Assignment 1 - BINF6210

Jesse Wolf - 0830233

2022-10-07

```
# Loading relevant packages and combining invisible with lapply
# to not print boolean statement of TRUE for each package being loaded
libs <- c("tidyverse", "ggplot2", "maps",
          "patchwork", "countrycode", "colorBlindness")
invisible (lapply(libs, require, character.only = TRUE))
```

Introduction

Members of the family *Sciuridae* are found in Asia, Africa, Europe, as well as both North and South America (Ferron 2015). In total, the *Sciuridae* (or squirrel) family comprises 262 species, with 22 of those species occurring in Canada (Ferron 2015). Among the species found in Canada, there are ground-dwelling, arboreal, and even flying species (Ferron 2015; Waterman et al. 2021). Squirrels live in habitats that vary in altitude, latitude, and can be both arid and arctic-adapted (Waterman et al. 2021). Due to the variety of environmental niches that squirrel species inhabit, I hypothesize that the geographic range of *Sciuridae* samples submitted to the BOLD (Barcode of Life Data System) database will be spread evenly across the 5 continents that they are known to inhabit. Additionally, due to the large number of species within the *Sciuridae* family and the confirmed phenomenon of hybridization in certain squirrel species (e.g., Garroway et al. 2009; Kapustina et al. 2018; Wolf et al. 2022), I hypothesize that the ratio of BINs (Barcode Index Numbers) to species will be greater than one across all continents. If confirmed, this may be indicative of possible cryptic or introgressed species. Determining the geographic distribution of *Sciuridae* species will help identify possible hybrid zones. These results will then be compared to the areas in which the BIN:species ratio exceeds 1.

Code Part 1 - Data Exploration

```
# Obtaining the raw data from the BOLD API. raw_data <-
# read_tsv('http://www.boldsystems.org/index.php/API_Public/combined?taxon=Sciuridae&format=tsv')
# write_tsv (raw_data, file = 'bold_data.txt') Importing
# raw data - using fill = TRUE as some cells are empty and
# read.csv doesn't like that, also telling read.csv which
# strings can be referred to as NA. I downloaded the data
# from BOLD but commented that line out so it wasn't
# reading in every time I ran the code.
raw_data <- as.data.frame(read.csv("Input/bold_data.txt", header = TRUE,
    sep = "\t", fill = TRUE, na.strings = c("", "NA")))

# Initially exploring the data.
str(raw_data)
```

```r
# Ensuring the Sciuridae dataset meets the criteria of >=10
# BINs and 100 records.
dim(raw_data)

# Checking to see how many unique BINs are present.
unique(raw_data$bin_uri)

# Filtering the dataset to only contain variables that will
# be used for downstream analyses and writing as a csv to
# make sure I have a file of my filtered data - this line
# has been commented out.
raw_filtered <- raw_data %>%
    select(c(species_name, genus_name, country, lat, lon, bin_uri))
# write.csv(raw_filtered,
# 'Output/Assignment1_Filtered_BOLD_data.csv')

# Exploring dataset for potential errors/biases - looks
# like there are a couple of entries in the country column
# that aren't a country (Exception - Culture and Exception
# - Zoological Park) as well as 349 NAs.
countries_bold <- raw_filtered %>%
    count(country, sort = TRUE)

# Checking all unique values of country in the filtered
# dataset.
unique(raw_filtered$country)

# Getting the lnumber of unique countries in the filtered
# dataset.
length(unique(raw_filtered$country))

# Removing Exception - Culture and Exception - Zoological
# Park from Country column and any individuals with any NA
# data.
raw_filtered_QC <- raw_filtered %>%
    filter(!grepl("Exception", country)) %>%
    drop_na()

# Exploring data to determine if our filtering step caught
# everything that may cause errors in our analyses - our
# filtering step removed 11 countries and the data has no
# NAs now.
summary(raw_filtered_QC)
str(raw_filtered_QC)
dim(raw_filtered_QC)
unique(raw_filtered_QC$country)
length(unique(raw_filtered_QC$country))

# Using the countrycode package, I created a new column
# named continent and used the package to take the country
# name and tell me which continent that country is located
# in.
raw_filtered_QC$continent <- countrycode(sourcevar = raw_filtered_QC$country,
```

```
    origin = "country.name", destination = "continent")
# Lets look at the data - we have four continents, but
# something is wrong with that; we should have five.
unique(raw_filtered_QC$continent)
length(unique(raw_filtered_QC$continent))

# One problem I can see is that the countrycode package did
# not differentiate between North and South America, so we
# can do that manually. Note, I changed Canada/USA/Mexico
# to North America FIRST, so that I can make use of the
# fact that the continent column for every country with
# 'Americas' in the continent column after the first step
# can now be changed to South America.
QC_data_continent <- within(raw_filtered_QC, continent[country ==
    "Canada" | country == "United States" | country == "Mexico"] <- "North America")
QC_data_continent <- within(QC_data_continent, continent[continent ==
    "Americas"] <- "South America")

# Let's check to make sure that worked - now we have 5
# continents, great!
unique(QC_data_continent$continent)
length(unique(QC_data_continent$continent))

# A couple of the variables (namely country/continent and
# species/genus name are being treated as characters, when
# they should be treated as factors if we are going to use
# them as grouping factors later).

# Coercing genus/species names and countries as factors
col_factors <- c("species_name", "genus_name", "country", "continent")
QC_data_continent[col_factors] <- lapply(QC_data_continent[col_factors],
    factor)

# Checking to see if the coercion to factor worked - it
# did! Now we know we have 61 species, 24 genera, 27
# countries, and 5 continents.
str(QC_data_continent)
```

Code Part 2 - Analysis - Figure 1

Creating a world map to look at distribution of data points across different continents.

```
# Creating a variable for the world map from the maps package.
world <- map_data("world")
# Using ggplot2 to create a map of the entire world and plotting all of the BOLD entries
# that passed our quality control and filtering.
figure_1 <- ggplot() +
  # Using geom_map to create a basemap for my data points.
  geom_map(
    data = world, map = world,
    aes(long, lat, map_id = region),
    # I set alpha = 0.5 for a softer/slightly more transparent look of the map.
```

```
    color = "white", fill = "lightgray", alpha = 0.5) +
  # Using geom_point to plot each data point.
  # I set alpha = 0.7 for a softer/slightly more transparent look.
  geom_point(
    data = QC_data_continent,
    aes(lon, lat),
    alpha = 0.7, size = 1.5)+
  # Removing the legend as it's not necessary.
 theme(legend.position = "none")+
  labs (y = "Latitude (°)", x = "Longitude (°)")

# Printing Figure 1 to the screen
figure_1
```

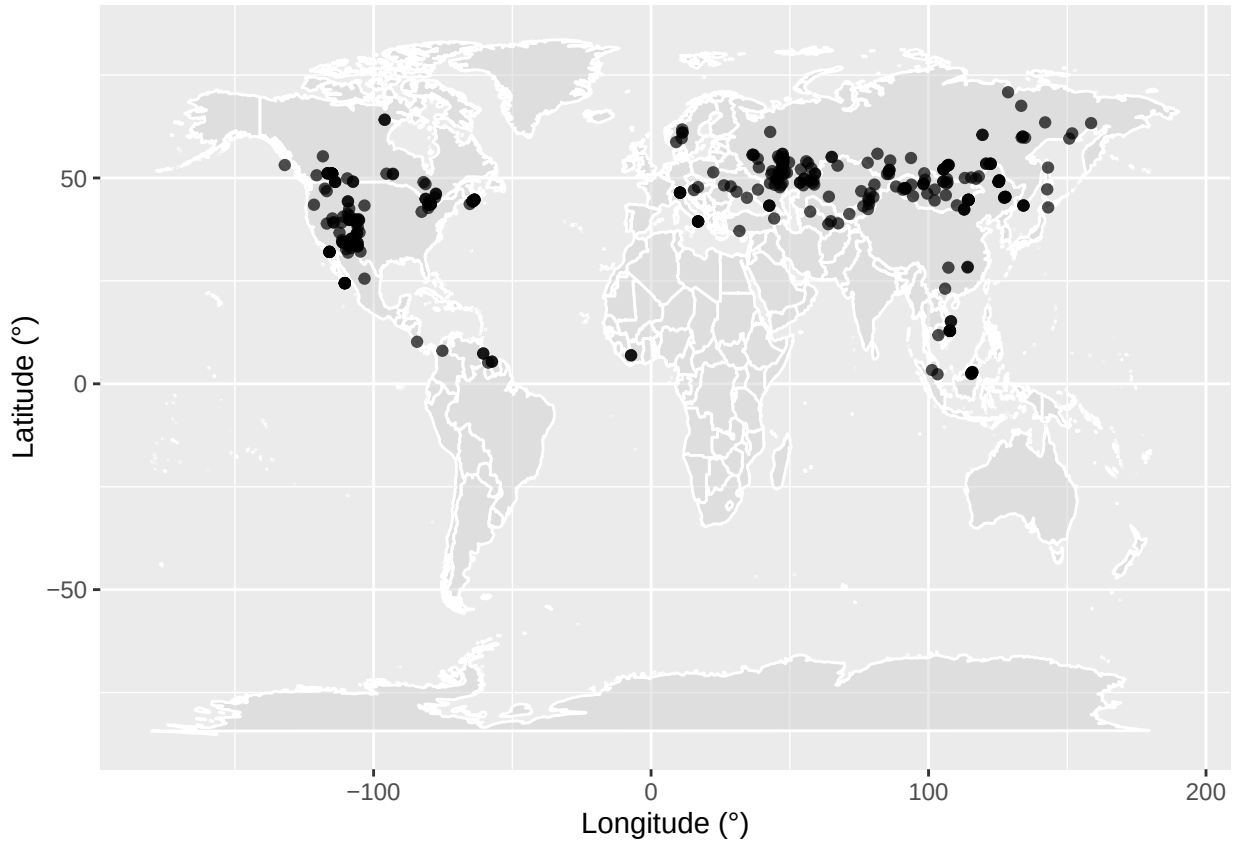

Figure 1: A world map with each of the *Sciuridae* individuals submitted to BOLD (The Barcode of Life Data System) that remained in the filtered and cleaned dataset (n=379).

Code Part 2 - Analysis - Figure 2

We can see that there is a relatively high proportion of data points in North America and Europe, with much less in Africa, Asia, and South America - but can we be certain? Maybe there's another way to view this data that is more quantitative.

```r
# Creating a data frame with number of data points per country.
QC_data_count<- QC_data_continent %>%
  count(continent, genus_name)

# Creating another data frame with number of data points per country and
# filtering out any genera with <5 observations.
QC_data_count_nfilter <- QC_data_continent %>%
  count(continent, genus_name) %>%
  filter (n>=5)

# Using ggplot to create a barplot that will show us the data more clearly.
ggbar_allgenera<- ggplot(data = QC_data_count,
                         aes(x = genus_name, y = n, fill = continent)) +
  geom_bar(stat = "identity")+
  # Manually setting the y-axis limits and breaks and extending the bar to touch
  # the x-axis line.
  scale_y_continuous(expand = c(0,0),
                     breaks = seq(0, 150, by = 15)) +
  theme(
    # Set background color to white
    panel.background = element_rect(fill = "white"),
    # Remove tick marks by setting their length to 0.
    axis.ticks.length = unit(0, "mm"),
    # Setting the X-axis and only the left line of the vertical axis to be displayed in black.
    axis.line.y.left = element_line(color = "black"),
    axis.line.x = element_line(color = "black"),
    # Customize labels for the horizontal axis and the y-axis title.
    axis.text.x = element_text(size = 8, vjust = 0.5, hjust=1),
    axis.text.y = element_text (size = 8),
    axis.title.x = element_text (size = 10))+
  # Filling using the discrete Set1 palette from the RColorBrewer package.
  scale_fill_brewer(palette = "Set1")+
    # setting my y-axis label, and fixing the legend so Continent is capitalized.
  labs (y = "Number of BOLD entries", x = "", fill = "Continent") +
  # Flipping the x and y-axis to improve aesthetic.
  coord_flip()+
  # Sorting the genus names in the x-axis alphabetically.
  scale_x_discrete(limits = c(sort(x = unique(QC_data_count$genus_name), decreasing = T)))

# Creating another graph that is identical, but using the dataframe with any genera
# that has an incidence of 5 or greater.
# To make sure the colours match the original figure, I pulled out the HEX
# codes from ggbar_allgenera and made a custom colour vector for the figure below.
allgenera_color <- c("Africa" = '#E41A1C', "Asia" = '#377EB8',
                     "Europe" = '#4DAF4A', "North America"= '#984EA3',
                     "South America" = '#FF7F00')

ggbar_nfilter<- ggplot(data = QC_data_count_nfilter,
                       aes(x = genus_name, y = n, fill = continent)) +
  geom_bar(stat = "identity")+
  # Manually setting the y-axis limits and breaks and extending the bar to touch
  # the x-axis line.
  scale_y_continuous(expand = c(0,0),
```

```
                    breaks = seq(0, 150, by = 15), ) +
  theme(
    # Set background color to white.
    panel.background = element_rect(fill = "white"),
    # Remove tick marks by setting their length to 0.
    axis.ticks.length = unit(0, "mm"),
    # X-axis and only the left line of the vertical axis is painted in black.
    axis.line.y.left = element_line(color = "black"),
    axis.line.x = element_line(color = "black"),
    # Customize labels for the horizontal axis and the y-axis title.
    axis.text.x = element_text(size = 7, vjust = 0.5, hjust=1),
    # Creating a border around the inset figure.
    panel.border = element_rect(colour = "black", fill=NA, size=1),
    axis.text.y = element_text (size = 7))+
  # Filling using the discrete palette I created to make sure it matches
  # ggbar_allgenera from Figure 2A.
  scale_fill_manual(values = allgenera_color)+
  # Suppressing the legend and axes labels as the combined figure does not need
  # 2 separate legends and including the axes again would be redundant.
  theme(legend.position = "none")+
  labs (y = "", x = "")+
  # Flipping the x and y-axis to improve aesthetic.
  coord_flip()+
  # Sorting the genus names in the x-axis alphabetically.
  scale_x_discrete(limits = c(sort(x = unique(QC_data_count_nfilter$genus_name),
                                   decreasing = T)))

# Combining both plots using the patchwork package and
# setting the theme for my size/typeface of the title and subtitle and
# center-aligning the text.
figure_2 <- ggbar_allgenera + inset_element(ggbar_nfilter, 0.1, 0.38, 1, 1) +
  plot_annotation(title = ("Figure 2: Sciuridae Genera by Continent"),
    subtitle = "Source: BOLD: The Barcode of Life Data System") &
  theme(
    plot.title = element_text(
      face = "bold",
      size = 12,
      hjust = 0.5),
    plot.subtitle = element_text(
      size = 10,
      hjust = 0.5))
# Printing Figure 2 to the screen.
figure_2
```

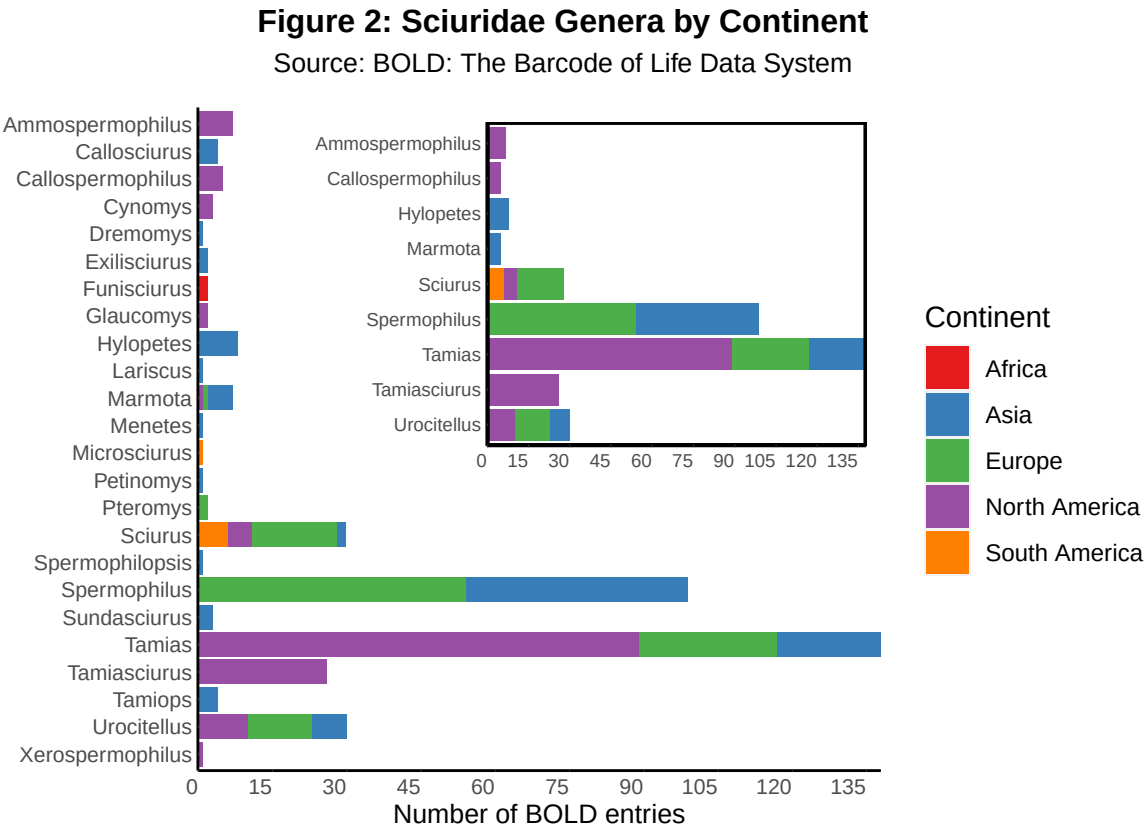

Figure 2: Each stacked bar represents a total count of individuals within a given genera that were submitted to BOLD (The Barcode of Life Data System) and is separated by colour, corresponding to a specific continent. The outset figure depicts the number of BOLD entries of the *Sciuridae* genera that passed our data filtering and quality control (n=379). The inset figure visualizes the number of BOLD entries from a genera that had a sample size of greater than or equal to 5 (n=346), thus giving us a better idea of the dominant genera within our data.

Code Part 2 - Analysis - Checking for accessibility (colour-blind friendly) of figure 2

```
# Using tool from package colorBlidness to simulate the
# color vision deficiency - CVD simulator.  This is a
# preliminary figure, purely for exploring the data.
cvdPlot(ggbar_allgenera)
```

```
# It looks like our figure is accessible for individuals
# with deuteranopia and protanopia (the most common forms
# of red-green colour blindness).
```

Code Part 2 - Analysis - Figure 3

Creating a map to visualize the BINs:species ratio across continents.

```r
# Reminding myself of the structure of our cleaned and filtered dataset.
str (QC_data_continent)

# Creating a new data frame from QC_data_continent that will include and
# summarize the BIN:species ratio by continent and generate the average coordinates for
# each continent using the n_distinct function in dplyr and finally create a data frame
# with one row per continent.
bin_species_ratio <- QC_data_continent %>%
  group_by (continent) %>%
  mutate (centroid_long = mean(lon))%>%
  mutate (centroid_lat = mean (lat)) %>%
  summarise (ratio = (n_distinct(bin_uri)/n_distinct(species_name)), across()) %>%
  distinct (continent, ratio, centroid_long, centroid_lat)

# Checking to make sure the summarise and mutate functions only give us
# 5 unique values each - one per continent.
unique(bin_species_ratio$ratio)
unique (bin_species_ratio$centroid_lat)
unique (bin_species_ratio$centroid_long)

# Generating a map that shows BIN:Species ratio by continent and
# symbolizes the size of the points based on the size of the ratio.
figure_3 <- ggplot() +
  # Using geom_map to create a basemap for my data points.
  geom_map(
    data = world, map = world,
    aes(long, lat, map_id = region),
    # I set alpha = 0.5 for a softer/slightly more transparent look of the map.
    color = "white", fill = "darkgray", alpha = 0.5) +
  # Using the same colour palette I generated for the inset for figure 2.
scale_colour_manual(values = allgenera_color)+
  # Using geom_point to plot each data point and using the BINs:species ratio to
  # change the size of the data point.
  # I set the colour to reflect the continent the data point is from and set alpha = 0.7
  # for a softer/slightly more transparent look.
  geom_point(
    data = bin_species_ratio,
    aes(x = centroid_long, y = centroid_lat,
        size = ratio,
        colour = continent,
        alpha = 0.7))+
  # Modifying the scale_size_binned parameter to make it more graphically clear the
  # variance in size of the ratio of BINs:Species.
  scale_size_binned(range = c(2,10), breaks = waiver(), n.breaks = 10)+
  # Adding a custom title and subtitle and setting my axes labels.
  labs(y = "Latitude (°)", x = "Longitude (°)",
       size = "BIN:Species ratio", colour = "Continent",
    title = ("Figure 3: Ratio of BINs to Species by Continent"),
    subtitle = "Source: BOLD: The Barcode of Life Data System (www.barcodinglife.org)") +
  # Setting the theme for my size/typeface of the title and subtitle and center-aligning the text
  theme(
    plot.title = element_text(
      face = "bold",
```

```
      size = 12,
      hjust = 0.5),
    plot.subtitle = element_text(
      size = 10,
      hjust = 0.5))+
  # Removing the size legend as it's not relevant in the scope of this research.
  guides (size = FALSE, alpha = FALSE)
# Printing figure 3 to the screen.
figure_3
```

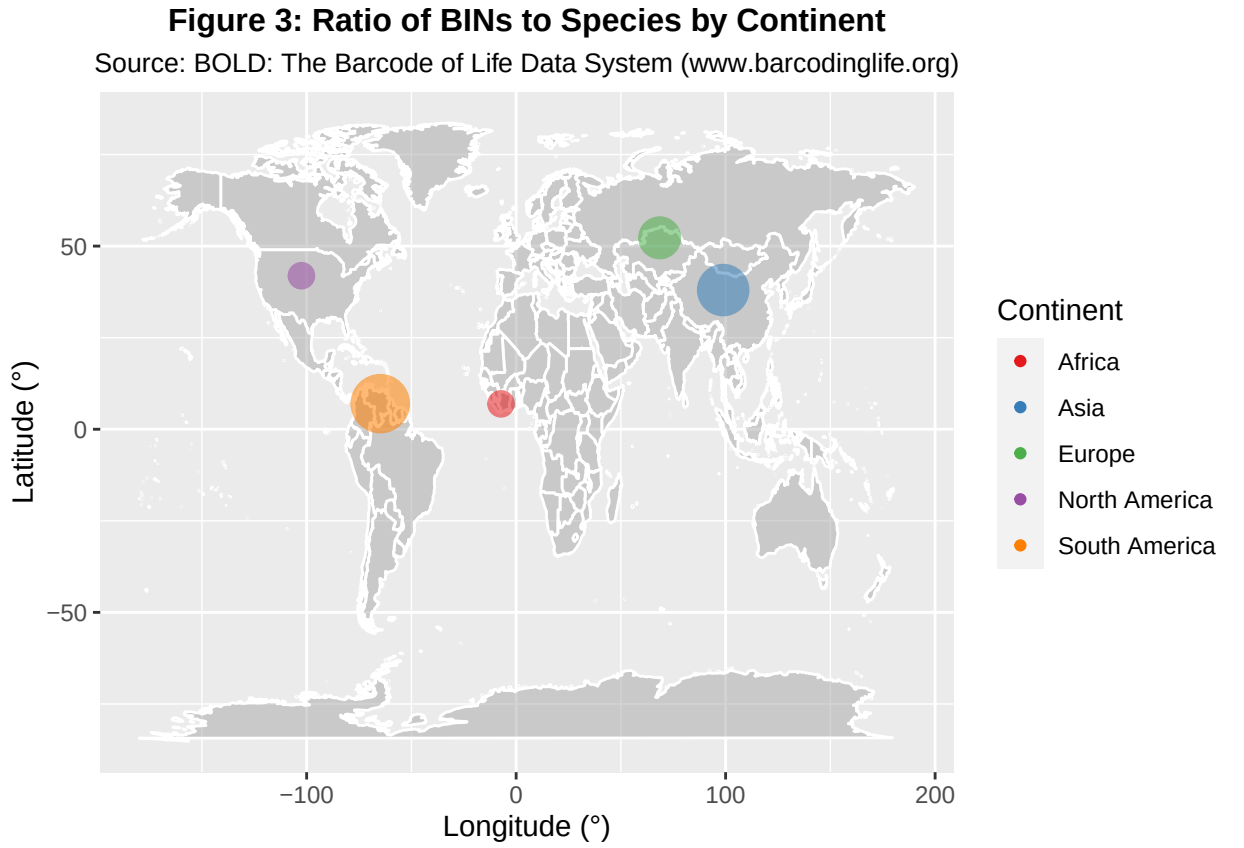

Figure 3: The ratio of unique BIN (Barcode Index Number) assignments to unique number of species are broken down by continent. The colour corresponds to the continent, while the size of the point corresponds to the size of the ratio (higher ratios indicate larger unique number of BINs relative to unique number of species.)

Discussion and Conclusion

Many sciurids are considered keystone species due to their ability to create new habitat niches, and as such, the identification of possible new species is an important area of research (Waterman et al. 2021). When looking at the 24 unique genera within the family *Sciuridae*, it is clear that while individuals are located on each of the 5 continents within their known range (Figure 1), there are a greater number of individuals submitted to BOLD from Asia, Europe, and North America (Figure 2). In fact, among genera with >=5 individuals submitted to BOLD, there is only one from South America (Genus *Sciurus*; Figure 2). As such,

this might support the conclusion that there is a larger diversity of genera in Asia, Europe, and North America.

Amori et al. (2009) discovered that the total diversity of the genera *Sciuridae* was inversely correlated to latitude. To assess the relationship between the number of genera and the number of BINs in the BOLD database relative to geographic location, I generated a ratio that divided the number of unique BINs to the number of unique species in a given continent. To display this data graphically, a map was generated depicting these results. Interestingly, we can see that while all 5 continents have a BIN:Species ratio of >=1, there are a gradient of values (Figure 3). It was noted that the highest BIN:Species ratio was seen in South America, followed by Asia and Europe respectively. It is important to note that the final dataset obtained via BOLD and used in this study only has a single genus from South America. As such, the large BIN:Species ratio may be a product of our small sample size. However, this is still possibly indicative of cryptic or introgressed individuals among South American *Sciuridae* genera and warrants investigation. The findings presented here provide further motivation for investigating the presence of cryptic or possibly introgressed regions and highlights the large-scale geographic areas that may require more focus than others. Identifying finer-scale areas that have relatively high BIN:Species ratios would help facilitate a more targeted approach in distributing resources to investigate possibly crpytic or introgressed species.

Acknowledgements

I would like to thank Isadora Bischoff Nunes, Alvaro De la Mora Pena, Linoy Jacobs, and Thomas Papp-Simon for their feedback and constructive criticism of the analyses and figures presented here.

References Cited

Amori G, Gippoliti S, Luisell L, & Battisti C. Are there latitudinal gradients in taxa turnover? A worldwide study with Sciuridae (Mammalia: Rodentia). Comm. Ecol. 2010;11(1): 22-26. doi: 10.1556/ComEc.11.2010.1.4

Broman K. Knitr with R Markdown. https://kbroman.org/knitr_knutshell/pages/Rmarkdown.html. Accessed 26 September 2022.

Ferron, J "Squirrel". The Canadian Encyclopedia, 04 March 2015, Historica Canada.

www.thecanadianencyclopedia.ca/en/article/squirrel. Accessed 26 September 2022.

Garroway C, Bowman J, Cascaden TJ, Holloway GL, Mahan CG, Malcolm JR, Steele MA, Turner G, and Wilson PJ. Climate change induced hybridization in flying squirrels. Glob. Chang. Biol. 2010;16: 113 - 121. doi: 10.1111/j.1365-2486.2009.01948.x.

Get continent name from country name in R. https://stackoverflow.com/questions/47510141/get-continent-name-from-country-name-in-r. Accessed 26 September 2022.

How to Filter Rows that Contain a Certain String Using dplyr. https://www.statology.org/filter-rows-that-contain-string-dplyr/. Accessed 26 September 2022.

How To Make World Map with ggplot2 in R? https://datavizpyr.com/how-to-make-world-map-with-ggplot2-in-r/. Accessed 26 September 2022.

Kapustina SY, Lyapunova EA, Adiya Y, & Brandler OV. Features of Interspecific Contacts and Hybridization of Ground Squirrels (Marmotinae, Sciuridae, Rodentia) in Mongolia. Dokl. Biochem. Biophys. 2018;482(1):275-278. doi: 10.1134/S1607672918050125.

Pedersen, TL. Create an inset to be added on top of the previous plot. https://patchwork.data-imaginist.com/reference/inset_element.html. Accessed 04 October 2022.

Ou J. colorBlindness Guide. https://cran.r-project.org/web/packages/colorBlindness/vignettes/colorBlindness.html. Accessed 26 September 2022.

Waterman JM, Gossmann TI, Brandler O and Koprowski JL (2021) Editorial: Ecological, Behavioral and Genomic Consequences in the Rodent Family Sciuridae: Why Are Squirrels So Diverse? Front. Ecol. Evol. 9:765558. doi: 10.3389/fevo.2021.765558

Wickham H, Navarro D, & Pedersen TL. ggplot2: Elegant Graphics for Data Analysis: Chapter 11: Colour scales and legends. https://ggplot2-book.org/scale-colour.html#brewer-scales. Accessed 26 September 2022.

Wolf JF, Bowman J, Keobouasone S, Taylor RS, & Wilson PJ. A de novo genome assembly and annotation of the southern flying squirrel (Glaucomys volans). G3 Genes. Genomes. Genet. 2022;12(1). 2022. jkab373, doi: 10.1093/g3journal/jkab373