# Assignment 2

Qingyuan Pei

2023-10-27

## Introduction

The genus *Streptopelia* is widely spread all around the globe and is also an important model animal for physiological studies (Johnson et al., 2001). To investigate the evolutionary relationship among the species from *Streptopelia*, one of the mitocondrial gene cytochrome c oxidase I (COI), which is generally used to identify birds (Hebert et al., 2004), are chose to be the marker gene and the sequences of sister species are downloaded from NCBI. Whether the distribution of these different species in the world will be diversified or not will also be explored in this assignment using occurrences data from GBIF and be presented through a geophylogenetic figure.

## Code Section 1

Packages that will be used in this project.

```
library(rentrez)
library(seqinr)
library(Biostrings)
library(tidyverse)
library(rgbif)
library(stringi)
library(ape)
library(RSQLite)
library(muscle)
library(DECIPHER)
library(dendextend)
library(ggplot2)
library(ggtree)
library(mapdata)
library(phytools)
```

Search for COI sequences from genus *Streptopelia* and check the hits.

```
NCBI_search <- entrez_search(
  db = "nuccore",
  term = "Streptopelia[ORGN] AND COI[Gene] AND 400:700[SLEN]",
  retmax = 200
  )
NCBI_search
```

Use entrez_fetch to download sequence data of in FASTA format and then write to the file called NCBI_fetch.fasta in the current directory.

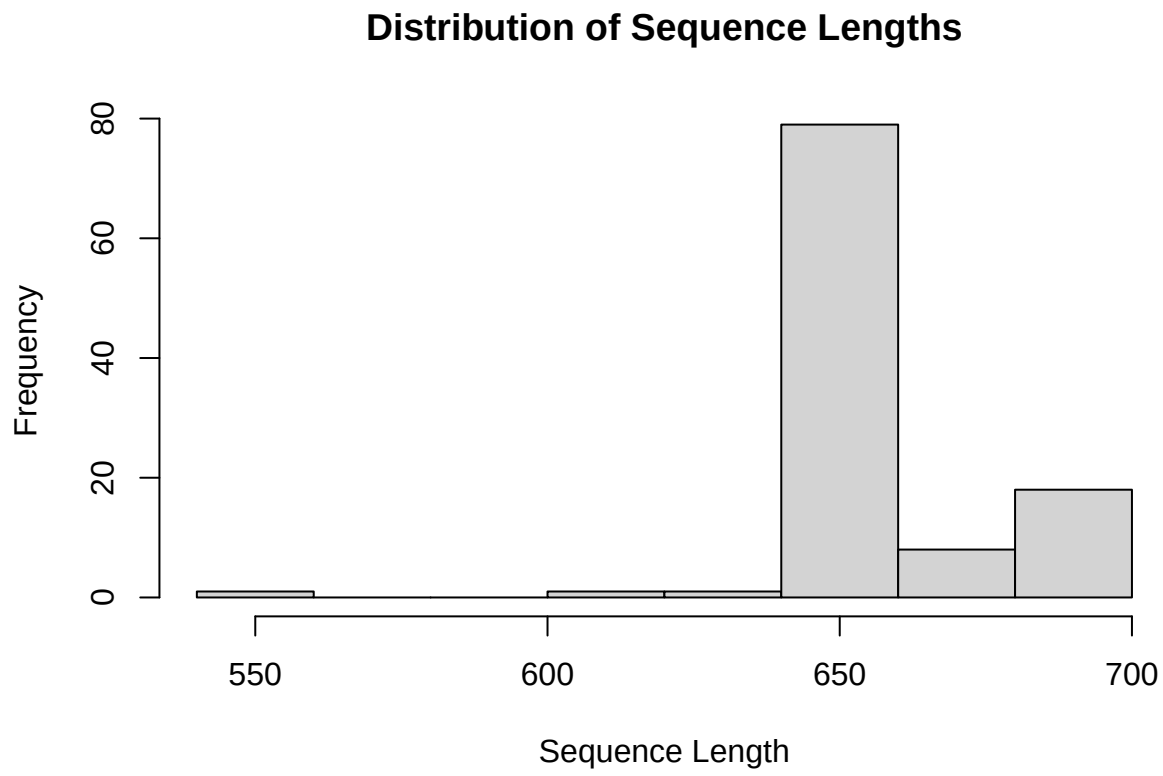
```
NCBI_fetch <- entrez_fetch(
  db = "nuccore",
  id = NCBI_search$ids,
  rettype = "fasta"
  )
write(NCBI_fetch, "NCBI_fetch.fasta", sep = "\n")
```

Read the FASTA file back to the environment and create a dataframe to save the names and sequences.

```
stringSet <- readDNAStringSet("NCBI_fetch.fasta")
dfNCBI_COI <- data.frame(COI_Title = names(stringSet),
                         COI_Sequence = paste(stringSet))

dfNCBI_COI$Species_Name <- word(dfNCBI_COI$COI_Title, 2L, 3L)
```

Draw a histogram of the distribution of COI sequence lengths, to see what sequences should be deleted.

```
dfNCBI_COI$Sequence_Length <- nchar(dfNCBI_COI$COI_Sequence)

hist((dfNCBI_COI$Sequence_Length),
     xlab = "Sequence Length",
     ylab =  "Frequency",
     main = "Distribution of Sequence Lengths")
```




Filter out the species with much shorter sequence lengths than all the majority and those with less than 3 samples.

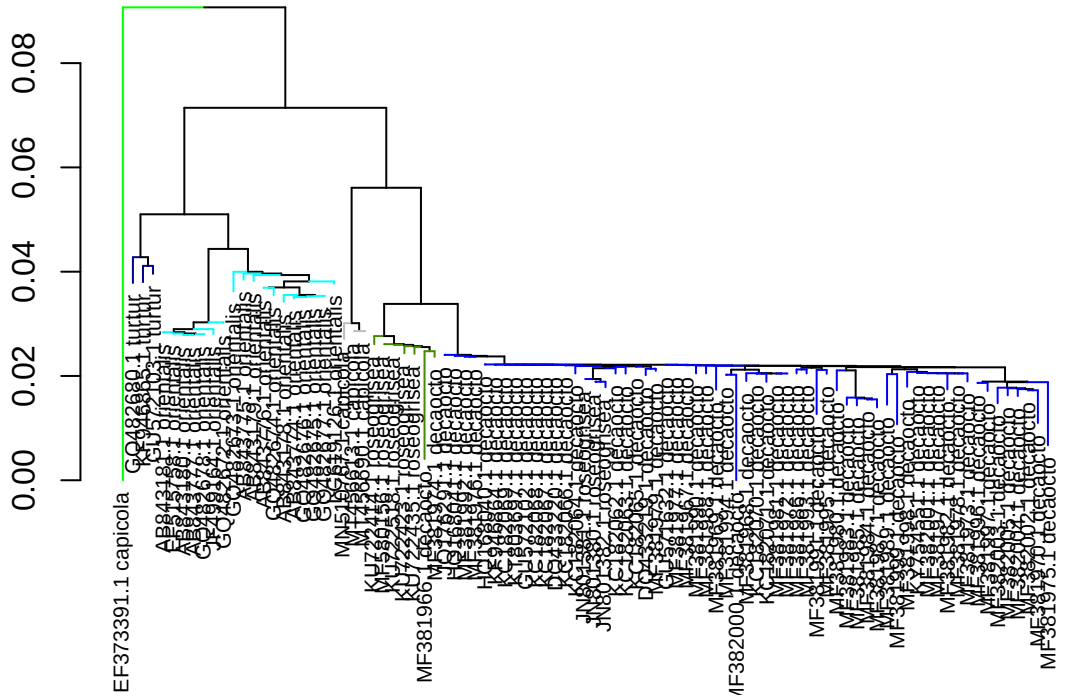
```
dfCOI_species_count <- dfNCBI_COI %>%
  group_by(Species_Name) %>%
  count()

dfCOI_filtered <- dfNCBI_COI %>%
  mutate(COI_Sequence = str_remove_all(COI_Sequence, "^N+|N+$|-")) %>%
  #remove the unverified sequences
  filter(!grepl('UNVERIFIED_ORG', COI_Title)) %>%
  #remove sequences with length shorter or longer than the median length -/+50
  filter(str_count(COI_Sequence) >= median(str_count(COI_Sequence)) - 50 &
           str_count(COI_Sequence) <= median(str_count(COI_Sequence)) + 50) %>%
  #remove all sequences with the proportion of N is bigger than 1%
  filter(str_count(COI_Sequence, "N") <= (0.01 * str_count(COI_Sequence))) %>%
  #remove species with less than 3 sequences
  group_by(Species_Name) %>%
  filter(n() >= 3) %>%
  arrange(Species_Name) %>%
  select(-Sequence_Length)
```

To add a Biostrings column using all the sequences in column COI_Sequence of dfCOI_filtered and a column for naming the nucleotides for better visualization of sequence alignment.

```
dfCOI_filtered1 <- as.data.frame(dfCOI_filtered)

dfCOI_filtered1$nucleotides <- DNAStringSet(dfCOI_filtered1$COI_Sequence)

dfCOI_filtered1$Alignment_id <- paste(word(dfCOI_filtered1$COI_Title, 1L),
                                      word(dfCOI_filtered1$Species_Name, 2L),
                                      sep = " ")
names(dfCOI_filtered1$nucleotides) <- dfCOI_filtered1$Alignment_id

dfCOI.alignment.test <- DNAStringSet(muscle::muscle(dfCOI_filtered1$nucleotides,
                                                    maxiters = 2), use.names = TRUE)
```

To cluster the aligned sequences see if there are any outliers or abnormal sequences.

```
dnaBin.COI.test <- as.DNAbin(dfCOI.alignment.test)

distanceMatrix.test <- dist.dna(dnaBin.COI.test, model = "TN93",
                                as.matrix = TRUE, pairwise.deletion = TRUE)

#use neighbour joining method
clustering.method <- "NJ"

#set the threshold to 0.035
clustering.threshold <- 0.035

clusters.COI.test <- DECIPHER::TreeLine(myDistMatrix = distanceMatrix.test,
                                        method = clustering.method,
                                        cutoff = clustering.threshold,
                                        showPlot = TRUE,
                                        type = "both",
                                        verbose = TRUE)
```




As it shows in the dendrogram, EF373391.1 from *Streptopelia capitol*, MF381966.1 and MF381974.1 from *Streptopelia decaocto*, and JN801380.1, JN801381.1 and JN801382.1 from *Streptopelia roseogrisea* are clustered into wrong group. After BLAST, it is showed that EF373391.1 is highly possible from another genus *Treron*, so this can be a mislabel considering these two genera look very different. For MF381966.1, MF381974.1, JN801380.1, JN801381.1 and JN801382.1, these two species are very close to each other and looks very similar, so they might be misclassificated.

Filter out these DNA sequences for making phylogenetic tree in the next section.

```
dfCOI_filtered2 <- dfCOI_filtered %>%
  filter(!grepl('EF373391.1|MF381966.1|MF381974.1|JN801380.1|JN801381.1|JN801382.1',
                COI_Title))
```

Use rgbif package functions to download occurance data of Streptopelia and write it into the environment.

```
#save usage Key to a vector
usagekey <- name_backbone("Streptopelia")$usageKey

#download the filtered data
GBIFdownload <- occ_download(
  pred("hasGeospatialIssue", FALSE),
  pred("hasCoordinate", TRUE),
  pred("occurrenceStatus","PRESENT"),
  pred_not(pred_in("basisOfRecord",c("FOSSIL_SPECIMEN",
                                     "LIVING_SPECIMEN",
                                     "HUMAN_OBSERVATION"))),
```

```
  pred("taxonKey", usagekey),
  format = "SIMPLE_CSV"
  )

#see the downloading process status
occ_download_wait(GBIFdownload)

#import the downloaded data into the environment
Streptopelia_ocurrence <- occ_download_get('0028663-231002084531237') %>%
  occ_download_import()
```

To filter out columns that are not needed and all the missing data.

```
#to check the names of all the columns in Streptopelia_ocurrence
names(Streptopelia_ocurrence)

#choose and retain the columns that are needed
dfGBIF_filtered <- Streptopelia_ocurrence %>%
  dplyr:::select.data.frame(species, decimalLatitude, decimalLongitude) %>%
  group_by(species) %>%
  filter(!is.na(decimalLatitude)) %>%
  filter(!is.na(decimalLongitude))

#find the common species shared by dfCOI_filtered and dfGBIF_filtered
dfCOI_unique_species <- unique(dfCOI_filtered$Species_Name)
dfGBIF_unique_species <- unique(dfGBIF_filtered$species)
common_species <- intersect(dfCOI_unique_species, dfGBIF_unique_species)

#subset dfGBIF_filtered with species only exist in the common_species
dfGBIF_subset <- subset(dfGBIF_filtered, species %in% common_species)
```

## Code Section 2

DNA sequence alignment of filtered data.

```
dfCOI_filtered3 <- as.data.frame(dfCOI_filtered2)

dfCOI_filtered3$nucleotides <- DNAStringSet(dfCOI_filtered3$COI_Sequence)

#use the sequence unique id and species name to name the nucleotides
dfCOI_filtered3$Alignment_id <- paste(word(dfCOI_filtered3$COI_Title, 1L),
                                      word(dfCOI_filtered3$Species_Name, 2L),
                                      sep = " ")

names(dfCOI_filtered3$nucleotides) <- dfCOI_filtered3$Alignment_id

dfCOI.alignment <- DNAStringSet(muscle::muscle(dfCOI_filtered3$nucleotides,
                                               maxiters = 2), use.names = TRUE)
```

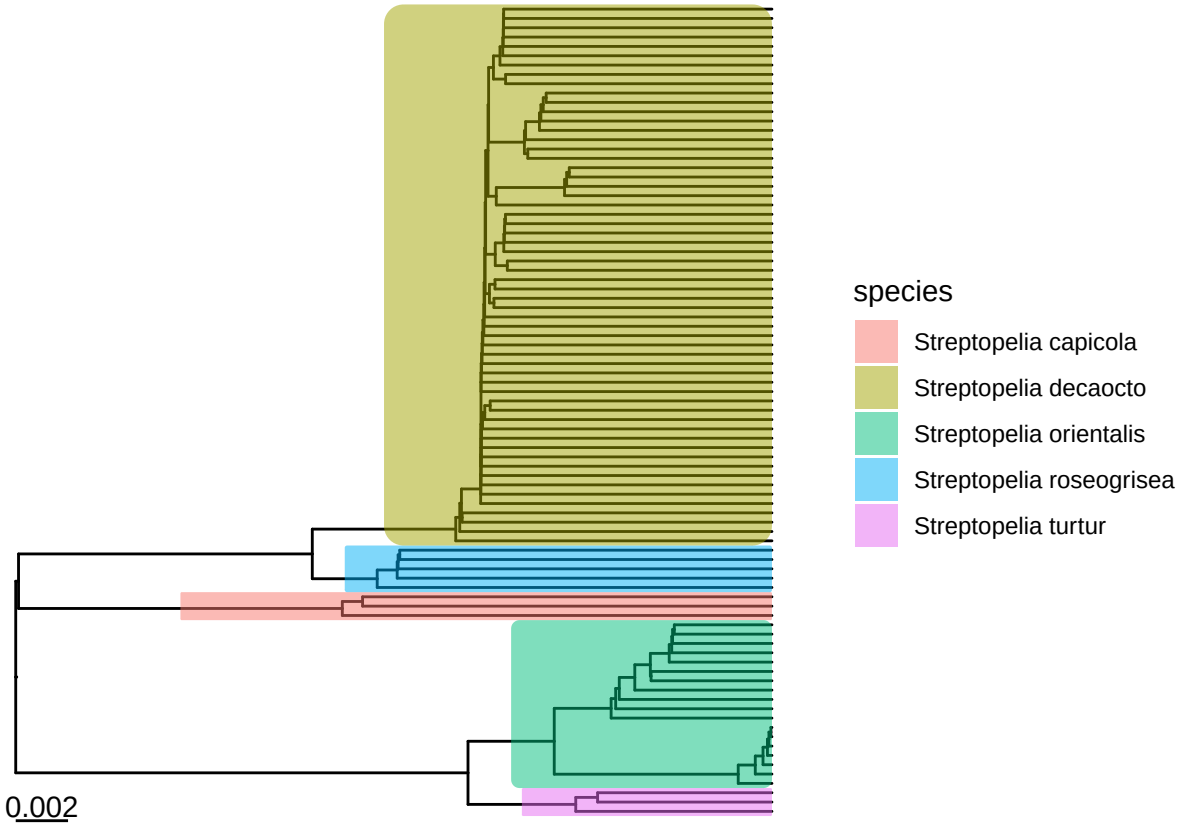
Create a dendrogram using package DECIPHER only with the species names for the next step to highlight the tree by different species names.

```
dnaBin.COI <- as.DNAbin(dfCOI.alignment)

distanceMatrix <- dist.dna(dnaBin.COI, model = "TN93",
                           as.matrix = TRUE, pairwise.deletion = TRUE)

clusters.COI <- DECIPHER::TreeLine(myDistMatrix = distanceMatrix,
                                   method = clustering.method,
                                   cutoff = clustering.threshold,
                                   showPlot = FALSE,
                                   collapse = -1,
                                   type = "both",
                                   verbose = TRUE)
```

Use the package dendextend to convert the dendrogram into phylo and then use package tidytree to get the information of each nodes as a tibble. Highlight each clades with different colors to represent different species using the nodes' number.

```
#change the dendrogram into a phylo for using ggtree package
COI_tree <- as.phylo(clusters.COI[[2]])

COI_treetibble <- tidytree::as_tibble(COI_tree)

#keep the node number of each species into different vectors
capicola <- COI_treetibble$label[grepl('capicola', COI_treetibble$label)]
capicola.MRCA <- MRCA(COI_treetibble, capicola)
capicola.node <- capicola.MRCA$node

decaocto <- COI_treetibble$label[grepl('decaocto', COI_treetibble$label)]
decaocto.MRCA <- MRCA(COI_treetibble, decaocto)
decaocto.node <- decaocto.MRCA$node

turtur <- COI_treetibble$label[grepl('turtur', COI_treetibble$label)]
turtur.MRCA <- MRCA(COI_treetibble, turtur)
turtur.node <- turtur.MRCA$node

roseogrisea <- COI_treetibble$label[grepl('roseogrisea', COI_treetibble$label)]
roseogrisea.MRCA <- MRCA(COI_treetibble, roseogrisea)
roseogrisea.node <- roseogrisea.MRCA$node

orientalis <- COI_treetibble$label[grepl('orientalis', COI_treetibble$label)]
orientalis.MRCA <- MRCA(COI_treetibble, orientalis)
orientalis.node <- orientalis.MRCA$node

#give the nodes the labels using species name
nodes.label <- data.frame(node=c(capicola.node, decaocto.node,
                                 turtur.node, roseogrisea.node, orientalis.node),
                          species=c("Streptopelia capicola",
                                    "Streptopelia decaocto",
                                    "Streptopelia turtur",
                                    "Streptopelia roseogrisea",
                                    "Streptopelia orientalis"))

#use ggtree to draw the phylogenetic tree
```

```
ggtree(COI_tree) +
  geom_treescale(x=0, y=0) +
  geom_hilight(data=nodes.label,
               aes(node=node, fill=species),
               type = "roundrect")
```




Find one representative sequence from each species to draw a phylogenetic tree only on species level for better visualization of their geographic distribution. Here I choose to find the median of the parent node of each species group and choose one sequence with the median parent node number to be the representative sequence.

```
COI_treetibble$species <- paste(word(COI_treetibble$label, 2L))
parent.median <- COI_treetibble %>%
  group_by(species) %>%
  summarize(Median = round(median(parent)))

rep.sequences <- COI_treetibble %>%
  inner_join(parent.median, by = "species") %>%
  filter(parent == Median) %>%
  group_by(species) %>%
  sample_n(1)

print(rep.sequences$label)
```

```
## [1] "MT456690.1 capicola"    "MF381991.1 decaocto"    "GQ482677.1 orientalis"
## [4] "KU722518.1 roseogrisea" "KF946865.1 turtur"
```

There are 4 species have two sequences match the condition, so I choose the first one of each species as the representative sequence: MT456690.1 capicola, MF381995.1 decaocto, GQ482677.1 orientalis, KU722518.1 roseogrisea, and GU572103.1 turtur.

```
COI_species <- dfCOI_filtered %>%
  filter(
    grepl('MT456690.1|MF381995.1|GQ482677.1|KU722518.1|GU572103.1', COI_Title))
```

DNA sequence alignment of the representative sequences.

```
dfCOI_species <- as.data.frame(COI_species)

dfCOI_species$nucleotides <- DNAStringSet(dfCOI_species$COI_Sequence)

names(dfCOI_species$nucleotides) <- dfCOI_species$Species_Name

dfCOI_species.alignment <- DNAStringSet(muscle::muscle(dfCOI_species$nucleotides,
                                                       maxiters = 2), use.names = TRUE)
```

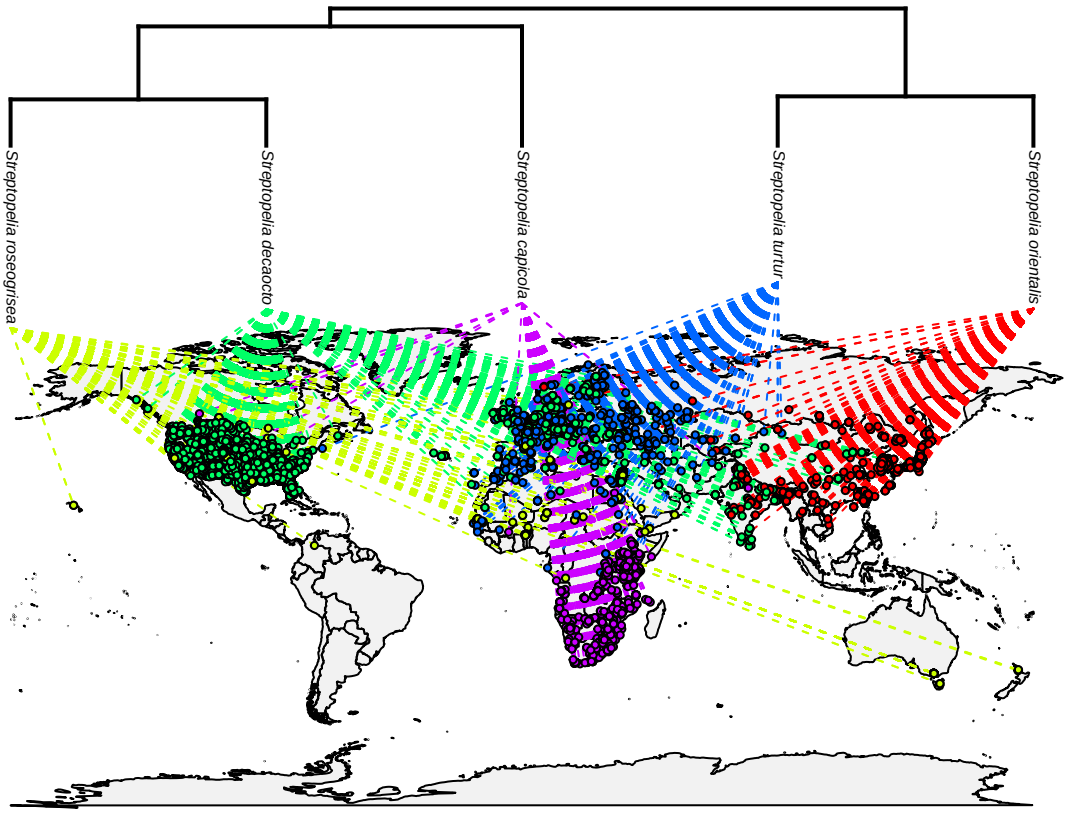
Draw the dedrogram only on species level.

```
dnaBin.COI_species <- as.DNAbin(dfCOI_species.alignment)

distanceMatrix_species <- dist.dna(dnaBin.COI_species, model = "TN93",
                                   as.matrix = TRUE, pairwise.deletion = TRUE)

clusters.COI_species <- DECIPHER::TreeLine(myDistMatrix = distanceMatrix_species,
                                   method = clustering.method,
                                   cutoff = clustering.threshold,
                                   showPlot = FALSE,
                                   collapse = -1,
                                   type = "both",
                                   verbose = FALSE)

COI_species.tree <- as.phylo(clusters.COI_species[[2]])
```

Use package phytools to project the phylogeny onto a world map.

```
dfGBIF_geo <- as.data.frame(dfGBIF_subset)
colnames(dfGBIF_geo)[2] <- "lat"
colnames(dfGBIF_geo)[3] <- "long"

#the geology data should be in matrix format
dfGBIF_geo <- as.matrix(dfGBIF_geo)

#species names should be row names instead of a column
species.names <- as.vector(dfGBIF_geo[,1])
rownames(dfGBIF_geo) <- species.names

#remove the column of species names
dfGBIF_geo <- dfGBIF_geo[, -1]
```

```
obj <- phylo.to.map(COI_species.tree, dfGBIF_geo, plot=FALSE)

cols<-setNames(sample(rainbow(n=Ntip(COI_species.tree))),
    COI_species.tree$tip.label)

plot(obj,colors=cols,ftype="i",fsize=0.5,cex.points=c(0,0.5))
```




## Results and Discussion

In the unrooted phylogenetic tree of the first figure in Code Section 2, *Streptopelia decaocto* and *Streptopelia roseogrisea* share the same parent who is the sister of *Streptopelia capicola*. As closely related siblings, *Streptopelia turtur* and *Streptopelia orientalis* are more genetically distinct to the other three species. For the geological distritubtion, the map does show certain diversification among the five species. In East Asia, *Streptopelia orientalis* is the major local species, while its sibling *Streptopelia turtur*(European Turtle Dove) mainly live in Europe, Middle East and North Africa, this species is reported that the number of it rapidly decreased in Europe over past decades(Brown etal., 2003, De Vries et al., 2022). The habitat of *Streptopelia capicola* is in the middle and south part of Africa. *Streptopelia roseogrisea*, African Collared Dove, is found almost in all continents but inhabit mostly in North Africa and North America. As it shows in the map, the most common species from *Streptopelia* in North America is *Streptopelia decaocto*(Eurasian Collared Dove) which is actually orignated from Europe and Asia. It is an invasive species in North America and its population rapidly growed since it was first observed in the early 1980s(Fujisaki, et al., 2010, Bled, et al., 2011). For the limitation of this project, the numbers of COI sequences in some species from NCBI are not large enough and using only one marker gene COI can influence the accuracy of the phylogenetic tree.

## Acknowledgements

I would like to express my deep gratitude to Jessica Castellanos-Labaarcena for her insightful advice when I was experiencing standstills while coding. Her first suggestion is to BLAST the abnormal sequences in my first dendrogram and delete them if possible for the next step. The second is choosing representative sequence from each species to draw phylogenetic tree on species level.

## References

Bled, F., Royle, J. A., & Cam, E. (2011). Hierarchical modeling of an invasive spread: the Eurasian Collared-Dove Streptopelia decaocto in the United States. Ecological Applications, 21(1), 290-302.

Browne, S. J., & Aebischer, N. J. (2003). Habitat use, foraging ecology and diet of Turtle Doves Streptopelia turtur in Britain. Ibis, 145(4), 572-582.

De Vries, E. H. J., Foppen, R. P., Van Der Jeugd, H., & Jongejans, E. (2022). Searching for the causes of decline in the Dutch population of European Turtle Doves (Streptopelia turtur). Ibis, 164(2), 552-573.

Fujisaki, I., Pearlstine, E. V., & Mazzotti, F. J. (2010). The rapid spread of invasive Eurasian Collared Doves Streptopelia decaocto in the continental USA follows human-altered habitats. Ibis, 152(3), 622-632.

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. PLoS biology, 2(10), e312.

Johnson, K. P., De Kort, S., Dinwoodey, K., Mateman, A. C., Ten Cate, C., Lessells, C. M., ... & Sheldon, F. (2001). A molecular phylogeny of the dove genera Streptopelia and Columba. The Auk, 118(4), 874-887.

sbha. (2018). *Stackoverflow* https://stackoverflow.com/questions/22249702/delete-rows-containing-specific-strings-in-r

bnaul. (2012). *Stackoverflow* https://stackoverflow.com/questions/3695677/how-to-find-common-elements-from-multiple-vectors

csgillespie. (2012). *Stackoverflow* https://stackoverflow.com/questions/9350025/filtering-a-data-frame-on-a-vector

Yu, G. (2022). *Data Integration, Manipulation and Visualization of Phylogenetic Trees.* CRC Press. https://yulab-smu.top/treedata-book/index.html

Revell, L (2019) *Projecting a phylogenetic tree onto a map with multiple geographic points per taxon* http://blog.phytools.org/2019/03/projecting-phylogenetic-tree-onto-map.html

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