

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.

1

Abstract

2

3 ()

INTRODUCTION

We need trees with both living and fossil taxa blablablabla

However, because of the nature of TEM matrices, we can have a lot of missing data blabalbalbal

Some missing has more effect on recovering the correct topology than others. Missing living taxa are important! Or they are available (because we have molecular data at least).

In this study we want to have an accurate overlook of the state of morphological availability in mammals.

Also, when data is available, we want to know how the data is distributed along the clade. Because it is unrealistic or unnecessary to sample the whole clade, we want the data to be at least randomly distributed (so no bias) or at the best evenly distributed (so every clade is sampled). However, what we don't want is that data to be clustered.

After looking at the data availability and its distribution, we propose several strategies to improve data coverage in weekly sampled orders.

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

DATA COLLECTION

We need to use the same protocol as the GMPD to make sure our data are comparable.

The original version used lots of online databases but luckily we can now just use

Google Scholar! Into the Google Scholar search enter the species Latin Binomial as

search keywords, followed by the search terms related to parasites. So for *Marmota*

monax:

parasite OR helminth OR virus OR bacteria OR fungi OR ectoparasite OR

arthropod OR protozoa "*Marmota monax*"

[Note that the quotation marks (" ") tell Google that you want the exact phrase

Marmota monax, and the OR means it will contain *Marmota monax* and parasite or

Marmota monax and helminth etc.]

You then need to search the results for appropriate papers. This is tedious so I

advise setting yourself a reasonable target, for example, do it for a few hours a day then

do something else!

You should also search for each squirrel genus independently; however I would

only do this for genera with few papers on the species. If you did this for *Marmota*

spp. you could be there forever and it's not clear what you could do with that data.

Finally you should search for common synonyms of the species. The original

compiler of the database did this very thoroughly. However, I suggest just using the

names from the 1993 taxonomy if they've changed between then and 2005. In theory the original compiler should have picked up all the old papers with data already, so you want to focus on the current taxonomy.

b) Saving the paper and reference Once you find a paper that looks useful download the paper and save it as a PDF. Also add the reference to EndNote. Not all the papers will end up being included in the database, but its quicker to add them to EndNote now and delete them later, than to add them to EndNote later.

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 `Latinname` files but with three

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_filesrepository](https://github.com/RossMounce/cladistic-data/nexus_filesrepository) (accessed on the 02/12/2014).

Graeme Lloyd

Selecting the downloadable matrices

Morphobank

56 *order*

57

Google scholars

58 20 first results since 2010 with the following key words: *order* ("morphology" OR
59 "morphological" OR "cladistic") AND characters matrix paleontology phylogeny Why
60 20 first results? Because rarefaction curve, check supplementaries

61

WRONG BINOMIAL NAMES AND TYPOS

62 I fixed the wrong binomial names format (e.g. *H. sapiens*) into the correct ones (e.g.
63 *Homo sapiens*) manually using the abbreviation list in the concerned publications.

64

RESULTS

65 Table 1: data availability

66 Table 2: randomness

67 Table 3: phylotargeting

68 Figure 1: example of good and bad coverage (artiodactyles and carnivores)?

69

DISCUSSION

70

ETHICS STATEMENT

71

DATA ACCESSIBILITY STATEMENT

72 All data is available and reproducible on github.

73

AUTHORS CONTRIBUTIONS STATEMENT

74 Conceived and designed the experiments: TG NC. Performed the experiments: TG.

75 Analyzed the data: TG. Wrote the paper: TG NC.

76

ACKNOWLEDGEMENTS

77 April Wright, David Bapst and Graeme Lloyd.

78

FUNDING STATEMENT

79 This work was funded by a European Commission CORDIS Seventh Framework

80 Programme (FP7) Marie Curie CIG grant (proposal number: 321696).