Binf 6210 Assignment 2:

Introduction:

Halicryptus spinulosus is an aquatic worm, which belongs to the priapulid phylum. In recent times aquatic worms under the mentioned phylum have gained evolutionary importance, sometimes called "living fossils" (Janssen *et al*, 2009). It has been found that existing priapulid worms resemble the morphology of early Scalidophorans of the Cambrian era (Nanglu *et al*, 2024). Scalidophorans were found in Burgess Shale-type fossil biotas and these fossils were discovered in Canada and China (Nanglu *et al*, 2024). Halicrptus spinulosus is primarily found in the Baltic Sea, contains the burrower that was found in early Cambrian era Scalidophorans suggesting that aquatic worms under priapulid phylum have the capability to preserve evolutionary traits from a Cambrian era worm that existed 500 million years ago (Kesidis *et al*, 2019). This is particularly interesting because the fossils were found in North America and Asia, in contrast Halicryptus spinulosus is found in Northern Europe. Performing bioinformatic analysis on genomic sequences of Priapulid worms, in particular Halicrptus spinulosus would give an insight into how evolutionary traits are preserved despite different geographic locations and possibly provide evolutionary relationship between other marine organisms.

Due to the increasing potential and importance of Priapulid worms, this project will focus on can Halicryptus spinulosus be found in different geographic areas, if yes? Are there any genetic variations? The COX gene will be used to compare the phylogenetic relationship between the different strains of Halicryptus spinulosus from various locations. The aquatic worm has shown strong preservation of Cambrian era traits; therefore, it is hypothesized that minimal genetic variations will occur even if the specie is found in different geographic locations.

Figures:

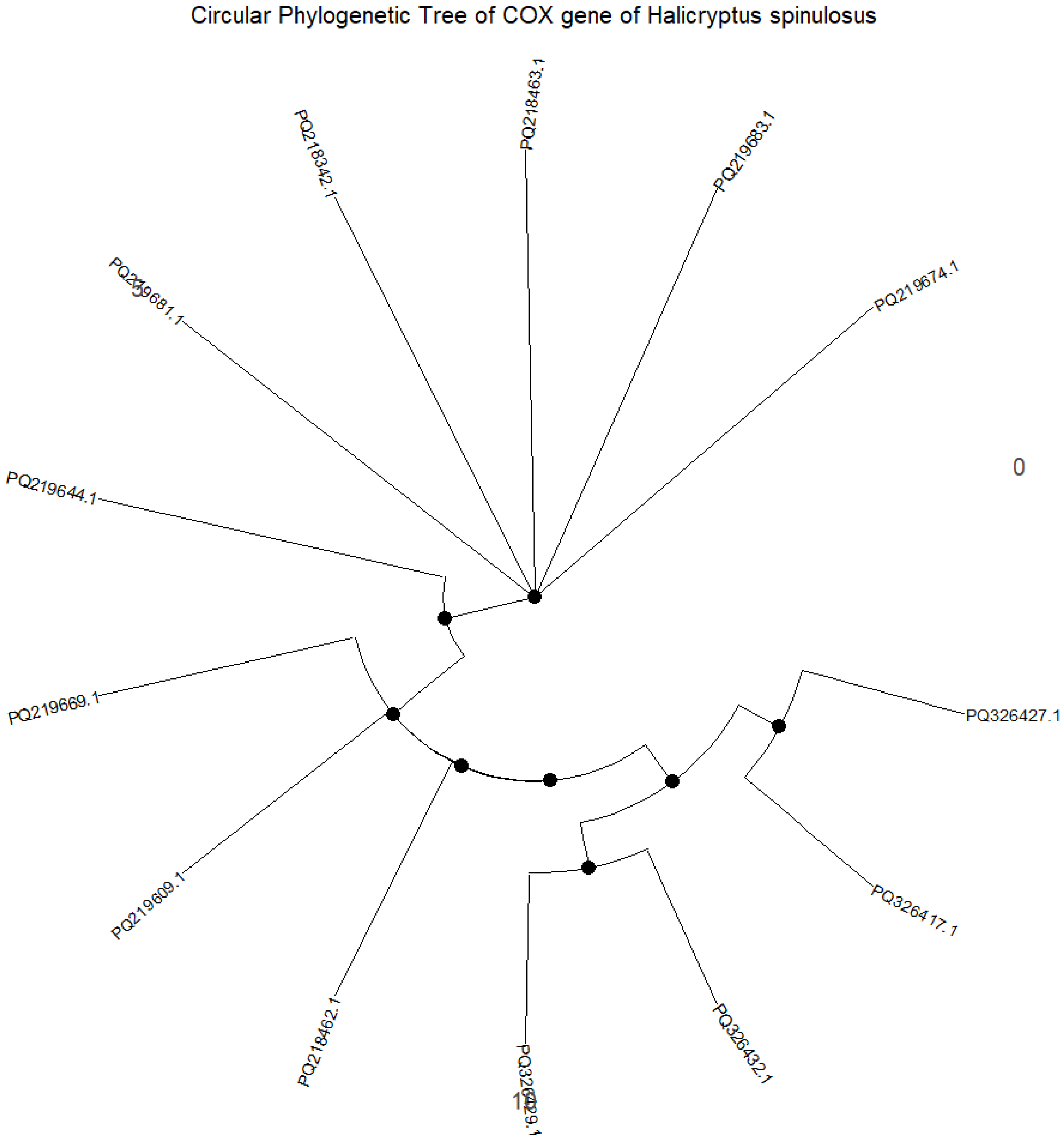


Figure 1:

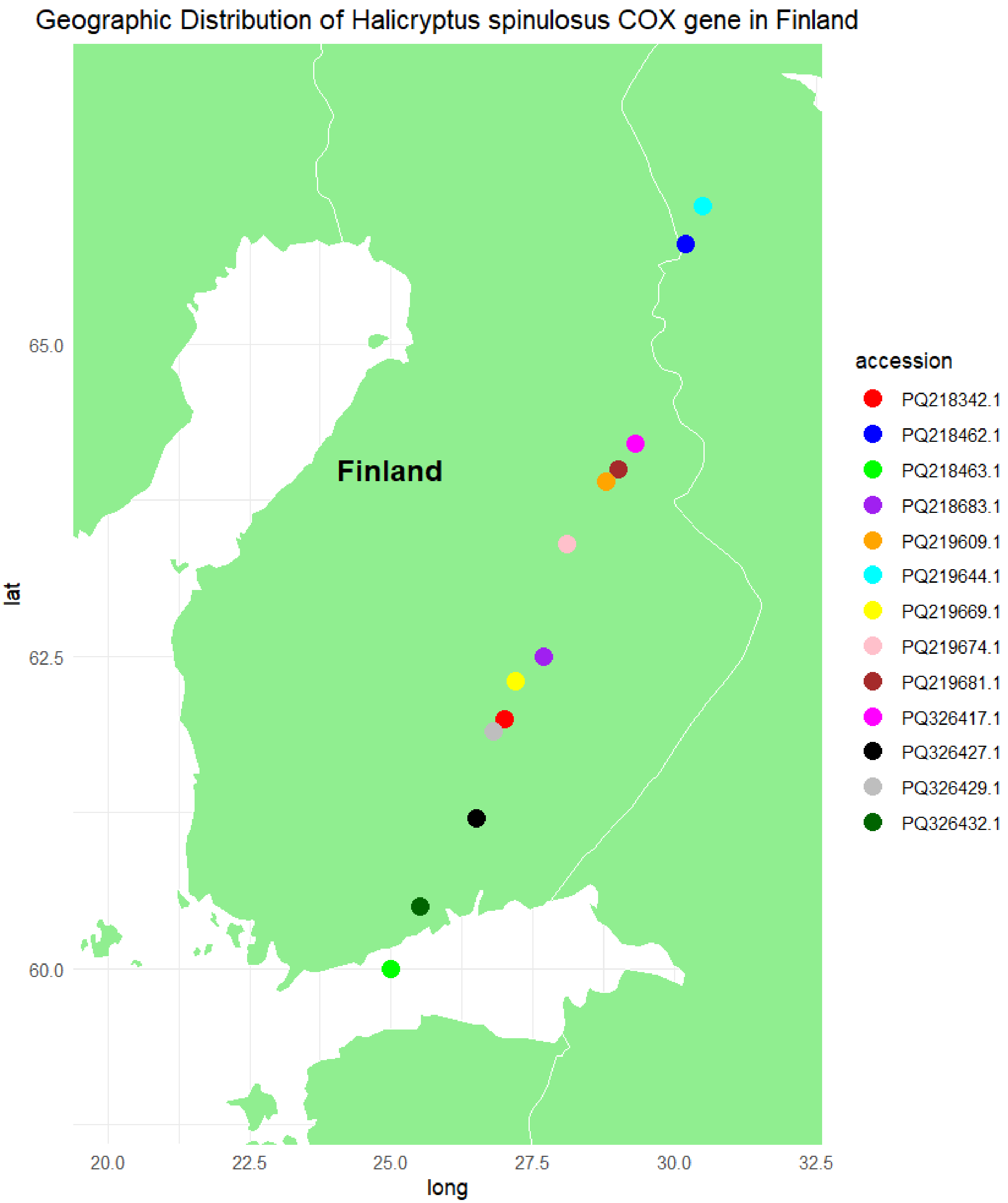
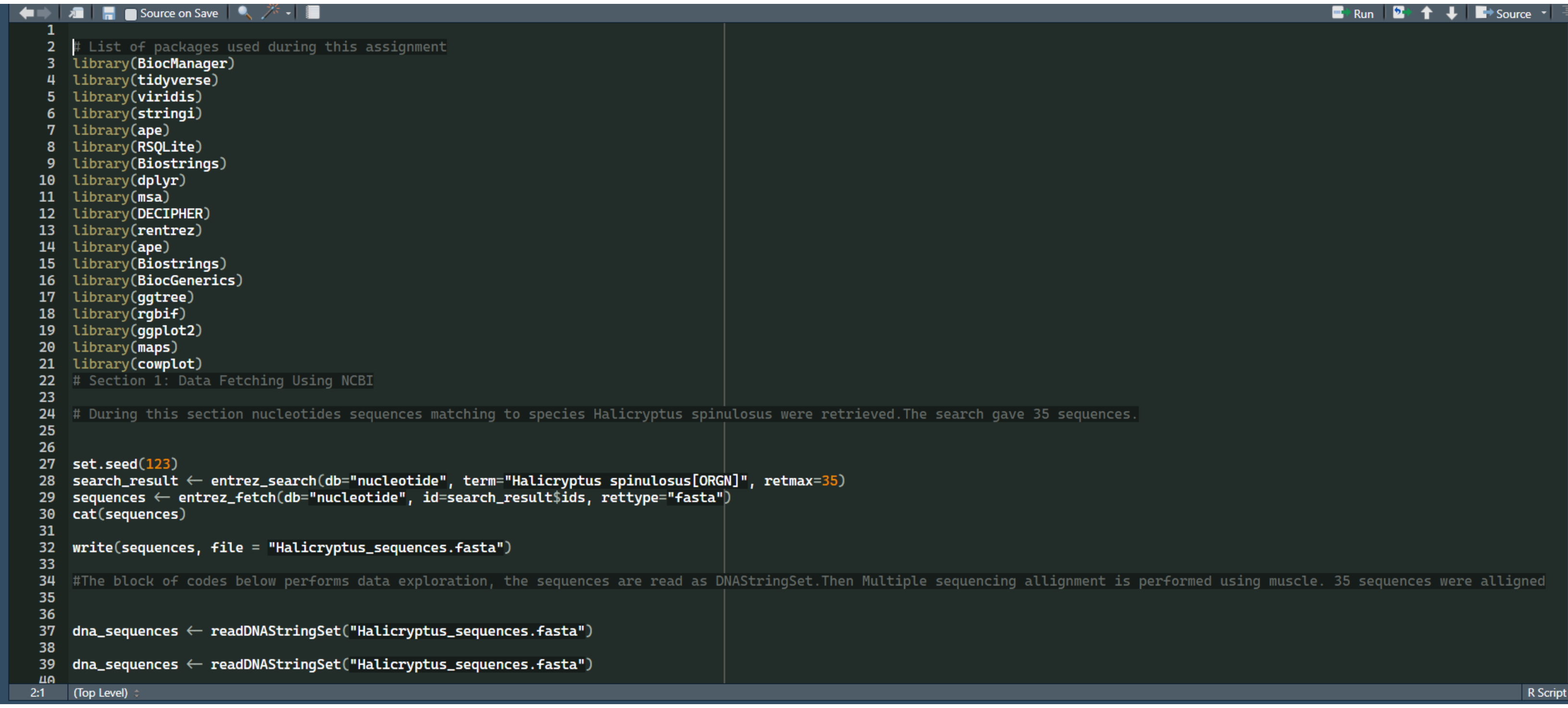
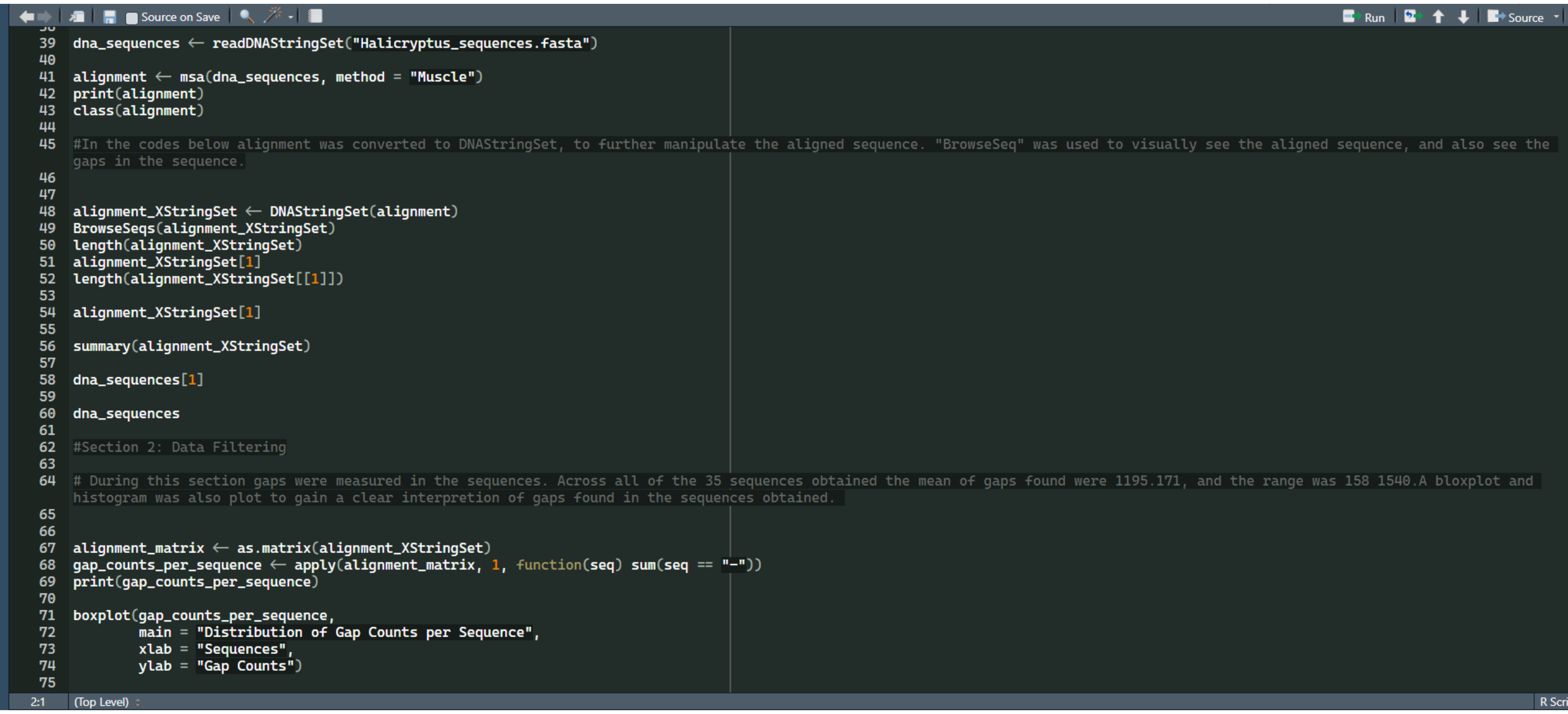

Geographic Distribution of Halicryptus spinulosus COX gene in Finland
Finland
lat
65.0
62.5
60.0
long
20.0
22.5
25.0
27.5
30.0
32.5
accession
● PQ218342.1
● PQ218462.1
● PQ218463.1
● PQ218683.1
● PQ219609.1
● PQ219644.1
● PQ219669.1
● PQ219674.1
● PQ219681.1
● PQ326417.1
● PQ326427.1
● PQ326429.1
● PQ326432.1

Figure 2:

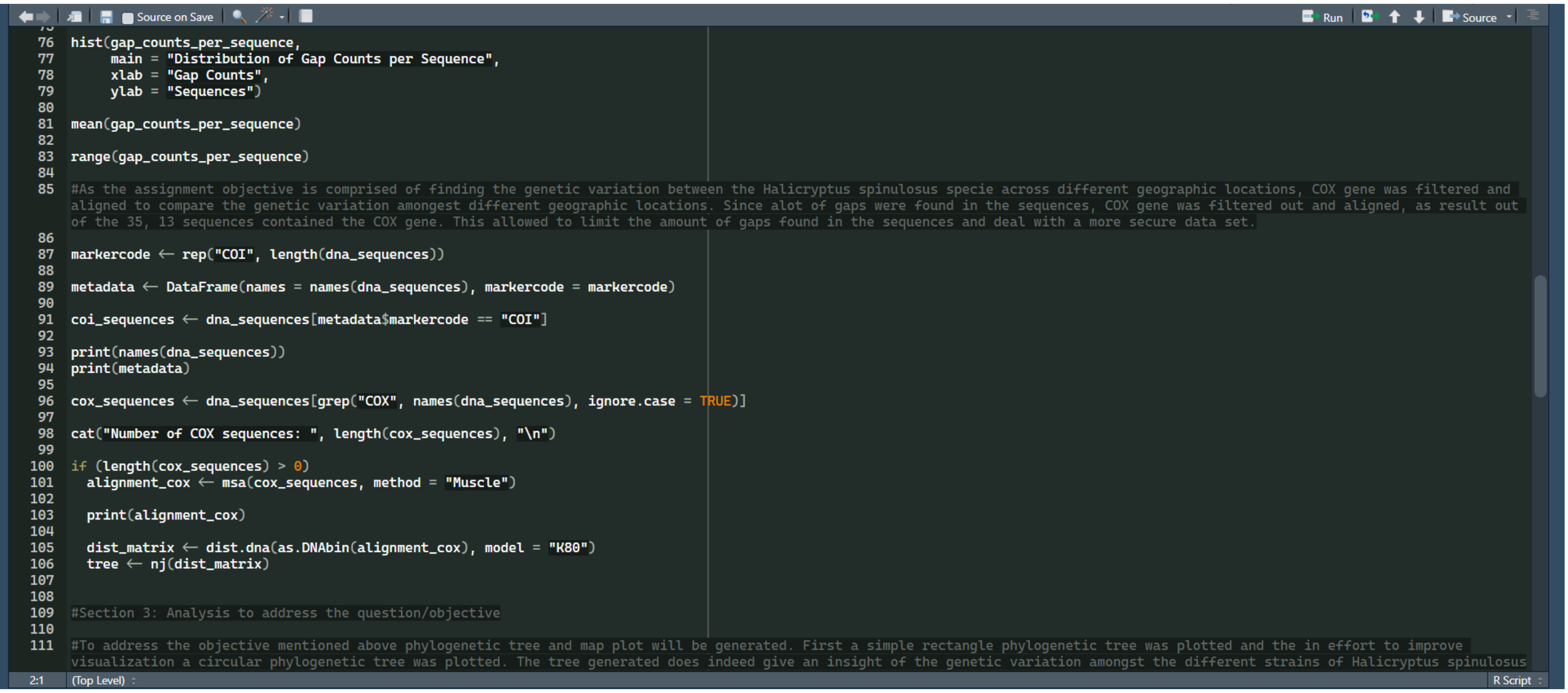
Coding Section:



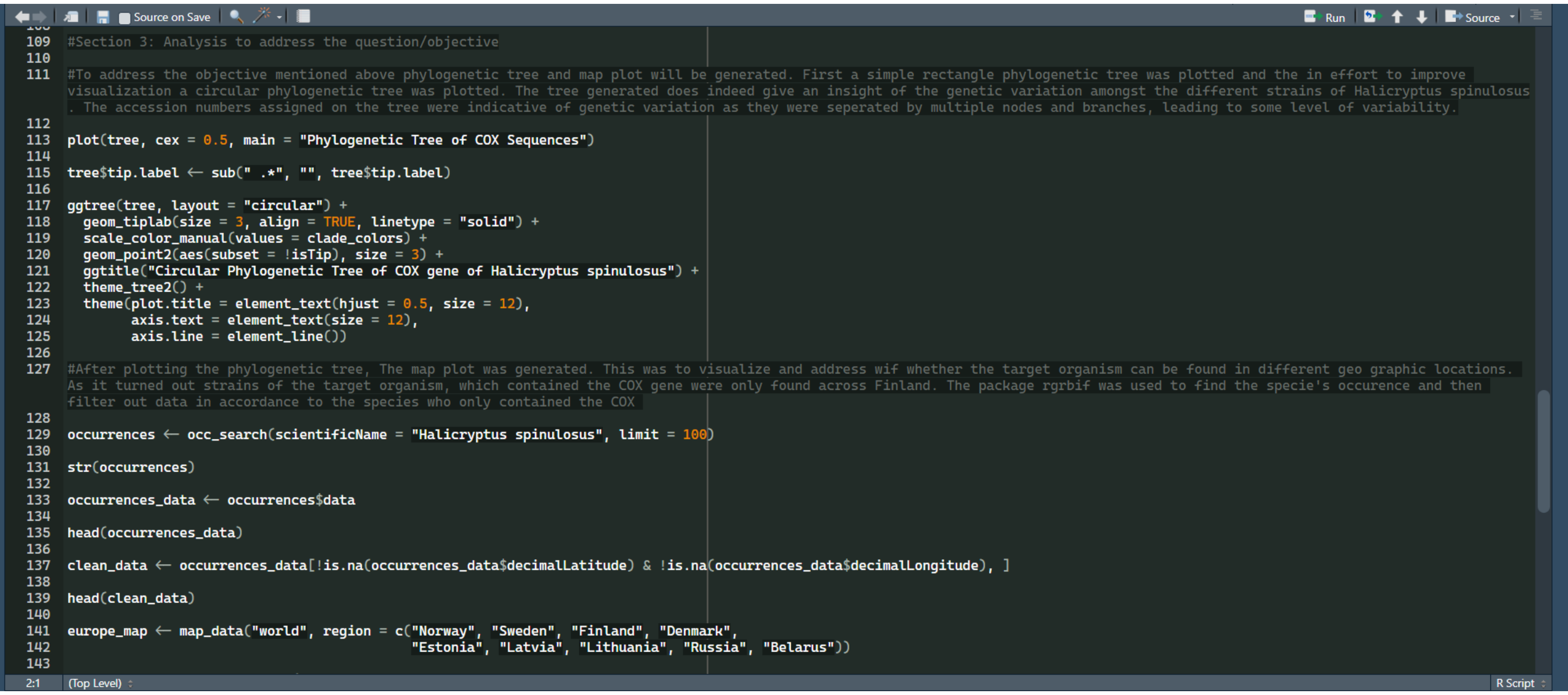
```r

# List of packages used during this assignment
library(BiocManager)
library(tidyverse)
library(viridis)
library(stringi)
library(ape)
library(RSQLite)
library(Biostrings)
library(dplyr)
library(msa)
library(DECIPHER)
library(rentrez)
library(ape)
library(Biostrings)
library(BiocGenerics)
library(ggtree)
library(rgbif)
library(ggplot2)
library(maps)
library(cowplot)
# Section 1: Data Fetching Using NCBI

# During this section nucleotides sequences matching to species Halicryptus spinulosus were retrieved.The search gave 35 sequences.


set.seed(123)
search_result <- entrez_search(db="nucleotide", term="Halicryptus spinulosus[ORGN]", retmax=35)
sequences <- entrez_fetch(db="nucleotide", id=search_result$ids, rettype="fasta")
cat(sequences)

write(sequences, file = "Halicryptus_sequences.fasta")

#The block of codes below performs data exploration, the sequences are read as DNAStringSet.Then Multiple sequencing allignment is performed using muscle. 35 sequences were alligned


dna_sequences <- readDNAStringSet("Halicryptus_sequences.fasta")

dna_sequences <- readDNAStringSet("Halicryptus_sequences.fasta")
```



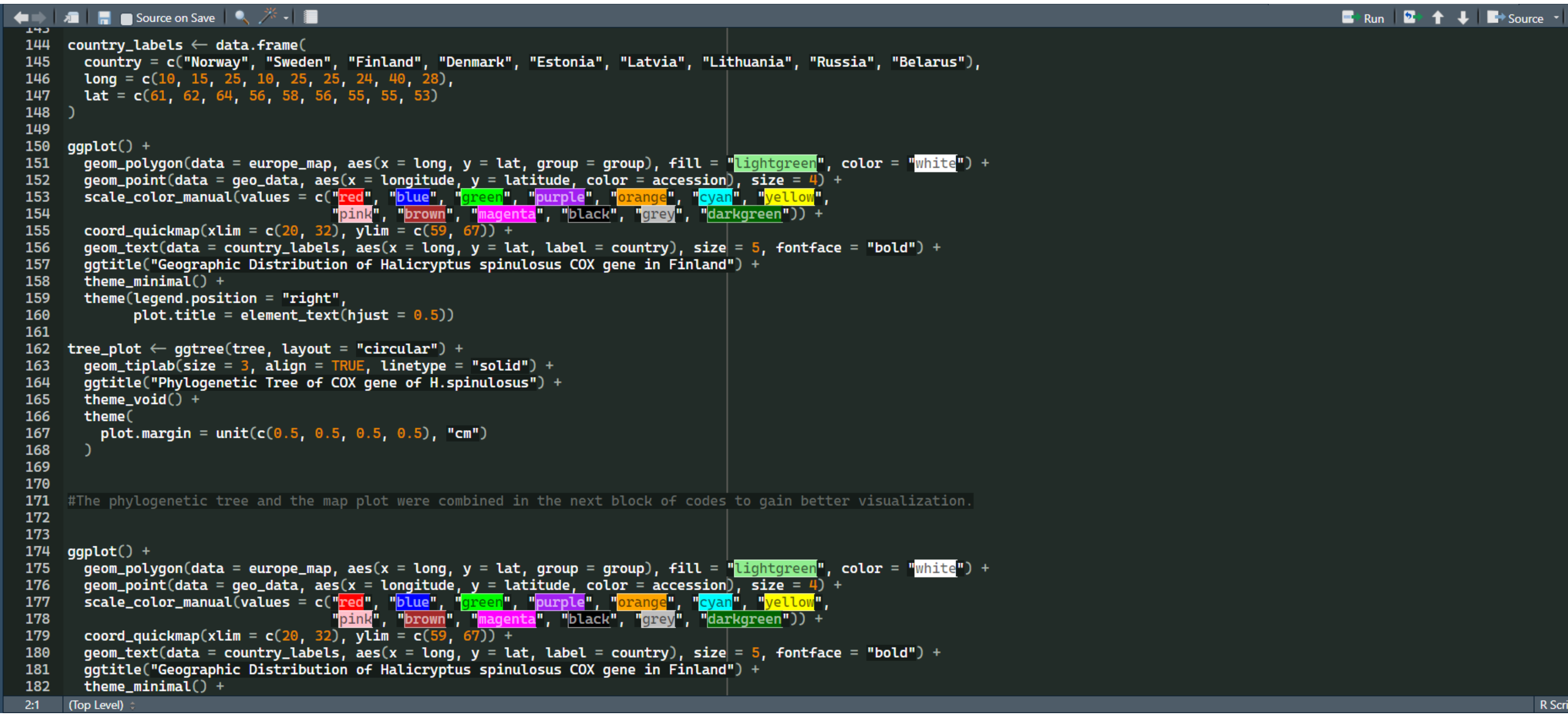
```r
dna_sequences <- readDNAStringSet("Halicryptus_sequences.fasta")

alignment <- msa(dna_sequences, method = "Muscle")
print(alignment)
class(alignment)

#In the codes below alignment was converted to DNAStringSet, to further manipulate the aligned sequence. "BrowseSeq" was used to visually see the aligned sequence, and also see the gaps in the sequence.


alignment_XStringSet <- DNAStringSet(alignment)
BrowseSeqs(alignment_XStringSet)
length(alignment_XStringSet)
alignment_XStringSet[1]
length(alignment_XStringSet[[1]])

alignment_XStringSet[1]

summary(alignment_XStringSet)

dna_sequences[1]

dna_sequences

#Section 2: Data Filtering

# During this section gaps were measured in the sequences. Across all of the 35 sequences obtained the mean of gaps found were 1195.171, and the range was 158 1540.A bloxplot and histogram was also plot to gain a clear interpretion of gaps found in the sequences obtained.


alignment_matrix <- as.matrix(alignment_XStringSet)
gap_counts_per_sequence <- apply(alignment_matrix, 1, function(seq) sum(seq == "-"))
print(gap_counts_per_sequence)

boxplot(gap_counts_per_sequence,
        main = "Distribution of Gap Counts per Sequence",
        xlab = "Sequences",
        ylab = "Gap Counts")

```



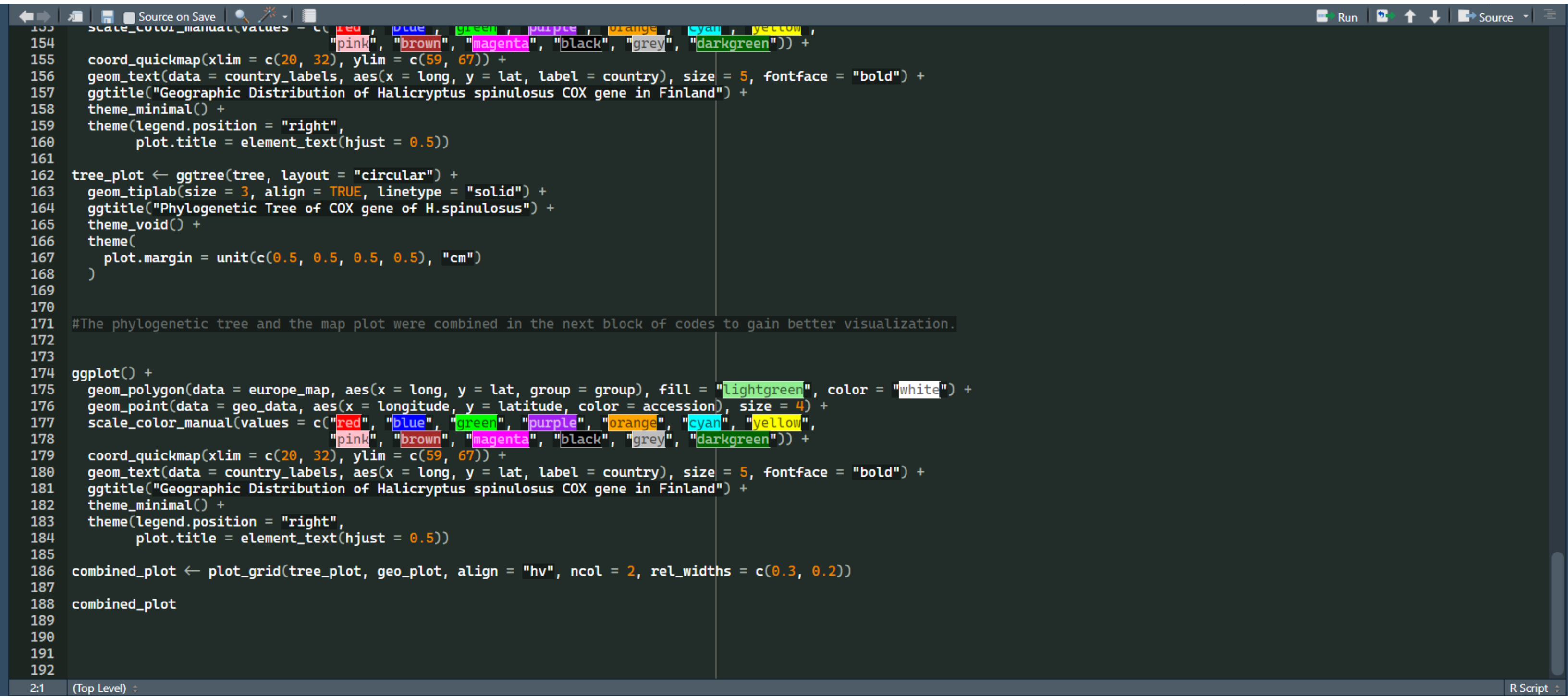
```r
hist(gap_counts_per_sequence,
     main = "Distribution of Gap Counts per Sequence",
     xlab = "Gap Counts",
     ylab = "Sequences")

mean(gap_counts_per_sequence)

range(gap_counts_per_sequence)

#As the assignment objective is comprised of finding the genetic variation between the Halicryptus spinulosus specie across different geographic locations, COX gene was filtered and aligned to compare the genetic variation amongest different geographic locations. Since alot of gaps were found in the sequences, COX gene was filtered out and aligned, as result out of the 35, 13 sequences contained the COX gene. This allowed to limit the amount of gaps found in the sequences and deal with a more secure data set.

markercode <- rep("COI", length(dna_sequences))

metadata <- DataFrame(names = names(dna_sequences), markercode = markercode)

coi_sequences <- dna_sequences[metadata$markercode == "COI"]

print(names(dna_sequences))
print(metadata)

cox_sequences <- dna_sequences[grep("COX", names(dna_sequences), ignore.case = TRUE)]

cat("Number of COX sequences: ", length(cox_sequences), "\n")

if (length(cox_sequences) > 0)
  alignment_cox <- msa(cox_sequences, method = "Muscle")

  print(alignment_cox)

  dist_matrix <- dist.dna(as.DNAbin(alignment_cox), model = "K80")
  tree <- nj(dist_matrix)


#Section 3: Analysis to address the question/objective

#To address the objective mentioned above phylogenetic tree and map plot will be generated. First a simple rectangle phylogenetic tree was plotted and the in effort to improve visualization a circular phylogenetic tree was plotted. The tree generated does indeed give an insight of the genetic variation amongst the different strains of Halicryptus spinulosus
```



```r
#Section 3: Analysis to address the question/objective

#To address the objective mentioned above phylogenetic tree and map plot will be generated. First a simple rectangle phylogenetic tree was plotted and the in effort to improve visualization a circular phylogenetic tree was plotted. The tree generated does indeed give an insight of the genetic variation amongst the different strains of Halicryptus spinulosus. The accession numbers assigned on the tree were indicative of genetic variation as they were seperated by multiple nodes and branches, leading to some level of variability.

plot(tree, cex = 0.5, main = "Phylogenetic Tree of COX Sequences")

tree$tip.label <- sub(" .*", "", tree$tip.label)

ggtree(tree, layout = "circular") +
  geom_tiplab(size = 3, align = TRUE, linetype = "solid") +
  scale_color_manual(values = clade_colors) +
  geom_point2(aes(subset = !isTip), size = 3) +
  ggtitle("Circular Phylogenetic Tree of COX gene of Halicryptus spinulosus") +
  theme_tree2() +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.text = element_text(size = 12),
        axis.line = element_line())

#After plotting the phylogenetic tree, The map plot was generated. This was to visualize and address wif whether the target organism can be found in different geo graphic locations. As it turned out strains of the target organism, which contained the COX gene were only found across Finland. The package rgrbif was used to find the specie's occurence and then filter out data in accordance to the species who only contained the COX

occurrences <- occ_search(scientificName = "Halicryptus spinulosus", limit = 100)

str(occurrences)

occurrences_data <- occurrences$data

head(occurrences_data)

clean_data <- occurrences_data[!is.na(occurrences_data$decimalLatitude) & !is.na(occurrences_data$decimalLongitude), ]

head(clean_data)

europe_map <- map_data("world", region = c("Norway", "Sweden", "Finland", "Denmark",
                                           "Estonia", "Latvia", "Lithuania", "Russia", "Belarus"))

```



```r
country_labels <- data.frame(
  country = c("Norway", "Sweden", "Finland", "Denmark", "Estonia", "Latvia", "Lithuania", "Russia", "Belarus"),
  long = c(10, 15, 25, 10, 25, 25, 24, 40, 28),
  lat = c(61, 62, 64, 56, 58, 56, 55, 55, 53)
)

ggplot() +
  geom_polygon(data = europe_map, aes(x = long, y = lat, group = group), fill = "lightgreen", color = "white") +
  geom_point(data = geo_data, aes(x = longitude, y = latitude, color = accession), size = 4) +
  scale_color_manual(values = c("red", "blue", "green", "purple", "orange", "cyan", "yellow",
                                "pink", "brown", "magenta", "black", "grey", "darkgreen")) +
  coord_quickmap(xlim = c(20, 32), ylim = c(59, 67)) +
  geom_text(data = country_labels, aes(x = long, y = lat, label = country), size = 5, fontface = "bold") +
  ggtitle("Geographic Distribution of Halicryptus spinulosus COX gene in Finland") +
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5))

tree_plot <- ggtree(tree, layout = "circular") +
  geom_tiplab(size = 3, align = TRUE, linetype = "solid") +
  ggtitle("Phylogenetic Tree of COX gene of H.spinulosus") +
  theme_void() +
  theme(
    plot.margin = unit(c(0.5, 0.5, 0.5, 0.5), "cm")
  )


#The phylogenetic tree and the map plot were combined in the next block of codes to gain better visualization.


ggplot() +
  geom_polygon(data = europe_map, aes(x = long, y = lat, group = group), fill = "lightgreen", color = "white") +
  geom_point(data = geo_data, aes(x = longitude, y = latitude, color = accession), size = 4) +
  scale_color_manual(values = c("red", "blue", "green", "purple", "orange", "cyan", "yellow",
                                "pink", "brown", "magenta", "black", "grey", "darkgreen")) +
  coord_quickmap(xlim = c(20, 32), ylim = c(59, 67)) +
  geom_text(data = country_labels, aes(x = long, y = lat, label = country), size = 5, fontface = "bold") +
  ggtitle("Geographic Distribution of Halicryptus spinulosus COX gene in Finland") +
  theme_minimal() +
```



```r
  scale_color_manual(values = c("red", "blue", "green", "purple", "orange", "cyan", "yellow",
                                "pink", "brown", "magenta", "black", "grey", "darkgreen")) +
  coord_quickmap(xlim = c(20, 32), ylim = c(59, 67)) +
  geom_text(data = country_labels, aes(x = long, y = lat, label = country), size = 5, fontface = "bold") +
  ggtitle("Geographic Distribution of Halicryptus spinulosus COX gene in Finland") +
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5))

tree_plot <- ggtree(tree, layout = "circular") +
  geom_tiplab(size = 3, align = TRUE, linetype = "solid") +
  ggtitle("Phylogenetic Tree of COX gene of H.spinulosus") +
  theme_void() +
  theme(
    plot.margin = unit(c(0.5, 0.5, 0.5, 0.5), "cm")
  )


#The phylogenetic tree and the map plot were combined in the next block of codes to gain better visualization.


ggplot() +
  geom_polygon(data = europe_map, aes(x = long, y = lat, group = group), fill = "lightgreen", color = "white") +
  geom_point(data = geo_data, aes(x = longitude, y = latitude, color = accession), size = 4) +
  scale_color_manual(values = c("red", "blue", "green", "purple", "orange", "cyan", "yellow",
                                "pink", "brown", "magenta", "black", "grey", "darkgreen")) +
  coord_quickmap(xlim = c(20, 32), ylim = c(59, 67)) +
  geom_text(data = country_labels, aes(x = long, y = lat, label = country), size = 5, fontface = "bold") +
  ggtitle("Geographic Distribution of Halicryptus spinulosus COX gene in Finland") +
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5))

combined_plot <- plot_grid(tree_plot, geo_plot, align = "hv", ncol = 2, rel_widths = c(0.3, 0.2))

combined_plot
```

Results and Discussion:

As depicted in Figure 2 above, it was found that Halicyrptus spinulosus is mostly located across Finland. The NCBI hits gave thirty-five sequences of the target specie and were further filtered as thirteen sequences contained the COX gene. Therefore, Halicryptus spinulosus sequences, which contained the COX gene were found in Finland. Even though, it cannot be stated that the aquatic worm was located in different geographic locations, however as depicted in Figure 2, the worm was found to be scattered all across Finland. Since Halicryptus spinulosus were located in different parts of Finland, this still gave an insight as to whether there were any genetic variations in the target specie. In Figure 2, accession number 219644.1 (light blue) and 218463.1 (light green), were found to be the furthest from each other. Figure 1 illustrates that the accession numbers mentioned above are separated by two branch nodes, suggesting that there is moderate genetic distance, thus noticeable genetic divergence. This could indicate that the COX genes compared may have evolved separately to certain extent, pertaining to some degree of genetic variability. The data accumulated during this project, has proved the opposite to the hypothesis presented in the introduction, in which geographic distance does indeed lead to genetic variation.

To further progress this study, different genes of Halicryptus spinulosus should be used to study their phylogeny and geographic locations. During the conduction of this project, it was discovered Priapulid worms also contain histone genes, thus the next step of this study could use histone genes found in Halicryptus spinulosus and compare their phylogeny with different geographic locations. The caveat of this study was the limited sample size, despite the evolutionary importance of Halicryptus spinulosus, there is very limited information regarding the organism, which may have influenced the results obtained.

Acknowledgments:

I would like to thank and acknowledge my classmates who have either helped during the completion of this assignment. I would like to give a special thanks to the TA, Brittany, for helping me throughout this assignment and giving me valuable insight.

References:

1. Janssen, R., Wennberg, S. A., & Budd, G. E. (2009). The hatching larva of the priapulid worm Halicryptus spinulosus. *Frontiers in zoology*, *6*, 8. https://doi.org/10.1186/1742-9994-6-8
2. Kesidis, G., Slater, B. J., Jensen, S., & Budd, G. E. (2019). Caught in the act: priapulid burrowers in early Cambrian substrates. *Proceedings. Biological sciences*, *286*(1894), 20182505. https://doi.org/10.1098/rspb.2018.2505
3. Wang, D., Vannier, J., Schumann, I., Wang, X., Yang, X. G., Komiya, T., Uesugi, K., Sun, J., & Han, J. (2019). Origin of ecdysis: fossil evidence from 535-million-year-old scalidophoran worms. *Proceedings. Biological sciences*, *286*(1906), 20190791. https://doi.org/10.1098/rspb.2019.0791
4. Nanglu, K., & Ortega-Hernández, J. (2024). Post-Cambrian survival of the tubicolous scalidophoran *Selkirkia*. *Biology letters*, *20*(3), 20240042. https://doi.org/10.1098/rsbl.2024.0042