Taxonomic level	Fraction of OTUs	Percentage of OTUs
Monotremata	Family	2/2	100
Monotremata	Genus	2/3	66.67
Monotremata	Species	2/4	50
Didelphimorphia	Family	1/1	100
Didelphimorphia	Genus	16/16	100
Didelphimorphia	Species	40/84	47.62
Paucituberculata	Family	1/1	100
Paucituberculata	Genus	2/3	66.67
Paucituberculata	Species	2/5	40
Microbiotheria	Family	1/1	100
Microbiotheria	Genus	1/1	100

Microbiotheria	Species	1/1	100
Notoryctemorphia	Family	1/1	100
Notoryctemorphia	Genus	1/1	100
Notoryctemorphia	**Species**	**0/2**	**0**
Dasyuromorphia	Family	2/2	100
Dasyuromorphia	Genus	7/22	31.82
Dasyuromorphia	**Species**	**8/64**	**12.5**
Peramelemorphia	Family	2/2	100
Peramelemorphia	Genus	7/7	100
Peramelemorphia	Species	16/18	88.89
Diprotodontia	Family	9/11	81.82
Diprotodontia	Genus	20/38	52.63
Diprotodontia	**Species**	**16/126**	**12.7**
Afrosoricida	Family	2/2	100
Afrosoricida	Genus	17/17	100
Afrosoricida	Species	23/42	54.76
Macroscelidea	Family	1/1	100
Macroscelidea	Genus	4/4	100
Macroscelidea	Species	5/15	33.33
Tubulidentata	Family	1/1	100
Tubulidentata	Genus	1/1	100
Tubulidentata	Species	1/1	100
Hyracoidea	Family	1/1	100
Hyracoidea	Genus	1/3	33.33
Hyracoidea	Species	1/4	25
Proboscidea	Family	1/1	100
Proboscidea	Genus	1/2	50
Proboscidea	Species	1/3	33.33
Sirenia	Family	2/2	100
Sirenia	Genus	2/2	100
Sirenia	Species	2/4	50
Cingulata	Family	1/1	100
Cingulata	Genus	8/9	88.89
Cingulata	**Species**	**6/25**	**24**
Pilosa	Family	3/5	60

Pilosa	Genus	3/5	60
Pilosa	**Species**	**3/29**	**10.35**
Scandentia	Family	2/2	100
Scandentia	Genus	2/5	40
Scandentia	**Species**	**2/20**	**10**
Dermoptera	Family	1/1	100
Dermoptera	Genus	1/2	50
Dermoptera	Species	1/2	50
Primates	Family	15/15	100
Primates	Genus	48/68	70.59
Primates	**Species**	**56/351**	**15.95**
Rodentia	Family	10/32	31.25
Rodentia	**Genus**	**20/451**	**4.43**
Rodentia	**Species**	**10/2095**	**0.48**
Lagomorpha	Family	1/2	50
Lagomorpha	**Genus**	**1/12**	**8.33**
Lagomorpha	**Species**	**1/86**	**1.16**
Erinaceomorpha	Family	1/1	100
Erinaceomorpha	Genus	10/10	100
Erinaceomorpha	Species	21/22	95.45
Soricomorpha	Family	3/4	75
Soricomorpha	Genus	19/43	44.19
Soricomorpha	**Species**	**19/392**	**4.85**
Chiroptera	Family	13/18	72.22
Chiroptera	Genus	68/202	33.66
Chiroptera	**Species**	**108/1054**	**10.25**
Pholidota	Family	1/1	100
Pholidota	Genus	1/1	100
Pholidota	Species	3/8	37.5
Carnivora	Family	11/15	73.33
Carnivora	**Genus**	**30/125**	**24**
Carnivora	**Species**	**42/283**	**14.84**
Perissodactyla	Family	3/3	100
Perissodactyla	Genus	6/6	100
Perissodactyla	Species	7/16	43.75

Cetartiodactyla	Family	20/21	95.24
Cetartiodactyla	Genus	76/128	59.38
Cetartiodactyla	Species	106/311	34.08

Available data structure

Among the orders containing at least one OTU with no available morphological data, only two orders are significantly clustered: Carnivora and Chiroptera both at the species and the genus level (table 2.

Table 2: Data structure for the orders with OTUs without morphological data per taxonomic level. When the Net Relatedness Index (NRI) is negative, the OTUs are more dispersed than expected by chance (random); when the NRI is positive, the OTUs are more clustered by expected by chance. The p-value indicates the significance in difference from the null model (random).

Order	Taxonomic level	Percentage of OTUs	NRI	p-value
Monotremata	Genus	66.667	-0.695	0.663
Monotremata	Species	50	-0.966	0.566
Didelphimorphia	Species	47.619	-1.33	0.915
Paucituberculata	Genus	66.667	-0.756	0.682
Paucituberculata	Species	40	-0.64	0.493
Dasyuromorphia	Genus	31.818	-1.102	0.894
Dasyuromorphia	Species	12.5	-1.098	0.93
Peramelemorphia	Species	88.889	-0.55	0.748
Diprotodontia	Family	81.818	-0.349	0.551
Diprotodontia	Genus	52.632	-0.31	0.595
Diprotodontia	Species	12.698	-0.975	0.849
Afrosoricida	Species	54.762	1.555	0.077
Macroscelidea	Species	33.333	-0.474	0.66

Sirenia	Species	50	-0.957	0.845
Cingulata	Genus	88.889	1.31	0.229
Cingulata	Species	24	0.648	0.223
Pilosa	Family	60	-0.603	0.891
Pilosa	Genus	60	-0.877	0.795
Pilosa	Species	10.345	-1.508	0.997
Scandentia	Genus	40	-0.747	0.639
Scandentia	Species	10	-1.259	0.984
Primates	Genus	70.588	-0.302	0.607
Primates	Species	15.954	-1.504	0.951
Rodentia	Family	31.25	0.113	0.395
Rodentia	Genus	4.435	-1.032	0.848
Rodentia	Species	0.477	-0.967	0.856
Erinaceomorpha	Species	95.455	-0.777	0.914
Soricomorpha	Family	75	-0.95	0.619
Soricomorpha	Genus	44.186	1.157	0.116
Soricomorpha	Species	4.847	-1.869	0.977
Chiroptera	Family	72.222	0.866	0.201
Chiroptera	**Genus**	**33.663**	**18.87**	**0.001**
Chiroptera	**Species**	**10.247**	**20.112**	**0.001**
Pholidota	Species	37.5	1.14	0.175
Carnivora	Family	73.333	0.517	0.285
Carnivora	**Genus**	**24**	**4.014**	**0.002**
Carnivora	**Species**	**14.841**	**17.954**	**0.001**
Perissodactyla	Species	43.75	0.733	0.199
Cetartiodactyla	Family	95.238	0.352	0.193
Cetartiodactyla	Genus	59.375	-3.221	1
Cetartiodactyla	Species	34.084	-2.768	1

Two contrasted results are shown on figures 2 and 3 with randomly distributed data in Cetartiodactyla (figure 2 and clustered available data in Carnivora (mainly

Canidae; figure 3.

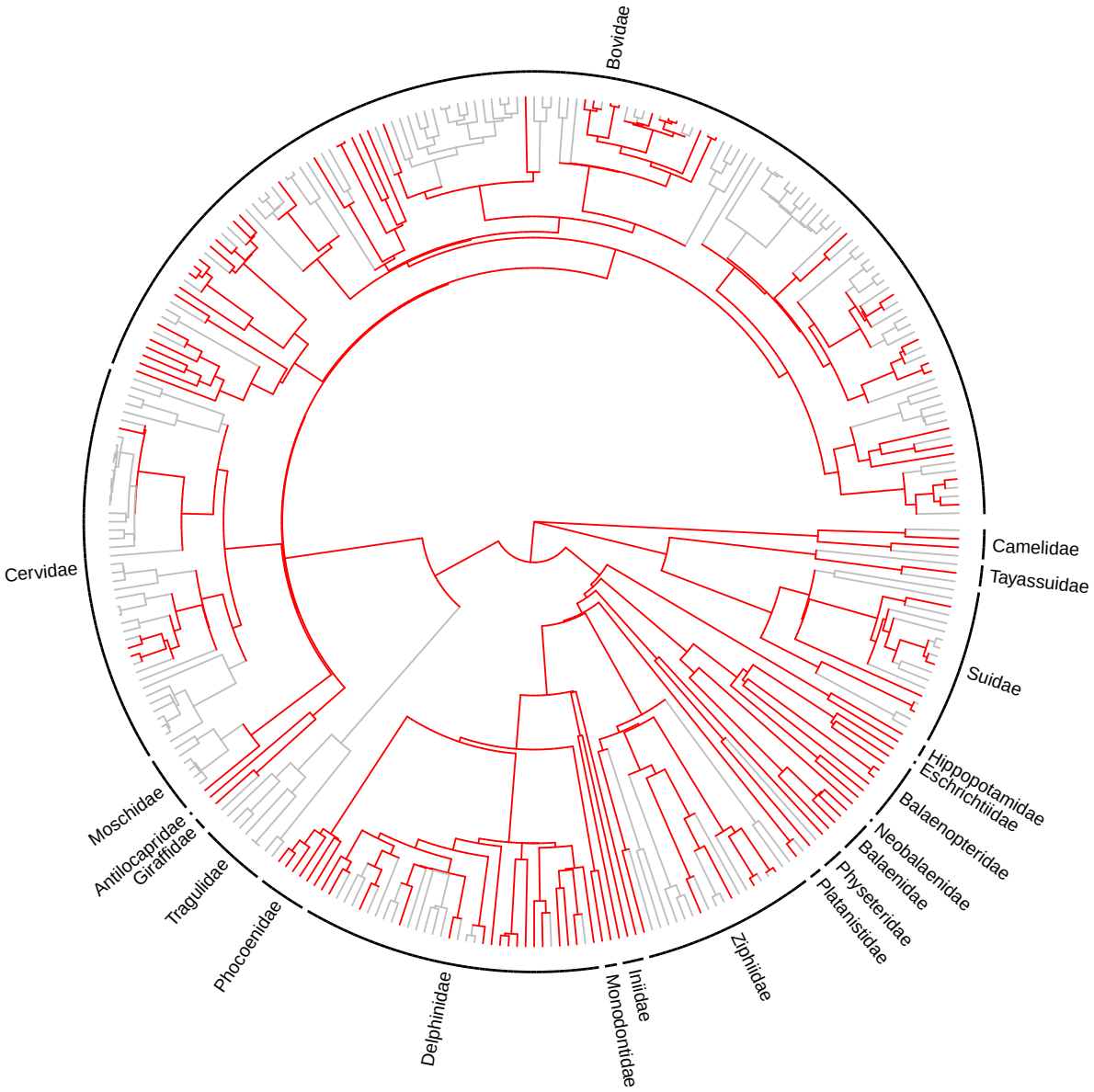

Figure 2: Distribution of available morphological data across Cetartiodactyla. Edges are colored in grey when no morphological data is available or in red when data is available.

Discussion

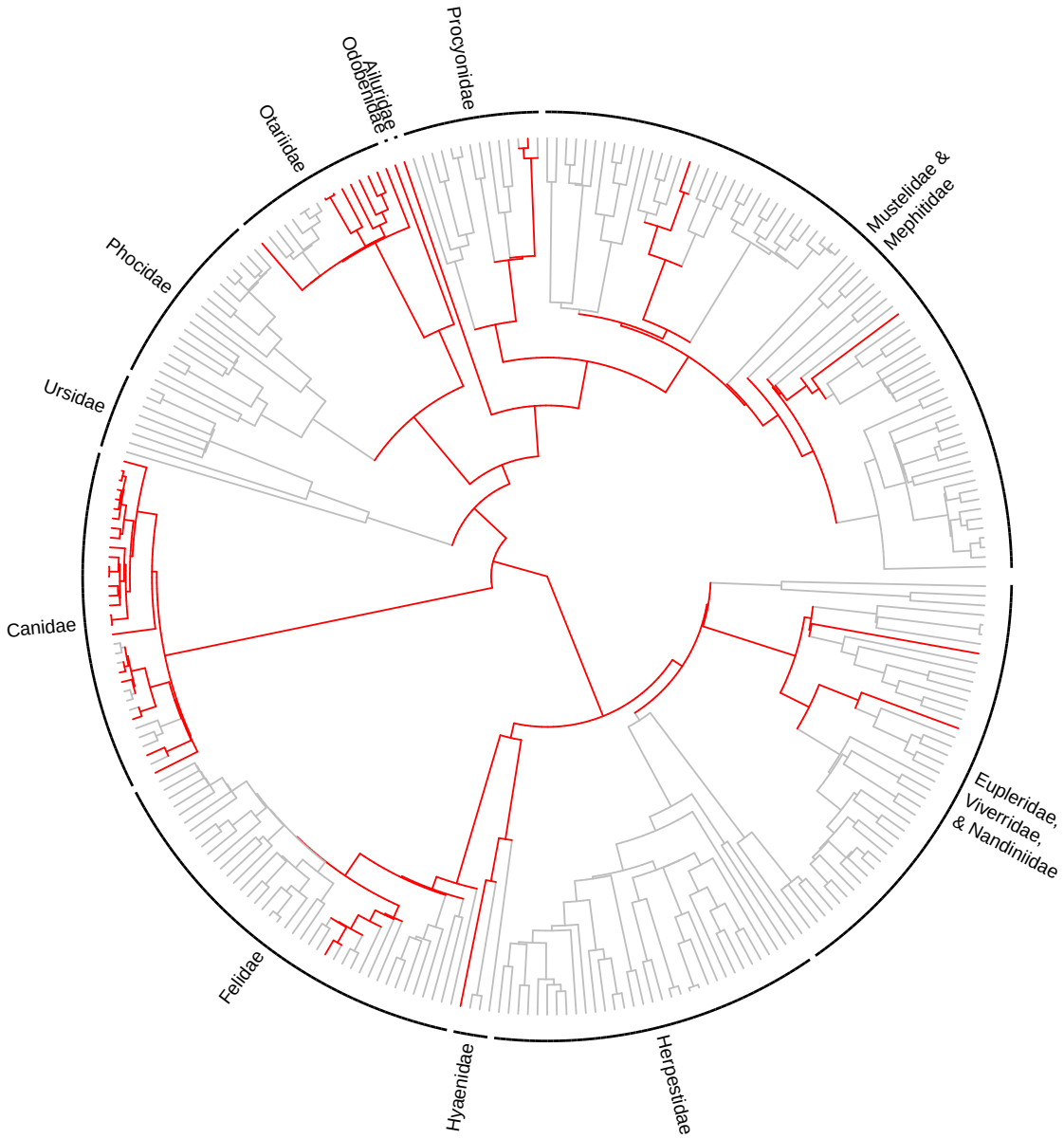

Figure 3: Distribution of available morphological data across Carnivora. Edges are colored in grey when no morphological data is available or in red when data is available.

Our results show that even though relations among living mammals are well known in a molecular framework (e.g. Fritz et al., 2009; Meredith et al., 2011; May-Collado et al., 2015) only little morphological data is available. This is primarily due to the fact that that morphological studies are mainly focusing on fossil taxa rather than living taxa (e.g. O'Leary et al., 2013; Ni et al., 2013). However, data availability varies in its structure as well as across taxonomic levels. Higher taxonomic levels are always better sampled than lower ones (table 1 and within these taxonomic levels, the structure of available data is mostly random apart for two orders (table 2.

The amount of living taxa with no available morphological data was surprisingly high at the species level: 24/28 orders have less than 75% taxa with available morphological data (and two of the 28 orders are monospecific). This threshold of 75% available data is the maximum amount of possible missing data (25%) before having a significant effect on the topology of Total Evidence trees (Guillerme and Cooper, 2015). Beyond this threshold, there is a noteworthy displacement of the wildcard taxa (*sensu* (Kearney, 2002) as well as a important decrease in the conservation of clades (Guillerme and Cooper, 2015). At the species level within most orders of mammals, it is therefore likely to see topological artefacts for the placement of fossil taxa. However, the data availability seems to be less an issue at higher taxonomic level (i.e. the Genus and the Family level). This point is important to consider from a practical point of view because of the slight discrepancy between neontological and palaeontological species. While neontological species are described on accurately measured morphology, genetic

distance, spacial distribution and even behaviour (e.g. Kelly et al., 2014); palaeontological data can be based only on morphological, spatial and temporal data (e.g. Ni et al., 2013). Because of this difference between neontological species (*sensu* reproductive isolates) and paleontological species (*sensu* morpho-species), most palaeontological studies are using the Genus level for their smallest taxonomic OTU (e.g. Ni et al., 2013; O'Leary et al., 2013). In this frame of palaeontological data usage, the data availability at the Genus level in living mammals should be our priority concern.

Regardless the low level of available data, it is encouraging that for most orders (except Carnivora and Chiroptera), it is randomly distributed across the phylogeny. Therefore, it is likely that the addition of fossil taxa within these orders will not be biased by oversampling but will probably branch near the taxa that has the most similar available data (as expected). When only few data is available, the ideal scenario will be that the data is over-dispersed (i.e. that there is data in at least every sub-clade) in order to maximize the possibilities of the fossil to branch in the right clade. In the second scenario, when the data is randomly distributed we expected no special bias in the placement of the fossil (Guillerme and Cooper, 2015). However, in the third scenario, when the data is clustered we expect two major biases to occur: (1) first, the fossil will not be able to branch within a clade containing no data (e.g. Herpestidae in Carnivora; figure 1 and 3, and (2) second, the fossil will have a higher probability, at random, to branch within the clade containing most of the available data (e.g. a

Carnivora fossil will have more chance to branch, randomly, to the Canidae clade than any other clade in Carnivora; figure 3.

On the other hand, any Carnivora fossil with uncertain phylogenetic affinities (*Incertae sedis* will have a higher probability of branching by chance within the Canidae because they are oversampled within the Carnivora (figure 3.

In this study, we treated all morphological matrices to be equal in a similar way that molecular matrices are: for example if a matrix A contains 100 characters for taxa X and Y, and a matrix B contains 50 characters for taxa X and Z, we assumed that both matrices can be combined in a supermatrix way leading to a matrix containing 150 characters for taxon X, 100 for taxon Y and 50 for taxon Z. However, it is clear that morphological data cannot be treated that way (Brazeau, 2011). In fact, for the same taxon, some characters might overlap (e.g. matrix A has a character coding for the shape of a particular morphological feature and matrix B as a first character coding for the presence of this morphological feature and a second one coding for its shape; in this case, these three characters are compound characters (Brazeau, 2011). However, in reasonably sized matrices ($>$ 100 characters (Guillerme and Cooper, 2015; Harrison and Larsson, 2014) it is more likely that a number of characters are consistently conserved among the different matrices (e.g. Ross et al., 1998; Seiffert et al., 2003; Marivaux et al., 2005; Seiffert et al., 2005; Bloch et al., 2007; Kay et al., 2008; Silcox, 2008; Seiffert et al., 2009; Tabuce et al., 2009; Boyer et al., 2010; Seiffert et al., 2010; Marivaux et al., 2013; Ni et al., 2013). A conservative approach to avoid compound characters could be to select

only the latest matrix for each taxon.

Following the recommendations in (Guillerme and Cooper, 2015), one should code morphological characters for a maximum of living species in order to improve the topology of Total Evidence trees containing both living and fossil taxa. Since the data for living mammals is usually easily available in world wide vast natural history collections, we propose that an increased effort in coding morphological characters from living species should be done via engaging in collaborative data collection projects through web portals such as *morphobank* (O'Leary and Kaufman, 2011).

Ethics statement

Data accessibility statement

All data is available and reproducible on GitHub.

Authors contributions statement

Conceived and designed the experiments: TG NC. Performed the experiments: TG. Analyzed the data: TG. Wrote the paper: TG NC.

Acknowledgements

Nick Matzke, April Wright, David Bapst and Graeme Lloyd.

Funding statement

This work was funded by a European Commission CORDIS Seventh Framework Programme (FP7) Marie Curie CIG grant (proposal number: 321696).

*

References

Beck, R. M. and M. S. Lee. 2014. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. Proceedings of the Royal Society B: Biological Sciences 281:1–10.

Bloch, J. I., M. T. Silcox, D. M. Boyer, and E. J. Sargis. 2007. New paleocene skeletons and the relationship of plesiadapiforms to crown-clade primates. Proceedings of the National Academy of Sciences 104:1159–1164.

Boyer, D. M., E. R. Seiffert, and E. L. Simons. 2010. Astragalar morphology of afradapis, a large adapiform primate from the earliest late eocene of egypt. American journal of physical anthropology 143:383–402.

Brazeau, M. D. 2011. Problematic character coding methods in morphology and their effects. Biological Journal of the Linnean Society 104:489–498.

Dietl, G. P. and K. W. Flessa. 2011. Conservation paleobiology: putting the dead to work. Trends in Ecology and Evolution 26:30–37.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with beauti and the beast 1.7. Molecular Biology and Evolution 29:1969–1973.

Eernisse, D. and A. Kluge. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molecular Biology and Evolution 10:1170–1195.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1 – 10.

Fritz, S. A., O. R. P. Bininda-Emonds, and A. Purvis. 2009. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. Ecology Letters 12:538–549.

Fritz, S. A., J. Schnitzler, J. T. Eronen, C. Hof, K. Bhning-Gaese, and C. H. Graham. 2013. Diversity in time and space: wanted dead and alive. Trends in Ecology and Evolution 28:509 – 516.

Guillerme, T. and N. Cooper. 2015. Effects of missing data on topological inference using a total evidence approach. PLoS ONE X:X.

Harrison, B., Luke and C. E. Larsson, Hans. 2014. Among-Character rate variation distributions in phylogenetic analysis of discrete morphological characters. Systematic biology .

Jackson, J. and D. Erwin. 2006. What can we learn about ecology and evolution from the fossil record? Trends in Ecology and Evolution 21:322–328.

Kay, R. F., J. Fleagle, T. Mitchell, M. Colbert, T. Bown, and D. W. Powers. 2008. The anatomy of dolichocebus gaimanensis, a stem platyrrhine monkey from argentina. Journal of Human Evolution 54:323–382.

Kearney, M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Systematic Biology 51:369–381.

Kelly, S. B. A., D. J. Kelly, N. Cooper, A. Bahrun, K. Analuddin, and N. M. Marples. 2014. Molecular and phenotypic data support the recognition of the Wakatobi flowerpecker (*Dicaeum kuehni*) from the unique and understudied Sulawesi region. PLoS ONE 9:e98694.

Kembel, S., P. Cowan, M. Helmus, W. Cornwell, H. Morlon, D. Ackerly, S. Blomberg, and C. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464.

Letcher, S. G. 2009. Phylogenetic structure of angiosperm communities during tropical forest succession. Proceedings of the Royal Society of London B: Biological Sciences .

Marivaux, L., P.-O. Antoine, S. R. H. Baqri, M. Benammi, Y. Chaimanee, J.-Y. Crochet, D. De Franceschi, N. Iqbal, J.-J. Jaeger, G. Métais, et al. 2005. Anthropoid primates from the oligocene of pakistan (bugti hills): data on early anthropoid evolution and biogeography. proceedings of the national Academy of Sciences of the United States of America 102:8436–8441.

Marivaux, L., A. Ramdarshan, E. M. Essid, W. Marzougui, H. K. Ammar, R. Lebrun,

B. Marandat, G. Merzeraud, R. Tabuce, and M. Vianey-Liaud. 2013. Djebelemur, a tiny pre-tooth-combed primate from the eocene of tunisia: a glimpse into the origin of crown strepsirhines. PloS one 8:e80778.

Matzke, N. J. 2014. Beastmaster: Automated conversion of nexus data to beast2 xml format, for fossil tip-dating and other uses. `http://phylo.wikidot.com/beastmaster`.

May-Collado, L. J., C. W. Kilpatrick, and I. Agnarsson. 2015. Mammals from down under: a multi-gene species-level phylogeny of marsupial mammals (mammalia, metatheria). PeerJ 3:e805.

Meredith, R., J. Janečka, J. Gatesy, O. Ryder, C. Fisher, E. Teeling, A. Goodbla, E. Eizirik, T. L. Simão, T. Stadler, D. Rabosky, R. Honeycutt, J. Flynn, C. Ingram, C. Steiner, T. Williams, T. Robinson, B. Angela, M. Westerman, N. Ayoub, M. Springer, and W. Murphy. 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.

Meseguer, A. S., J. M. Lobo, R. Ree, D. J. Beerling, and I. Sanmartn. 2015. Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: The case of hypericum (hypericaceae). Systematic Biology 64:215–232.

Ni, X., D. L. Gebo, M. Dagosto, J. Meng, P. Tafforeau, J. J. Flynn, and K. C. Beard. 2013.

The oldest known primate skeleton and early haplorhine evolution. Nature 498:60–64.

O'Leary, M. A., J. I. Bloch, J. J. Flynn, T. J. Gaudin, A. Giallombardo, N. P. Giannini, S. L. Goldberg, B. P. Kraatz, Z.-X. Luo, J. Meng, X. Ni, M. J. Novacek, F. A. Perini, Z. S. Randall, G. W. Rougier, E. J. Sargis, M. T. Silcox, N. B. Simmons, M. Spaulding, P. M. Velazco, M. Weksler, J. R. Wible, and A. L. Cirranello. 2013. The placental mammal ancestor and the postK-Pg radiation of placentals. Science 339:662–667.

O'Leary, M. A. and S. Kaufman. 2011. Morphobank: phylophenomics in the cloud. Cladistics 27:529–537.

Pyron, R. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology 60:466–481.

Quental, T. and C. Marshall. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology and Evolution 25:434–441.

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. Murray, and A. Rasnitsyn. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Systematic Biology 61:973–999.

Ross, C., B. Williams, and R. F. Kay. 1998. Phylogenetic analysis of anthropoid relationships. Journal of Human Evolution 35:221–306.

Schrago, C., B. Mello, and A. Soares. 2013. Combining fossil and molecular data to date

the diversification of New World Primates. Journal of Evolutionary Biology 26:2438–2446.

Seiffert, E. R., J. M. Perry, E. L. Simons, and D. M. Boyer. 2009. Convergent evolution of anthropoid-like adaptations in eocene adapiform primates. Nature 461:1118–1121.

Seiffert, E. R., E. L. Simons, and Y. Attia. 2003. Fossil evidence for an ancient divergence of lorises and galagos. Nature 422:421–424.

Seiffert, E. R., E. L. Simons, D. M. Boyer, J. M. Perry, T. M. Ryan, and H. M. Sallam. 2010. A fossil primate of uncertain affinities from the earliest late eocene of egypt. Proceedings of the National Academy of Sciences 107:9712–9717.

Seiffert, E. R., E. L. Simons, W. C. Clyde, J. B. Rossie, Y. Attia, T. M. Bown, P. Chatrath, and M. E. Mathison. 2005. Basal anthropoids from egypt and the antiquity of africa's higher primate radiation. Science 310:300–304.

Silcox, M. T. 2008. The biogeographic origins of primates and euprimates: east, west, north, or south of eden? Pages 199–231 *in* Mammalian Evolutionary Morphology. Springer.

Slater, G. J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the cretaceous-palaeogene boundary. Methods in Ecology and Evolution 4:734–744.

Slater, G. J. and L. J. Harmon. 2013. Unifying fossils and phylogenies for comparative

analyses of diversification and trait evolution. Methods in Ecology and Evolution 4:699–702.

Swenson, N. G. 2009. Phylogenetic resolution and quantifying the phylogenetic diversity and dispersion of communities. PLoS ONE 4:e4390.

Tabuce, R., L. Marivaux, R. Lebrun, M. Adaci, M. Bensalah, P.-H. Fabre, E. Fara, H. G. Rodrigues, L. Hautier, J.-J. Jaeger, et al. 2009. Anthropoid versus strepsirhine status of the african eocene primates algeripithecus and azibius: craniodental evidence. Proceedings of the Royal Society B: Biological Sciences Page rspb20091339.

Wagner, P. J. 2000. Exhaustion of morphologic character states among fossil taxa. Evolution 54:365–386.

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and community ecology. Annual review of ecology and systematics Pages 475–505.

Wilson, D. E. and D. M. Reeder. 2005. Mammal species of the world: a taxonomic and geographic reference vol. 1. JHU Press.

Wood, H. M., N. J. Matzke, R. G. Gillespie, and C. E. Griswold. 2013. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Systematic Biology 62:264–284.

Supplementary Material

Data collection

1- Data collection: key words, clade (ordinal) metacharacters, Google Search terms, Google Search protocol, Google Search rarefaction curve.

Public repositories

We downloaded available matrices containing fossil and/or living mammal taxa from the three following data bases using the following list of keywords:

`Mammalia; Monotremata; Marsupialia; Placentalia; Macroscelidea; Afrosoricida; Tubulidentata; Hyracoidea; Proboscidea; Sirenia; Pilosa; Cingulata; Scandentia; Dermoptera; Primates; Lagomorpha; Rodentia; Erinaceomorpha; Soricomorpha; Cetacea; Artiodactyla; Cetartiodactyla; Chiroptera; Perissodactyla; Pholidota; Carnivora; Didelphimorphia; Paucituberculata; Microbiotheria; Dasyuromorphia; Peramelemorphia; Notoryctemorphia; Diprotodontia.`

Details about each public repository specific search option is listed below. Note that some matrices have been downloaded from more than one database but that it is not an issue since we are interested in the total number of living OTUs and that if some where present in more than one matrix, they still only counted as a unique OTU.

Morphobank.— We accessed the Morphobank repository (`http://www.morphobank.org/`) on the 5th of December 2014 and used the keywords listed above in the search menue. We downloaded the data associated with each project matching with the keyword.

Graeme Lloyd.— We accessed Graeme Lloyd's website repository

(http://graemetlloyd.com/) on the 5th of December 2014 and downloaded all the matrices that were available with a direct download link in the mammal data section of the website (http://graemetlloyd.com/matrmamm.html).

Ross Mounce.— We accessed Ross Mounce's GitHub repository (https://github.com/rossmounce) on the 2nd of December 2014 and downloaded every 601 matrix. We then ran a shell script to select only the matrices that had any text element that match with one of the search terms. To make the matrix selection more thorough, we ignored the keywords case as well as the latin suffix (*ia*, *ata*, *ea*, and *a*).

Google scholars

To make sure we didn't miss any extra matrix that wasn't available on one of these repository, we ran a Google Scholar search on the 5th of January. We used the following key words:

order ("morphology" OR "morphological" OR "cladistic") AND characters matrix paleontology phylogeny

were *order* was replaced by all the keywords listed above. For each 33 keywords, we selected the 20 first papers to match the Google search published since 2010 resulting in 660 papers. Among these papers, not all contained relevant data (discrete morphological characters AND mammalian data). We selected only the 20 first results per search term to avoid downloading articles that were to irrelevant. Among the 660 papers, only 50 contained a total of 425 extra living OTUs (figure S1). Also we decided

to select only the articles published since 2010 because nearly every one of the recent published matrix contains both a fraction of morphological characters and OTUs from previous studies. For example in primates the character *p7* coded first by Ross et al. (1998) is reused with the same living species in Seiffert et al. (2003), Marivaux et al. (2005), Seiffert et al. (2005), Bloch et al. (2007), Bloch et al. (2007), Kay et al. (2008), Silcox (2008), Seiffert et al. (2009), Tabuce et al. (2009), Boyer et al. (2010), Seiffert et al. (2010), Marivaux et al. (2013) and Ni et al. (2013).

Standardising the matrices

We transformed all the non-nexus matrices (tnt, word, excel, jpeg) to nexus format manually. We then cleaned the nexus matrices by removing any extra information (trees, continuous characters, morphological characters description, molecular data) to end up with nexus matrices containing only the discrete morphological data. We then manually fixed the wrong bionomial names format (e.g. *H. sapiens*) into the correct ones (e.g. *Homo sapiens*) using the abbreviation list in the concerned publications.

Selecting the living OTUs

Finally we applied a taxonomic matching algorithm to classify the OTUs as either living or fossil. The algorithm is matching every OTU name from every matrix with one of the following taxonomic references: the list of taxa from the Fritz *et al.* supertree (2009) Fritz et al. (2009); the taxonomic list from the Wilson and Reeder's Mammals Species of

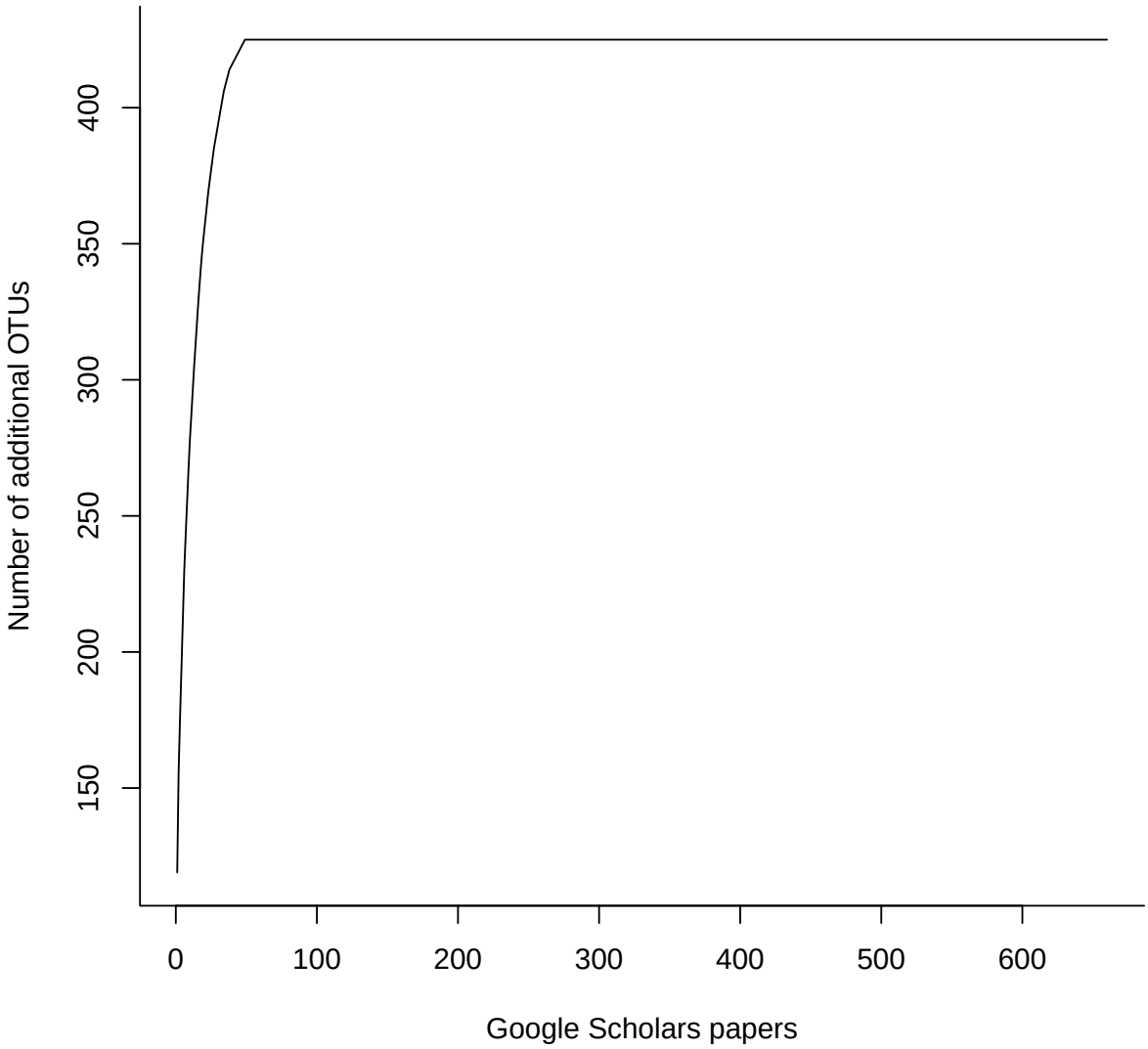

Figure S1: Google searches additional OTUs rarefaction curve. The x axis represent the number of google scholar matches (papers, books or abstracts) and the y axis represents the cumulative number of additional living OTUs per google scholar match.

the World (2005) Wilson and Reeder (2005) and the list of all the mammal fossil from the Paleobio Database (`http://paleobiodb.org/cgi-bin/bridge.pl?a=login`) accessed on the 13th of Janurary 2015. The OTUs that matched with one of the two first references were considered as living OTUs, the OTUs matching with the third reference were considered as fossil OTUs, finally, the OTUs matching with non of the references were discarded (figure S2).

Supplementary results

The following section contains supplementary results to the main body: the available data structure using the NTI and the PD metric; the proportion of available data and the data structure for all the matrices (including the matrices with less than 100 characters); and phylogenetical representation of the data availability per order (excluding Cetartiodactyla and Carnivora, present in the main body).

Table S1: Data structure for the orders with OTUs without morphological data per taxonomic level. When the Nearest Taxon Index (NTI) is negative, the OTUs are more dispersed than expected by chance (random); when the NTI is positive, the OTUs are more clustered by expected by chance. The p-value indicates the significance in difference from the null model (random).

Order	Taxonomic level	Fraction of OTUs	Percentage of OTUs	NTI	p-value
Monotremata	Genus	2/3	66.667	-0.711	0.668
Monotremata	Species	2/4	50	-1.012	0.591
Didelphimorphia	Species	42/84	50	-1.531	0.937
Dasyuromorphia	Genus	8/22	36.364	-1.326	0.89
Dasyuromorphia	Species	9/64	14.062	-0.899	0.813
Peramelemorphia	Species	16/18	88.889	0.482	0.211

Diprotodontia	Genus	25/38	65.789	-0.937	0.812
Diprotodontia	Species	31/126	24.603	-2.521	0.998
Afrosoricida	Species	23/42	54.762	-2.229	0.993
Macroscelidea	Species	12/15	80	-0.401	0.649
Proboscidea	Species	2/3	66.667	-0.697	0.664
Sirenia	Species	2/4	50	-0.904	0.83
Cingulata	Genus	8/9	88.889	-1.663	0.893
Cingulata	Species	9/25	36	0.469	0.325
Pilosa	Family	4/5	80	1.767	0.095
Pilosa	Genus	4/5	80	-0.623	0.805
Pilosa	Species	5/29	17.241	0.381	0.337
Scandentia	Genus	2/5	40	-0.755	0.645
Scandentia	Species	3/20	15	-0.965	0.866
Primates	Genus	48/68	70.588	-1.262	0.905
Primates	Species	57/351	16.239	-2.459	0.99
Rodentia	Family	16/32	50	0.827	0.201
Rodentia	**Genus**	**63/451**	**13.969**	**2.271**	**0.012**
Rodentia	**Species**	**76/2095**	**3.628**	**4.25**	**0.001**
Lagomorpha	Genus	5/12	41.667	-0.887	0.661
Lagomorpha	Species	12/86	13.953	-2.559	0.992
Erinaceomorpha	Species	21/22	95.455	-0.202	0.349
Soricomorpha	Family	3/4	75	-0.951	0.594
Soricomorpha	Genus	19/43	44.186	-1.033	0.843
Soricomorpha	Species	21/392	5.357	-2.976	0.999
Chiroptera	Family	15/18	83.333	0.527	0.376
Chiroptera	Genus	77/202	38.119	-0.65	0.741
Chiroptera	**Species**	**155/1054**	**14.706**	**1.973**	**0.025**
Pholidota	Species	4/8	50	-0.304	0.593
Carnivora	Family	14/15	93.333	0.534	0.279
Carnivora	Genus	54/125	43.2	1.342	0.095
Carnivora	Species	76/283	26.855	0.484	0.318
Perissodactyla	Species	10/16	62.5	-2.849	1
Cetartiodactyla	Family	20/21	95.238	1.05	0.271
Cetartiodactyla	Genus	99/128	77.344	-0.37	0.648
Cetartiodactyla	Species	150/311	48.232	-1.589	0.94

Table S2: Data structure for the orders with OTUs without morphological data per taxonomic level. When the Faith's Phylogenetic Distance (PD). The p-value indicates the significance in difference from the null model (random).

Order	Taxonomic level	Fraction of OTUs	Percentage of OTUs	PD	p-value
Monotremata	Genus	2/3	66.667	-0.694	0.663
Monotremata	Species	2/4	50	-1.066	0.621
Didelphimorphia	Species	42/84	50	-1.258	0.897
Notoryctemorphia	Species	0/2	0	NaN	0.5
Dasyuromorphia	Genus	8/22	36.364	-1.078	0.857
Dasyuromorphia	Species	9/64	14.062	-1.209	0.888
Peramelemorphia	Species	16/18	88.889	-0.101	0.268
Diprotodontia	Genus	25/38	65.789	0.305	0.355
Diprotodontia	Species	31/126	24.603	-1.178	0.888
Afrosoricida	Species	23/42	54.762	-2.248	0.988
Macroscelidea	Species	12/15	80	-0.721	0.692
Proboscidea	Species	2/3	66.667	-0.67	0.655
Sirenia	Species	2/4	50	-0.916	0.744
Cingulata	Genus	8/9	88.889	-0.539	0.553
Cingulata	Species	9/25	36	1.508	0.07
Pilosa	Family	4/5	80	0.461	0.515
Pilosa	Genus	4/5	80	-1.007	0.799
Pilosa	Species	5/29	17.241	-0.48	0.599
Scandentia	Genus	2/5	40	-0.771	0.658
Scandentia	Species	3/20	15	-1.51	0.894
Primates	Genus	48/68	70.588	-1.305	0.924
Primates	Species	57/351	16.239	-3.499	1
Rodentia	Family	16/32	50	0.671	0.241
Rodentia	Genus	63/451	13.969	1.238	0.111
Rodentia	**Species**	**76/2095**	**3.628**	**6.803**	**0.001**
Lagomorpha	Genus	5/12	41.667	-1.05	0.661
Lagomorpha	Species	12/86	13.953	-2.321	0.989

Erinaceomorpha	Species	21/22	95.455	-1.205	0.865
Soricomorpha	Family	3/4	75	-0.96	0.626
Soricomorpha	Genus	19/43	44.186	-1.559	0.944
Soricomorpha	Species	21/392	5.357	-3.346	1
Chiroptera	Family	15/18	83.333	0.439	0.352
Chiroptera	Genus	77/202	38.119	1.155	0.125
Chiroptera	**Species**	**155/1054**	**14.706**	**2.172**	**0.011**
Pholidota	Species	4/8	50	-0.234	0.485
Carnivora	Family	14/15	93.333	0.47	0.348
Carnivora	**Genus**	**54/125**	**43.2**	**2.866**	**0.005**
Carnivora	**Species**	**76/283**	**26.855**	**2.849**	**0.005**
Perissodactyla	Species	10/16	62.5	-2.274	0.995
Cetartiodactyla	Family	20/21	95.238	0.921	0.25
Cetartiodactyla	Genus	99/128	77.344	-1.037	0.85
Cetartiodactyla	Species	150/311	48.232	-2.007	0.983

Table S3: Proportion of available OTUs with morphological data per order and per taxonomic level (Character threshold = 1). We highlighted in bold the orders that have more than 75% of missing data for each taxonomic level. Note that it is possible that more data is available at a higher taxonomic level (Genus > Species) since if the species name for an OTU was not or miss specified, we still counted the OTU for higher taxonomic level analysis.

Order	Taxonomic level	Fraction of OTUs	Percentage of OTUs
Monotremata	Family	2/2	100
Monotremata	Genus	2/3	66.67
Monotremata	Species	2/4	50
Didelphimorphia	Family	1/1	100
Didelphimorphia	Genus	16/16	100
Didelphimorphia	Species	42/84	50
Paucituberculata	Family	1/1	100
Paucituberculata	Genus	3/3	100
Paucituberculata	Species	5/5	100

Microbiotheria	Family	1/1	100
Microbiotheria	Genus	1/1	100
Microbiotheria	Species	1/1	100
Notoryctemorphia	Family	1/1	100
Notoryctemorphia	Genus	1/1	100
Notoryctemorphia	**Species**	**0/2**	**0**
Dasyuromorphia	Family	2/2	100
Dasyuromorphia	Genus	8/22	36.36
Dasyuromorphia	**Species**	**9/64**	**14.06**
Peramelemorphia	Family	2/2	100
Peramelemorphia	Genus	7/7	100
Peramelemorphia	Species	16/18	88.89
Diprotodontia	Family	11/11	100
Diprotodontia	Genus	25/38	65.79
Diprotodontia	**Species**	**31/126**	**24.6**
Afrosoricida	Family	2/2	100
Afrosoricida	Genus	17/17	100
Afrosoricida	Species	23/42	54.76
Macroscelidea	Family	1/1	100
Macroscelidea	Genus	4/4	100
Macroscelidea	Species	12/15	80
Tubulidentata	Family	1/1	100
Tubulidentata	Genus	1/1	100
Tubulidentata	Species	1/1	100
Hyracoidea	Family	1/1	100
Hyracoidea	Genus	1/3	33.33
Hyracoidea	Species	1/4	25
Proboscidea	Family	1/1	100
Proboscidea	Genus	2/2	100
Proboscidea	Species	2/3	66.67
Sirenia	Family	2/2	100
Sirenia	Genus	2/2	100
Sirenia	Species	2/4	50
Cingulata	Family	1/1	100
Cingulata	Genus	8/9	88.89

Cingulata	Species	9/25	36
Pilosa	Family	4/5	80
Pilosa	Genus	4/5	80
Pilosa	**Species**	**5/29**	**17.24**
Scandentia	Family	2/2	100
Scandentia	Genus	2/5	40
Scandentia	**Species**	**3/20**	**15**
Dermoptera	Family	1/1	100
Dermoptera	Genus	1/2	50
Dermoptera	Species	1/2	50
Primates	Family	15/15	100
Primates	Genus	48/68	70.59
Primates	**Species**	**57/351**	**16.24**
Rodentia	Family	16/32	50
Rodentia	**Genus**	**63/451**	**13.97**
Rodentia	**Species**	**76/2095**	**3.63**
Lagomorpha	Family	2/2	100
Lagomorpha	Genus	5/12	41.67
Lagomorpha	**Species**	**12/86**	**13.95**
Erinaceomorpha	Family	1/1	100
Erinaceomorpha	Genus	10/10	100
Erinaceomorpha	Species	21/22	95.45
Soricomorpha	Family	3/4	75
Soricomorpha	Genus	19/43	44.19
Soricomorpha	**Species**	**21/392**	**5.36**
Chiroptera	Family	15/18	83.33
Chiroptera	Genus	77/202	38.12
Chiroptera	**Species**	**155/1054**	**14.71**
Pholidota	Family	1/1	100
Pholidota	Genus	1/1	100
Pholidota	Species	4/8	50
Carnivora	Family	14/15	93.33
Carnivora	Genus	54/125	43.2
Carnivora	Species	76/283	26.86
Perissodactyla	Family	3/3	100

Perissodactyla	Genus	6/6	100
Perissodactyla	Species	10/16	62.5
Cetartiodactyla	Family	20/21	95.24
Cetartiodactyla	Genus	99/128	77.34
Cetartiodactyla	Species	150/311	48.23

Table S4: Data structure for the orders with OTUs without morphological data per taxonomic level (Character threshold = 1). When the Net Relatedness Index (NRI) is negative, the OTUs are more dispersed than expected by chance (random); when the NRI is positive, the OTUs are more clustered by expected by chance. The p-value indicates the significance in difference from the null model (random).

Order	Taxonomic level	Fraction of OTUs	Percentage of OTUs	NRI	p-value
Monotremata	Genus	2/3	66.667	-0.695	0.663
Monotremata	Species	2/4	50	-0.966	0.566
Didelphimorphia	Species	42/84	50	-1.96	0.991
Dasyuromorphia	Genus	8/22	36.364	-0.747	0.768
Dasyuromorphia	Species	9/64	14.062	-0.641	0.789
Peramelemorphia	Species	16/18	88.889	-0.514	0.742
Diprotodontia	**Genus**	**25/38**	**65.789**	**2.305**	**0.021**
Diprotodontia	**Species**	**31/126**	**24.603**	**2.006**	**0.042**
Afrosoricida	Species	23/42	54.762	1.553	0.089
Macroscelidea	Species	12/15	80	-1.023	0.832
Proboscidea	Species	2/3	66.667	-0.727	0.673
Sirenia	Species	2/4	50	-0.94	0.833
Cingulata	Genus	8/9	88.889	1.366	0.215
Cingulata	Species	9/25	36	1.821	0.055
Pilosa	Family	4/5	80	-0.247	0.48
Pilosa	Genus	4/5	80	-1.21	0.798
Pilosa	Species	5/29	17.241	-1.015	0.861
Scandentia	Genus	2/5	40	-0.785	0.669
Scandentia	Species	3/20	15	-1.462	0.898
Primates	Genus	48/68	70.588	-0.353	0.617

Primates	Species	57/351	16.239	-1.586	0.941
Rodentia	Family	16/32	50	0.956	0.155
Rodentia	Genus	63/451	13.969	-1.614	0.961
Rodentia	**Species**	**76/2095**	**3.628**	**5.184**	**0.001**
Lagomorpha	Genus	5/12	41.667	-1.078	0.661
Lagomorpha	Species	12/86	13.953	-1.288	0.954
Erinaceomorpha	Species	21/22	95.455	-0.808	0.916
Soricomorpha	Family	3/4	75	-0.941	0.611
Soricomorpha	Genus	19/43	44.186	1.202	0.11
Soricomorpha	Species	21/392	5.357	-2.298	0.996
Chiroptera	Family	15/18	83.333	0.047	0.434
Chiroptera	**Genus**	**77/202**	**38.119**	**14.216**	**0.001**
Chiroptera	**Species**	**155/1054**	**14.706**	**11.347**	**0.001**
Pholidota	Species	4/8	50	-0.034	0.482
Carnivora	Family	14/15	93.333	0.671	0.363
Carnivora	**Genus**	**54/125**	**43.2**	**4.624**	**0.001**
Carnivora	**Species**	**76/283**	**26.855**	**7.448**	**0.001**
Perissodactyla	Species	10/16	62.5	-0.042	0.474
Cetartiodactyla	Family	20/21	95.238	0.461	0.166
Cetartiodactyla	Genus	99/128	77.344	-1.616	0.954
Cetartiodactyla	Species	150/311	48.232	-0.901	0.81

Table S5: Data structure for the orders with OTUs without morphological data per taxonomic level (Character threshold = 1). When the Nearest Taxon Index (NTI) is negative, the OTUs are more dispersed than expected by chance (random); when the NTI is positive, the OTUs are more clustered by expected by chance. The p-value indicates the significance in difference from the null model (random).

Order	Taxonomic level	Fraction of OTUs	Percentage of OTUs	NTI	p-value
Monotremata	Genus	2/3	66.667	-0.711	0.668
Monotremata	Species	2/4	50	-1.012	0.591
Didelphimorphia	Species	42/84	50	-1.531	0.937
Dasyuromorphia	Genus	8/22	36.364	-1.326	0.89

Dasyuromorphia	Species	9/64	14.062	-0.899	0.813
Peramelemorphia	Species	16/18	88.889	0.482	0.211
Diprotodontia	Genus	25/38	65.789	-0.937	0.812
Diprotodontia	Species	31/126	24.603	-2.521	0.998
Afrosoricida	Species	23/42	54.762	-2.229	0.993
Macroscelidea	Species	12/15	80	-0.401	0.649
Proboscidea	Species	2/3	66.667	-0.697	0.664
Sirenia	Species	2/4	50	-0.904	0.83
Cingulata	Genus	8/9	88.889	-1.663	0.893
Cingulata	Species	9/25	36	0.469	0.325
Pilosa	Family	4/5	80	1.767	0.095
Pilosa	Genus	4/5	80	-0.623	0.805
Pilosa	Species	5/29	17.241	0.381	0.337
Scandentia	Genus	2/5	40	-0.755	0.645
Scandentia	Species	3/20	15	-0.965	0.866
Primates	Genus	48/68	70.588	-1.262	0.905
Primates	Species	57/351	16.239	-2.459	0.99
Rodentia	Family	16/32	50	0.827	0.201
Rodentia	**Genus**	**63/451**	**13.969**	**2.271**	**0.012**
Rodentia	**Species**	**76/2095**	**3.628**	**4.25**	**0.001**
Lagomorpha	Genus	5/12	41.667	-0.887	0.661
Lagomorpha	Species	12/86	13.953	-2.559	0.992
Erinaceomorpha	Species	21/22	95.455	-0.202	0.349
Soricomorpha	Family	3/4	75	-0.951	0.594
Soricomorpha	Genus	19/43	44.186	-1.033	0.843
Soricomorpha	Species	21/392	5.357	-2.976	0.999
Chiroptera	Family	15/18	83.333	0.527	0.376
Chiroptera	Genus	77/202	38.119	-0.65	0.741
Chiroptera	**Species**	**155/1054**	**14.706**	**1.973**	**0.025**
Pholidota	Species	4/8	50	-0.304	0.593
Carnivora	Family	14/15	93.333	0.534	0.279
Carnivora	Genus	54/125	43.2	1.342	0.095
Carnivora	Species	76/283	26.855	0.484	0.318
Perissodactyla	Species	10/16	62.5	-2.849	1
Cetartiodactyla	Family	20/21	95.238	1.05	0.271

Cetartiodactyla	Genus	99/128	77.344	-0.37	0.648
Cetartiodactyla	Species	150/311	48.232	-1.589	0.94

Table S6: Data structure for the orders with OTUs without morphological data per taxonomic level (Character threshold = 1). When the Faith's Phylogenetic Distance (PD). The p-value indicates the significance in difference from the null model (random).

Order	Taxonomic level	Fraction of OTUs	Percentage of OTUs	PD	p-value
Monotremata	Genus	2/3	66.667	-0.694	0.663
Monotremata	Species	2/4	50	-1.066	0.621
Didelphimorphia	Species	42/84	50	-1.258	0.897
Notoryctemorphia	Species	0/2	0	NaN	0.5
Dasyuromorphia	Genus	8/22	36.364	-1.078	0.857
Dasyuromorphia	Species	9/64	14.062	-1.209	0.888
Peramelemorphia	Species	16/18	88.889	-0.101	0.268
Diprotodontia	Genus	25/38	65.789	0.305	0.355
Diprotodontia	Species	31/126	24.603	-1.178	0.888
Afrosoricida	Species	23/42	54.762	-2.248	0.988
Macroscelidea	Species	12/15	80	-0.721	0.692
Proboscidea	Species	2/3	66.667	-0.67	0.655
Sirenia	Species	2/4	50	-0.916	0.744
Cingulata	Genus	8/9	88.889	-0.539	0.553
Cingulata	Species	9/25	36	1.508	0.07
Pilosa	Family	4/5	80	0.461	0.515
Pilosa	Genus	4/5	80	-1.007	0.799
Pilosa	Species	5/29	17.241	-0.48	0.599
Scandentia	Genus	2/5	40	-0.771	0.658
Scandentia	Species	3/20	15	-1.51	0.894
Primates	Genus	48/68	70.588	-1.305	0.924
Primates	Species	57/351	16.239	-3.499	1
Rodentia	Family	16/32	50	0.671	0.241
Rodentia	Genus	63/451	13.969	1.238	0.111
Rodentia	**Species**	**76/2095**	**3.628**	**6.803**	**0.001**

Lagomorpha	Genus	5/12	41.667	-1.05	0.661
Lagomorpha	Species	12/86	13.953	-2.321	0.989
Erinaceomorpha	Species	21/22	95.455	-1.205	0.865
Soricomorpha	Family	3/4	75	-0.96	0.626
Soricomorpha	Genus	19/43	44.186	-1.559	0.944
Soricomorpha	Species	21/392	5.357	-3.346	1
Chiroptera	Family	15/18	83.333	0.439	0.352
Chiroptera	Genus	77/202	38.119	1.155	0.125
Chiroptera	**Species**	**155/1054**	**14.706**	**2.172**	**0.011**
Pholidota	Species	4/8	50	-0.234	0.485
Carnivora	Family	14/15	93.333	0.47	0.348
Carnivora	**Genus**	**54/125**	**43.2**	**2.866**	**0.005**
Carnivora	**Species**	**76/283**	**26.855**	**2.849**	**0.005**
Perissodactyla	Species	10/16	62.5	-2.274	0.995
Cetartiodactyla	Family	20/21	95.238	0.921	0.25
Cetartiodactyla	Genus	99/128	77.344	-1.037	0.85
Cetartiodactyla	Species	150/311	48.232	-2.007	0.983

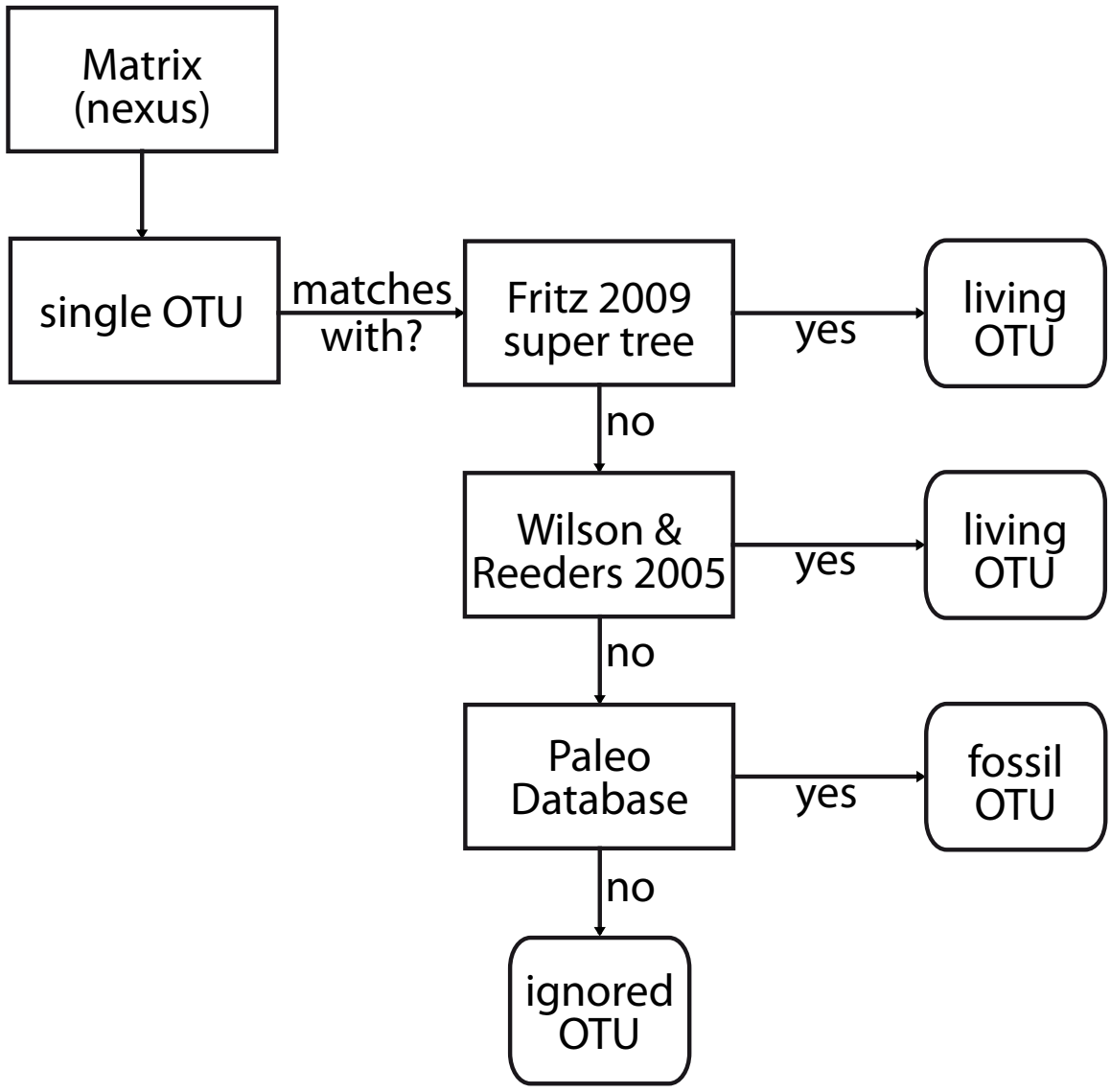

Figure S2: Taxonomic matching algorithm used in this study. For each matrix, each operational taxonomic units (OTU) is matched with the super tree from Fritz 2009. If the OTU matches, then it is classified as living. Else it is matched with the Wilson & Reeders 2005 taxonomy list. If the OTU matches, then it is classified as living. Else it is matched with the Paleo Database list of mammals. If the OTU matches, then it is classified as fossil. Else it is ignored.

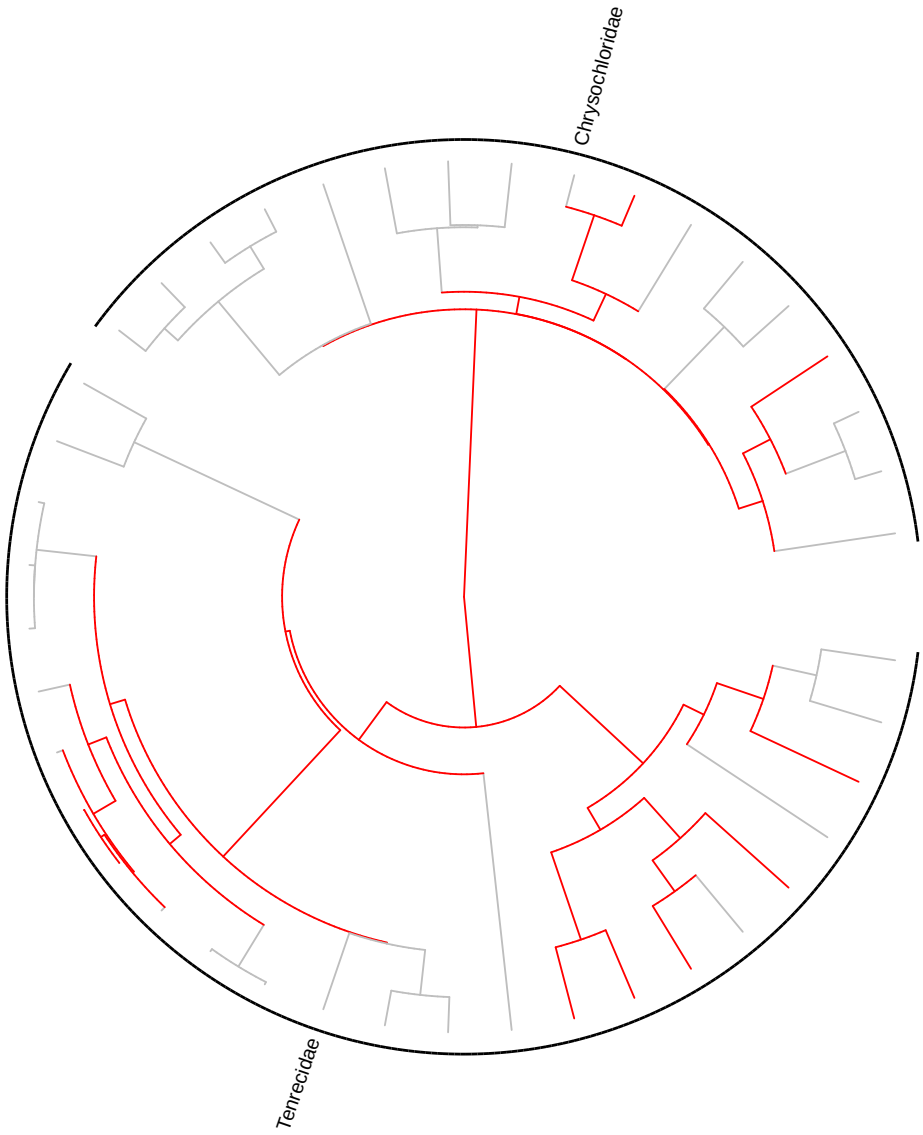

Figure S3: Distribution of available morphological data across Afrosoricida. Edges are colored in grey when no morphological data is available or in red when data is available.

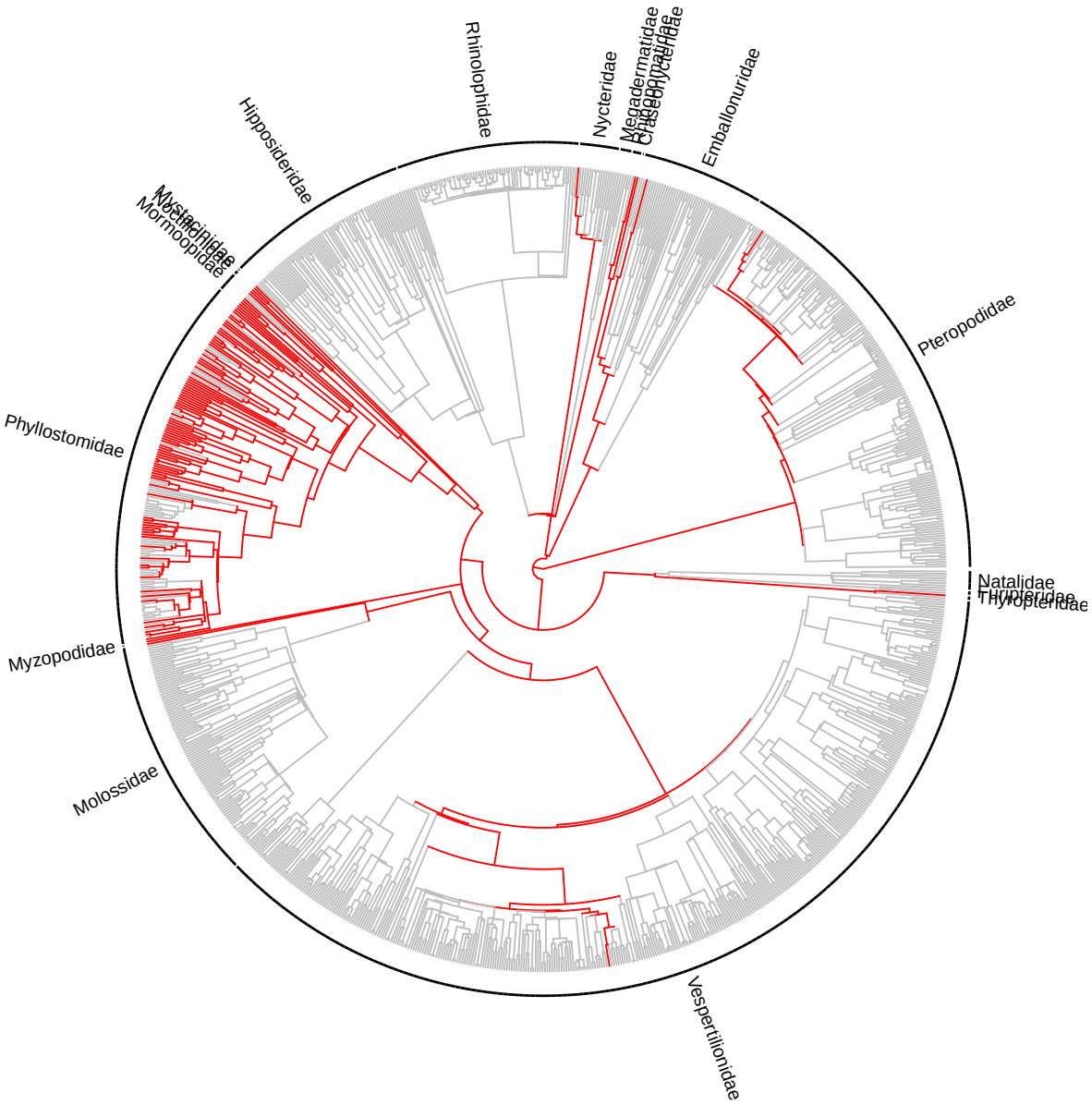

Figure S4: Distribution of available morphological data across Chiroptera. Edges are colored in grey when no morphological data is available or in red when data is available.

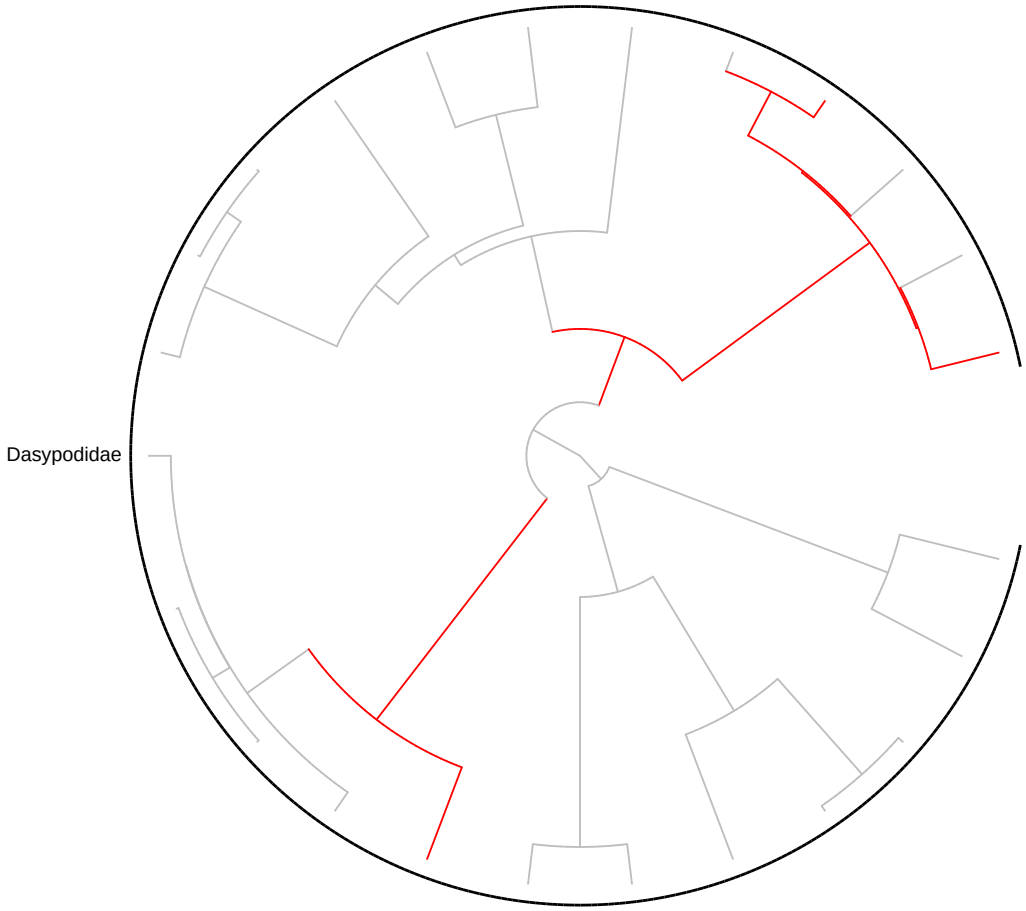

Figure S5: Distribution of available morphological data across Cingulata. Edges are colored in grey when no morphological data is available or in red when data is available.

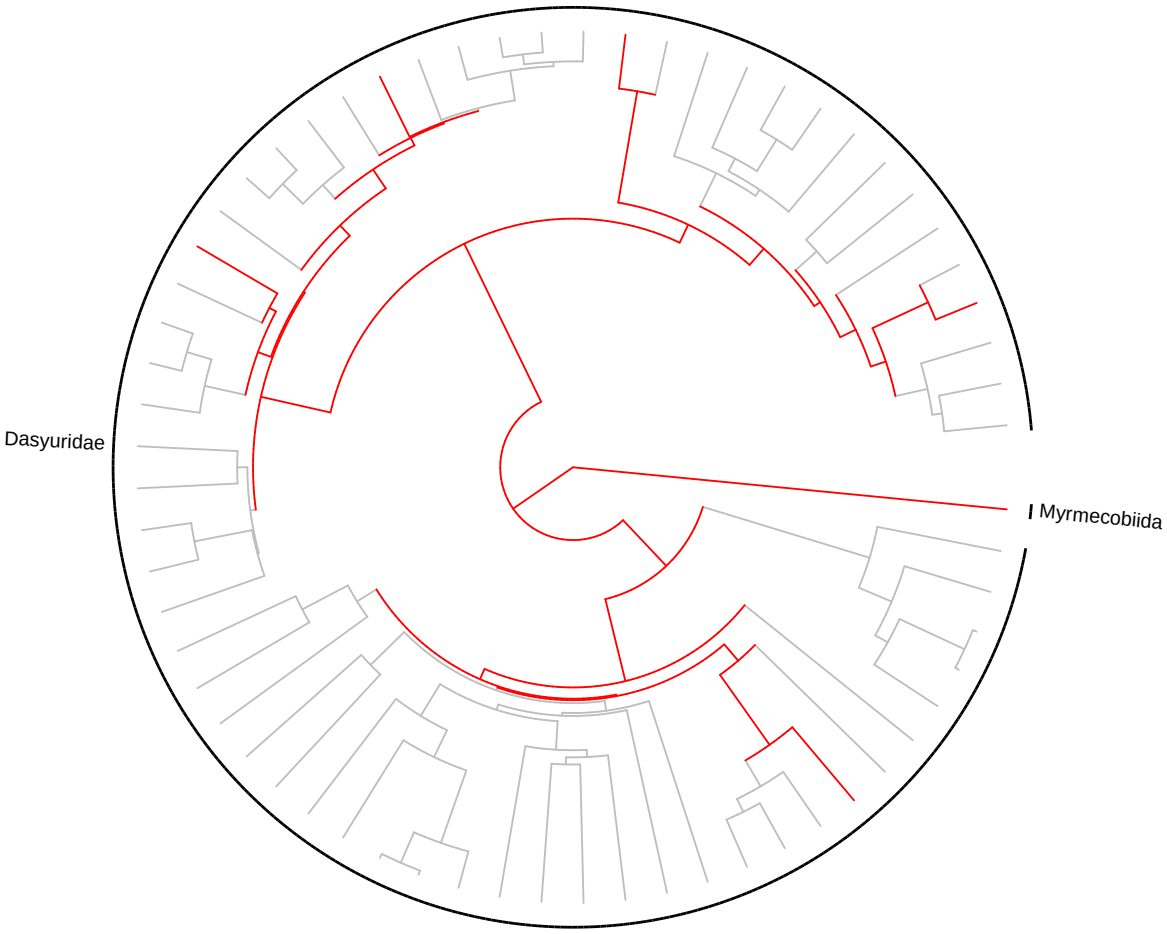

Figure S6: Distribution of available morphological data across Dasyuromorphia. Edges are colored in grey when no morphological data is available or in red when data is available.

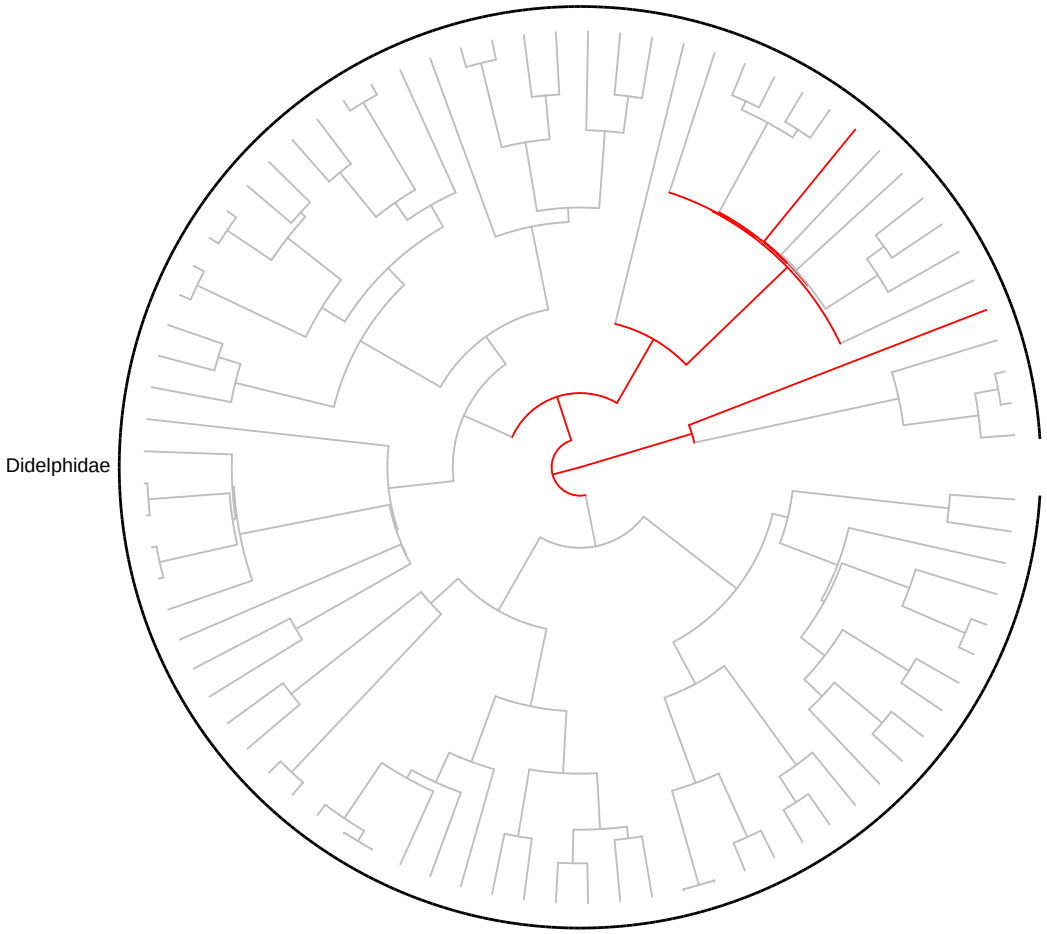

Figure S7: Distribution of available morphological data across Didelphimorphia. Edges are colored in grey when no morphological data is available or in red when data is available.

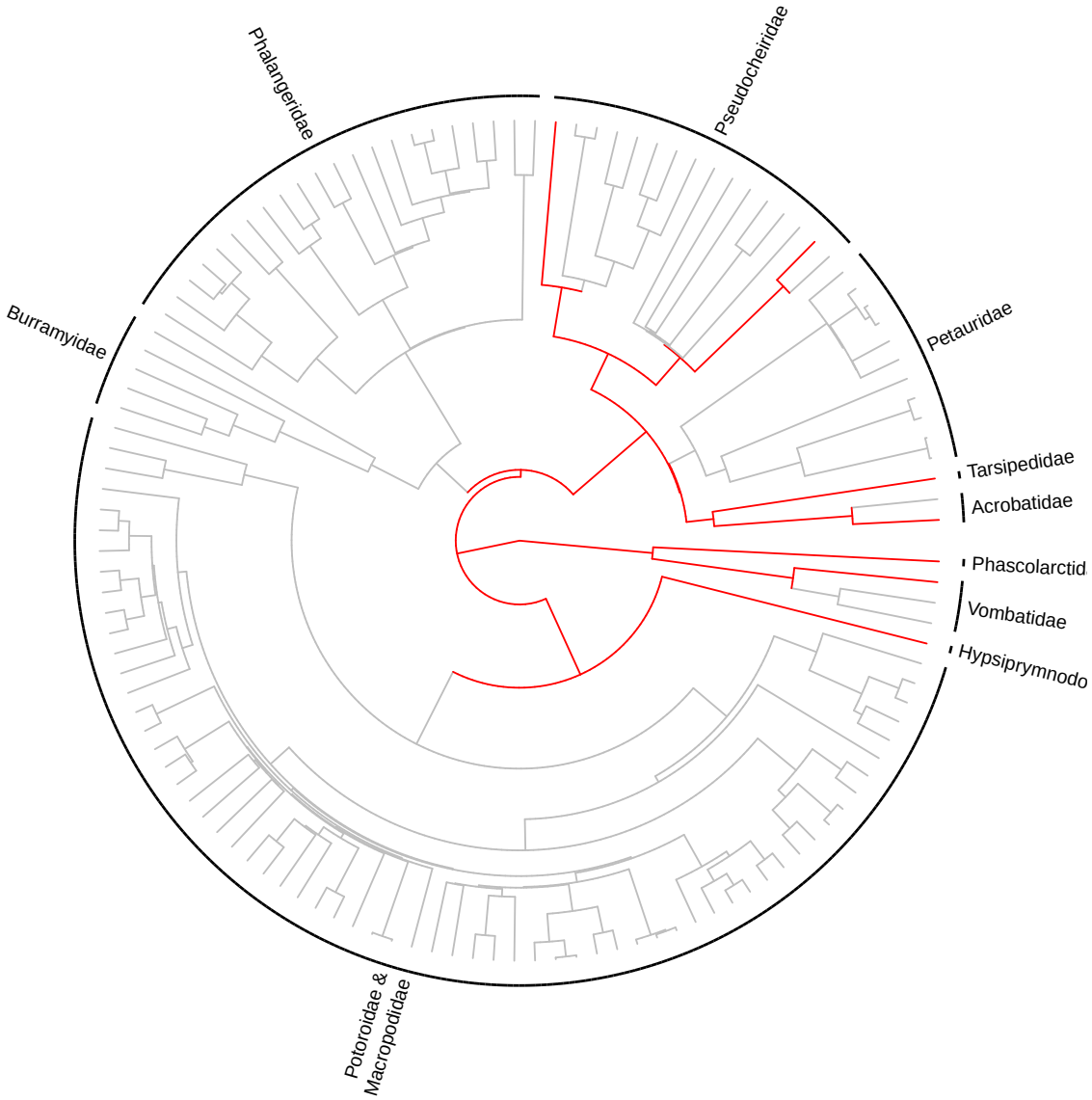

Figure S8: Distribution of available morphological data across Diprotodontia. Edges are colored in grey when no morphological data is available or in red when data is available.

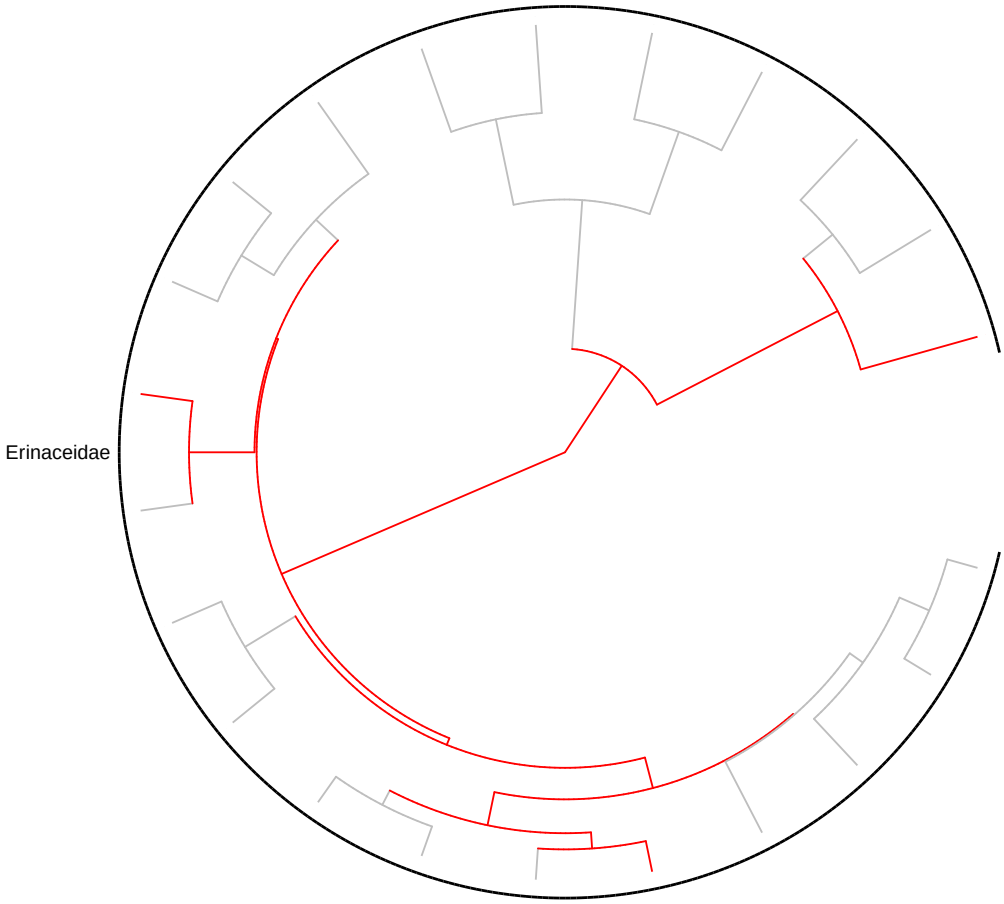

Figure S9: Distribution of available morphological data across Erinaceomorpha. Edges are colored in grey when no morphological data is available or in red when data is available.

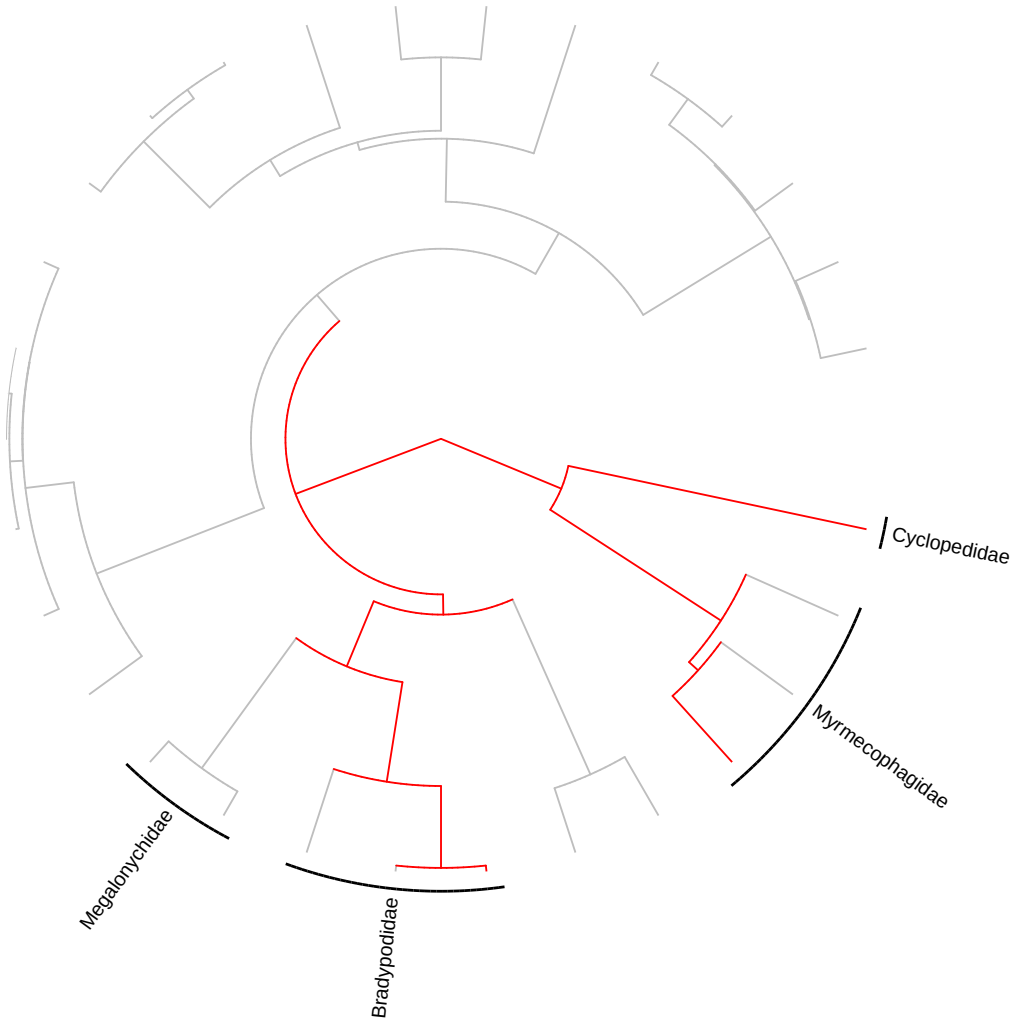

Figure S10: Distribution of available morphological data across Pilosa. Edges are colored in grey when no morphological data is available or in red when data is available.

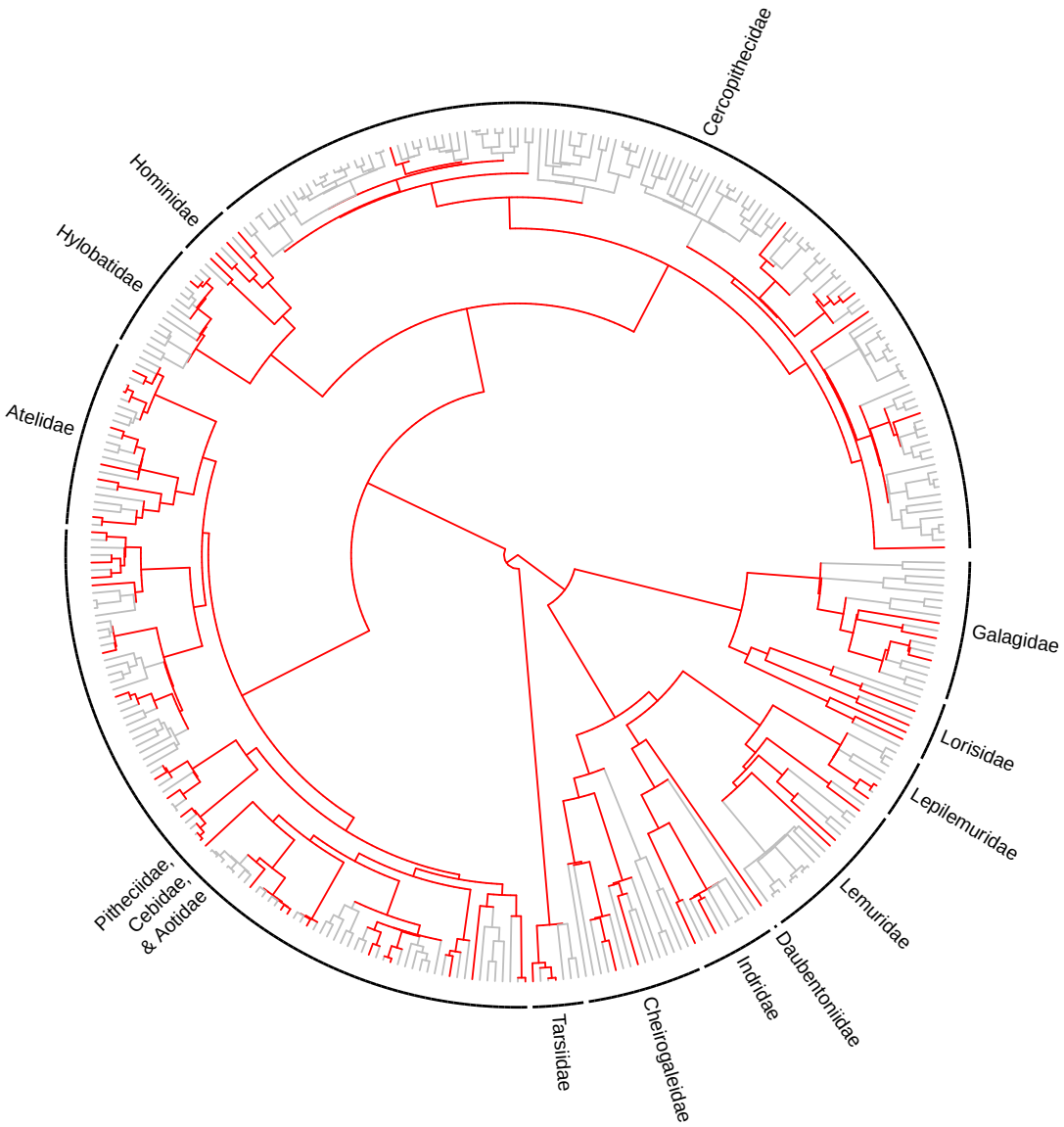

Figure S11: Distribution of available morphological data across Primates. Edges are colored in grey when no morphological data is available or in red when data is available.

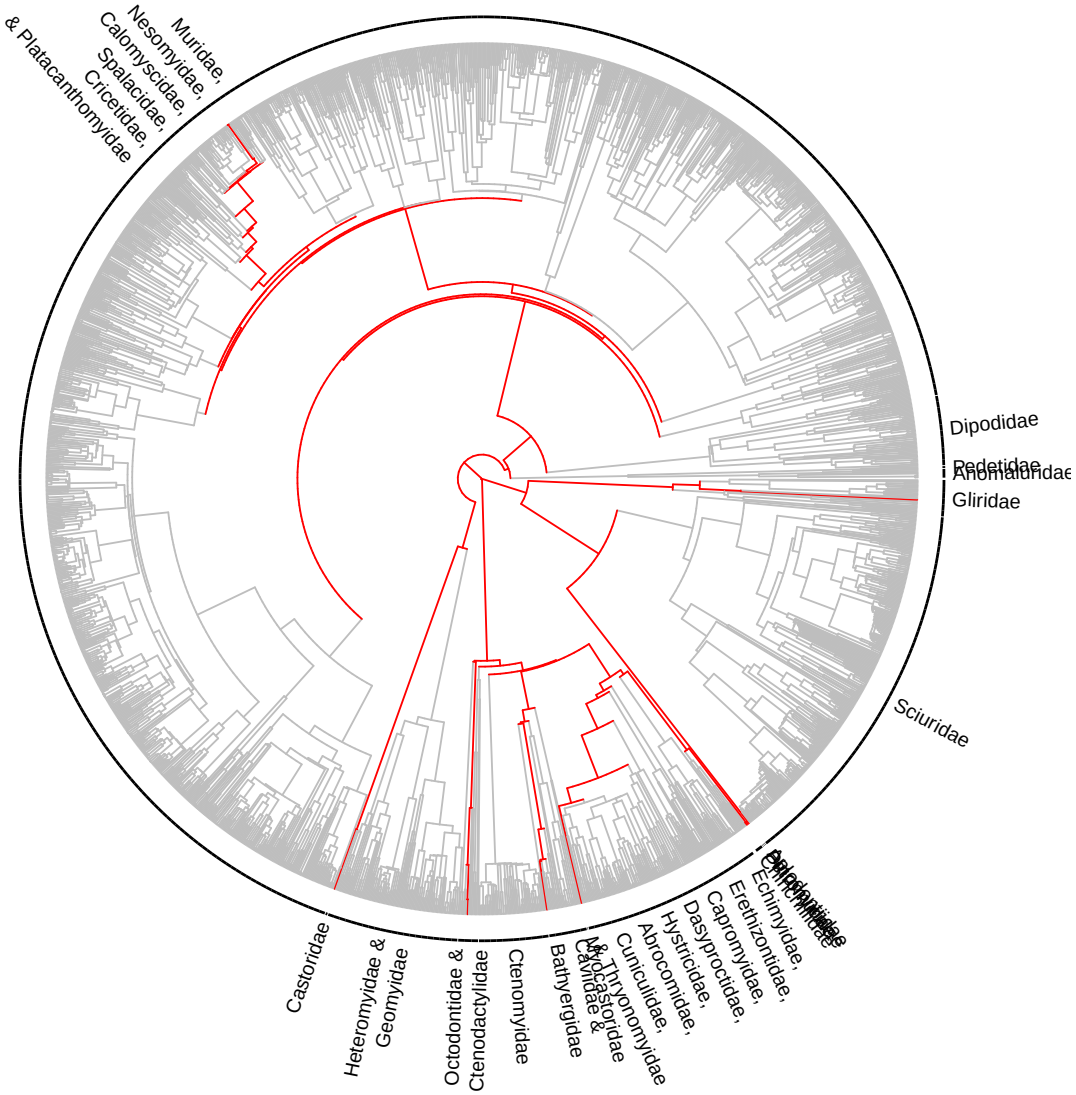

Figure S12: Distribution of available morphological data across Rodentia. Edges are colored in grey when no morphological data is available or in red when data is available.

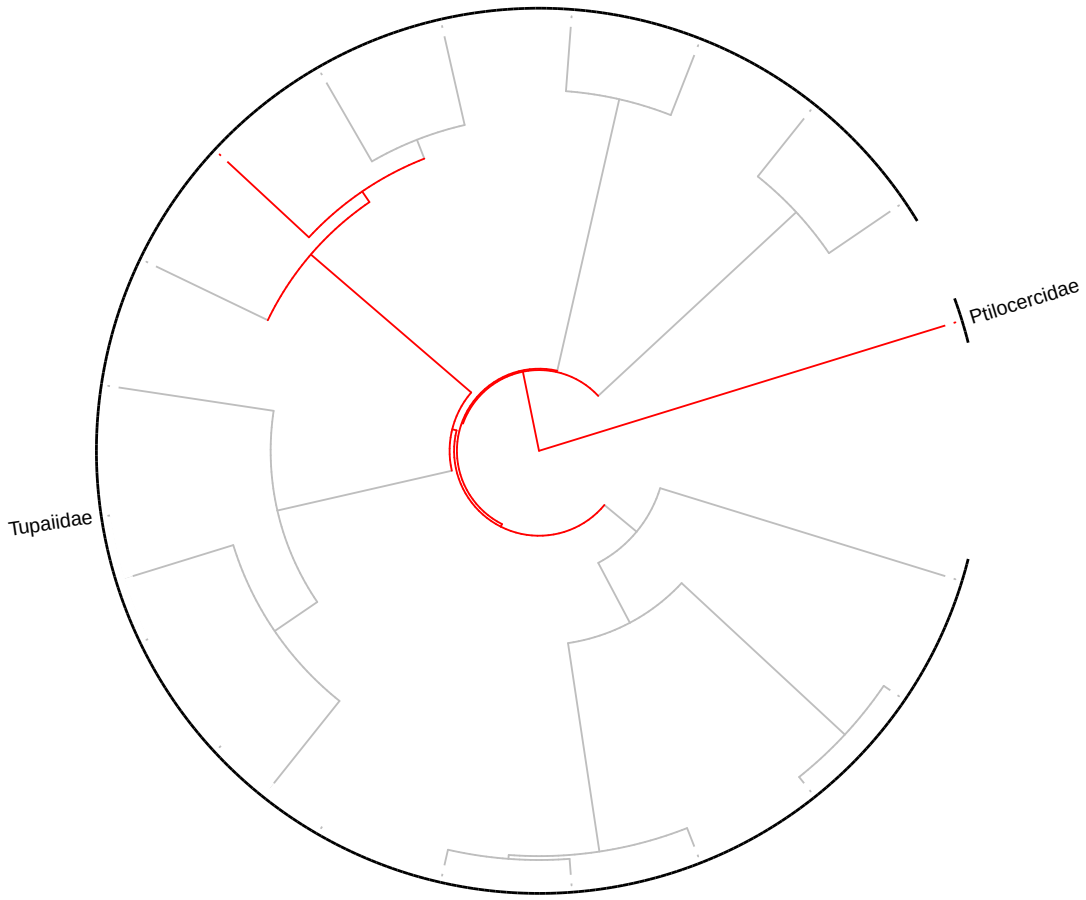

Figure S13: Distribution of available morphological data across Scandentia. Edges are colored in grey when no morphological data is available or in red when data is available.

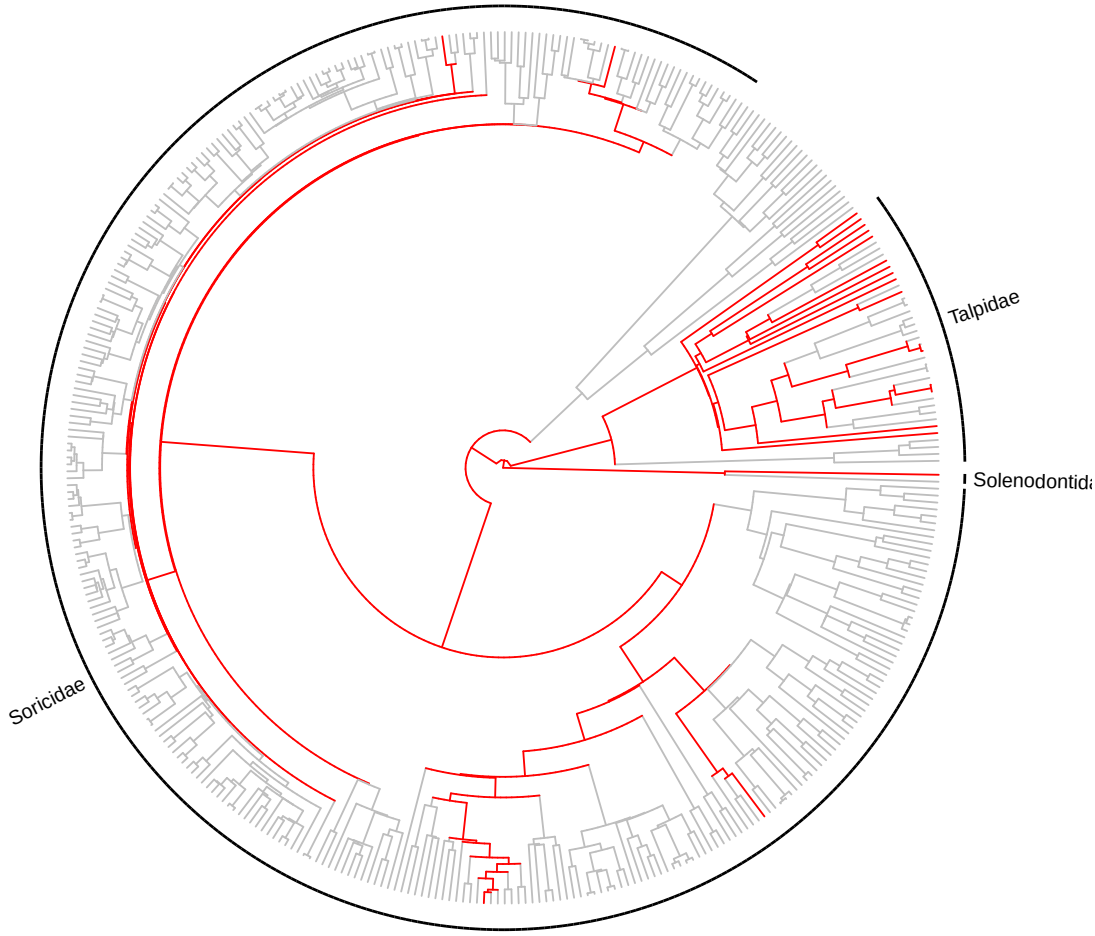

Figure S14: Distribution of available morphological data across Soricomorpha. Edges are colored in grey when no morphological data is available or in red when data is available.