



# phylogatR

# PHYLOGEOGRAPHIC DATA AGGREGATION AND REPURPOSING



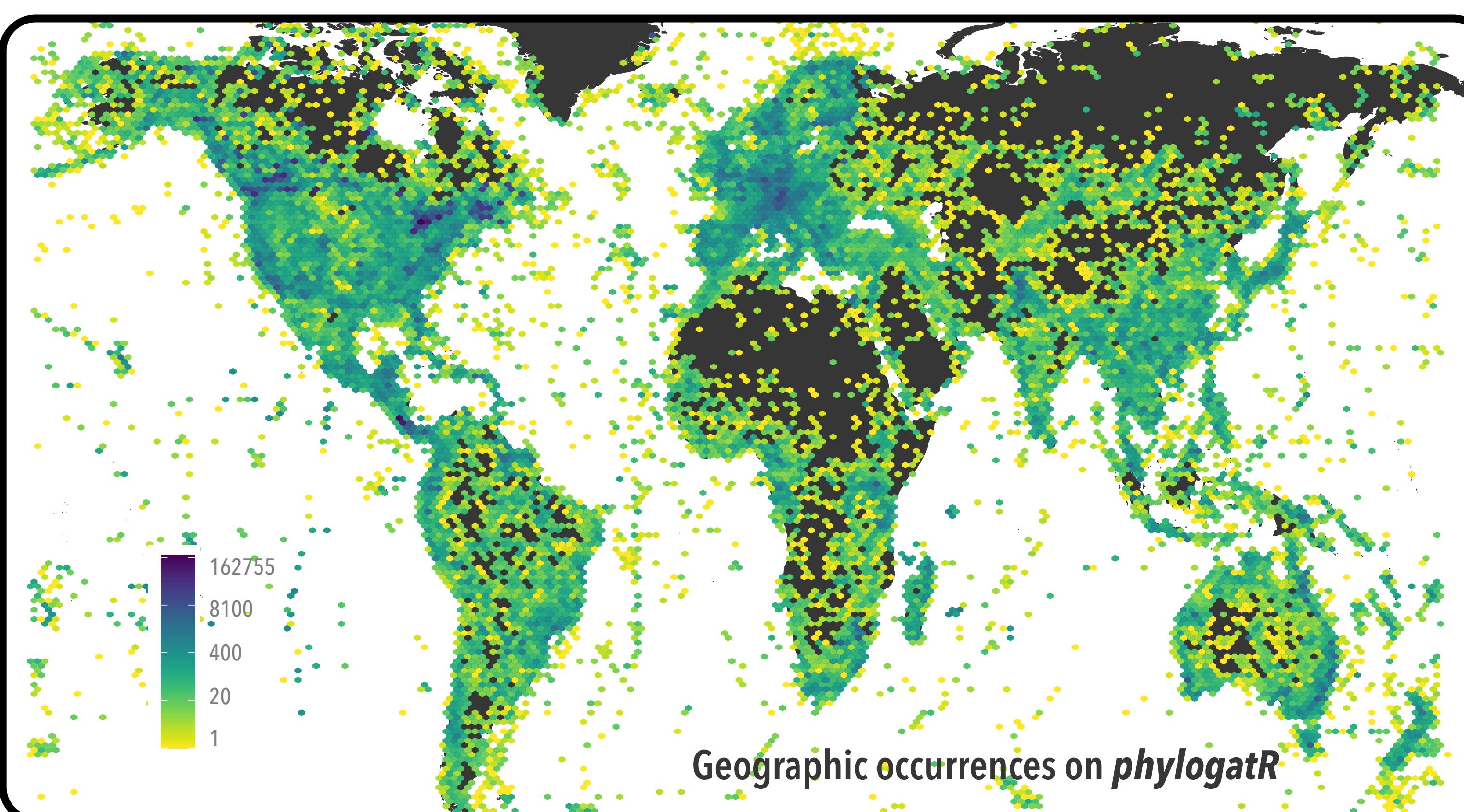
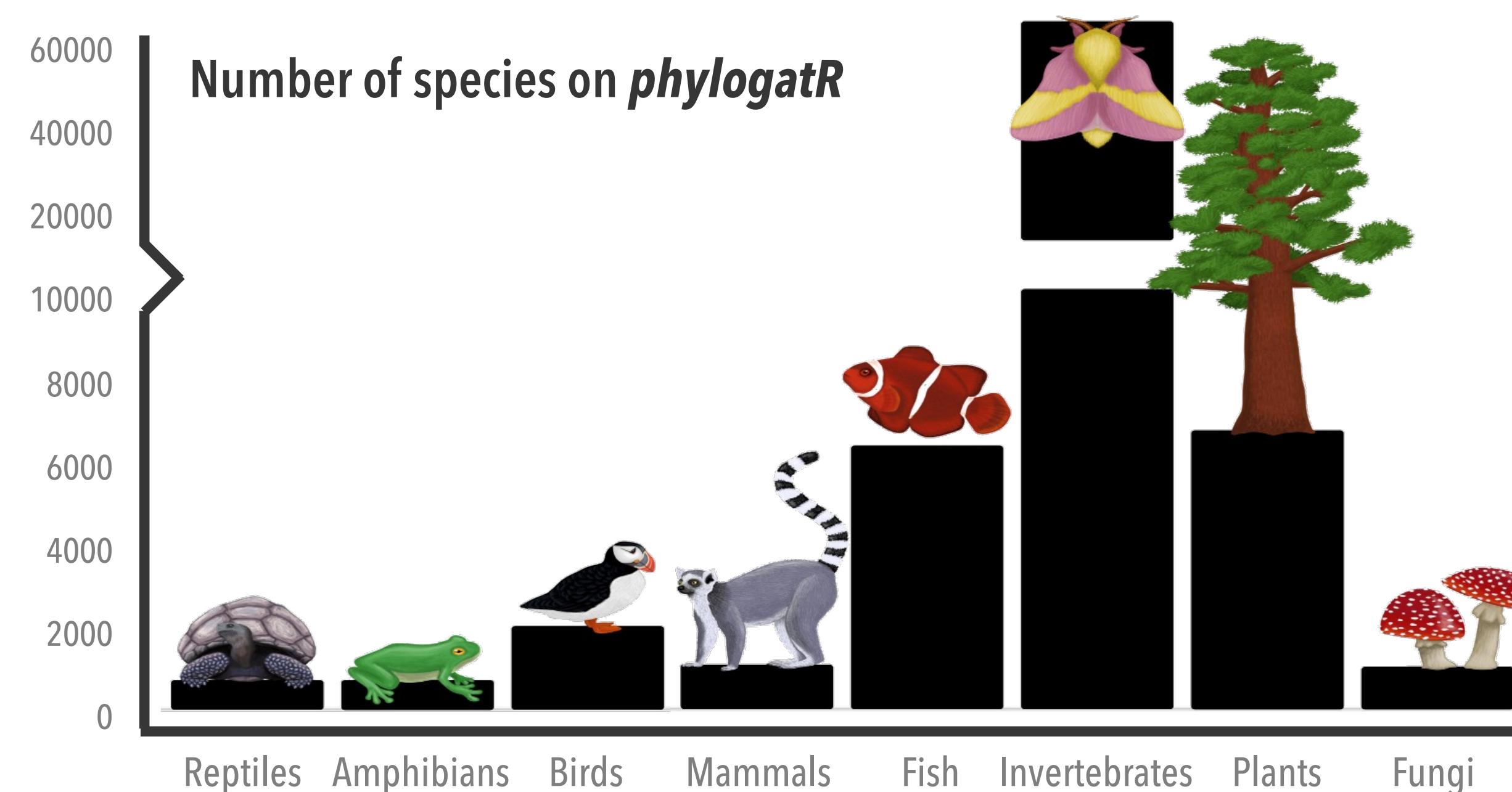
Danielle J. Parsons<sup>1</sup> | Sydney K. Decker<sup>1</sup> | Tara A. Pelletier<sup>2</sup> | Bryan C. Carstens<sup>1</sup> [1.Ohio State University 2.Radford University]

## INTRODUCTION

Patterns of genetic diversity within species contain information about the history of that species, including how it responded to historical climate change. Researchers in many disciplines collect DNA sequence data from hundreds of samples and deposit these in open-source online data repositories, such as GenBank and BOLD. The existence of georeferenced DNA sequence data in databases can enable novel comparative analyses in ecology and evolution.

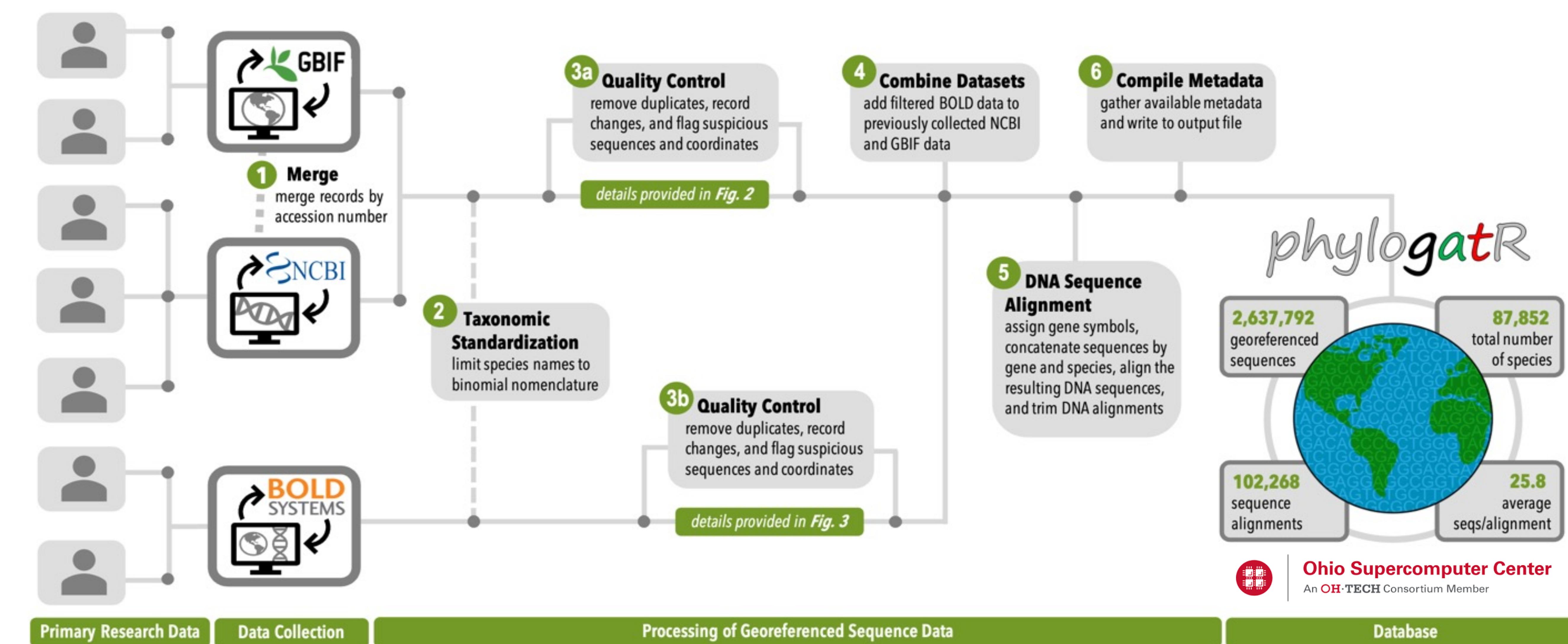
In order to facilitate these types of analyses on the largest possible scale from thousands of species, we developed software that parses data from several repositories of geographic and genetic data, organizes them under a taxonomic hierarchy, and produces data that are analysis ready.

## DATABASE AT A GLANCE



Data analysis can be conducted using R scripts or R Shiny apps from *phylogatR*. The database *phylogatR* is freely available via the Ohio Supercomputer Center

## PIPELINE



## DATABASE GOALS

[ to 1) empower students to actively learn about computer code, genetics, and biodiversity science by repurposing genetic and geographic data and 2) enable basic scientific research in a discipline that is fundamentally about global change ]

## RESULTS

### RESEARCH

MOLECULAR ECOLOGY RESOURCES

PNAS RESEARCH ARTICLE | EVOLUTION Analysis of biodiversity data suggests that mammal species are hidden in predictable places

BIOLOGY LETTERS rsl.royalsocietypublishing.org Research article Geographical range size and latitude predict population genetic structure in a global survey

Journal of Biogeography RESEARCH PAPER A global analysis of bats using automated comparative phylogeography uncovers a surprising impact of Pleistocene glaciation

Research using *phylogatR* is ongoing and includes several published projects.

### EDUCATION

Sequence alignments: what to watch out for

In the first example, I have edited the alignment to contain sequences that have been entered in reverse (Inv. 1-4) and inserted them into the alignment in the wrong position. This causes the sequences to be aligned in the wrong order and are lacking large regions of gaps. However, upon visual inspection they clearly do not line up with the original alignment. In the second example, I have added a sequence that is clearly not related to the rest of the alignment. This is a common mistake when creating alignments. PhylogatR takes steps to minimize these issues (see here), but it can never hurt to check your data.

Number of individuals per sampling locality

Educational modules and teaching resources for coding and data analysis are available.

### OUTREACH

Calculate Genetic Diversity

Now that we have a better understanding of what our dataset looks like, let's use a tool to look at our data more closely. We will use the phylogatR tool to calculate genetic diversity. In order to do this, we will need to know what is in our dataset. The first step is to upload our dataset to phylogatR. Once our dataset is uploaded, we can calculate the average value of  $\pi$  for all of the species belonging to each taxonomic category. We can then examine the results in the graph displayed below.

Number of individuals per sampling locality

Interactive web applications are available to facilitate learning at all skill levels.