



ANALYSIS OF BIODIVERSITY DATA SUGGESTS THAT MAMMAL SPECIES ARE HIDDEN IN PREDICTABLE PLACES

Danielle J. Parsons, Tara A. Pelletier, Jamin G. Wieringa, Drew J. Duckett, Bryan C. Carstens
The Ohio State University Museum of Biological Diversity, Department of EEOB

WHAT DO WE KNOW ABOUT HIDDEN DIVERSITY?

- Only an estimated ~1-10% of extant species are thought to be formally described ^{1,2}
- Even in mammals, hidden diversity continues to be found
- Previous attempts to identify large-scale patterns of hidden diversity have produced conflicting results

ELSEVIER **Review** *TRENDS in Ecology and Evolution* Vol.22 No.3 Full text provided by www.sciencedirect.com
ScienceDirect

Cryptic species as a window on diversity and conservation

David Bickford¹, David J. Lohman¹, Navjot S. Sodhi¹, Peter K.L. Ng¹, Rudolf Meier¹, Kevin Winker², Krista K. Ingram³ and Indraneil Das⁴

BMC Evolutionary Biology BioMed Central

Research article **Open Access**

Cryptic animal species are homogeneously distributed among taxa and biogeographical regions

Markus Pfenninger* and Klaus Schwenk

1. Heywood & Watson, 1995

2. Costello et al., 2013

WHAT DO WE KNOW ABOUT HIDDEN DIVERSITY?



Undescribed species are thought to be common, but understanding this phenomenon is inhibited by a lack of information regarding the traits that make a clade likely to contain hidden species.

INTEGRATING SPECIMEN DATA AND BIOINFORMATICS

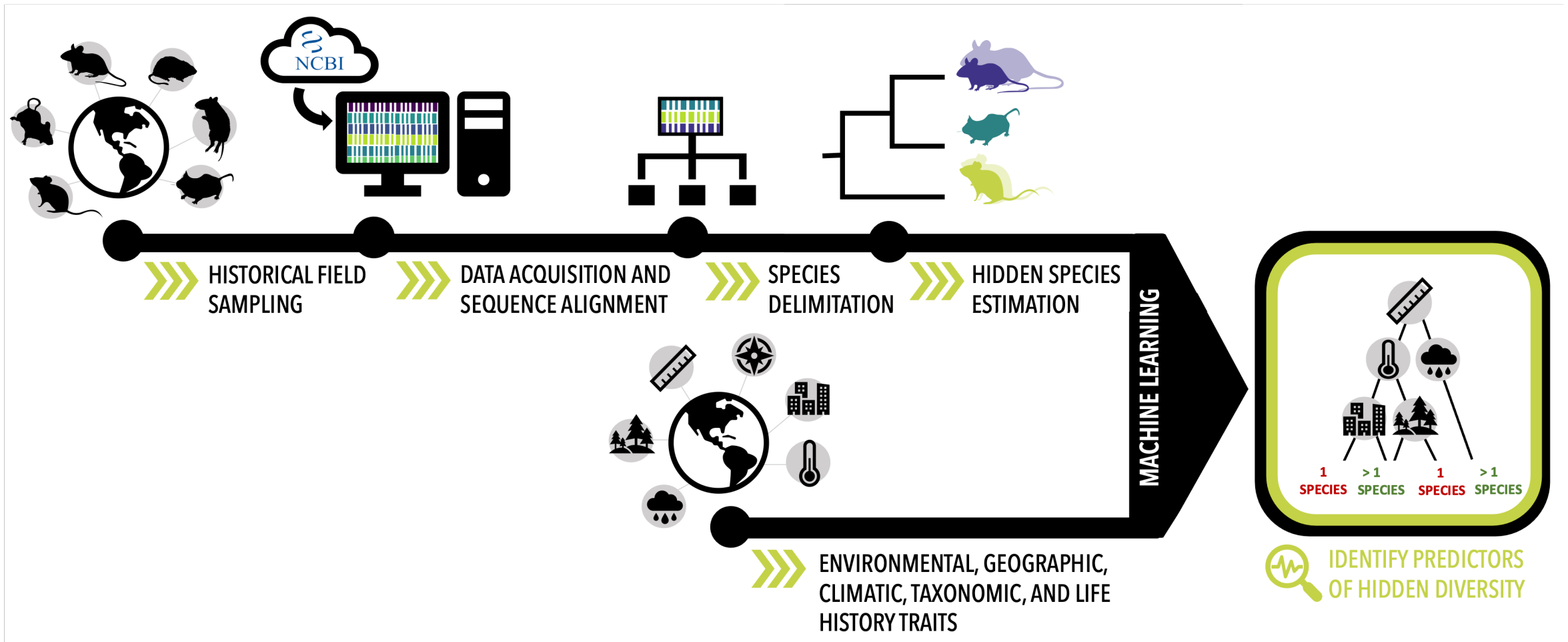


How much hidden diversity is present in mammals?

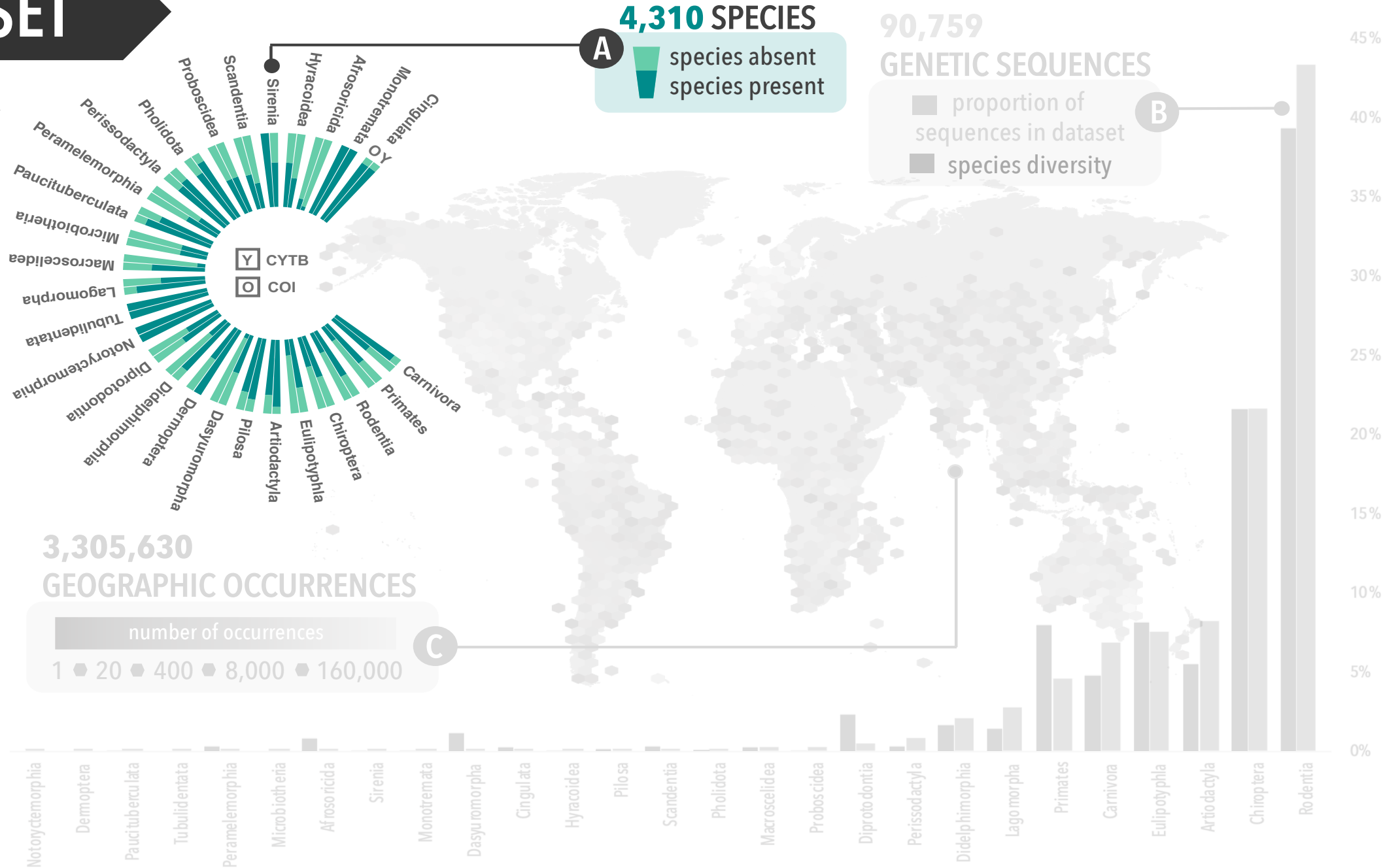


What attributes make a clade likely to contain hidden species?

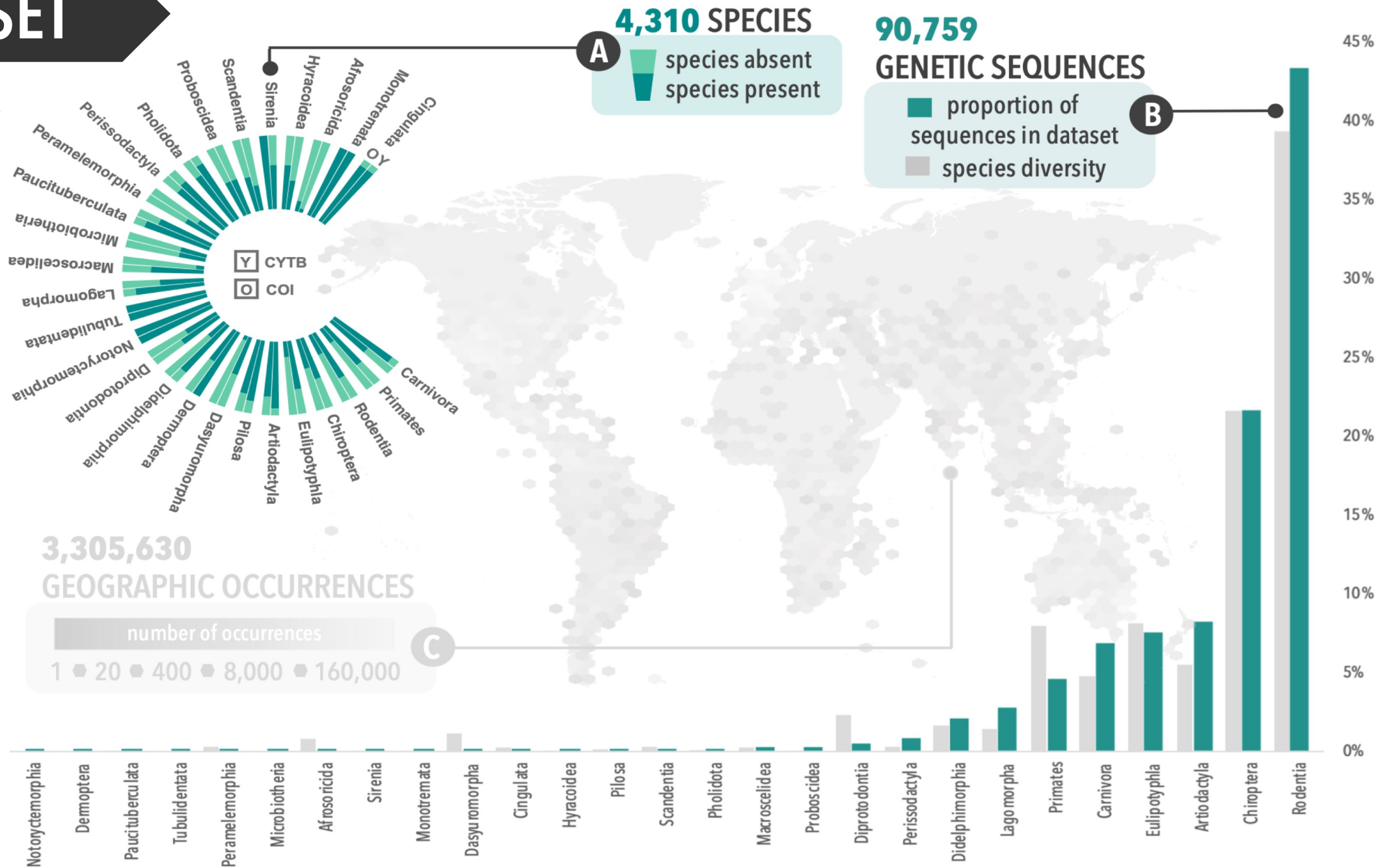
Develop a predictive framework to **identify clades likely to contain hidden species** and **pinpoint specific trait complexes** that indicate where this hidden diversity is likely to be found.



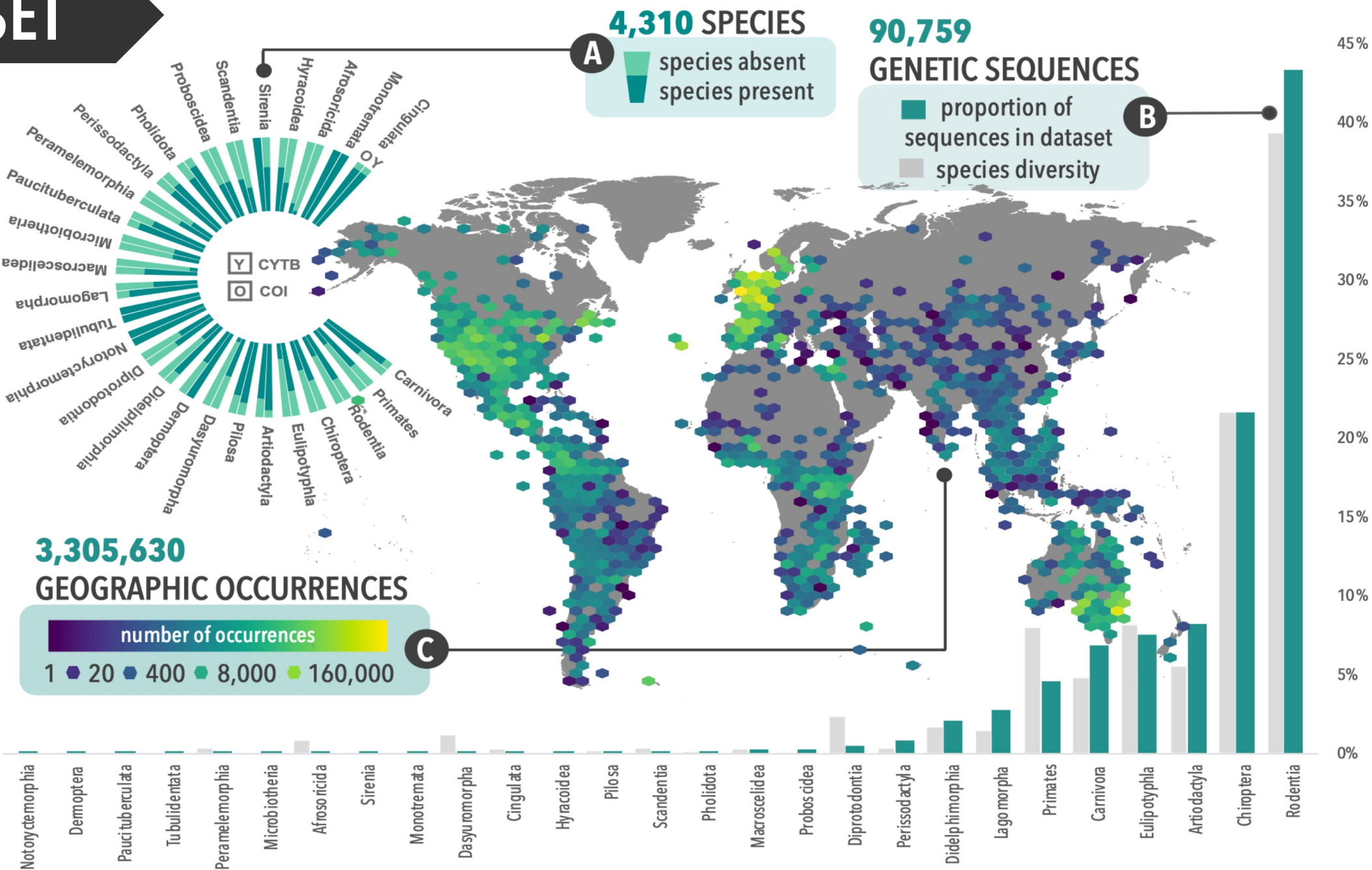
DATASET



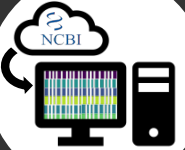
DATASET



DATASET



I. HIDDEN SPECIES ESTIMATION



STANDARDIZATION AND SEQUENCE ALIGNMENT

- Standardize taxonomy
 - Mammal Diversity Database
- Generate family-level alignments
 - *MUSCLE v3.5*
- Visually check alignments for errors
- Model nucleotide substitution
 - *jModelTest2*



AUTOMATED SPECIES DELIMITATION

- Generalized Mixed Yule Coalescent
- Automatic Barcode Gap Discovery

GMYC COI

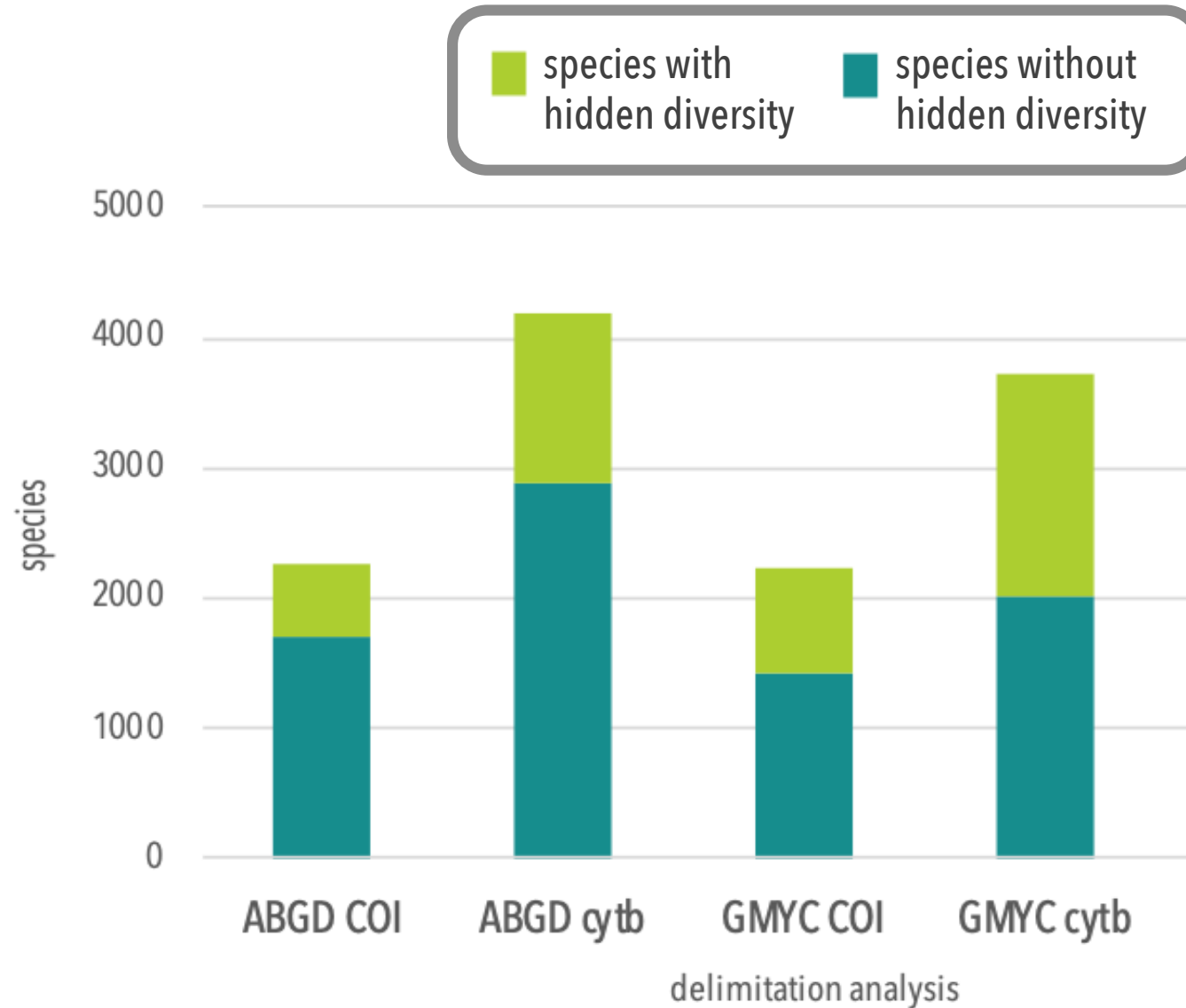
GMYC cytb

ABGD COI

ABGD cytb

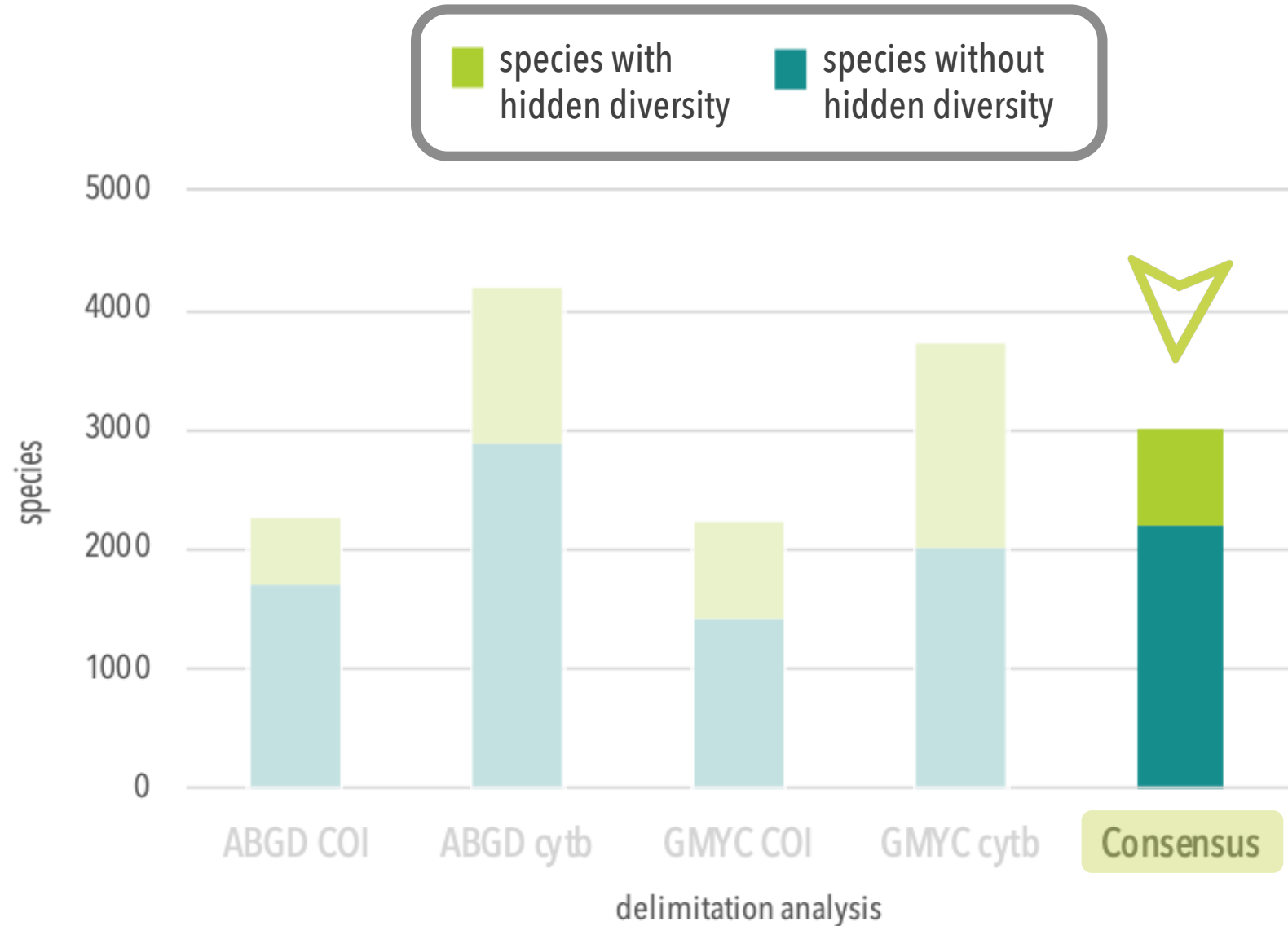
I. RESULTS

How much hidden diversity does class Mammalia have?



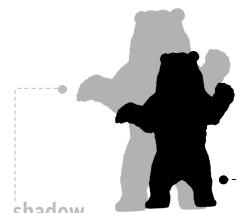
I. RESULTS

How much hidden diversity does class Mammalia have?



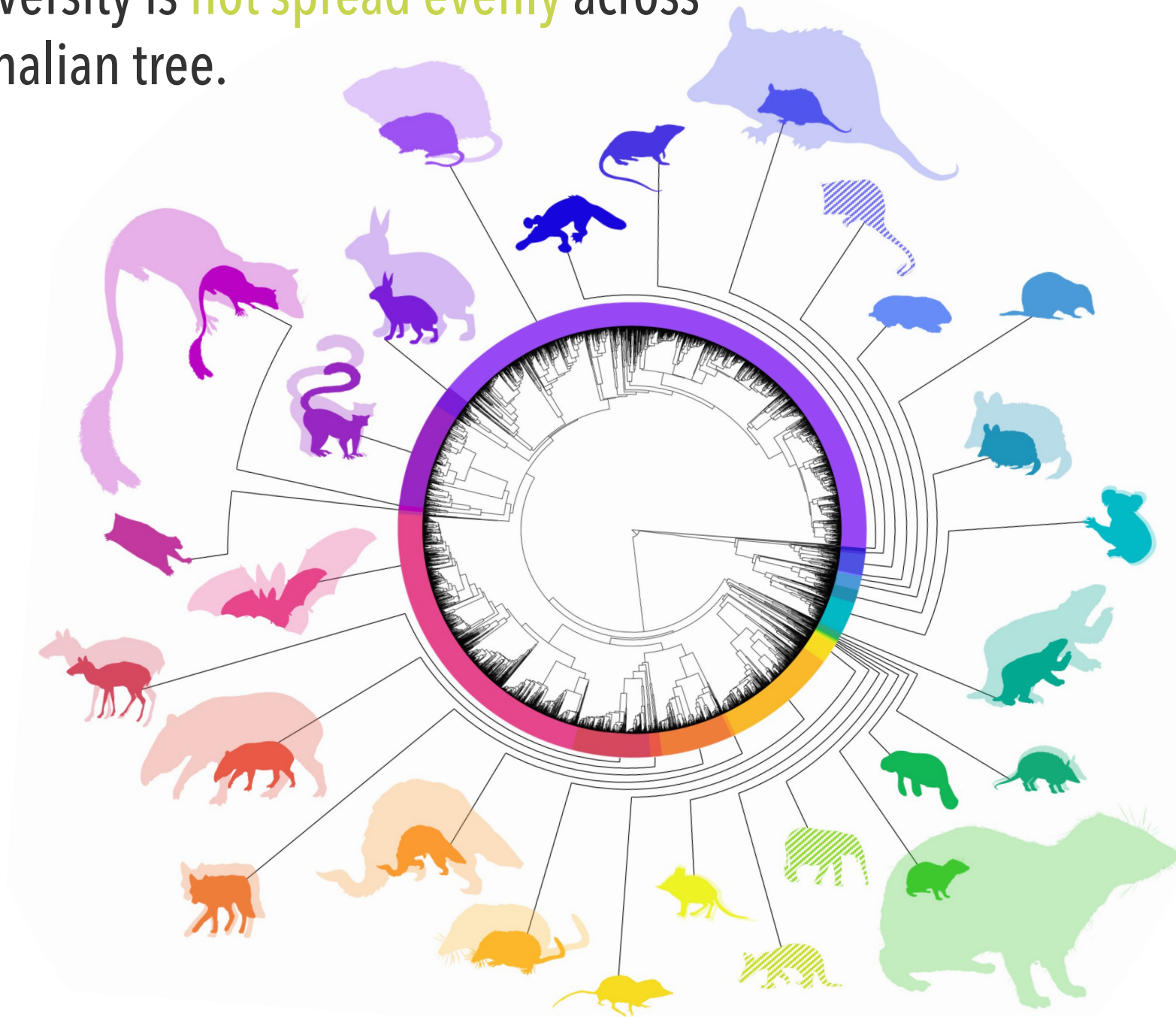
I. RESULTS

Hidden diversity is **not spread evenly** across the mammalian tree.



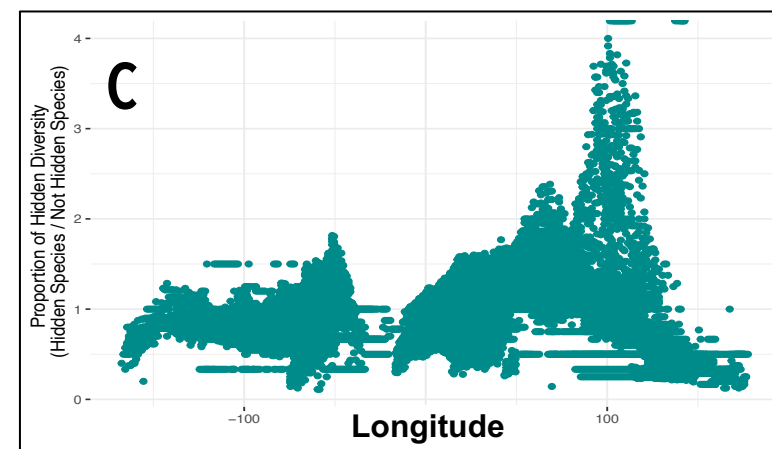
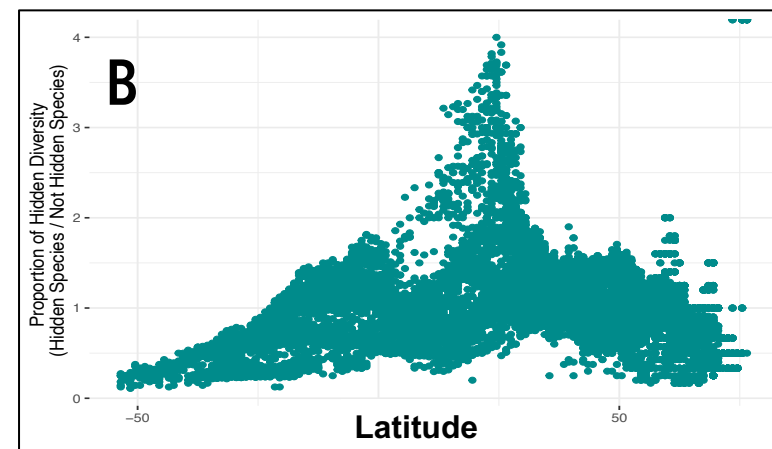
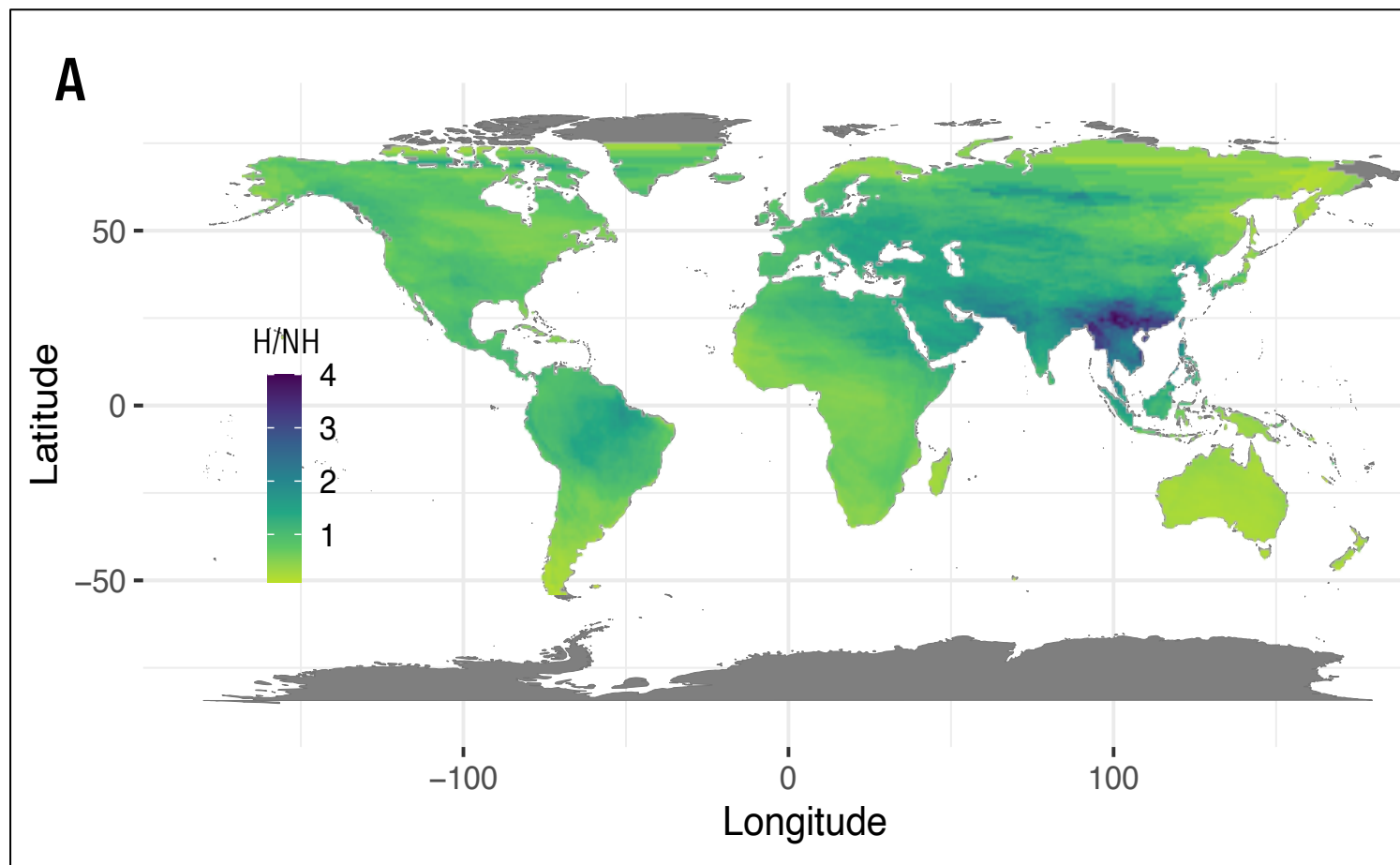
shadow **PREDICTED SPECIES** solid **RECOGNIZED SPECIES**

RODENTIA (2446 / 1319)	TUBULIDENTATA (--- / ---)
LAGOMORPHA (106 / 56)	PROBOSCIDEA (--- / ---)
PRIMATES (283 / 219)	HYRACOIDEA (6 / 1)
SCANDENTIA (20 / 8)	SIRENIA (3 / 3)
DERMOPTERA (1 / 1)	CINGULATA (19 / 17)
CHIROPTERA (1061 / 596)	PILOSA (16 / 7)
ARTIODACTYLA (244 / 164)	DIPROTODONTIA (81 / 74)
PERISSODACTYLA (11 / 5)	DASYUROMORPHA (65 / 38)
CARNIVORA (224 / 176)	PERAMELEMORPHA (6 / 6)
PHOLIDOTA (12 / 5)	NOTORYCTEMORPHA (1 / 1)
EULIPOTYPHILA (381 / 237)	MICROBIOTHERIA (--- / ---)
AFROSORICIDA (7 / 7)	DIDELPHIMORPHA (195 / 58)
MACROSCELIDEA (10 / 9)	PAUCITUBERCULATA (6 / 6)
	MONOTREMATA (2 / 2)



I. RESULTS

Mammalian hidden diversity is **not spread evenly** across the globe.

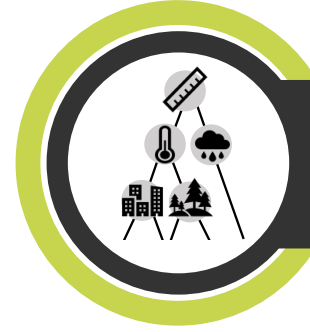


II. IDENTIFYING PREDICTORS OF HIDDEN DIVERSITY



PREDICTIVE TRAIT DATABASE

- 117 total variables
- **Geographic:** GBIF-based metrics
- **Environmental:** GIS data-layers
- **Life History:** PanTHERIA-based metrics
- **Climatic:** BioClim variables
- **Taxonomic Effort:** publication-based metrics (Web of Science)



MACHINE LEARNING MODEL

- Random Forest classification
 - 1000 decision trees
 - Training set: 80% of data
 - Test set: 20% of data
 - 10-fold cross validation x 5
- Variable importance measurements: MDA and Gini

II. RESULTS

Predictive models can be used to accurately identify mammal species containing hidden diversity.

B MODEL EVALUATION	ABGD COI	ABGD cytb	GMYC COI	GMYC cytb	Consensus
Model Accuracy	0.737	0.68	0.6429	0.6517	0.781
Accuracy (95% CI)	(0.6802, 0.7885)	(0.6333, 0.7241)	(0.5821, 0.7004)	(0.6014, 0.6996)	(0.7273, 0.8285)
Pos Predictive Value	0.56667	0.6304	0.5571	0.6624	0.807
Neg Predictive Value	0.75833	0.6937	0.6735	0.6345	0.7742

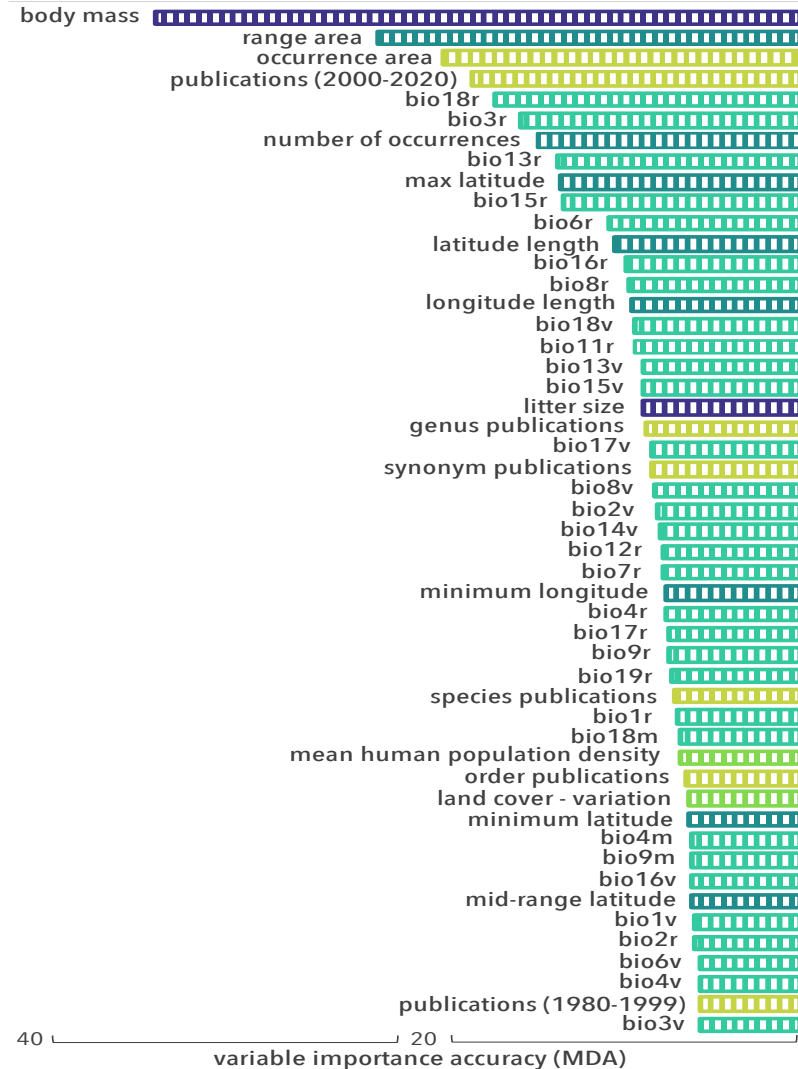
II. RESULTS

Predictive models can be used to accurately identify mammal species containing hidden diversity.

B MODEL EVALUATION	ABGD COI	ABGD cytb	GMYC COI	GMYC cytb	Consensus
Model Accuracy	0.737	0.68	0.6429	0.6517	0.781
Accuracy (95% CI)	(0.6802, 0.7885)	(0.6333, 0.7241)	(0.5821, 0.7004)	(0.6014, 0.6996)	(0.7273, 0.8285)
Pos Predictive Value	0.56667	0.6304	0.5571	0.6624	0.807
Neg Predictive Value	0.75833	0.6937	0.6735	0.6345	0.7742

II. RESULTS

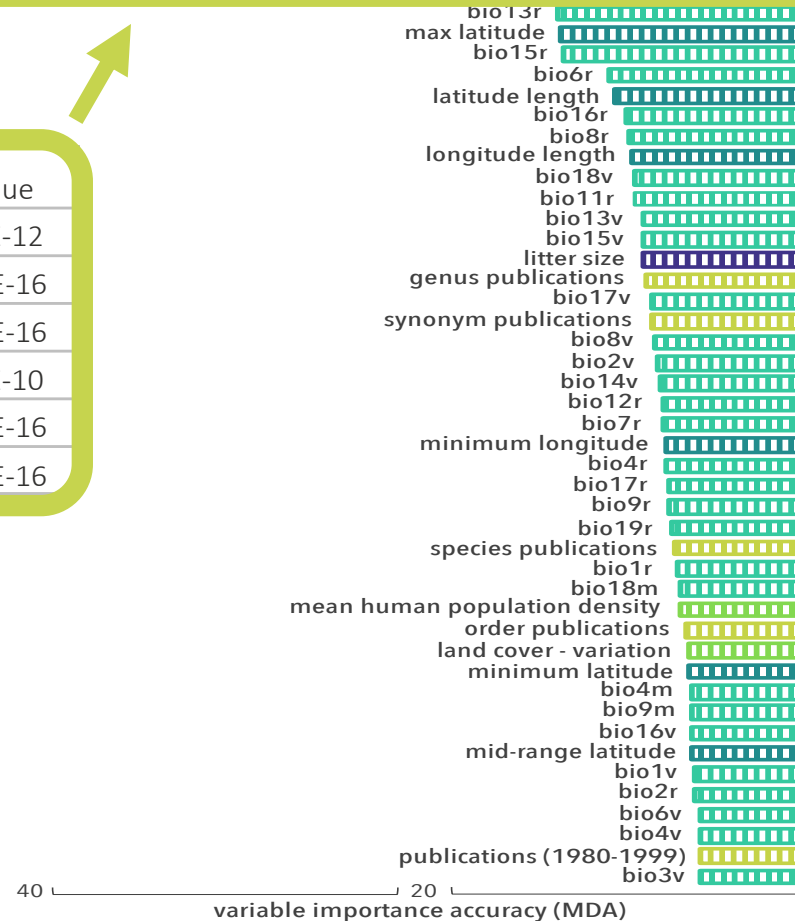
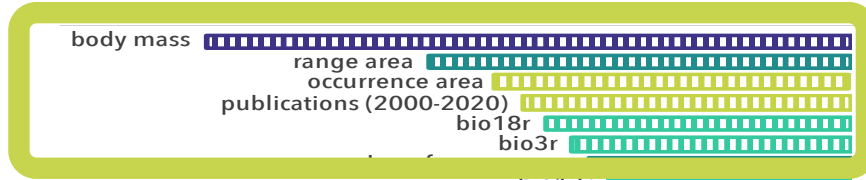
Specific trait complexes distinguish taxa harboring potentially undescribed diversity.



II. RESULTS

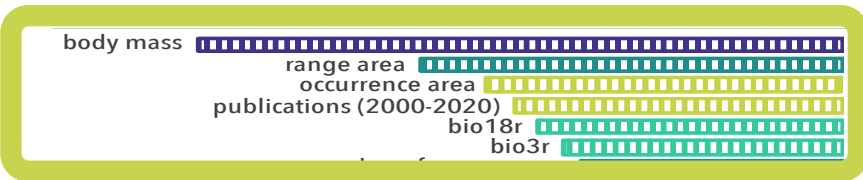
Specific trait complexes distinguish taxa harboring potentially undescribed diversity.

Kruskal-Wallis Test	p-value
Adult Body Mass (g)	2.33E-12
Range Area (km ²)	< 2.2E-16
Occurrence Area (km ²)	< 2.2E-16
Recent Publications	1.39E-10
bio18r (precipitation mm)	< 2.2E-16
bio3r (isothermality)	< 2.2E-16

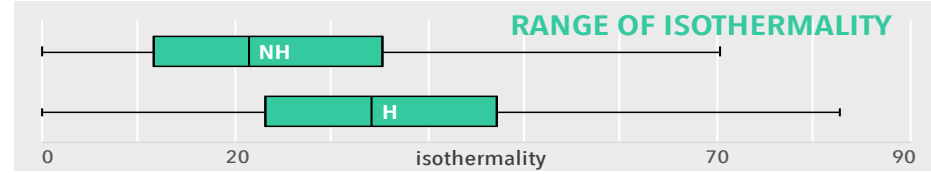
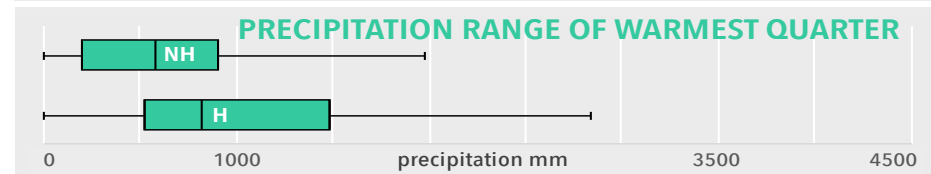
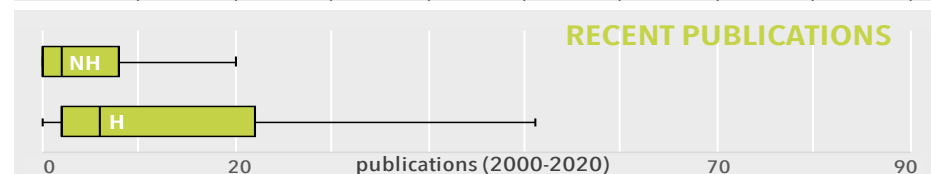
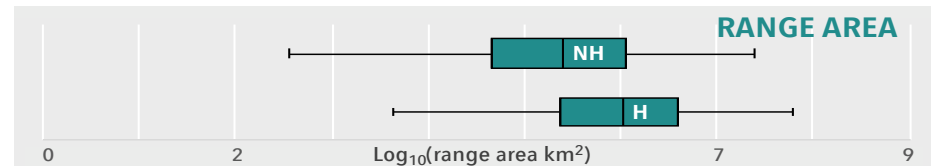
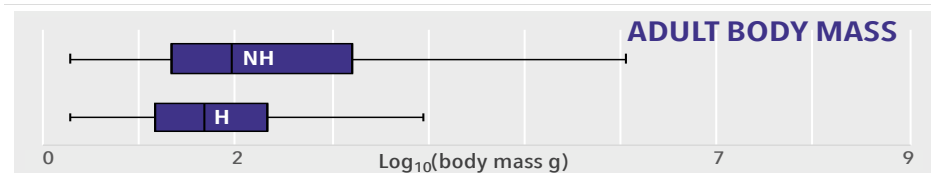
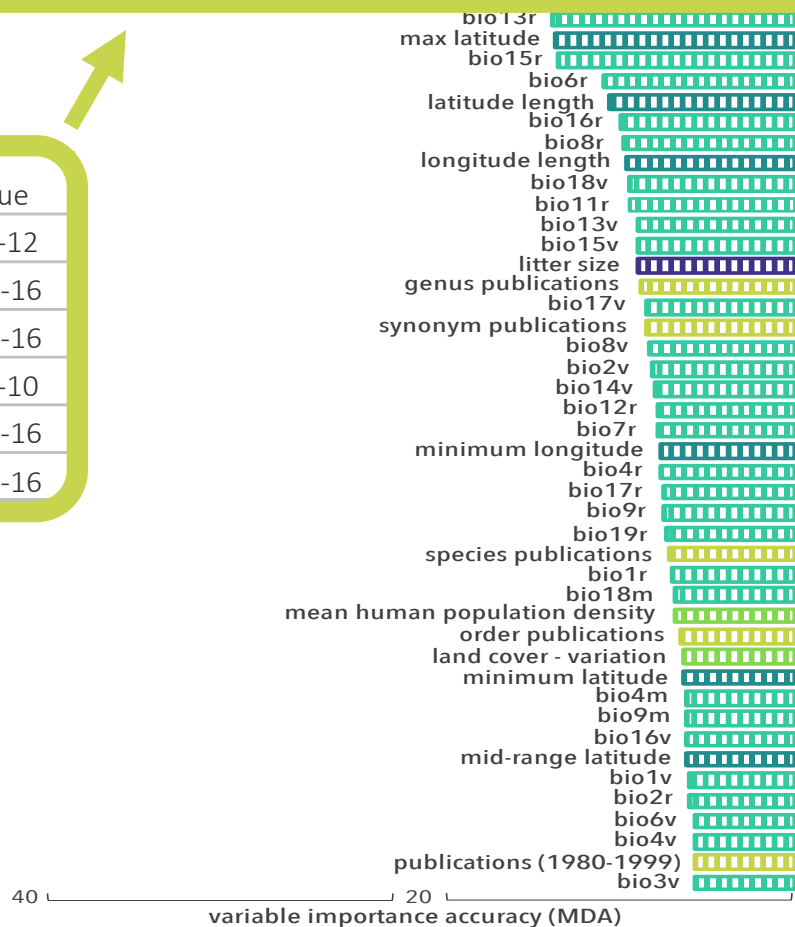


II. RESULTS

Specific trait complexes distinguish taxa harboring potentially undescribed diversity.



Kruskal-Wallis Test	p-value
Adult Body Mass (g)	2.33E-12
Range Area (km ²)	< 2.2E-16
Occurrence Area (km ²)	< 2.2E-16
Recent Publications	1.39E-10
bio18r (precipitation mm)	< 2.2E-16
bio3r (isothermality)	< 2.2E-16



CONCLUSIONS

Delimitation results suggest roughly 30% of recognized mammal species are likely to contain hidden diversity

Support for previous taxonomic research suggesting hidden species are likely to be found in insular systems and areas of high endemism.

Identification of specific trait complexes traits associated with the presence of hidden diversity in mammals.

↓ adult body mass

↑ range size

↑ recent publications

↑ precip. of warmest quarter

↑ range of isothermality

CONCLUSIONS

Delimitation results suggest roughly 30% of recognized mammal species are likely to contain hidden diversity

(~43%: *Baker & Bradley, 2006*)

Support for previous taxonomic research suggesting hidden species are likely to be found in insular systems and areas of high endemism.

(*Reeder, Helgen & Wilson, 2007*)

Identification of specific trait complexes traits associated with the presence of hidden diversity in mammals.

↓ adult body mass

↑ range size

↑ recent publications

↑ precip. of warmest quarter

↑ range of isothermality

IMPLICATIONS

"BIG DATA" AND BIODIVERSITY RESEARCH



Integration of bioinformatics and specimen data = more comprehensive understanding of biodiversity



IMPLICATIONS

"BIG DATA" AND BIODIVERSITY RESEARCH



Integration of bioinformatics and specimen data = more comprehensive understanding of biodiversity

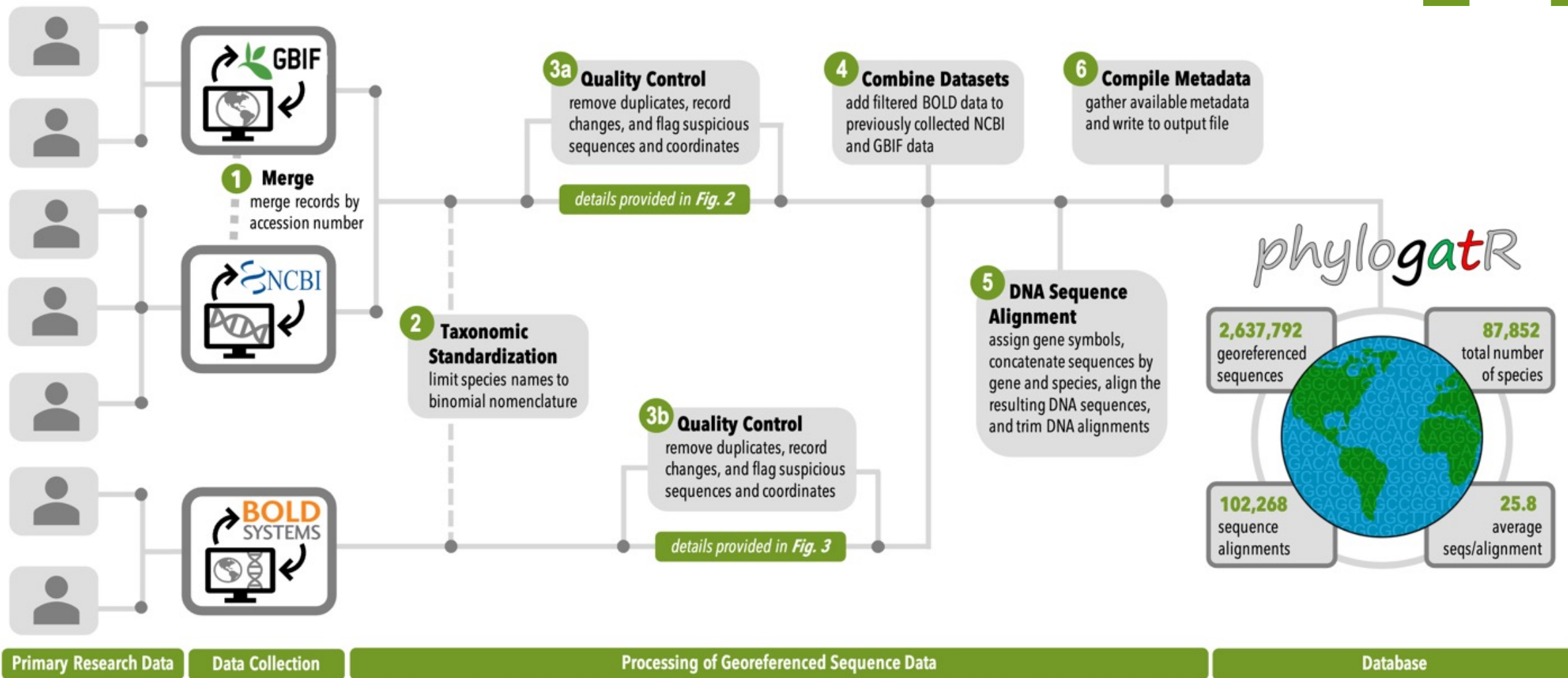
requires



Data accessibility and standardization



Cultivation of computational and technological competency



Primary Research Data

Data Collection

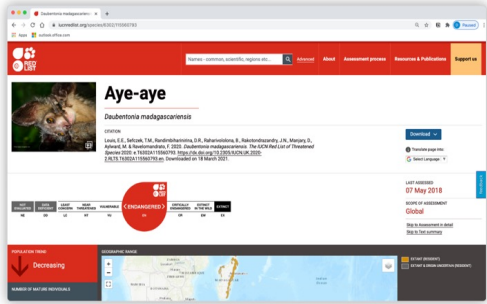
Processing of Georeferenced Sequence Data

Database



IUCN Red List: Conservation Data

In this module, we're interested in evaluating the link between genetic diversity and species conservation status. That means, once we've collected our genetic data, our next step is to collect complementary conservation data. In order to do this, we'll turn to the International Union for Conservation of Nature (IUCN) Red List of Threatened Species, or IUCN Red List for short. Established in the year 1964, the IUCN Red List is the most comprehensive inventory of global species conservation status in the world. And fortunately for both us and the many species it protects, the IUCN Red List and its underlying information exist as an online database available to the public. A central focus of the IUCN Red List is to provide information that can help guide actions to conserve biological diversity. By making their data open and available, the IUCN enables the world to combat threats to biodiversity on all sides, from policy and decision-making, to conservation biology research, to public engagement, and even to us!



The IUCN Red List uses a set of specific criteria when evaluating the extinction risk faced by species. Because the IUCN assesses conservation on a global scale, evaluation criteria are relevant to all species and all regions of the world. When evaluating a species risk of extinction, the IUCN Red List considers information such as population size, rate of population decline, size of geographic distribution, and degree of population fragmentation. After assessing all available information, the IUCN assigns each species a conservation status. The following list describes each of the conservation categories currently used by the IUCN Red List.

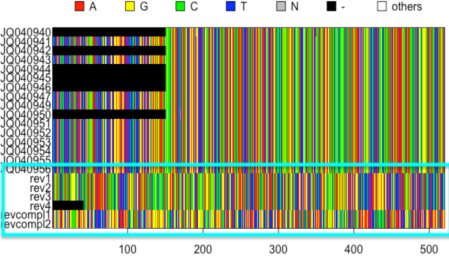
- Key Point:** The IUCN Red List is the most comprehensive inventory of global species conservation status in the world
- Key Point:** The IUCN Red List evaluates the extinction risk of species using a set of specific criteria that are relevant to all species across all regions of the globe
- Key Point:** When evaluating a species risk of extinction, the IUCN Red List considers criteria such as population size, rate of population decline, size of geographic distribution, and degree of population fragmentation

Throughout the remainder of this module, you'll be guided through the process of data collection, visualization, analysis, and interpretation. After completing each of these steps, you should have all of the information that you need to answer our main question. Good Luck!

Sequence alignments: what to watch out for

The following alignments are meant to show you potential errors you might run into when inspecting your sequence alignments. PhylogatR takes steps to minimize these issues (see [here](#)), but it can never hurt to check your data.

In the first example, I have edited the alignment to contain sequences that have been entered in reverse (rev1-4) and reverse complement (revcomp1-2). As you can see, these sequences are comparable in length to the rest of the alignment and are lacking large regions of gaps. However, upon visual inspection they clearly do not line up with the rest of the alignment. These sequences can be checked and edited manually in a sequence viewing program (e.g., UGENE or mesquite).




In this example, I have added parasitic sequences (parasite1-3) to the alignment. These sequences contain long regions of gaps and the regions of nucleotides present do not line up with the rest of the alignment. If your alignment has questionable sequences such as these, you can do a search of their accession numbers [here](#) to ensure they are from the correct species and locus.

Geographic coordinates: what to watch out for

The following map is meant to show you potential errors you might encounter when inspecting your geographic data.

You'll want to look out for any data points that are obviously outside of your species expected range. Data points that are very far away from the rest of the points (i.e., outliers) as well as data points in areas physiologically unlikely for your species (e.g., a terrestrial shrew in the middle of the ocean) should warrant further inspection. If you are unsure of the extent of your species range, you can search for your species [here](#).



5. Spatial Principal Component Analysis (sPCA)

An sPCA is designed to investigate non-random spatial distributions of genetic variation. In other words, is the genetic variation that we observe in our species due to population structure, or is it the result of random mating?

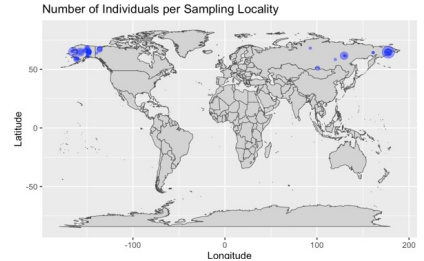
We will use the `sPCA` function to perform a spatial principle component analysis (sPCA) on our data. This analysis implements a connection network in order to incorporate spatial data. We will use `Delanury triangulation` to generate a connection network by setting the argument `type=1`. The connection network type can be changed by providing a different value for the `type` argument, and more information regarding the different types of connection networks available can be found with the command `?connection`.

Our `genind` object "seq_genind" will act as input for the analysis, which we will save in R as "mySPca". For now, we will retain the first positive axis (by setting `nfpos=1`) and the first negative axis (by setting `nfneg=1`). The figure returned is a plot of our connection network.

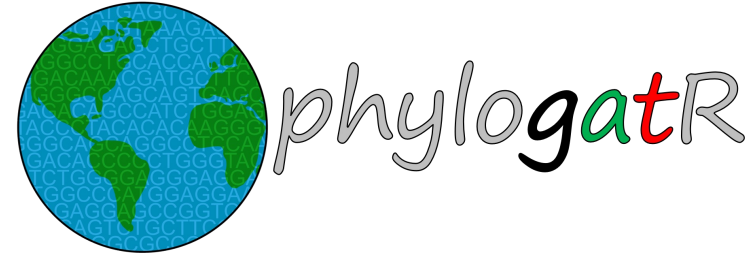
```
#performs a sPCA using the Delanury triangulation as connection network (type=1)
mySPca <- sPCA(seq_genind, type=1, ask=FALSE, scann=FALSE, nfpos=1, nfneg=1)
```

Now that we have completed our sPCA, we can use the results to determine how many principal component (PC) axes we want to retain for further analysis. Our goal here is to retain only the axes that explain the largest portion of the spatial genetic structure of our species. We will evaluate axis contribution in two ways: 1) plotting the sPCA eigenvalues and 2) plotting the spatial and genetic variance components of each eigenvalue.

First, we will use the function `barplot` to display a barplot of eigenvalues from our sPCA. Positive eigenvalues, which are located on the left side of the plot in warmer colors, indicate the presence of global structure. Global structures display positive spatial autocorrelation, which is typically observed when populations are split into patches or located along clines. Negative eigenvalues, which are located on the right side of the plot in cooler colors, indicate the presence of local structure. Local structures display negative spatial autocorrelation, which is typically observed when neighboring individuals are more genetically distinct than expected at random. For more information on global and local structure, see Jombart et al. (2008).⁴



Acknowledgements



CARSTENS  LAB

- Carstens Lab
- Tara Pelletier
- Ryan Norris
- Andreas Chavez
- Andrew Hope
- Meg Daly

Databases

- GBIF
- GenBank
- PanTHERIA
- ASM Mammal Diversity Database
- Ohio Supercomputer Center

Graphics

- PhyloPic
- VertNet
- Natural History Collections housing the specimens studied
- All original researchers who made their data accessible

