Version dated: February 25, 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

()

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 [1, 2, 3, 4, 5]. Combining the global biodiversity (living and fossil) in such studies allows for example, to improve accuracy of the timing of diversification events (e.g. [6]), to better understand relationships among lineages (e.g. [7]) or to infer biogeographical patterns through time (e.g. [8]). In order to do so, one must efficiently combine living and fossil taxa data in macroevolutionary models. One trending method, called the total evidence method, allows to combine molecular data from living taxa and morphological data from both living and fossil taxa in a supermatrix (e.g. [9, 6, 10, 4, 7, 8]). This method allows to use all the available data and to treat fossil as tips rather than nodes and allows integrative phylogenetic inference methods such as tip-dating [6, 11, 12].

However, the total evidence method requires a lot of data, both molecular and morphological for both living and fossil taxa. 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 when molecular data is available (e.g. [13] *vs.* [14]). Therefore, the ability of this method to recover correct topology is expected to decrease with missing data and especially when few morphological data overlaps between the living and the fossil taxa [15]. The effect of missing data is most important on topology when too few living taxa have available morphological data [15]. For example, if there is no morphological available for any living taxa of a clade, it is impossible to link a fossil 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 can 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 clearly does not belong to this clade (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 artefacts: (1) the impossibility for a fossil to branch in the right clade if there is no morphological data available in this clade; 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.

In this study, we assess the level of available data in living mammals in order to highlight the two caveats described above. We extracted living mammals taxa with available morphological data from 256 phylogenetic matrices available online and measured the proportion of morphological data availability for each mammalian order (1). Additionally, in the mammalian orders where data was missing, we estimated the structure of the available data to detect if the available data was biases towards certain

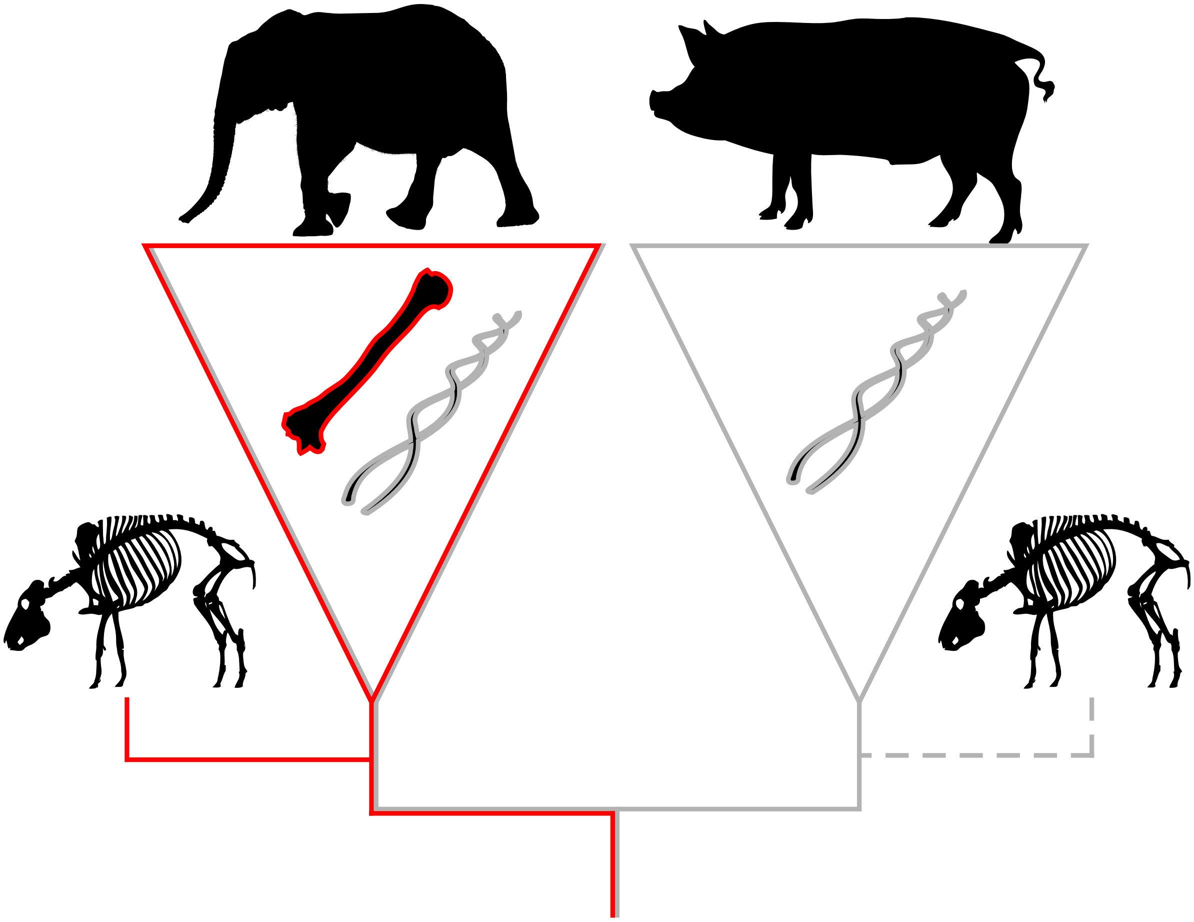

Figure 1: Example of topological errors due to missing morphological data in living taxa. If a phylogeny contains two clades, for example Proboscidea and Cetartiodactyla, with molecular data for both but only morphological data for Proboscidea. If an additional Cetartiodactyla fossil (with no molecular data) will be added to the phylogeny, it will erroneously branch with the Proboscidea clade instead of the Cetartiodactyla one.

clades only or if it was randomly distributed (2).

Material and Methods

Matrices search

To investigate the available living taxa with morphological data, we downloaded morphological matrices from three main public databases: morphobank (`http://www.morphobank.org/`), 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 data bases. 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 eventual additional morphological matrices from any of the 20 first papers matching with our selected key words 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).

We then transformed all the matrices to be in the same nexus format. We then 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 [16] or [17]. 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 to Wilson Reeders taxonomy at any taxonomic level (Family, Genus, *Genus_species*, etc.). 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 had at least 100 morphological characters. This arbitrary threshold of a minimal number of morphological characters was chosen to be in adequacy with [15] and [18]. Also this threshold avoids biases towards small matrices that could be either not informative (e.g. too few characters not allowing character overlap among matrices; see discussion) or made of non-applicable characters (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, Genera 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 [15] and therefore highly probable of suffering from the effect of missing data.

Available data structure.— For the orders with no morphological data for all its OTUs at the three different taxonomic levels (Family, Genera 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 [19]. 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 [20]. 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) [21, 22] but we also calculated the two other common phylogenetic structure indices: Faith's Phylogenetic Distance (PD) [23] and the Nearest Taxon Index (NTI) [20]. Both metrics are available in the Supplementary results.

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. After removing all the matrices with less than 100 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	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

Available data structure

Among the orders containing at least one OTU with no available morphological data, only two orders (Carnivora and Chiroptera) are significantly clustered both at the species and the genus level but not at the family 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	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

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

Discussion

We know we might have missed some data but we performed a thorough and

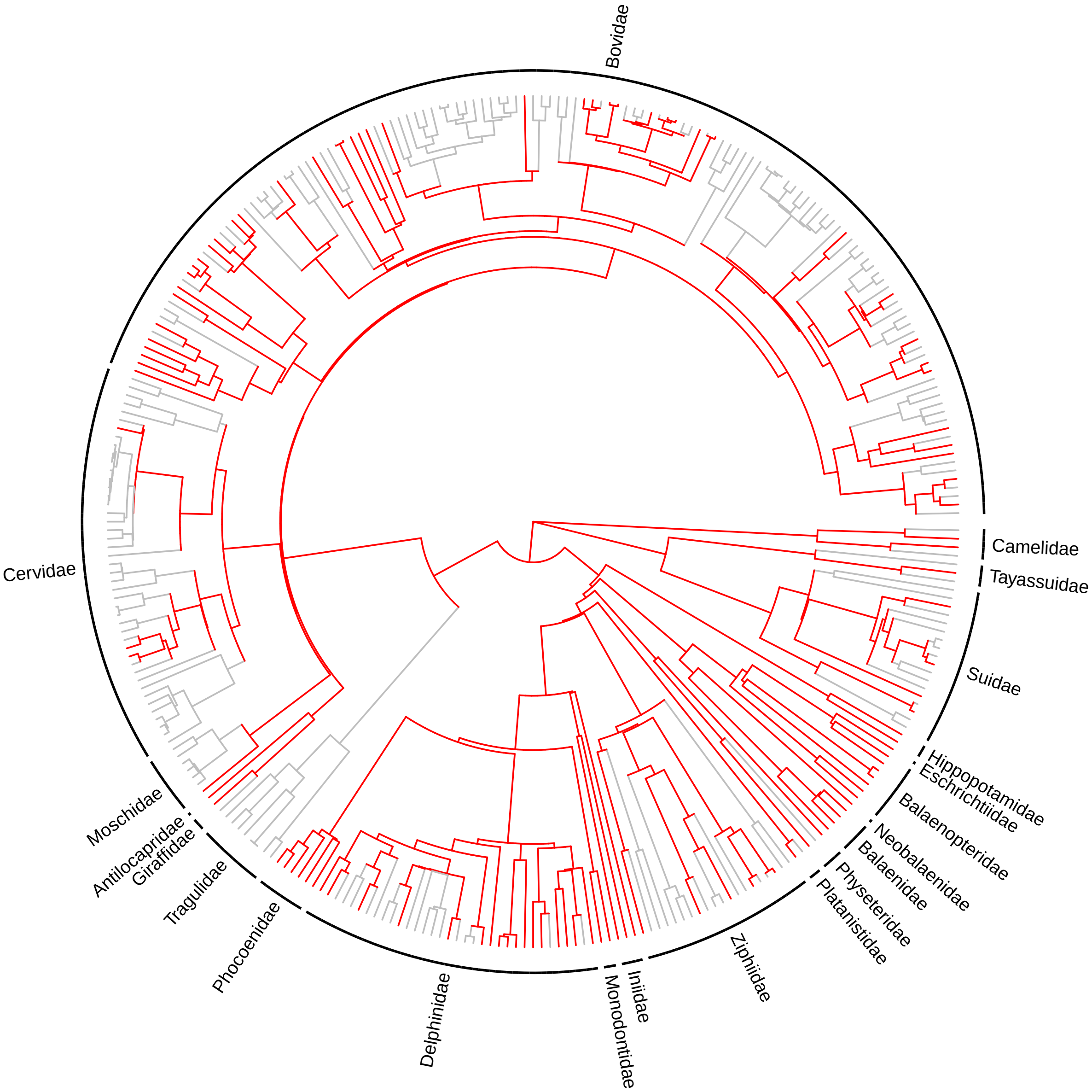

Figure 2: Distribution of available morphological data across Cetartiodactyla. Edges are colored in grey when no morphological data is available or in red when data is available.

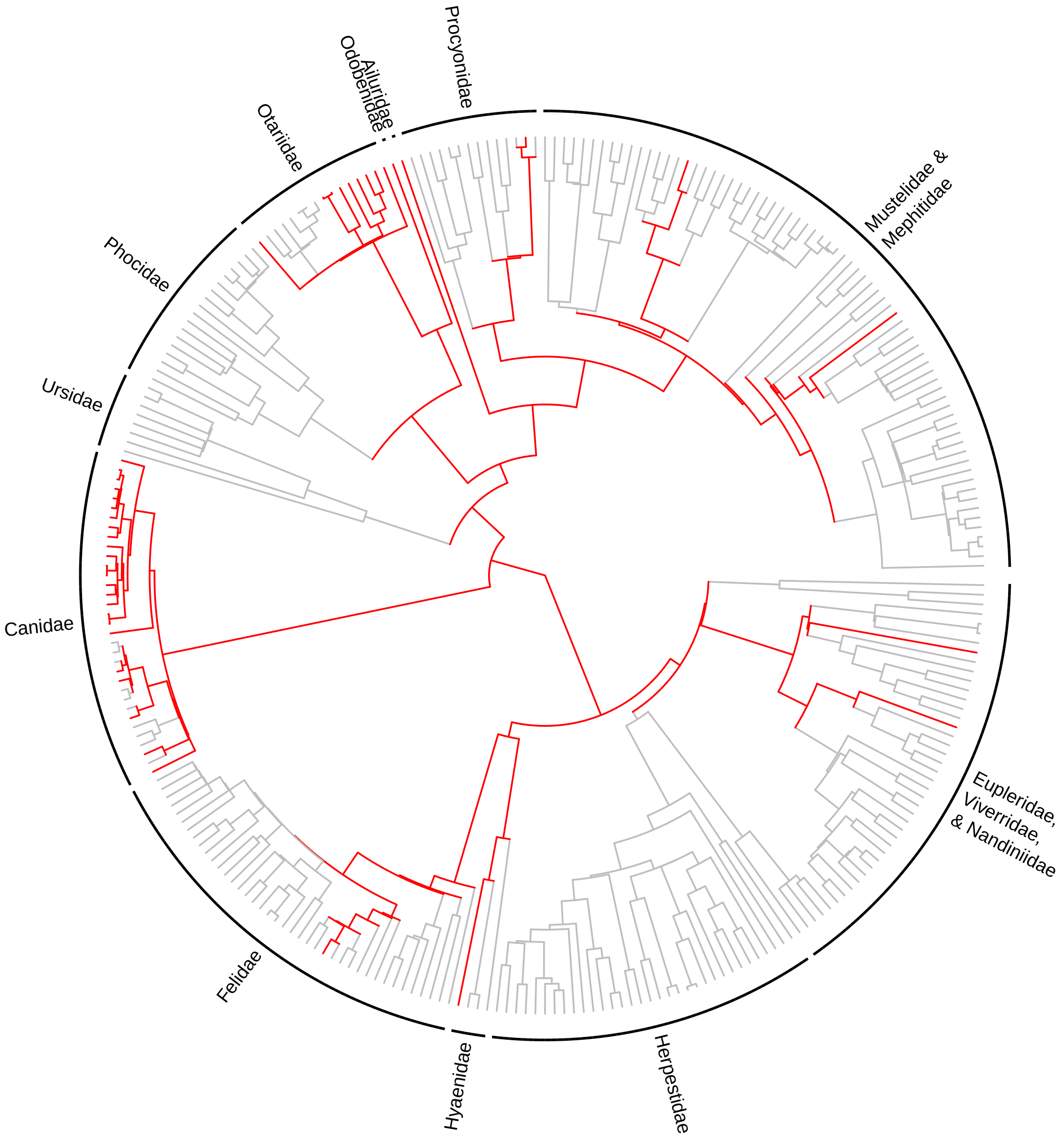

Figure 3: Distribution of available morphological data across Carnivora. Edges are colored in grey when no morphological data is available or in red when data is available.

repeatable search and still managed to collected 1422 OTUs which represents 38% of all living mammals

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) which can lead to topological misplacements of fossils in Total Evidence trees. The effect of missing living taxa with morphological data has no significant impact on inferring the correct topology up to 25% of missing data, beyond that threshold, both the position of ”flying” taxa and the clades are significantly different from the correct topology [15]. At the species level within mammals, it is therefore likely to see topological artefacts for the placement of fossil taxa. However, the data availability is more promising at higher taxonomic level (i.e. the Genus and the Family level). This value is probably more important to consider since there is a slight discrepancy between neontological species and paleontological species. While the second ones are described based on accurately measured morphology, genetic distance, spacial distribution and even behaviour, the second one are usually estimate based on morphological, spatial and temporal differences only. Because of the lack of behavioural and genetic data, it is likely that cryptic species are often ignored in palaeontological data. Therefore, most palaeontological studies are using the Genus level for their smallest taxonomic OTU. In this frame of palaeontological data usage, the data availablity at the Genus level in living mammals should be our priority concern.

However, it is encouraging that for most orders (except Carnivora and

Chiroptera), the few available data is randomly distributed, therefore the addition of fossil taxa is likely to 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 [15]. 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 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. Canidae in Carnivora; figure 3).

In this study we only used the raw number of characters and not there similarity. This might not work in reality since the data may actually not overlap. It might be that actually the amount of available data here is lower due to non overlap of characters....

Since the only way to fix missing data issues is to collect the missing data, it queit encouraging to know that the morphological data is available through natural history museum collections and that online plateforms and tools exist and are opperational to facilitate this giant task through collaboration and sharing open source and traceable data!

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

[1] Jackson J, Erwin D. What can we learn about ecology and evolution from the fossil record? Trends in Ecology and Evolution. 2006;21(6):322–328. Available from: http://dx.doi.org/10.1016/j.tree.2006.03.017.

[2] Quental T, Marshall C. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology and Evolution. 2010;25(8):434–441. Available from: http://dx.doi.org/10.1016/j.tree.2010.05.002.

[3] Dietl GP, Flessa KW. Conservation paleobiology: putting the dead to work. Trends in Ecology and Evolution. 2011;26(1):30–37. Available from: http://www.sciencedirect.com/science/article/pii/S0169534710002375.

[4] Slater GJ, Harmon LJ. Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. Methods in Ecology and Evolution. 2013;4(8):699–702. Available from: http://dx.doi.org/10.1111/2041-210X.12091.

[5] Fritz SA, Schnitzler J, Eronen JT, Hof C, Bhning-Gaese K, Graham CH. Diversity in time and space: wanted dead and alive. Trends in Ecology and Evolution. 2013;28(9):509 – 516. Available from: http://www.sciencedirect.com/science/article/pii/S0169534713001110.

[6] Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray D, Rasnitsyn A. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Systematic Biology. 2012;61(6):973–999. Available from: http://dx.doi.org/10.1093/sysbio/sys058.

[7] Beck RM, Lee MS. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. Proceedings of the Royal Society B: Biological

Sciences. 2014;281(20141278):1–10. Available from: http://dx.doi.org/10.1098/rspb.2014.1278.

[8] Meseguer AS, Lobo JM, Ree R, Beerling DJ, Sanmartn I. Integrating Fossils, Phylogenies, and Niche Models into Biogeography to Reveal Ancient Evolutionary History: The Case of Hypericum (Hypericaceae). Systematic Biology. 2015;64(2):215–232. Available from: http://sysbio.oxfordjournals.org/content/64/2/215.abstract.

[9] Pyron R. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology. 2011;60(4):466–481. Available from: http://dx.doi.org/10.1093/sysbio/syr047.

[10] Schrago C, Mello B, Soares A. Combining fossil and molecular data to date the diversification of New World Primates. Journal of Evolutionary Biology. 2013;26(11):2438–2446. Available from: http://dx.doi.org/10.1111/jeb.12237.

[11] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution. 2012;29(8):1969–1973. Available from: http://mbe.oxfordjournals.org/content/29/8/1969.abstract.

[12] Matzke NJ. BEASTmasteR: Automated conversion of NEXUS data to BEAST2 XML format, for fossil tip-dating and other uses.; 2014.

http://phylo.wikidot.com/beastmaster. Available from: http://phylo.wikidot.com/beastmaster.

[13] O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, et al. The placental mammal ancestor and the postK-Pg radiation of placentals. Science. 2013;339(6120):662–667. Available from: http://www.sciencemag.org/content/339/6120/662.abstract.

[14] Meredith R, Janečka J, Gatesy J, Ryder O, Fisher C, Teeling E, et al. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science. 2011;334(6055):521–524. Available from: http://dx.doi.org/10.1126/science.1211028.

[15] Effects of missing data on topological inference using a Total Evidence approach, author=Guillerme, Thomas and Cooper, Natalie, journal=in review, year=2015,;.

[16] Fritz SA, Bininda-Emonds ORP, Purvis A. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. Ecology Letters. 2009;12(6):538–549. Available from: http://dx.doi.org/10.1111/j.1461-0248.2009.01307.x.

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

[18] Harrison B Luke, Larsson CE Hans. Among-Character Rate Variation Distributions in Phylogenetic Analysis of Discrete Morphological Characters. Systematic biology. 2014;Available from: http://dx.doi.org/10.1093/sysbio/syu098.

[19] Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et al. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26:1463–1464.

[20] Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. Phylogenies and community ecology. Annual review of ecology and systematics. 2002;p. 475–505.

[21] Letcher SG. Phylogenetic structure of angiosperm communities during tropical forest succession. Proceedings of the Royal Society of London B: Biological Sciences. 2009;.

[22] Swenson NG. Phylogenetic Resolution and Quantifying the Phylogenetic Diversity and Dispersion of Communities. PLoS ONE. 2009 02;4(2):e4390. Available from: http://dx.doi.org/10.1371%2Fjournal.pone.0004390.

[23] Faith DP. Conservation evaluation and phylogenetic diversity. Biological Conservation. 1992;61(1):1 – 10. Available from: http://www.sciencedirect.com/science/article/pii/0006320792912013.

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 4). 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 [?] is reused with the same living species in [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?] and [?].

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) [16]; the taxonomic list from the Wilson and Reeder's Mammals Species of the World (2005) [17] 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

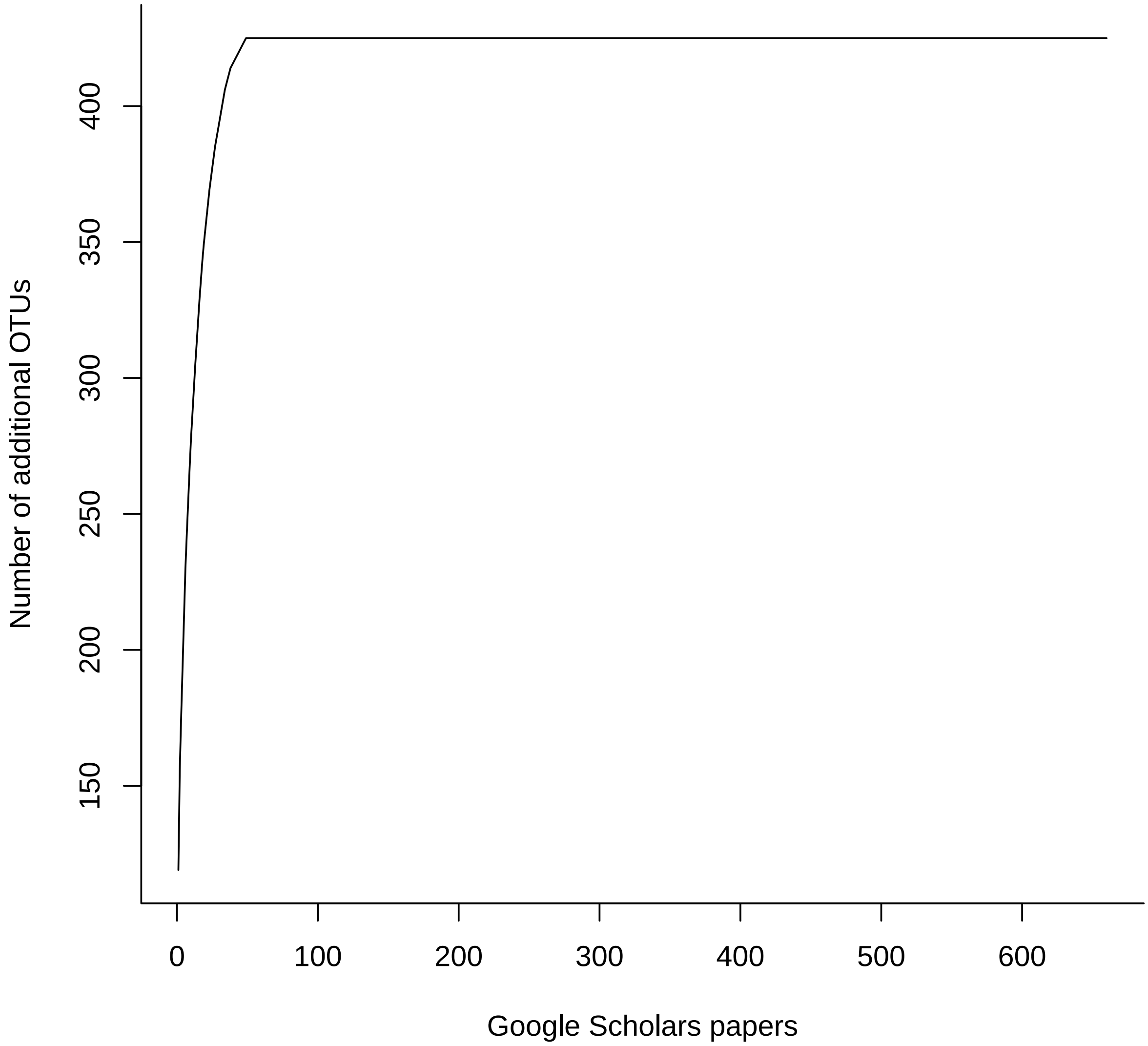

Figure 4: Google searches additional OTUs rarefaction curve. The x axis represent the number of google scholar matches (papers, books or abstracts) and the y axis represents the cumulative number of additional living OTUs per google scholar match.

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 5).

*

References

[1] Jackson J, Erwin D. What can we learn about ecology and evolution from the fossil record? Trends in Ecology and Evolution. 2006;21(6):322–328. Available from: `http://dx.doi.org/10.1016/j.tree.2006.03.017`.

[2] Quental T, Marshall C. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology and Evolution. 2010;25(8):434–441. Available from: `http://dx.doi.org/10.1016/j.tree.2010.05.002`.

[3] Dietl GP, Flessa KW. Conservation paleobiology: putting the dead to work. Trends in Ecology and Evolution. 2011;26(1):30–37. Available from: `http://www.sciencedirect.com/science/article/pii/S0169534710002375`.

[4] Slater GJ, Harmon LJ. Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. Methods in Ecology and Evolution. 2013;4(8):699–702. Available from: `http://dx.doi.org/10.1111/2041-210X.12091`.

[5] Fritz SA, Schnitzler J, Eronen JT, Hof C, Bhning-Gaese K, Graham CH. Diversity in time and space: wanted dead and alive. Trends in Ecology and Evolution.

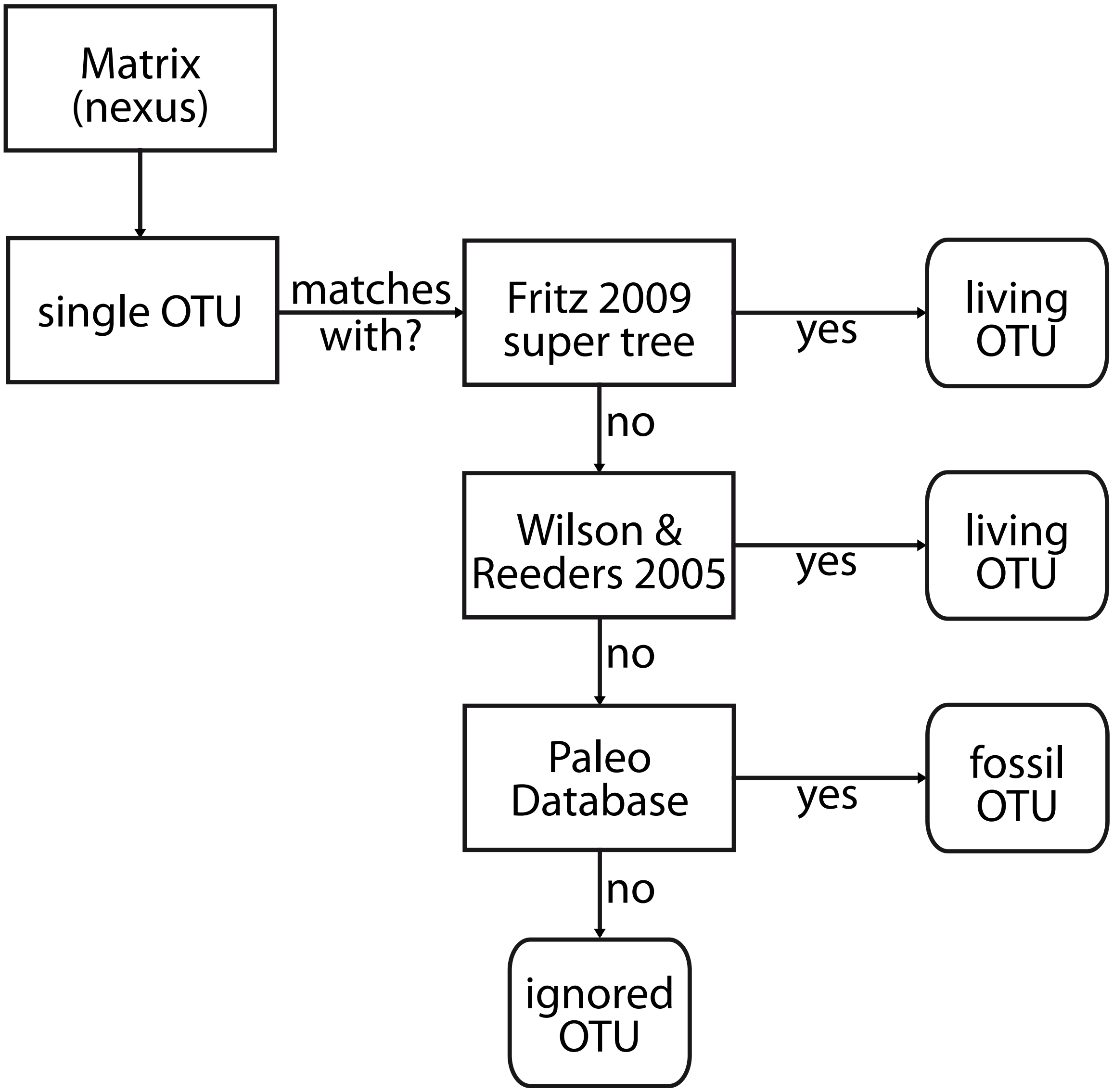

Figure 5: Taxonomic matching algorithm used in this study. For each matrix, each operational taxonomic units (OTU) is matched with the super tree from Fritz 2009. If the OTU matches, then it is classified as living. Else it is matched with the Wilson & Reeders 2005 taxonomy list. If the OTU matches, 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.

2013;28(9):509 – 516. Available from:
http://www.sciencedirect.com/science/article/pii/S0169534713001110.

[6] Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray D, Rasnitsyn A. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Systematic Biology. 2012;61(6):973–999. Available from:
http://dx.doi.org/10.1093/sysbio/sys058.

[7] Beck RM, Lee MS. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. Proceedings of the Royal Society B: Biological Sciences. 2014;281(20141278):1–10. Available from:
http://dx.doi.org/10.1098/rspb.2014.1278.

[8] Meseguer AS, Lobo JM, Ree R, Beerling DJ, Sanmartn I. Integrating Fossils, Phylogenies, and Niche Models into Biogeography to Reveal Ancient Evolutionary History: The Case of Hypericum (Hypericaceae). Systematic Biology. 2015;64(2):215–232. Available from:
http://sysbio.oxfordjournals.org/content/64/2/215.abstract.

[9] Pyron R. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology. 2011;60(4):466–481. Available from:
http://dx.doi.org/10.1093/sysbio/syr047.

[10] Schrago C, Mello B, Soares A. Combining fossil and molecular data to date the diversification of New World Primates. Journal of Evolutionary Biology.

2013;26(11):2438–2446. Available from: http://dx.doi.org/10.1111/jeb.12237.

[11] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution. 2012;29(8):1969–1973. Available from: http://mbe.oxfordjournals.org/content/29/8/1969.abstract.

[12] Matzke NJ. BEASTmasteR: Automated conversion of NEXUS data to BEAST2 XML format, for fossil tip-dating and other uses.; 2014. http://phylo.wikidot.com/beastmaster. Available from: http://phylo.wikidot.com/beastmaster.

[13] O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, et al. The placental mammal ancestor and the postK-Pg radiation of placentals. Science. 2013;339(6120):662–667. Available from: http://www.sciencemag.org/content/339/6120/662.abstract.

[14] Meredith R, Janečka J, Gatesy J, Ryder O, Fisher C, Teeling E, et al. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science. 2011;334(6055):521–524. Available from: http://dx.doi.org/10.1126/science.1211028.

[15] Effects of missing data on topological inference using a Total Evidence approach, author=Guillerme, Thomas and Cooper, Natalie, journal=in review, year=2015,;.

[16] Fritz SA, Bininda-Emonds ORP, Purvis A. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. Ecology Letters. 2009;12(6):538–549. Available from: http://dx.doi.org/10.1111/j.1461-0248.2009.01307.x.

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

[18] Harrison B Luke, Larsson CE Hans. Among-Character Rate Variation Distributions in Phylogenetic Analysis of Discrete Morphological Characters. Systematic biology. 2014;Available from: http://dx.doi.org/10.1093/sysbio/syu098.

[19] Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et al. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26:1463–1464.

[20] Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. Phylogenies and community ecology. Annual review of ecology and systematics. 2002;p. 475–505.

[21] Letcher SG. Phylogenetic structure of angiosperm communities during tropical forest succession. Proceedings of the Royal Society of London B: Biological Sciences. 2009;.

[22] Swenson NG. Phylogenetic Resolution and Quantifying the Phylogenetic Diversity and Dispersion of Communities. PLoS ONE. 2009 02;4(2):e4390. Available from: http://dx.doi.org/10.1371%2Fjournal.pone.0004390.

[23] Faith DP. Conservation evaluation and phylogenetic diversity. Biological Conservation. 1992;61(1):1 – 10. Available from: http://www.sciencedirect.com/science/article/pii/0006320792912013.

Data structure

Supplementary results