Arvind Srinivas
1301477
BINF*6210

Uncovering the Link Between Geographic Speciation and Evolutionary Relatedness Within the Genus Anas

Introduction:

Ever since Darwin introduced the concept of evolution, the field of phylogeny has taken great strides in uncovering natural history and the relatedness between species. Moreover, the development of genomic sequencing techniques has reconstructed evolutionary history (Sleator, 2011). Phylogenetic analyses allow scientists to reconstruct the historical branching patterns of species, elucidating the common ancestry and shared evolutionary history that underlie the diversity of life on Earth. This information is fundamental for numerous scientific disciplines, including evolutionary biology, ecology, and genetics. Furthermore, the study of phylogeny is integral to biodiversity conservation and can guide efforts to preserve evolutionary distinctiveness and ecosystems (Rolland et al., 2012).

Phylogeny is applied to various other fields, such as geography. Looking at how species evolve in a geographic lens may explain speciation events (Abellan & Ribera., 2011). For instance, allopatric speciation is when species separately evolve due to geographic isolation and/or physical barriers (Yamaguchi & Iwasa, 2013). Conversely, there would be instances where sister species for a particular taxon inhabit the same geographic region, fostering close ecological interactions. This nuanced exploration of geographic associations provides a comprehensive perspective on the interconnectedness of evolutionary processes and spatial dynamics.

The objective of this project is to examine whether sister species in the Anas genus (ducks) live in the same geographic region or in different geographic regions? Following this, what can we infer about the evolution of this genus? If sister species inhabit the same region, it suggests factors such as ecological interactions and shared environments influenced their evolution. On the other hand, if they occupy different regions, this could imply large-scale allopatric divergence, where geographic isolation played a significant role in shaping evolutionary pathways. The marker gene Cytochrome B (CytB) was used for the Anas dataset, as it is widely used in phylogenetics for vertebrates (Farias et al., 2001). A phylogenetic relationship will be formed, along with mapping it to a world map to show regions in which each sister species inhabits.

Description of Datasets:

Arvind Srinivas
1301477
BINF*6210

Dataset of NCBI Nucleotide Data from the Genus Anas (CytB): Nucleotide data for the CytB gene in Anas was retrieved through functions from the "rentrez" package. R was used to extract this information on Dec 1, 2023. The data contained the ID and sequence for each species within the NCBI Anas data. Eventually, these sequences were filtered and had many quality control checks on it. There were initially 104 hits, but after the filtering steps, 11 sequences from sister species were retained. This is a modest dataset that can make any visualizations easy to interpret. The nucleotide data was needed to perform sequence alignments and hierarchical clustering, to construct phylogenetic relationships.

BOLD Dataset to Match Geographical Data to Species Name: A BOLD API tool was used to extract data from the BOLDSystems website. The Anas data frame was retrieved on Dec 1, 2023 and had dimensions of 500x80. Only 3 variables are of interest from the data frame: "lat", "lon", and "species_name". The latitude and longitude values were used to construct map points, while the species names from this data frame were matched with the NCBI data frame.

Code Section 1: Data Acquisition, Exploration, Filtering, and Quality Control

```
#Load packages.
library(tidyverse)
library(ape)
library(vegan)
library(rentrez)
library(Biostrings)
library(ggplot2)
library(picante)
library(muscle)
library(phyloseq)
library(ggtree)
library(maps)
library(phytools)
library(DECIPHER)


####CODE PART 1: DATA ACQUISITION, EXPLORATION, FILTERING, AND QUALITY
CONTROL ----


#PART 1.1: DATA ACQUISITION ----

#Search for any categories called "nuccore" to confirm it is present in NCBI.
entrez_db_searchable("nuccore")
```

```
#Obtain NCBI data for taxonomy group "Anas" for the marker gene CytB.
search_query <- "Anas[ORGN] AND CytB[GENE]"
search_result <- entrez_search(db = "nuccore", term = search_query)
search_result

#Data exploration:
class(search_result)
str(search_result)

#Counting the number of hits for the search and assigning it to a variable
called max_hits.
max_hits <- search_result$count
max_hits

#Returning max number of hits for search_query in the nuccore database.
search_result1 <- entrez_search(db = "nuccore", term = search_query, retmax =
max_hits)

search_result1

#Data exploration:
class(search_result1)
str(search_result1)

#Obtaining the records for all IDs and getting a fasta file of the data as an
output.
sequence_data <- entrez_fetch(db = "nuccore", id = search_result1$ids,
rettype = "fasta")

#Writing data to the present working directory.
write(sequence_data, file = "Anas_Nuc_Data.fasta", sep = "\n")


#PART 1.2: FILTERING DATASET FROM NCBI ----

#From Biostrings package, transform file into a DNAStringSet object.
Anas_Nuc_Stringset <- readDNAStringSet("Anas_Nuc_Data.fasta")

#Data exploration:
class(Anas_Nuc_Stringset)
str(Anas_Nuc_Stringset)
head(Anas_Nuc_Stringset)

#Transforming the data into a data frame.
Df_Anas <- data.frame(Title = names(Anas_Nuc_Stringset), Sequence =
paste(Anas_Nuc_Stringset))
```

```r
view(Df_Anas) #Viewing data frame.

#Adding species column.
Df_Anas$Species_Name <- word(Df_Anas$Title, 2L, 3L)

#Create a new data frame to contain only the columns of interest for
downstream analysis.
Df_Anas <- Df_Anas[, c("Title", "Species_Name", "Sequence")]
view(Df_Anas)


#PART 1.3: FILTERING THE DATAFRAME AND PERFORMING QUALITY CONTROL ----
#It is important to filter the data frame. Filtering a DNA data frame is
crucial for isolating pertinent genetic information, ensuring accuracy in
genomic analyses. It enhances computational efficiency, focusing analyses on
specific genetic markers or sequences of interest, thereby optimizing
insights derived from genomic datasets.

#Filtering the Df_Anas data frame for missing nucleotide data.
#Below tidyverse code block selects specific columns, filters out rows with
missing genetic sequences, cleans up sequences by removing leading, trailing,
and hyphen characters, and further filters out sequences with more than 1%
ambiguous nucleotides ('N').
#The resulting data frame is refined, containing clean and relevant genetic
information suitable for downstream genomic analyses.
Df_Anas_CytB <- Df_Anas %>%
  filter(!is.na(Sequence)) %>%
  mutate(Sequence2 = str_remove_all(Sequence, "^N+|N+$|-")) %>%
  filter(str_count(Sequence2, "N") <= (0.01 * str_count(Sequence))) %>%
  mutate(Stringcount = str_count(Sequence2))

#A comprehensive exploration of the "Df_Anas_CytB" data frame.
summary(Df_Anas_CytB)
glimpse(Df_Anas_CytB)
dim(Df_Anas_CytB)
unique(Df_Anas_CytB$Species_Name)
sum(is.na(Df_Anas_CytB$Sequence2))
view(Df_Anas_CytB)

#Checking if there are Ns and "-".
sum(str_count(Df_Anas$Sequence2, "N"))
sum(str_count(Df_Anas$Sequence2, "-"))

#Creating a box plot and histogram for sequence length distribution. First
thing that was done is filtering out complete sequences as the CytB sequence
is about 800-1200 bases long.
#We don't want the plot to be less visually appealing to the eye due to
```

```
distribution skews.
max(Df_Anas_CytB$Stringcount)

Filtered_Df_Anas <- subset(Df_Anas_CytB, Stringcount < 18000)

ggplot(Filtered_Df_Anas, aes(x = Stringcount)) +
  geom_histogram(binwidth = 700, fill = "orange", color = "black", alpha =
0.8) +
  labs(x = "Sequence Length (bp)", y = "Frequency", title = "Sequence Length
Distribution for CytB Gene in Anas") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.text = element_text(size = 12),
    axis.title = element_text(size = 14, face = "bold")
  )

#The plot looks relatively good with most of the distribution between 1000-
1200 bp (which is the average size of a cds for CytB).
# There are some outliers that we can investigate, such as sequences >2000
and <1000 base pairs.
Df_Long_Sequences_Anas <- Df_Anas_CytB %>%
  filter(Stringcount > 2000)

view(Df_Long_Sequences_Anas)

Df_Short_Sequences_Anas <- Df_Anas_CytB %>%
  filter(Stringcount < 1000)

view(Df_Short_Sequences_Anas)

#It appears that there are some sequences that are outliers.
#Upon viewing both the Df_Long_Sequences_Anas and Df_Short_Sequences_Anas
data frames, it can be seen that there are the complete mitochondrial genome
for Df_Long_Sequences_Anas and partial sequences for Df_Short_Sequences_Anas.

#Making a data frame of filtered data containing sequences from 1010-1200bp.
Df_Anas_Sequences <- Df_Anas_CytB %>%
  filter(Stringcount >= 1010 & Stringcount <= 1200)

#Data exploration:
dim(Df_Anas_Sequences)
head(Df_Anas_Sequences)
view(Df_Anas_Sequences)
summary(nchar(Df_Anas_Sequences$Sequence2))

#Plotting a new histogram to observe if the data has been filtered between
```

```
1010bp and 1200 bp.
ggplot(Df_Anas_Sequences, aes(x = as.factor(Stringcount))) +
  geom_histogram(stat = "count", fill = "orange", color = "black", alpha =
0.8, width = 0.1) +
  labs(x = "Sequence Length (bp)", y = "Frequency", title = "Sequence Length
Distribution for CytB Gene in Anas") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.text = element_text(size = 12),
    axis.title = element_text(size = 14, face = "bold")
  )
```

Main Software Tools Used:

In my project, I employed the "phytools" software tool for visualizing Anas species distributions. This choice was driven by the desire to integrate phylogenetic trees with species distribution, aligning seamlessly with the project's objectives. "phytools" is a versatile R package, offering an array of functions, including phylogenetic mapping and geophylogenies. Its notable strength lies in the flexibility to tailor plots and create insightful geophylogenetic visualizations. However, a limitation surfaced in the absence of clear instructions from the author, Liam Revell, on creating lines from phylogeny tips to individual points. This feature would have facilitated assessing whether species evolve in similar geographic areas. I consulted the vignette, but it claims that the function "linklabel" only works if the direction of the phylogeny and map are either rightwards or leftwards. This was problematic as my phylogeny would be cut off if it was rightwards or leftwards. Alternative packages, such as "geojsonio", was considered but a dendrogram could not be plotted alongside the map. Despite this, the "phytools" vignette remains a valuable resource, which is cited in the references section.

Code Section 2: Data Analysis

```
####CODE PART 2: DATA ANALYSIS ----

#PART 2.1: SEQUENCE ALIGNMENT ----

#Converting the data frame "Df_Anas_Sequences" to a new data frame named
"Df_Anas_Sequences," then transforming the "Sequence2" column into a
DNAStringSet object, and finally assigning names to the elements of the
DNAStringSet based on the values in the "Species_Name" column.
Df_Anas_Sequences <- as.data.frame(Df_Anas_Sequences)
Df_Anas_Sequences$Sequence2 <- DNAStringSet(Df_Anas_Sequences$Sequence2)
names(Df_Anas_Sequences$Sequence2) <- Df_Anas_Sequences$Species_Name
```

```r
#Aligning sequences using muscle package. Gap penalty of -10000 is applied.
Anas_Alignment <- DNAStringSet(muscle::muscle(Df_Anas_Sequences$Sequence2,
gap.open = -10000))

#Writing to file to check for premature stop codons on MEGA.
writeXStringSet(Anas_Alignment, file = "Anas_Alignment.fas", format =
"fasta")

#Viewing alignment in browser.
BrowseSeqs(Anas_Alignment)


#PART 2.2: CLUSTERING AND DENDROGRAMS ----

#DNA sequence alignment dataset "Anas_Alignment" is converted into a DNA
sequence binary format ("DNA_Bin_Anas") using the "as.DNAbin" function.
Subsequently, a pairwise genetic distance matrix ("Distance_Matrix_Anas") is
computed from the binary DNA sequences using the Kimura 2-parameter model
("K80") with pairwise deletion of missing data.
DNA_Bin_Anas <- as.DNAbin(Anas_Alignment)

Distance_Matrix_Anas <- dist.dna(DNA_Bin_Anas, model = "K80", as.matrix =
TRUE, pairwise.deletion = TRUE)

#Creating a hierarchical clustering object and plotting the dendrogram.
Hc_Result <- hclust(as.dist(Distance_Matrix_Anas), method = "single")

plot(Hc_Result, main = "Cluster Dendrogram for Anas Species Based on the CytB
Gene")

#Converting the hierarchical clustering object to a phylo type object.
Dendrogram_Anas <- as.phylo(Hc_Result)

#Generating a circular phylogenetic tree visualization ("Circular_Tree_Anas")
from a dendrogram ("Dendrogram_Anas") using the 'ggtree' package in R. The
tree is adorned with tip labels, tip points in orange, and an overall title,
presenting a circular layout for Anas sequences based on the CytB gene.
Circular_Phylogram <- ggtree(Dendrogram_Anas, layout = "circular") +
  geom_tiplab(aes(label = label), size = 3.5, angle = 0, hjust = -0.13) +
  geom_tippoint(size = 1) +
  geom_tree(aes(color = label)) +
  ggtitle("Circular Phylogram for Anas Sequences (CytB Gene)") +
  theme(plot.title = element_text(hjust = 0.8)) +
  theme(legend.position = "bottom")

print(Circular_Phylogram)
```

```
#PART 2.3: OBTAINING GEOGRAPHIC DATA FOR ANAS FROM BOLD ----

#Obtain Anas data from BOLD.
#Using the API tool.
url <-
'http://www.boldsystems.org/index.php/API_Public/combined?taxon=Anas&format=tsv'

#Reading the file and storing it into a variable.
Anas_BOLD <- read_tsv(url)
head(Anas_BOLD)

#Checking class and dimensions.
class(Anas_BOLD)
dim(Anas_BOLD)

#Checking if relevant columns are present.
Column_Check <- c("lat", "lon")
stopifnot(all(Column_Check %in% colnames(Anas_BOLD))) #Shows that these
column names are present.

#Writing to a file.
write_tsv(Anas_BOLD, "Df_Anas_BOLD.tsv")

#Merging two data frames, "Df_Anas_Sequences" and "Anas_BOLD," using the
"Species_Name" column from the former ("by.x") and the "species_name" column
from the latter ("by.y"). The result, stored in the "Merged_Data" variable,
combines information from both data frames based on matching values in their
specified columns, creating a new data frame with a unified set of variables.
Merged_Data <- merge(Df_Anas_Sequences, Anas_BOLD, by.x = "Species_Name",
by.y = "species_name")

#Data exploration.
class(Merged_Data)
view(Merged_Data)

#Creating data frame called"Relevant_Data" by extracting specific columns
("Species_Name," "lat," and "lon") from the previously merged data frame
"Merged_Data." The resulting data frame retains only the relevant information
related to species names, latitude ("lat"), and longitude ("lon").
Relevant_Data <- Merged_Data[, c("Species_Name", "lat", "lon")]

#Removing NA values.
Relevant_Data1 <- Relevant_Data[complete.cases(Relevant_Data[, c("lat",
"lon")]), ]
```

```
view(Relevant_Data1)

#Using the 'ggplot2' package in R to create a world map visualization. I
defined the world map data using "map_data("world")". Then, I plotted the
world map with white polygons and black borders. Next, it overlays points on
the map using latitude ("lat") and longitude ("lon") coordinates from the
"Relevant_Data1" data frame, coloring the points based on the "Species_Name"
variable. IT IS IMPORTANT TO NOTE THAT THREE SPECIES ARE MISSING FROM THIS
MAP: Anas puna, Anas hottentota, and Anas sparsa. This is because BOLD did
not have data for these species. This is the link to the maps package:
https://www.rdocumentation.org/packages/maps/versions/3.4.1.1.
world <- map_data("world")

ggplot() + geom_polygon(data = world, aes(x = long, y= lat, group = group),
fill = "lightgray", color = "black") +
  geom_point(data = Relevant_Data1, aes(x = lon, y = lat, color =
Species_Name), size = 3, alpha = 0.7) + scale_color_discrete(name =
"Species") + labs(title = "Distribution of Anas Species Around the World",
xlab = "Longitude", ylab = "Latitude", ) + theme(legend.position = "bottom")
+
  theme(plot.title = element_text(hjust = 0.5))


#PART 2.4: CREATING A GEOPHYLOGENY PLOT ----

#Confirming missing tips from the BOLD data. Identifying species names
present in the phylogenetic tree but not in the provided data frame, storing
them in the variable Missing_Tips.
Missing_Tips <- setdiff(Dendrogram_Anas$tip.label,
Relevant_Data1$Species_Name)

#Dropping the missing tips from the tree.
Dendrogram_Anas <- drop.tip(Dendrogram_Anas, Missing_Tips)

#Extracting latitude and longitude columns from Relevant_Data1.
Lat_Lon <- Relevant_Data1[, c("lat", "lon")]

#Converting the data frame Lat_Lon into a matrix. The Phytools package needs
this data in a matrix format.
Lat_Lon_Matrix <- as.matrix(Lat_Lon)

#Assigning the values in the Species_Name column of the data frame
`Relevant_Data1` as row names to the matrix Lat_Lon_Matrix`.
rownames(Lat_Lon_Matrix) <- Relevant_Data1$Species_Name

#Using the Viridis package to create a color palette; I chose color scheme
```

Arvind Srinivas
1301477
BINF*6210

```
"H". More information can be found here: http://blog.phytools.org/2019/03/a-follow-up-comment-on-phylotomap-for.html
Cols <- setNames(sample(viridis(n = Ntip(Dendrogram_Anas), option = "H")), Dendrogram_Anas$tip.label)

#Generating a geographical map based on the phylogenetic tree Dendrogram_Anas and the matrix of latitude and longitude values (Lat_Lon_Matrix).
Map <- phylo.to.map(Dendrogram_Anas, Lat_Lon_Matrix, plot = FALSE)

#Plotting the map with species-specific colours.
plot(Map, colors = Cols, cex.points = c(0, 1), lwd = c(2,1), ftype = "i")
title("Geophylogeny Plot for Anas Species", line = -1.5)

#Extracting species names from hex color codes. Using the gsub() command. More information can be found here: https://www.codecademy.com/learn/working-with-data-in-r-skill-path/modules/learn-r-data-cleaning/cheatsheet#.
species_names <- gsub("#[A-Fa-f0-9]+$", "", names(cols))

#Adding a legend to plot.
legend("bottomleft", legend = species_names, fill = Cols, pch = 16, cex = 0.6, title = "Species")
```

Results:

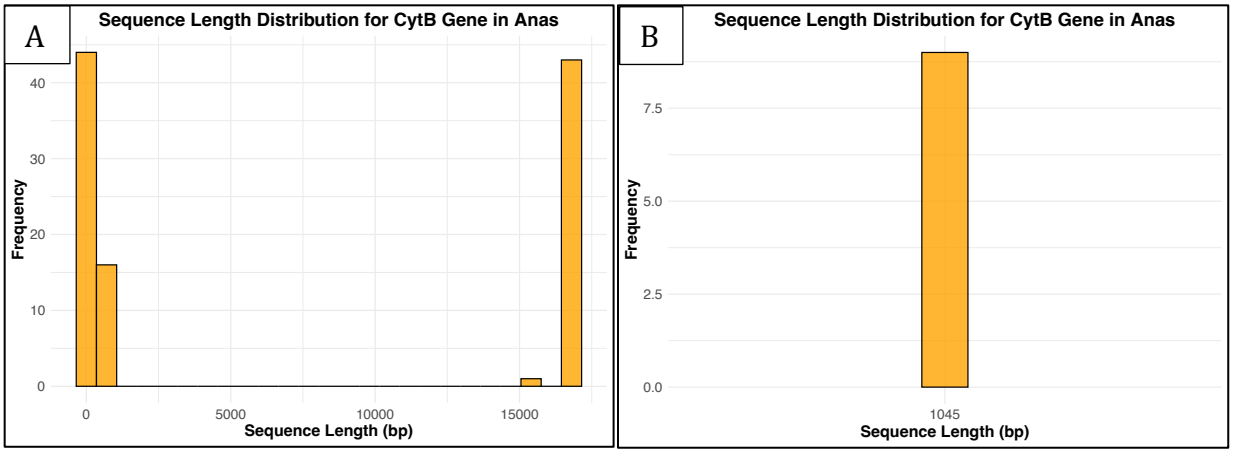

Figure 1: Comparing the sequence length distribution for the CytB gene in Anas before and after sequence filtering. Outliers were present in figure 1.A, meaning that there were complete mitochondrial sequences (>15000bp) and partial coding sequences (<1010bp). Figure 1.B demonstrates how 9 sequences were left after filtering the data for only complete coding sequences (a sequence length of 1045bp).

Arvind Srinivas
1301477
BINF*6210

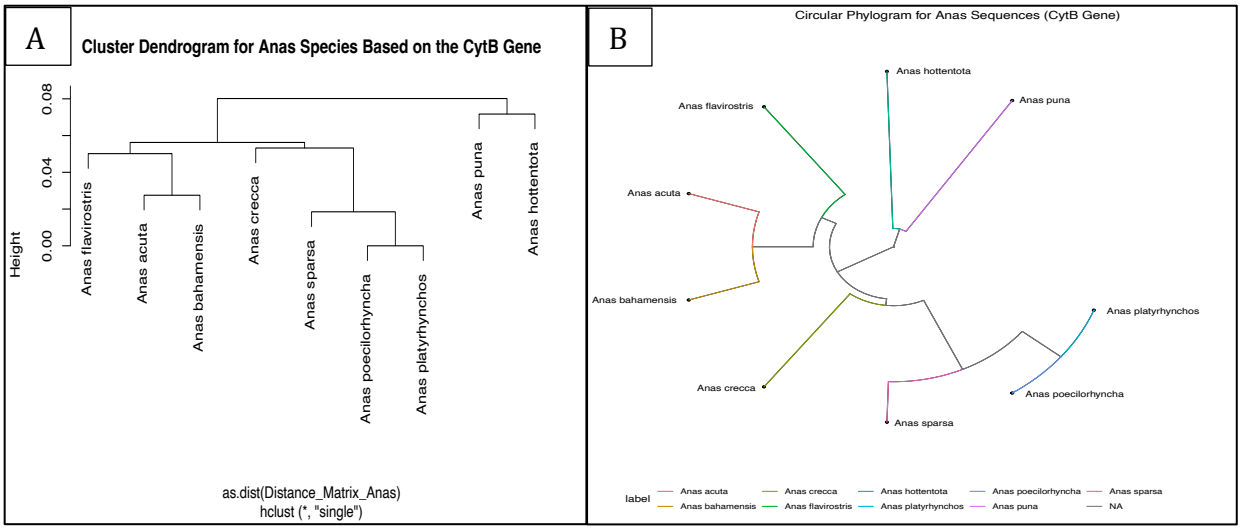

Figure 2: Different types of dendrograms to showcase the genetic relatedness between nine species of Anas. Figure 2.A shows a cluster dendrogram based on the CytB gene (using the K80 model). Figure 2.B displays a circular phylogram showcasing the same phylogenetic relationships.

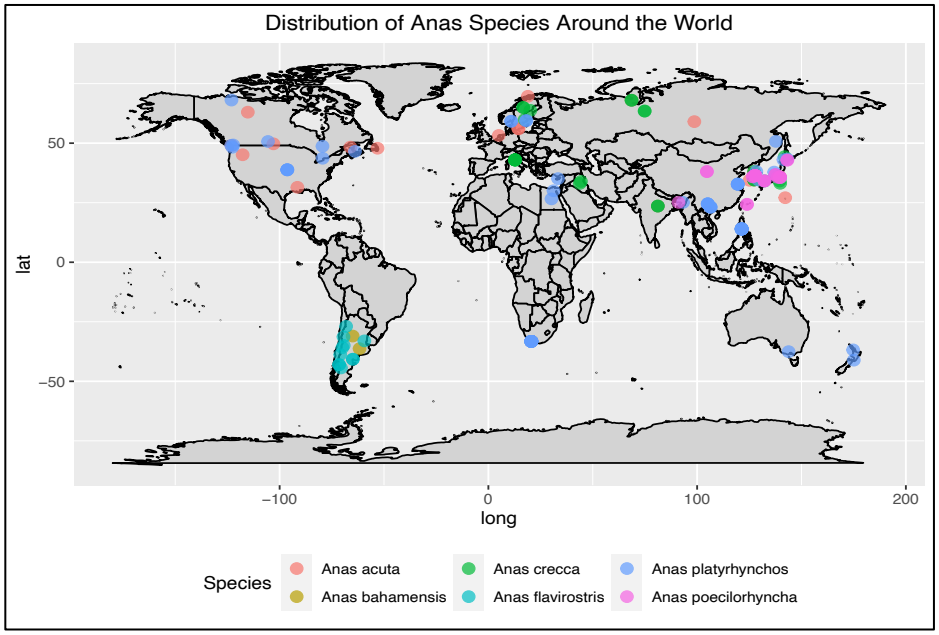

Figure 3: World map distribution of the Anas Species. There are only 6 species listed here as BOLD didn't have enough data to contain the other species. Plot was created using the "ggplot" package, with the world map being downloaded.

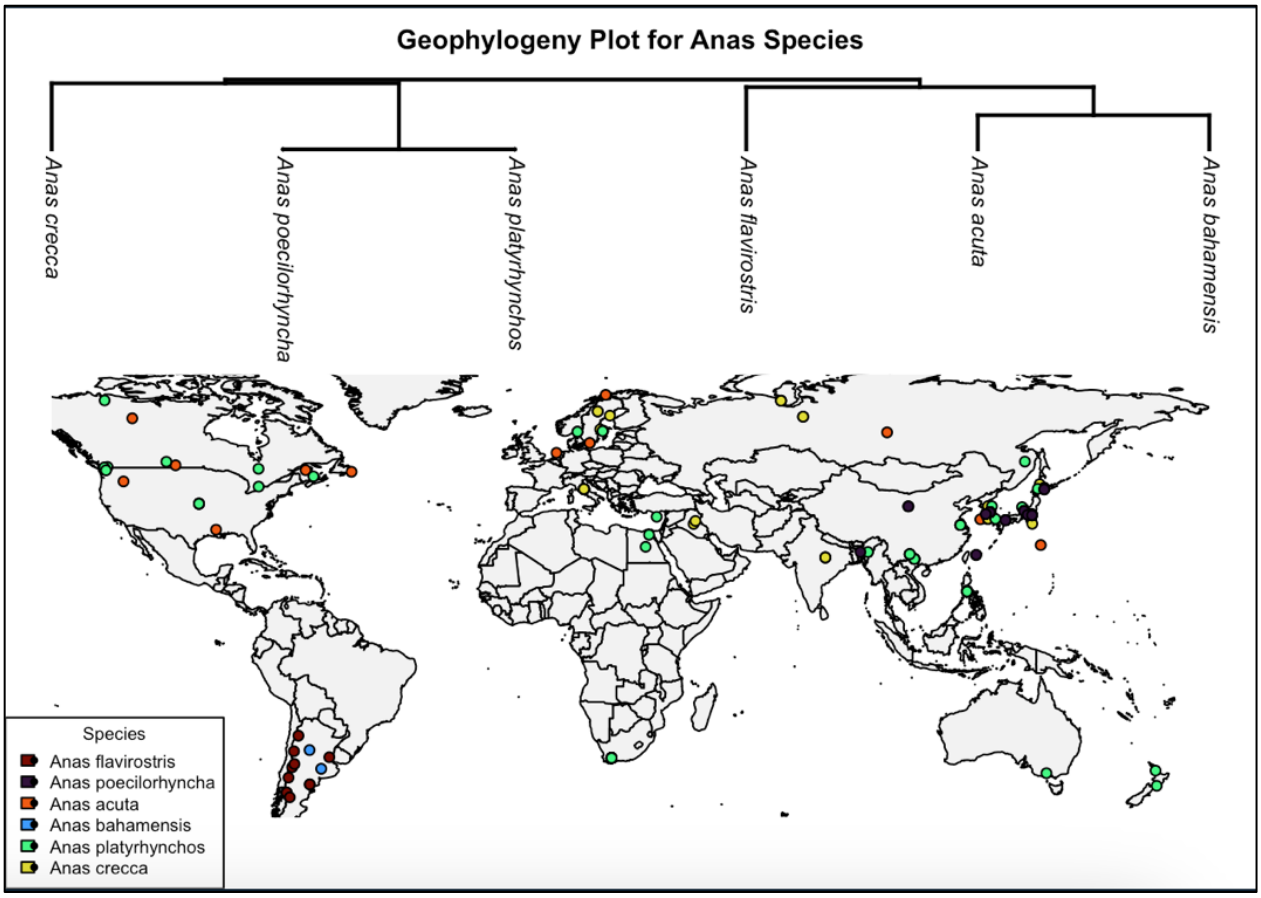

Figure 4: Geophylogeny plot for the Anas Species. Phylogeny was filtered to contain only the species that BOLD had spatial data for. The legend depicts the color scheme for each species on the map.

Discussion:

This project aimed to determine whether the duck genus, Anas, had species that were in the same geographic region or not. Looking at species distribution along with phylogenetics can offer an indicator on whether Anas species evolved through speciation or through large-scale allopatric speciation. From figures 4 and 5, it appears that most species within the Anas genus live in different geographic regions. This would show that Anas species have mostly went through allopatric speciation, which involves the Anas species to reproduce in isolation due to geographical barriers (Spaulding et al., 2023). This is to be expected as ducks inhabit a diverse range of environments and most of them can migrate to different geographic regions. However, there were some intriguing results from figures 4 and 5. For instance, some related Anas species cluster near certain regions, like

near Argentina. Anas flavirostris and Anas bahamensis are spatially located together in one specific region. According to the phylogenetic tree, they are related. This shows a historical connection, and possible signs of heteropatric speciation (Winker et al., 2013). Another discovery is that some species are distributed all over the globe, such as Anas acuta and Anas platyrhynchos. Though they are a bit distantly related, this global distribution shows successful dispersal and colonization strategies through time.

There were some limitations to this investigation. For instance, there were a limited number of species available from NCBI and BOLD. In terms of genetic data for alignment, NCBI only gave a provided 15 unique species from the genus, and after filtering steps, only 9 was left. Another caveat is that there is a huge assumption that samples were taken from the same site (NCBI and BOLD). Three species were missing from the final geophylogeny graph, showing that sample collection is not consistent between NCBI and BOLD. Lastly, NCBI doesn't have a GPS feature to provide latitude and longitude values. Therefore, mapping data from one database to the other can be difficult, especially if samples are different from each database.

With more time, the next steps would involve conducting a detailed analysis of the geophylogenetic patterns observed in the Anas species distributions. This could include statistical methods to quantify the degree of spatial clustering or dispersion among species. Additionally, exploring environmental variables and their influence on species distribution could provide insights into ecological factors shaping the observed patterns. Lastly, a comparative analysis with other avian groups or regions might offer a broader context for interpreting the observed geophylogenetic patterns and understanding the evolutionary implications of coexistence or segregation among Anas species.

Reflection:

This assignment provided an exciting opportunity to explore R within a novel biological context, particularly delving into geography, a realm I had initially hesitated to engage with in earlier assignments. Challenging myself to navigate this new topic prompted extensive research into unfamiliar packages. Morevoer, completing the assignment served as a good measure of my progress over the semester—from having no prior R experience to creating geophylogenetic plots, classifiers, and various visualizations. A valuable lesson learned was the importance of consulting vignettes and R package manuals during challenging moments in coding. It takes a lot of resilience and patience to code. I look forward to carrying this knowledge into future projects, notably for the summer project, which relates to evolutionary biology and geography. I really enjoyed this course and I'm going to practice even more during the break.

Arvind Srinivas
1301477
BINF*6210

References:

Abellán, P., Ribera, I. (2011). Geographic location and phylogeny are the main determinants of the size of the geographical range in aquatic beetles. BMC Evol Biol 11, 344. https://doi.org/10.1186/1471-2148-11-344

Code Academy. (n.d). Learn R: Data Cleaning (gsub() R Function).https://www.codecademy.com/learn/working-with-data-in-r-skill-path/modules/learn-r-data-cleaning/cheatsheet#

Deckmyn, A. (2023). Maps R Documentation. https://www.rdocumentation.org/packages/maps/versions/3.4.1.1

Farias, I. P., Ortí, G., Sampaio, I., Schneider, H., & Meyer, A. (2001). The cytochrome b gene as a phylogenetic marker: the limits of resolution for analyzing relationships among cichlid fishes. Journal of molecular evolution, 53(2), 89–103. https://doi.org/10.1007/s002390010197

Revell. L. (2019). Projecting a phylogeny onto a geographic map showing species ranges in R. http://blog.phytools.org/2019/03/projecting-phylogeny-onto-geographic.html

Revell, L. (2023). Package 'phytools'. https://cran.rproject.org/web/packages/phytools/phytools.pdf

Rolland, J., Cadotte, M. W., Davies, J., Devictor, V., Lavergne, S., Mouquet, N., … & Morlon, H. (2012). Using phylogenies in conservation: new perspectives.

Sleator, R.D. 2011. Phylogenetics. Arch Microbiol 193, 235–239. https://doi.org/10.1007/s00203-011-0677-x

Spaulding, F., McLaughlin, J. F., Cheek, R. G., McCracken, K. G., Glenn, T. C., & Winker, K. (2023). Population genomics indicate three different modes of divergence and speciation with gene flow in the green-winged teal duck complex. Molecular Phylogenetics and Evolution, 182, 107733.

Winker, K., McCracken, K. G., Gibson, D. D., & Peters, J. L. (2013). Heteropatric speciation in a duck, Anas crecca. Molecular ecology, 22(23), 5922–5935. https://doi.org/10.1111/mec.12525

Yamaguchi, R., & Iwasa, Y. (2013). First passage time to allopatric speciation. Interface focus, 3(6), 20130026.

Arvind Srinivas
1301477
BINF*6210

Acknowledgements:

Thank you, Robin Zutshi, for helping me to solve an error regarding my geophylogenetic plot. The error message said that my dendrogram had more species compared to my Relevant_Data1 data frame. Robin showed me how to use the drop_tip command to remove tips that were not included in my merged data frame. This positively impacted my project in terms of generating a phylogenetic plot alongside a map.