phylogatR

Biodiversity informatics in the classroom: blending biology and computer science using real data via open-source repositories

**phylo**geographic data aggre**gat**ion and **r**epurposing

RADFORD UNIVERSITY

THE OHIO STATE UNIVERSITY

Tara A Pelletier[1], Danielle J Parsons[2], Sydney K Decker[2], Stephanie Crouch[1]   [1.Radford University 2.Ohio State 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.
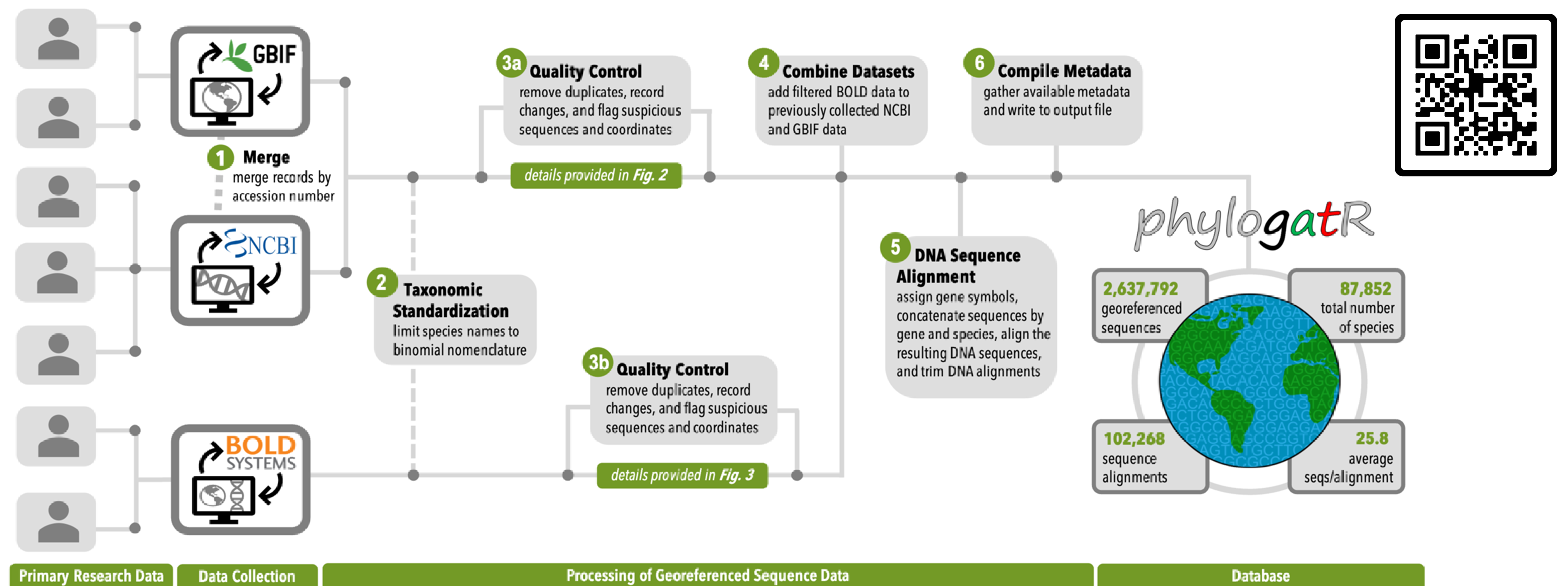
In order to facilitate these types of analyses on the largest possible scale from thousands of species, we developed software that parses data from several repositories of geographic and genetic data, organizes them under a taxonomic hierarchy, and produces data that are almost analysis ready.

Number of species on *phylogatR*
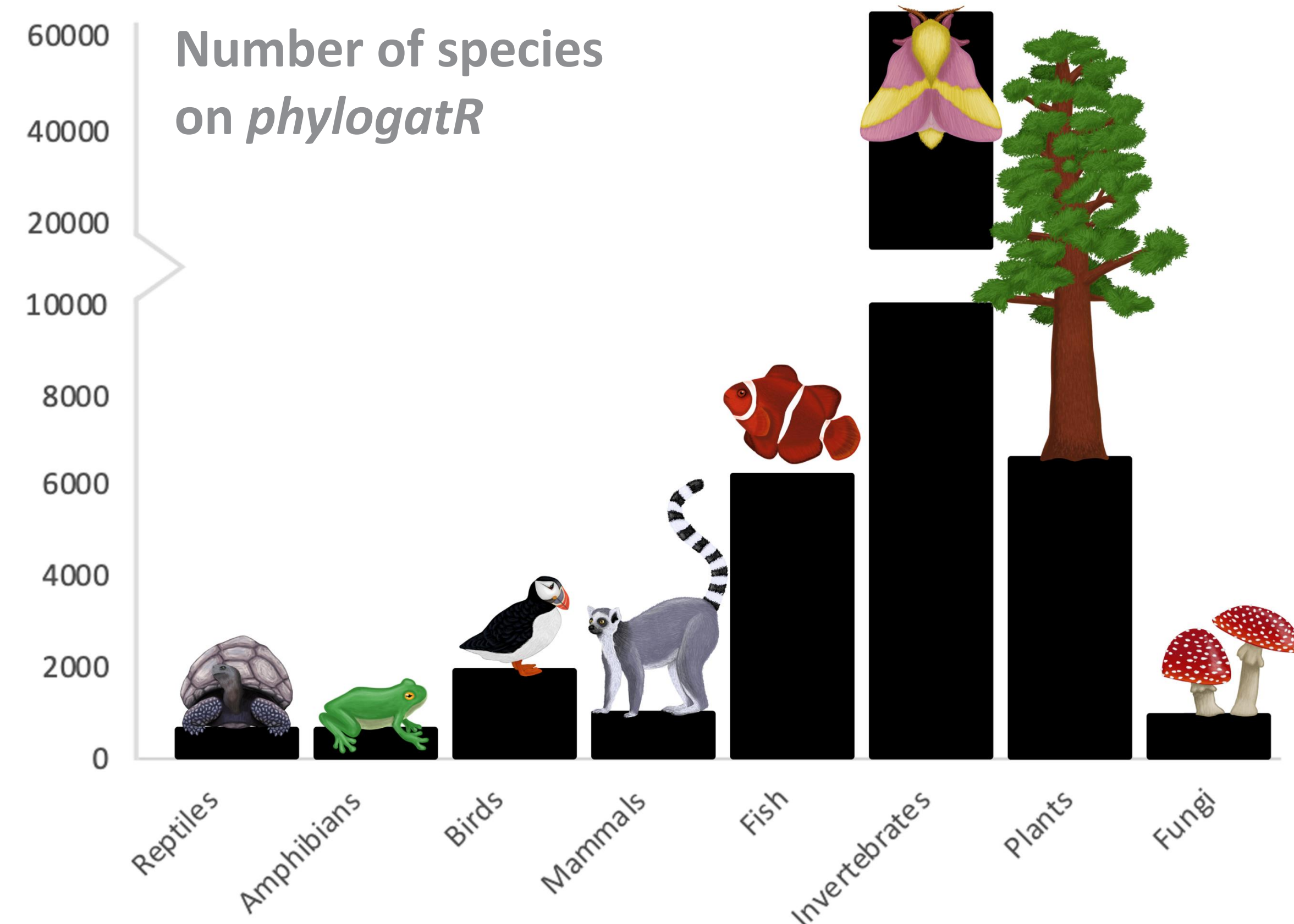
Analysis of these data can be conducted on any personal or university computer using R scripts or R Shiny apps provided by *phylogatR*. The database, *phylogatR* is freely available via the Ohio Supercomputer Center (OSC): https://phylogatr.org/

DATABASE GOALS: to **1)** empower students to actively learn about computer code, genetics, and biodiversity by repurposing genetic and climatic data and **2)** highlight basic scientific research in a discipline that is fundamentally about global change.

PIPELINE

1 Merge merge records by accession number

2 Taxonomic Standardization limit species names to binomial nomenclature

3a Quality Control remove duplicates, record changes, and flag suspicious sequences and coordinates

details provided in *Fig. 2*

3b Quality Control remove duplicates, record changes, and flag suspicious sequences and coordinates

details provided in *Fig. 3*

4 Combine Datasets add filtered BOLD data to previously collected NCBI and GBIF data

5 DNA Sequence Alignment assign gene symbols, concatenate sequences by gene and species, align the resulting DNA sequences, and trim DNA alignments

6 Compile Metadata gather available metadata and write to output file

phylogatR

2,637,792 georeferenced sequences
87,852 total number of species
102,268 sequence alignments
25.8 average seqs/alignment

Primary Research Data | Data Collection | Processing of Georeferenced Sequence Data | Database

IN THE CLASSROOM

Making data easy to access and reuse is important for researchers and educators who do not have the skills or resources for large-scale projects, or expensive and time-consuming field and lab work, increasing participation from underprivileged groups and persons from historically minoritized groups. The iterative process and data management skills necessary for handling large datasets has increased student confidence in problem solving and organizational skills.

"I believe that this class has many strengths. It allowed all of the students, regardless of background, to grasp the concepts and utilize them in a real research setting."

"This course provides excellent analytical skills that all science majors should have."

"Learning how to code has not been easy, but it will help me in my future job and will always look good on a resume`."

UNDERSTAND

APPLY

EVALUATE