Version dated: February 23, 2015

RH: Missing morphological data in living mammals

Missing morphological data 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 macrovelutionary studies is becoming increasingly common among evolutionary biologists [1, 2, 3, 4, 5]. One trending method, called Total Evidence allows to combine molecular data from living taxa and morphological data from both living and fossil taxa (e.g. [6, 7, 8, 4, 9, 10]). This promising method allows to apply integrative phylogenetic inference methods such as tip-dating [7, 11, 12]. However, because the Total Evidence method requires a lot of data (both molecular and morphological for both living and fossil taxa), this method has been shown to be sensible to missing morphological data [13].

In fact, [13] demonstrates that when few living taxa coded, topological recovery is bad. For example, a given phylogeny containing two clades A and B containing only living taxa, when using a Total Evidence method, molecular data is available for both clades but morphological data is only available for clade A for any reason. Then, the addition of any fossil taxa X related to clade B will lead to a wrong topological placement of this fossil taxa X, branching it somewhere in the clade A instead of the clade B because no morphological data is available to support the placement of the fossil taxa X in clade B.

If there is no overlapping characters between a living clade and a fossil one, then it is impossible to branch and fossil taxa to that clade. This can be due to evolutionary history (i.e. a fossil angiosperm has no overlapping characters with a living mammal) and is expected. However, this can also be due to missing data (i.e. a fossil primates has

no overlapping data with living primates because no data is available for living primates) and will produce artefactual wrong phylogenies (i.e. the fossil primate NOT branching in the primate clade).

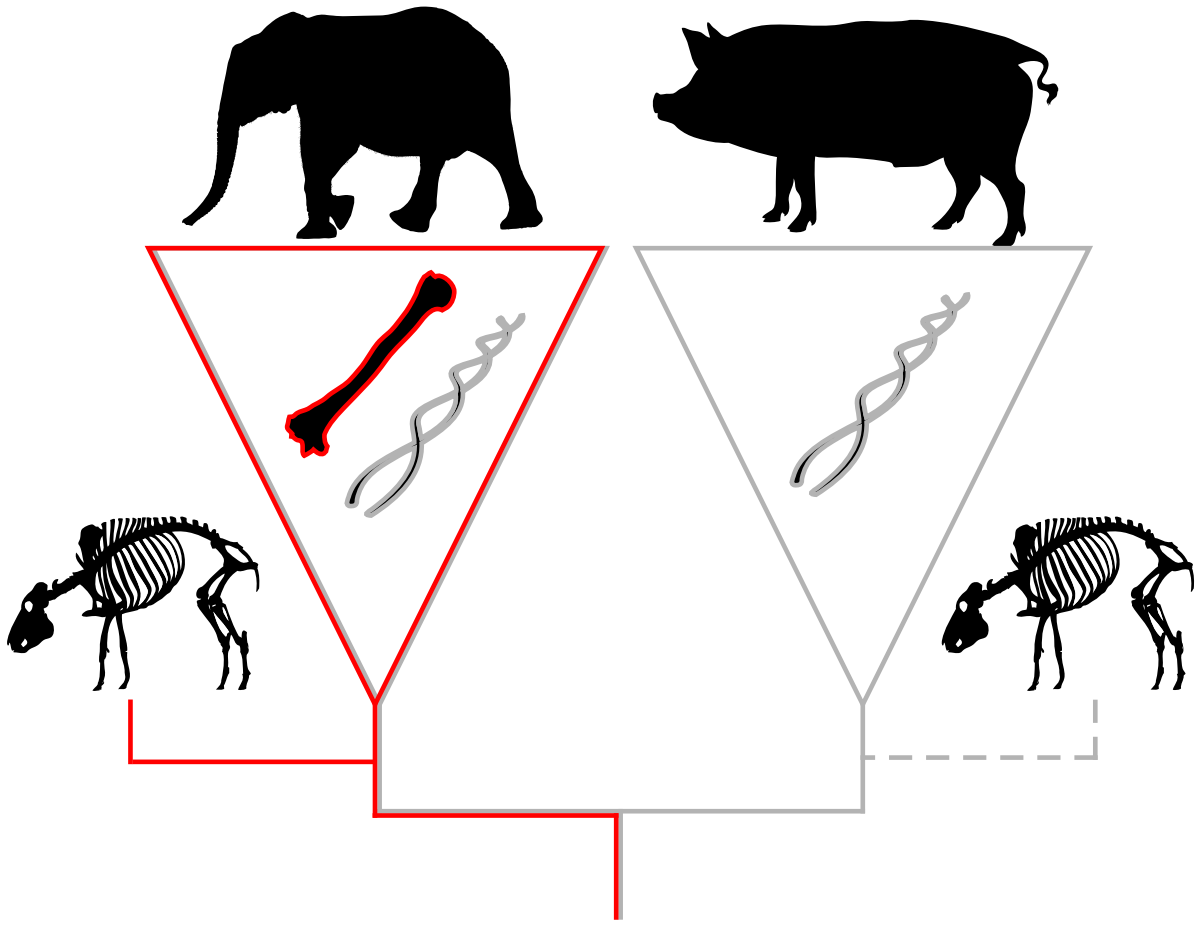

Figure 1: Example of topological errors due to missing morphological data in living taxa. If a phylogeny contains two clades, for example Proboscidea and Cetartiodactyla, with molecular data for both but only morphological data for Proboscidea. If an additional Cetartiodactyla fossil (with no molecular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

In this study we investigate the amount of living mammal taxa with available morphological data to assess the potential caveats in building Total Evidence mammal phylogenetic trees. As well as the data availability, we calculate the structure of the data to make sure that clades with a relatively high amount of data are not only containing all taxa from a single clade and none from another clade.

Questions:

1. -How many taxa with morphological data are available among each living mammals order?
2. -Within each order, how is this data distributed?
3. -How can we improve the data coverage in taxa with non random distributed data

Material and Methods

Matrices search

To investigate the available living taxa with morphological data, we downloaded morphological matrices from three main public databases: morphobank, graeamlloyd.com and rossmounce's github. We downloaded all the matrices containing any fossil or living mammal taxa from these data bases. Additionally we ran a thorough search for matrices that might not have been uploaded on the previously cited data bases through a Google Scholar search. 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 considered living all the OTUs that where either present in Fritz supper tree or in Wilson Reeders taxonomy. We considered fossil 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 not applicable (NA).

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 number of morphological characters was chosen to be in adequacy with [13] and [14]. Also this threshold avoids biases towards small matrices that could be either not informative (e.g. too few characters) 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 level (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 [13] and therefore highly probable of suffering from the effect of missing data.

Available data structure.— For the order with no morphological data for all 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 [15] where we compared the structure of the available data for each order to the structure of a potentially fully sampled data (i.e. only the OTUs with available morphological data *vs.* all the OTUs). For each orders and taxonomic level 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 [16]. 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) [17, 18] but we also calculated the two other common phylogenetic structure indices: Faith's Phylogenetic Distance (PD)[19] and the Nearest Taxon Index (NTI) [16]. 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

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% .

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 OTUs with no 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 .

Order	Taxonomic level	Fraction of OTUs	Percentage of OTUs	PD	p-value
Monotremata	Genus	2/3	66.667	-0.695	0.663
Monotremata	Species	2/4	50	-0.966	0.566
Didelphimorphia	Species	40/84	47.619	-1.33	0.915
Paucituberculata	Genus	2/3	66.667	-0.756	0.682
Paucituberculata	Species	2/5	40	-0.64	0.493
Dasyuromorphia	Genus	7/22	31.818	-1.102	0.894
Dasyuromorphia	Species	8/64	12.5	-1.098	0.93
Peramelemorphia	Species	16/18	88.889	-0.55	0.748
Diprotodontia	Family	9/11	81.818	-0.349	0.551
Diprotodontia	Genus	20/38	52.632	-0.31	0.595
Diprotodontia	Species	16/126	12.698	-0.975	0.849
Afrosoricida	Species	23/42	54.762	1.555	0.077
Macroscelidea	Species	5/15	33.333	-0.474	0.66
Sirenia	Species	2/4	50	-0.957	0.845
Cingulata	Genus	8/9	88.889	1.31	0.229
Cingulata	Species	6/25	24	0.648	0.223
Pilosa	Family	3/5	60	-0.603	0.891
Pilosa	Genus	3/5	60	-0.877	0.795
Pilosa	Species	3/29	10.345	-1.508	0.997
Scandentia	Genus	2/5	40	-0.747	0.639
Scandentia	Species	2/20	10	-1.259	0.984
Primates	Genus	48/68	70.588	-0.302	0.607
Primates	Species	56/351	15.954	-1.504	0.951
Rodentia	Family	10/32	31.25	0.113	0.395
Rodentia	Genus	20/451	4.435	-1.032	0.848
Rodentia	Species	10/2095	0.477	-0.967	0.856
Erinaceomorpha	Species	21/22	95.455	-0.777	0.914
Soricomorpha	Family	3/4	75	-0.95	0.619
Soricomorpha	Genus	19/43	44.186	1.157	0.116
Soricomorpha	Species	19/392	4.847	-1.869	0.977
Chiroptera	Family	13/18	72.222	0.866	0.201
Chiroptera	**Genus**	**68/202**	**33.663**	**18.87**	**0.001**
Chiroptera	**Species**	**108/1054**	**10.247**	**20.112**	**0.001**
Pholidota	Species	3/8	37.5	1.14	0.175

Carnivora	Family	11/15	73.333	0.517	0.285
Carnivora	**Genus**	**30/125**	**24**	**4.014**	**0.002**
Carnivora	**Species**	**42/283**	**14.841**	**17.954**	**0.001**
Perissodactyla	Species	7/16	43.75	0.733	0.199
Cetartiodactyla	Family	20/21	95.238	0.352	0.193
Cetartiodactyla	Genus	76/128	59.375	-3.221	1
Cetartiodactyla	Species	106/311	34.084	-2.768	1

Two different results are shown on figure 2 with randomly distributed data in Cetartiodactyla (Fig. 2A) and clustered available data in Carnivora (mainly Canida;e Fig. 2B).

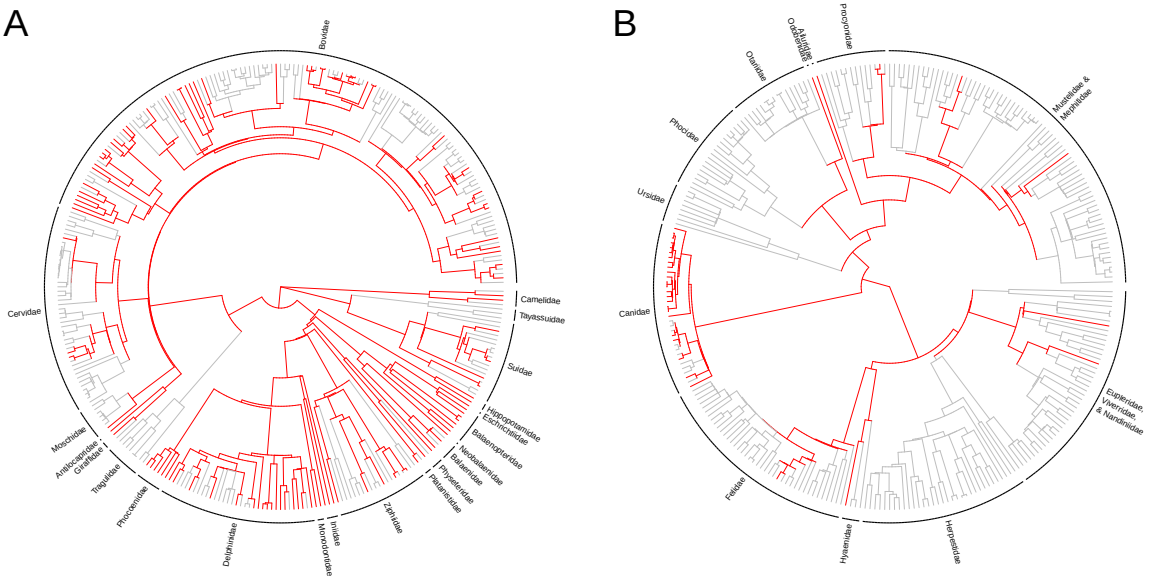

Figure 2: Distribution of available morphological data across cetartiodactyles (**A**) and carnivores (**B**). Edges are colored in grey when no morphological data is available or in red when data is available.

Discussion

It's pretty bad

But good news! It's at least mainly random

However, we counted only the raw number of characters here and not there similarity. It might be that actually the amount of available data here is lower due to non overlap of characters....

So let's go coding some data and using opensource traceable data bases like morphobank!

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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[14] 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.

[15] 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.

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

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

[18] 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.

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

SOM

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

SEARCH TERMS

Mammalian orders terms

The searched mammalian order terms are available in $search_terms_latin.txt$ or $search_terms_meta.txt$. The file containing the meta names is the Latin name files but with

replacing the latin suffixes ([ia—ata—ea—a]) by a joker character (*) and by replacing the first letter by a upper/lower case meta character (e.g. [Aa]).

Ross Mounce data set.— I selected all the matrices containing at least one of the mammalian orders names from Ross Mounce GitHub $cladistic - data/nexus_{f}iles$ repository (accessed on the 02/12/2014).

Graeme Lloyd

Selecting the downloadable matrices

Morphobank

order

Google scholars

20 first results since 2010 with the following key words: *order* (”morphology” OR ”morphological” OR ”cladistic”) AND characters matrix paleontology phylogeny Why 20 first results? Because rarefaction curve, check supplemetaries

Wrong Bionmial names and typos

I fixed the wrong bionomial names format (e.g. H. sapiens) into the correct ones (e.g. Homo sapiens) manually using the abreviation list in the concerned publications.

Supplementary results