Version dated: June 25, 2015

Assessment of cladistic data availability for living mammals

SUPPLEMENTARY MATERIAL

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

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

[2] *Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK.*

***Corresponding author.** *guillert@tcd.ie*

1 - Data collection

Public repositories

We downloaded available matrices containing fossil and/or living mammal taxa from the three 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 unique living OTUs and that if some where present in more than one matrix, they still only counted as one single OTU.

Morphobank.— We accessed the Morphobank repository (`http://www.morphobank.org/`) on the 10th of June 2015 and used the keywords listed above in the search menu. 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 10th of June 2015 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 11th of June 2015 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 (accessed 11th of June 2015)

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 downloaded the additional cladistic matrices from the 20 first search results matching with our selected keywords and with any of the 35 taxonomic levels (mammals Orders, Infraclasses and Class). 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 1). 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 [1] is reused with the same living species in [2], [3], [4], [5], [5], [6], [7], [8], [9], [10], [11], [12] and [13].

The list of all the 286 download matrices is available on The matrices contained a total of 11010 operational taxonomic units (OTUs) of which 5228 were unique. In this study, we refer to OTUs rather than species since the entries in the downloaded matrices were not standardised and ranged from specific individual specimen names (i.e. the name of a collection item) to the family-level. Where possible, we considered OTUs at their lowest valid taxonomic level (i.e. species) but some OTUs were only valid at a higher taxonomic level (e.g. genus or family). Therefore for some orders, we sampled more genera than species.

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

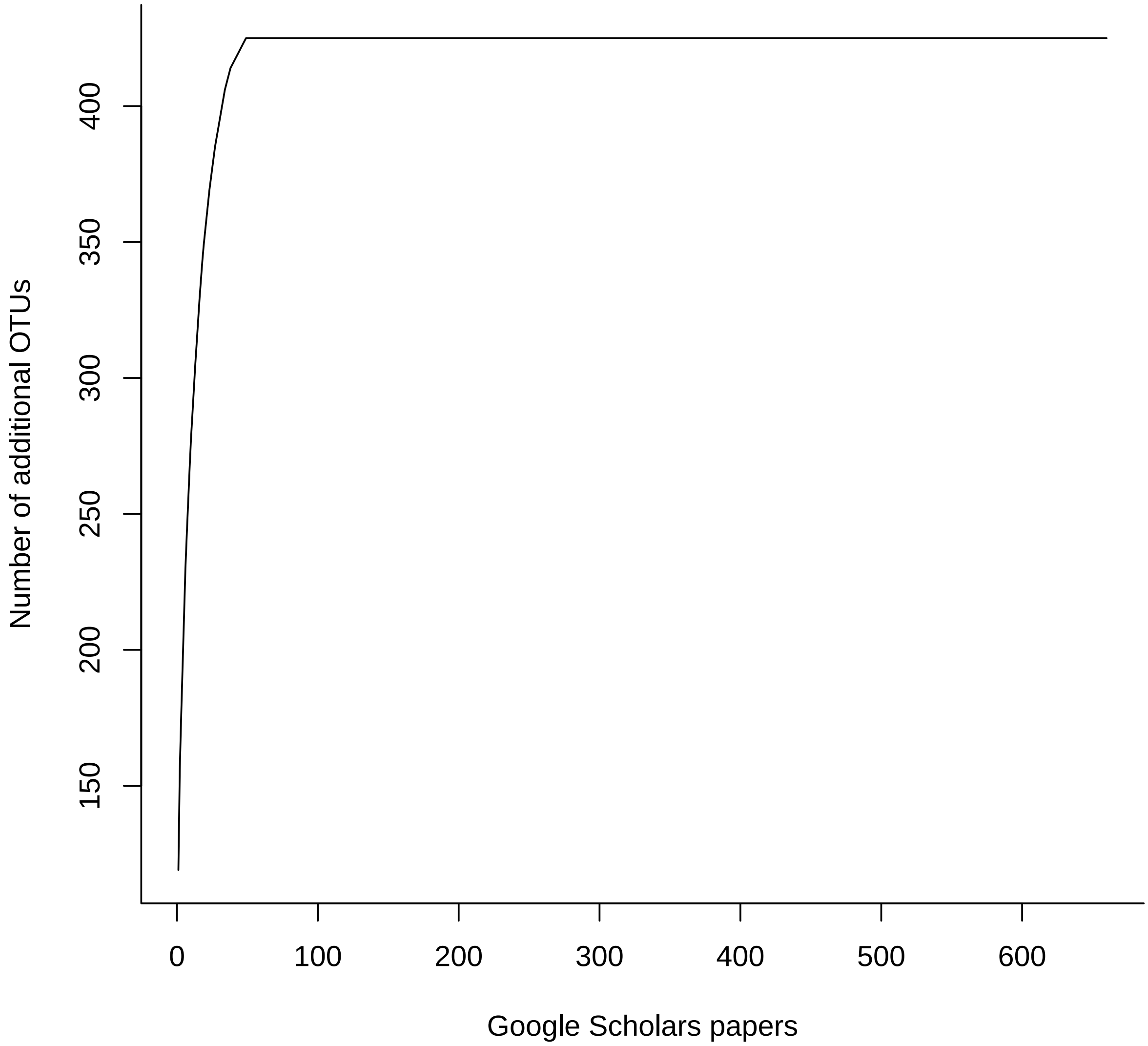


Figure 1: Google searches additional OTUs rarefaction curve. The x axis represent the number of google scholar matches (papers, books or abstracts) and the y axis represents the cumulative number of additional living OTUs per google scholar match.

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. All the standardised matrices are available on

Selecting the living OTUs

To select the lowest valid taxonomic level for each OTU, we standardised their taxonomy by correcting species names so they matched standard taxonomic nomenclature (e.g., *H. sapiens* was transformed to *Homo sapiens*). We designated as "living" all OTUs that were either present in the phylogeny of [14] or the taxonomy of [15], and designated as "fossil" all OTUs that were present in the Paleobiology database (`https://paleobiodb.org/`). For OTUs that did not appear in these three sources, we first decomposed the name (i.e. *Homo sapiens* became *Homo* and *sapiens*) and tried to match the first element with a higher taxonomic level (family, genus etc.). Any OTUs that still had no matches in the sources above were designated as non-applicable (NA; Figure 2).

2 - Data reproducibility

Every step of the analysis (apart from downloading and standardisation of the matrices) is entirely repeatable via GitHub (`https://github.com/TGuillerme/Missing_living_mammals`).

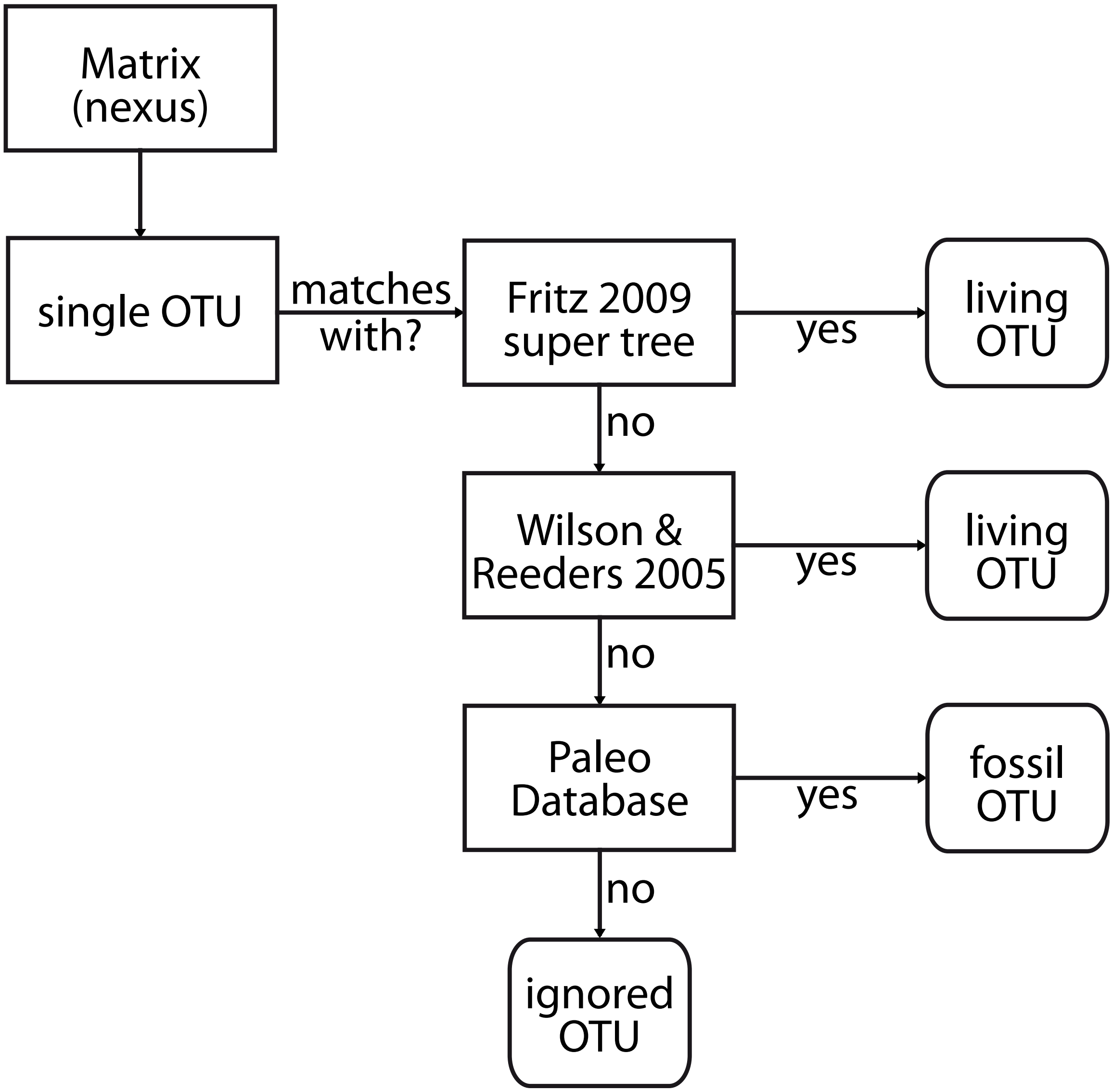


Figure 2: Taxonomic matching algorithm used in this study. For each matrix, each operational taxonomic units (OTU) is matched with the super tree from Fritz 2009. If the OTU matches, then it is classified as living. Else it is matched with the Wilson & Reeders 2005 taxonomy list. If the OTU matches, 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.

*

References

[1] Ross C, Williams B, Kay RF. Phylogenetic analysis of anthropoid relationships. J Hum Evol. 1998;35(3):221–306.

[2] Seiffert ER, Simons EL, Attia Y. Fossil evidence for an ancient divergence of lorises and galagos. Nature. 2003;422(6930):421–424.

[3] Marivaux L, Antoine PO, Baqri SRH, Benammi M, Chaimanee Y, Crochet JY, et al. Anthropoid primates from the Oligocene of Pakistan (Bugti Hills): data on early anthropoid evolution and biogeography. Proc Nat Acad Sci. 2005;102(24):8436–8441.

[4] Seiffert ER, Simons EL, Clyde WC, Rossie JB, Attia Y, Bown TM, et al. Basal anthropoids from Egypt and the antiquity of Africa's higher primate radiation. Science. 2005;310(5746):300–304.

[5] Bloch JI, Silcox MT, Boyer DM, Sargis EJ. New Paleocene skeletons and the relationship of plesiadapiforms to crown-clade primates. Proc Nat Acad Sci. 2007;104(4):1159–1164.

[6] Kay RF, Fleagle J, Mitchell T, Colbert M, Bown T, Powers DW. The anatomy of Dolichocebus gaimanensis, a stem platyrrhine monkey from Argentina. J Hum Evol. 2008;54(3):323–382.

[7] Silcox MT. The biogeographic origins of Primates and Euprimates: east, west, north, or south of Eden? In: Mammalian Evolutionary Morphology. Springer; 2008. p. 199–231.

[8] Seiffert ER, Perry JM, Simons EL, Boyer DM. Convergent evolution of anthropoid-like adaptations in Eocene adapiform primates. Nature. 2009;461(7267):1118–1121.

[9] Tabuce R, Marivaux L, Lebrun R, Adaci M, Bensalah M, Fabre PH, et al. Anthropoid versus strepsirhine status of the African Eocene primates Algeripithecus and Azibius: craniodental evidence. P Roy Soc B-Biol Scis. 2009;p. rspb20091339.

[10] Boyer DM, Seiffert ER, Simons EL. Astragalar morphology of Afradapis, a large adapiform primate from the earliest late Eocene of Egypt. Am J Phys Anthropol. 2010;143(3):383–402.

[11] Seiffert ER, Simons EL, Boyer DM, Perry JM, Ryan TM, Sallam HM. A fossil primate of uncertain affinities from the earliest late Eocene of Egypt. Proc Nat Acad Sci. 2010;107(21):9712–9717.

[12] Marivaux L, Ramdarshan A, Essid EM, Marzougui W, Ammar HK, Lebrun R, et al. Djebelemur, a tiny pre-tooth-combed primate from the Eocene of Tunisia: a glimpse into the origin of crown strepsirhines. PloS ONE. 2013;8(12):e80778.

[13] Ni X, Gebo DL, Dagosto M, Meng J, Tafforeau P, Flynn JJ, et al. The oldest known primate skeleton and early haplorhine evolution. Nature. 2013;498(7452):60–64.

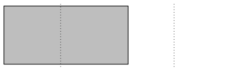
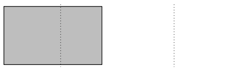
[14] Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, et al. The delayed rise of present-day mammals. Nature. 2007;446(7135):507–512. Available from: `http://dx.doi.org/10.1038/nature05634`.

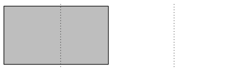
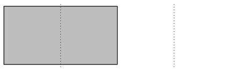
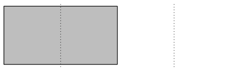
[15] Wilson DE, Reeder DM. Mammal species of the world: a taxonomic and geographic reference. vol. 1. JHU Press; 2005.

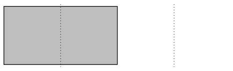
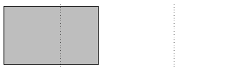
3 - Supplementary results

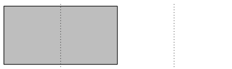
The following section contains supplementary results to the main body: the available data structure using the NTI and the PD metric; the proportion of available data and the data structure for all the matrices (including the matrices with less than 100 characters); and phylogenetical representation of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

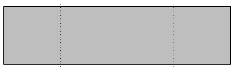
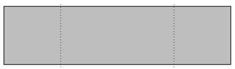
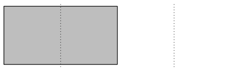
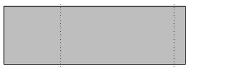
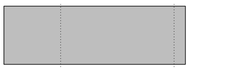
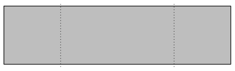
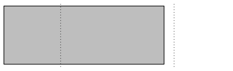
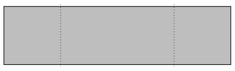
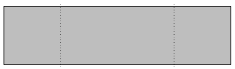
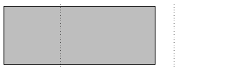
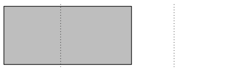
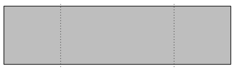
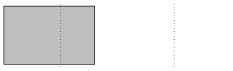
Table 1: Number of taxa with available cladistic data for mammalian orders at three taxonomic levels (without any character threshold; results from the 286 matrices). The coverage represents the proportion of taxa with available morphological data. The left vertical bar represents 25% of available data ("low" coverage if <25%); The right vertical bar represents 75% of available data ("high" coverage if >75%). When the Net Relatedness Index (NRI) and the Nearest Taxon Index (NTI) are negative, taxa are more phylogenetically dispersed than expected by chance; when NRI or NTI are positive, taxa are more phylogenetically clustered by expected by chance. Significant NRI or NTI are highlighted in bold. One star (*) represents a p-value between 0.05 and 0.005; two starts between 0.005 and 0.0005 and three stars a p-value less than 0.0005.

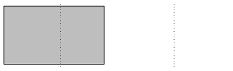
Order	Taxonomic level	Proportion of taxa	Coverage	NRI	NTI
Afrosoricida	family	2/2			
Afrosoricida	genus	17/17			
Afrosoricida	species	23/42		1.75	1.08
Carnivora	family	14/15		0.63	0.6
Carnivora	**genus**	**54/125**		**4.81****	**1.78***
Carnivora	**species**	**76/283**		**7.66****	**0.85**

Cetartiodactyla	family	21/21			
Cetartiodactyla	genus	100/128		0.85	0.94
Cetartiodactyla	**species**	**171/310**		**1.92***	**-0.46**
Chiroptera	family	15/18		-0.28	0.56
Chiroptera	**genus**	**93/202**		**13.47****	**1.1**
Chiroptera	**species**	**215/1053**		**8.82****	**1.22**
Cingulata	family	1/1			
Cingulata	genus	8/9		1.51	-1.57
Cingulata	**species**	**9/29**		**1.9***	**0.11**
Dasyuromorphia	family	2/2			
Dasyuromorphia	genus	8/22		-0.75	-1.07
Dasyuromorphia	species	9/64		-0.88	-0.34
Dermoptera	family	1/1			
Dermoptera	genus	1/2			
Dermoptera	species	1/2			
Didelphimorphia	family	1/1			
Didelphimorphia	genus	16/16			

Didelphimorphia	species	42/84		-1.65	0.2
Diprotodontia	family	11/11			
Diprotodontia	genus	25/38		-1.13	-1.31
Diprotodontia	species	31/126		0.48	-1.77
Erinaceomorpha	family	1/1			
Erinaceomorpha	genus	10/10			
Erinaceomorpha	species	21/22		-1.07	-0.2
Hyracoidea	family	1/1			
Hyracoidea	genus	1/3			
Hyracoidea	species	1/4			
Lagomorpha	family	2/2			
Lagomorpha	genus	5/12		-1.06	-0.95
Lagomorpha	species	12/86		-0.62	-1.88
Macroscelidea	family	1/1			
Macroscelidea	genus	4/4			
Macroscelidea	species	12/15		-1.3	-1.06
Microbiotheria	family	1/1			

Microbiotheria	genus	1/1			
Microbiotheria	species	1/1			
Monotremata	family	2/2			
Monotremata	genus	2/3		-0.72	-0.69
Monotremata	species	2/4		-0.97	-0.97
Notoryctemorphia	family	1/1			
Notoryctemorphia	genus	1/1			
Notoryctemorphia	species	0/2			
Paucituberculata	family	1/1			
Paucituberculata	genus	3/3			
Paucituberculata	species	5/5			
Peramelemorphia	family	2/2			
Peramelemorphia	genus	7/7			
Peramelemorphia	species	16/18		-0.13	0.97
Perissodactyla	family	3/3			
Perissodactyla	genus	6/6			
Perissodactyla	species	10/16		-0.07	-2.63

Pholidota	family	1/1			
Pholidota	genus	1/1			
Pholidota	species	4/8		1.18	0.94
Pilosa	family	4/5		1.87	2
Pilosa	genus	4/5		-0.96	0.36
Pilosa	**species**	**5/29**		**1.28**	**2.38***
Primates	family	15/15			
Primates	genus	48/68		-0.35	-1.33
Primates	species	64/351		-0.67	-1.27
Proboscidea	family	1/1			
Proboscidea	genus	2/2			
Proboscidea	species	2/3		-0.69	-0.69
Rodentia	family	18/32		0.66	-0.98
Rodentia	genus	82/450		-1.66	1.55
Rodentia	**species**	**90/2094**		**2.76***	**2.34***
Scandentia	family	2/2			
Scandentia	genus	2/5		-0.74	-0.74

Scandentia	species	3/20		-1.88	-0.84
Sirenia	family	2/2			
Sirenia	genus	2/2			
Sirenia	species	4/4			
Soricomorpha	family	3/4		-0.98	-0.99
Soricomorpha	**genus**	**19/43**		**7.11****	**2.59****
Soricomorpha	**species**	**21/392**		**10.65****	**3.56****
Tubulidentata	family	1/1			
Tubulidentata	genus	1/1			
Tubulidentata	species	1/1			

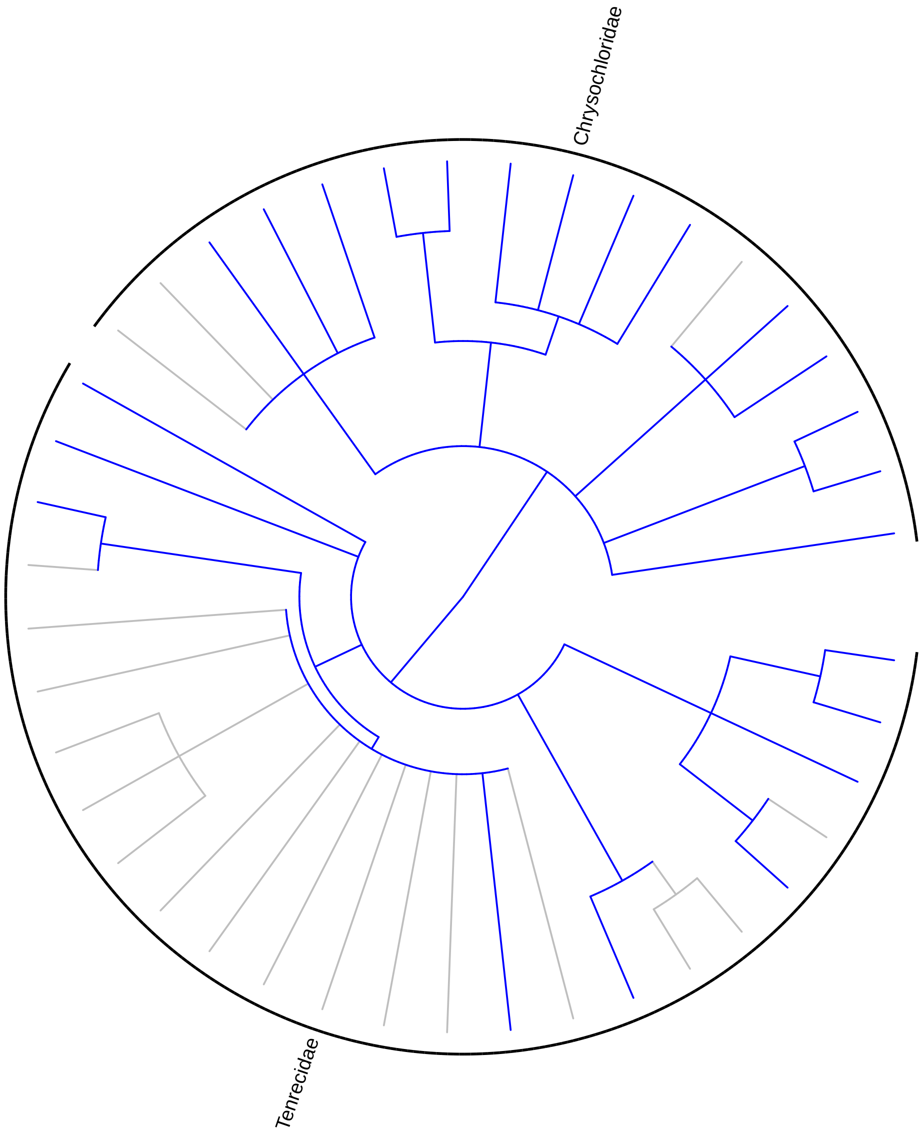


Figure 3: Distribution of available morphological data across Afrosoricida. Edges are colored in grey when no morphological data is available or in blue when data is available.

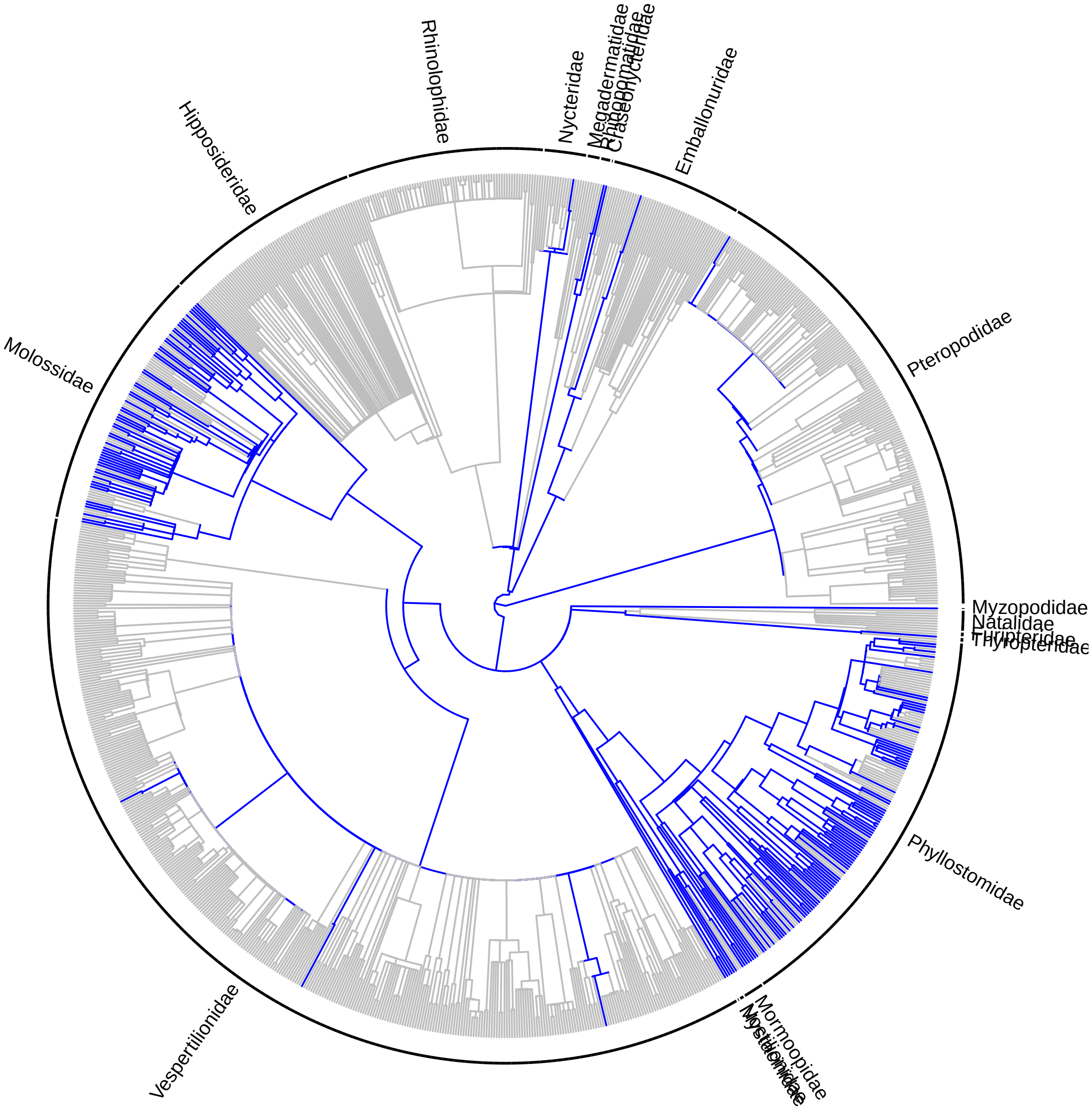


Figure 4: Distribution of available morphological data across Chiroptera. Edges are colored in grey when no morphological data is available or in blue when data is available.

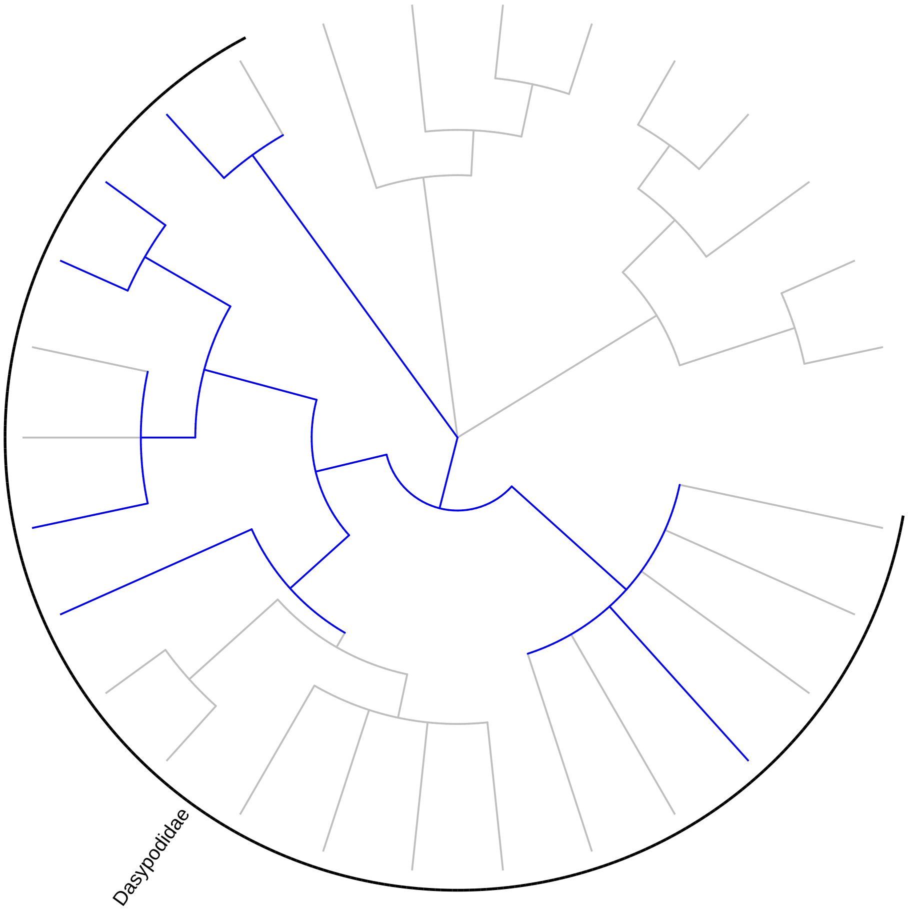


Figure 5: Distribution of available morphological data across Cingulata. Edges are colored in grey when no morphological data is available or in blue when data is available.

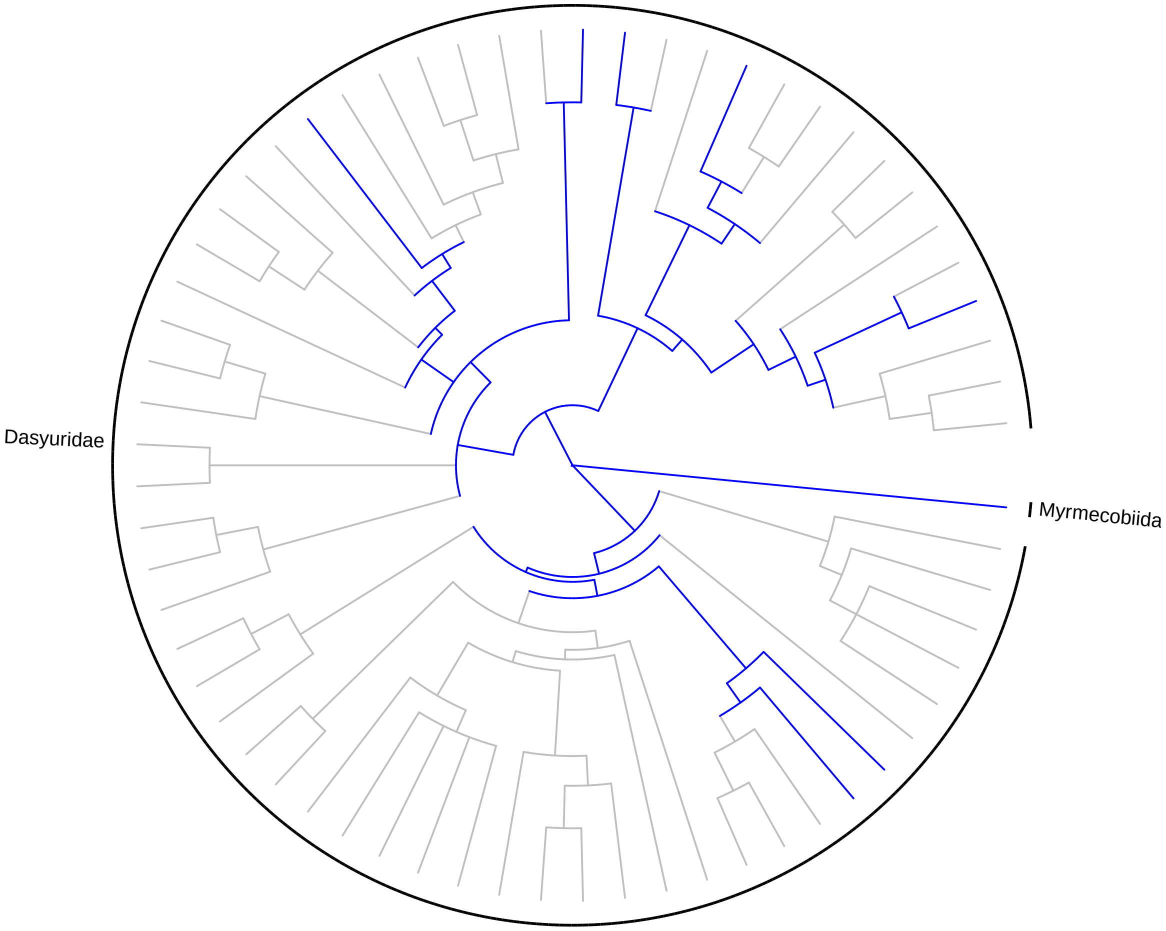


Figure 6: Distribution of available morphological data across Dasyuromorphia. Edges are colored in grey when no morphological data is available or in blue when data is available.

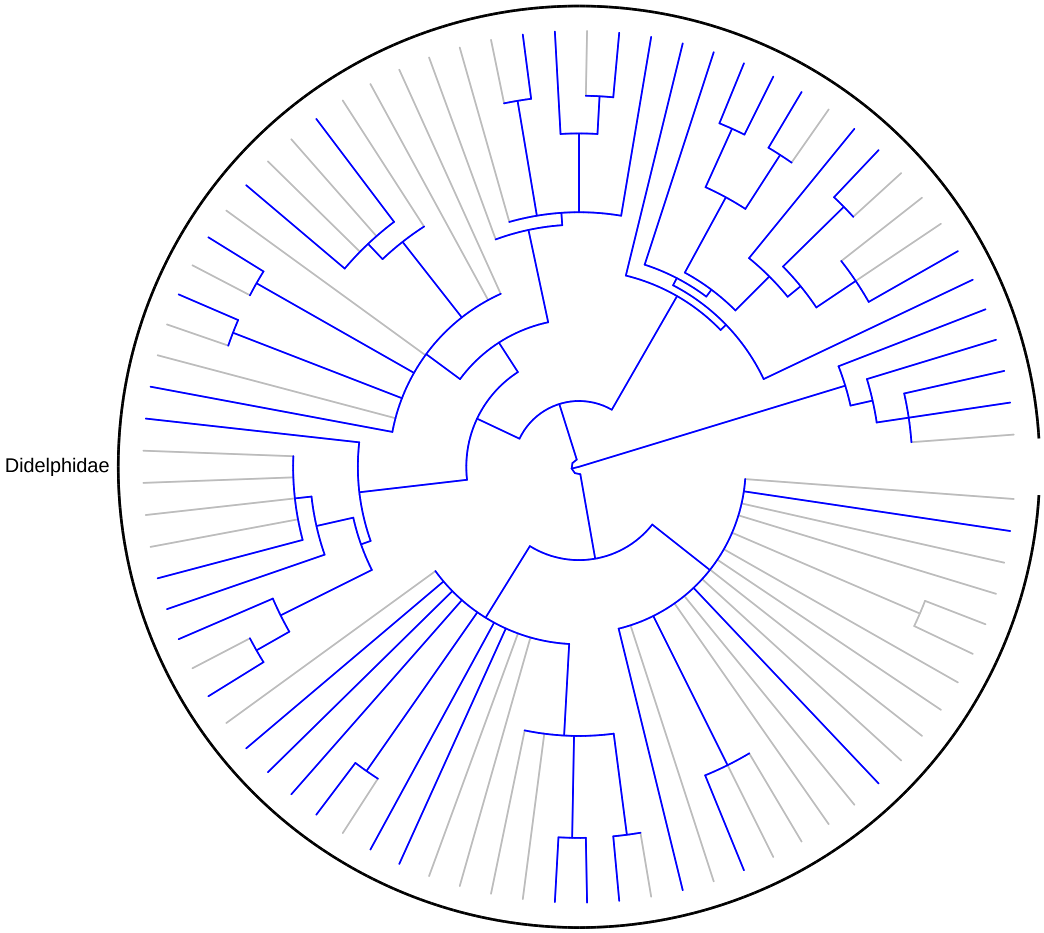


Figure 7: Distribution of available morphological data across Didelphimorphia. Edges are colored in grey when no morphological data is available or in blue when data is available.

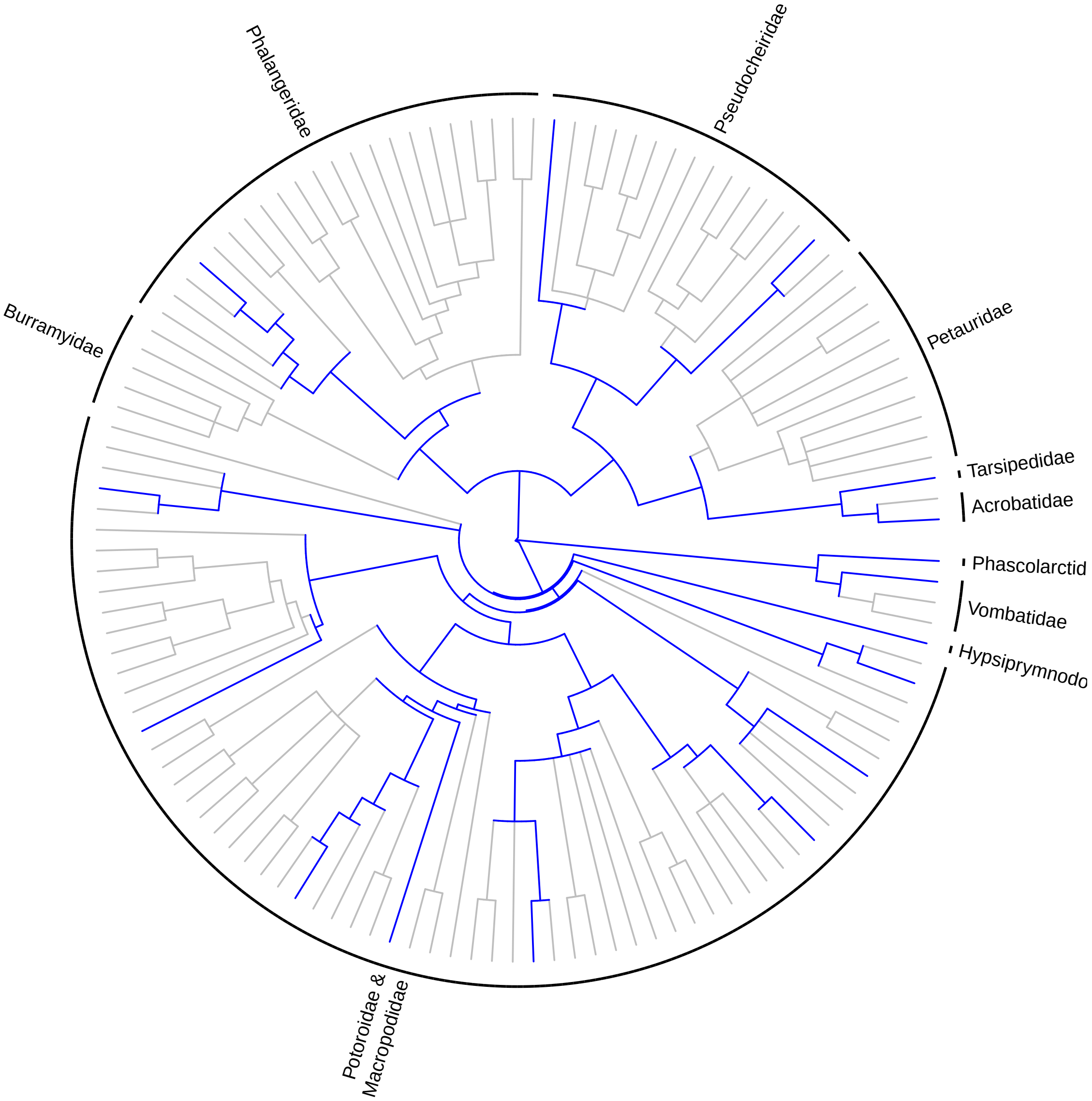


Figure 8: Distribution of available morphological data across Diprotodontia. Edges are colored in grey when no morphological data is available or in blue when data is available.

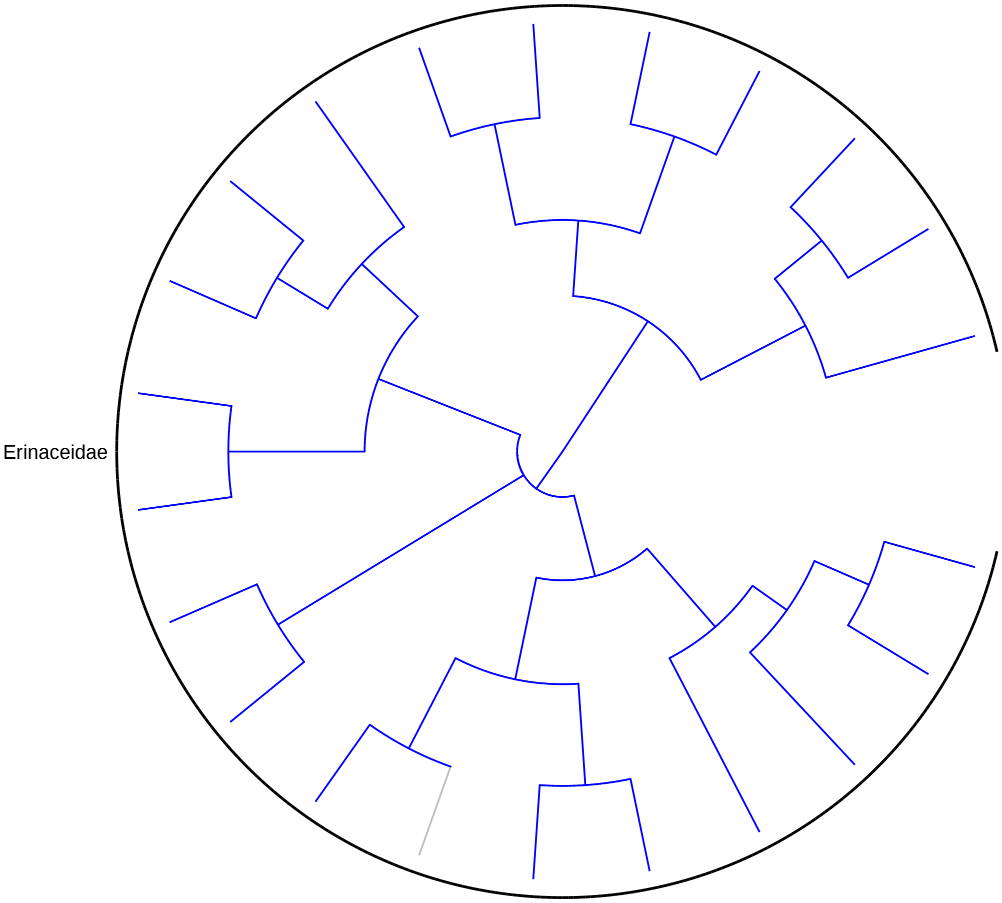


Figure 9: Distribution of available morphological data across Erinaceomorpha. Edges are colored in grey when no morphological data is available or in blue when data is available.

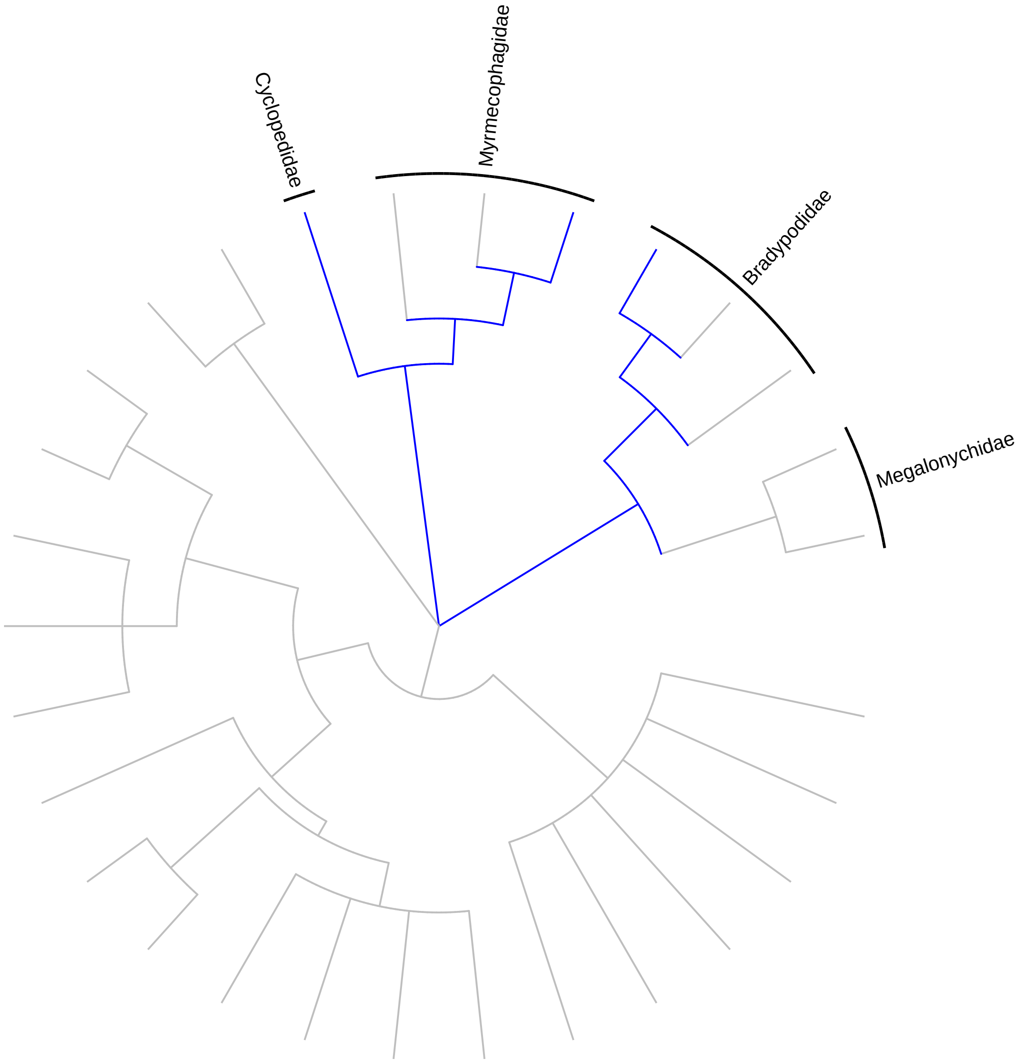


Figure 10: Distribution of available morphological data across Pilosa. Edges are colored in grey when no morphological data is available or in blue when data is available.

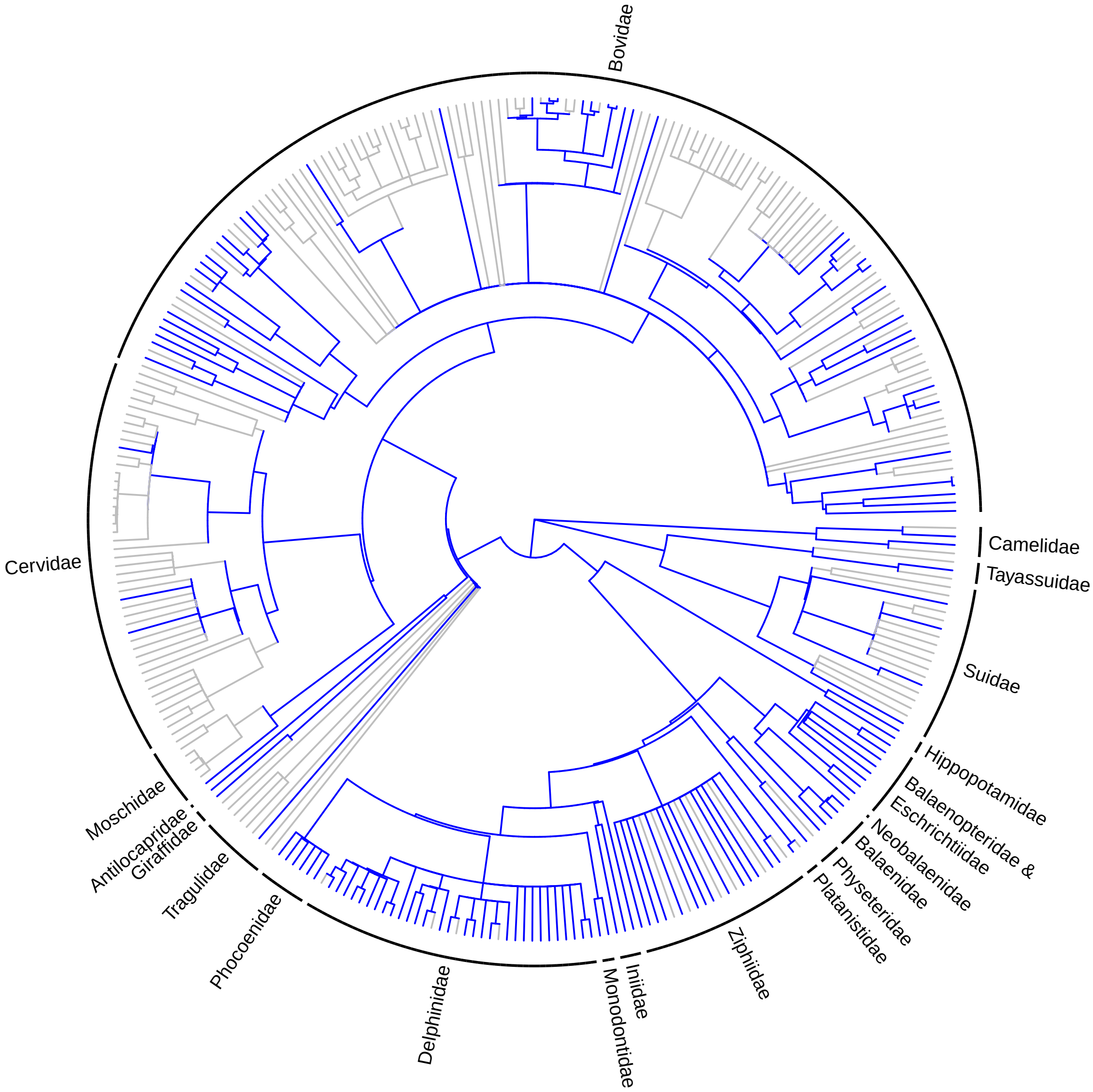


Figure 11: Distribution of available morphological data across Cetartiodactyla. Edges are colored in grey when no morphological data is available or in blue when data is available.

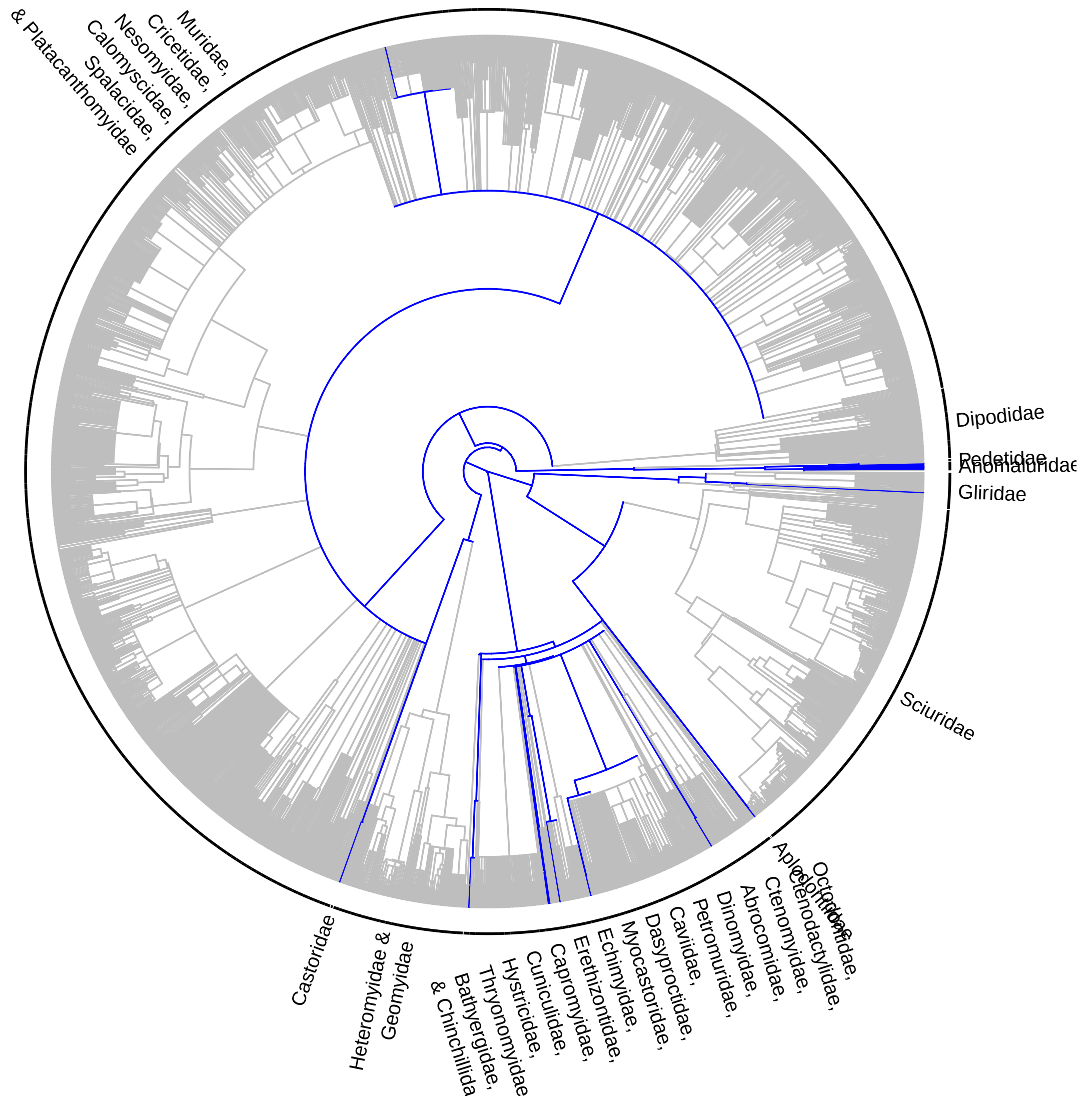


Figure 12: Distribution of available morphological data across Rodentia. Edges are colored in grey when no morphological data is available or in blue when data is available.

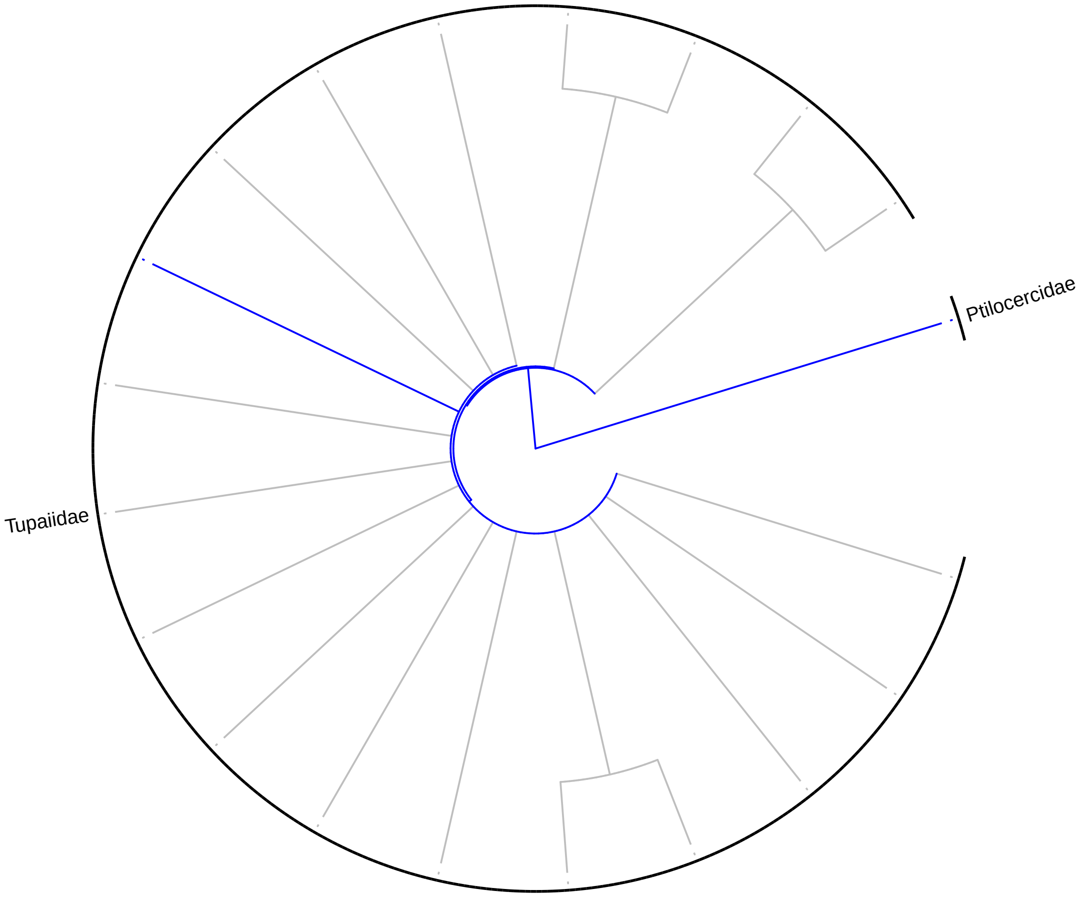


Figure 13: Distribution of available morphological data across Scandentia. Edges are colored in grey when no morphological data is available or in blue when data is available.

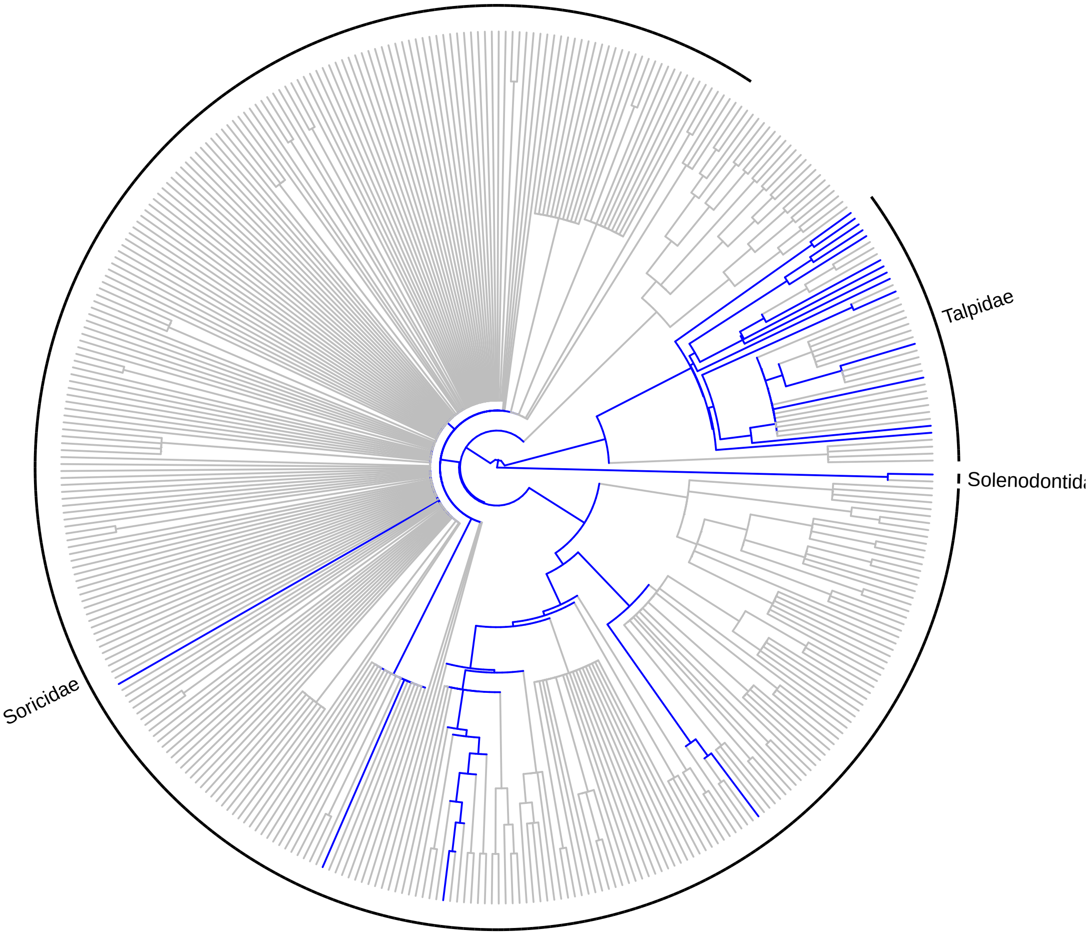


Figure 14: Distribution of available morphological data across Soricomorpha. Edges are colored in grey when no morphological data is available or in blue when data is available.