Assignment 2

Amanda Vanderzanden

2024-10-28

Introduction

The *Apis* genus, commonly referred to as bees, encompasses a large number of key pollinators from across the world that are involved in anywhere from 50-75% of pollination of globally significant crops (Requier et al., 2023). A decline in the bee population can have far reaching consequences in the food crop and biodiversity spheres of the global ecosystem (Arpaia et al., 2021; Bartomeus et al., 2011). An understanding of the genetic diversity and pressures on the species can allow a better understanding of their diversity and to guide conservation efforts in future developments (Ilyasov et al., 2021; Requier et al., 2023). Cytochrome oxidase I (COI) and NADH dehydrogenase subunit 1 (ND1) are two mitochondrial genes which can be utilized to understand evolutionary patterns and species diversity in the *Apis* genus (Chuang & Hsu, 2013; Ilyasov et al., 2021). COI is widely used in the study of DNA barcoding as it is highly conserved across species and ND1 acts as a similar but distinct marker which can allow differing perspectives on the genetic significance of different bee species (Chuang & Hsu, 2013; Ilyasov et al., 2021). In this investigation, I aim to identify the utility of these genes for effective cluster analysis and if sequences are more effective at higher grouping numbers.

The Code

Library Setup

```
library(tidyverse)
conflicted::conflicts_prefer(dplyr::filter())
library(Biostrings)
library(viridis)
library(muscle)
library(DECIPHER)
library(rentrez)
library(seqinr)
library(BiocManager)
library(pacman)
library(clValid)
library(factoextra)
library(ggplot2)
library(ape)
library(dendextend)
library(cowplot)
```

Finding the Data

```
#nd_search <- entrez_search(db = "nuccore", term = "Apis[ORGN]
#AND ND2[Gene] AND 300:800[SLEN]", retmax = 100)

#COI_search <- entrez_search(db = "nuccore", term = "Apis[ORGN]
#AND COI[Gene] AND 300:800[SLEN]", retmax = 100)

#The initial search results included the whole genome in some results so the search was
#modified to only include sizes 300-800bp as the target genes
#appear to be around 600bp in both cases.

#Bee_COI_Fetch <- entrez_fetch(db = "nuccore", id = COI_search$ids, rettype = "fasta")
#write(Bee_COI_Fetch, "Bee_COI.fasta", sep = "\n")

#Bee_ND_Fetch <- entrez_fetch(db = "nuccore", id = nd_search$ids, rettype = "fasta")
#write(Bee_ND_Fetch, "Bee_ND.fasta", sep = "\n")
```

Data initially retrieved as a StringSet to be manipulated properly in downstream analysis. Dataframes for each gene are generated for later use with headings that have consistent descriptions.

```
COI_StringSet <- readDNAStringSet("Bee_COI.fasta")
#change target for file as needed for your computer
names(COI_StringSet) <- substr(names(COI_StringSet), 1, 10) #shorten IDS
COI_DF <- data.frame(COI_Title = names(COI_StringSet), COI_Sequence = paste(COI_StringSet))
#convert to dataframe
COI_DF$Species_Name <- word(COI_DF$COI_Title, 2L, 3L)
COI_DF <- COI_DF[, c("COI_Title", "Species_Name", "COI_Sequence")] #consistent headings

ND_StringSet <- readDNAStringSet("Bee_ND.fasta")
#change target for file as needed for your computer
names(ND_StringSet) <- substr(names(ND_StringSet), 1, 10) #shorten IDs
ND_DF <- data.frame(ND_Title = names(ND_StringSet), ND_Sequence = paste(ND_StringSet))
#convert to dataframe
ND_DF$Species_Name <- word(ND_DF$ND_Title, 2L, 3L)
ND_DF <- ND_DF[, c("ND_Title", "Species_Name", "ND_Sequence")] #consistent headings
shortened_labelsND <- substr(ND_DF$ND_Title, 1, 10)

sum(is.na(COI_DF$COI_Sequence))
sum(is.na(ND_DF$ND_Sequence)) #checking for empty rows
sum(sapply(COI_DF$COI_Sequence, function(seq) {
     sum(as.character(seq) == "-")
 }))
sum(sapply(ND_DF$ND_Sequence, function(seq) {
     sum(as.character(seq) == "-")
 })) #checking for gaps in the gathered sequences, none found for either gene
sum(sapply(COI_DF$COI_Sequence, function(seq) {
     sum(as.character(seq) == "N")
 }))
sum(sapply(ND_DF$ND_Sequence, function(seq) {
     sum(as.character(seq) == "N")
 })) #checking for Ns in the gathered sequences, none found for either gene
```

Data is now modified and verified as not having gaps or missing nucleotides for downstream analysis.

Analysis

```
COI_DF$COI_Sequence <- DNAStringSet(COI_DF$COI_Sequence)
names(COI_DF$COI_Sequence) <- COI_DF$COI_Title
#ensure the column labels are conserved
COI.alignment <- DNAStringSet(muscle::muscle(COI_DF$COI_Sequence))
#sequence alignment to ensure duplicates and
#irregular sequences are identified and caught
dnaBin.COI <- as.DNAbin(COI.alignment)
#modify alignment for distance matrix
distanceMatrix.COI <- dist.dna(dnaBin.COI, model = "TN93", as.matrix = TRUE, pairwise.deletion = TRUE)
clusters.COI <- hclust(as.dist(distanceMatrix.COI), method = "single") #clustering of the segments

ND_DF$ND_Sequence <- DNAStringSet(ND_DF$ND_Sequence)
names(ND_DF$ND_Sequence) <- ND_DF$ND_Title
#ensure the column labels are conserved
ND.alignment <- DNAStringSet(muscle::muscle(ND_DF$ND_Sequence))
#sequence alignment to ensure duplicates and irregular sequences are identified and caught
dnaBin.ND <- as.DNAbin(ND.alignment)
#modify alignment for distance matrix
distanceMatrix.ND <- dist.dna(dnaBin.ND, model = "TN93", as.matrix = TRUE, pairwise.deletion = TRUE)
#generate a distance matrix from the alignment
clusters.ND <- hclust(as.dist(distanceMatrix.ND), method = "single")
#clustering of the segments

#cluster analysis was chosen over kmer as even with most robust search terms,
#sequence numbers were lower (less than 1000 individuals per gene)
```

```
COI_dend <- as.dendrogram(clusters.COI) |>
dendextend::set("labels_cex", 0.3)
ND_dend <- as.dendrogram(clusters.ND) |>
dendextend::set("labels_cex", 0.3)
par(mar = c(3,3,1,1))
nf <- layout(matrix(c(1,2), nrow=2))
COI_dend |>
  set("branches_k_color") |>
  plot(main = "Dendrogram of COI gene in Apis", ylim = c(0,0.05))
ND_dend |>
  set("branches_k_color") |>
  plot(main = "Dendrogram of ND1 gene in Apis", ylim = c(0,0.05))
#generate dendogram plots of both cluster analyses, cutting the y axis
#to 0.05 to be able to see the clustering better, one cluster on both
#had a massive height and made it impossible to see any separation of
#the lower clusters
```

```
COI.cut.2 <- cutree(clusters.COI, 2)
COI.cut.3 <- cutree(clusters.COI, 3) #cut clustering data into 2 and 3 groupings, respectively
COI.sil.2 <- silhouette(COI.cut.2, dist(distanceMatrix.COI))
COI.sil.3 <- silhouette(COI.cut.3, dist(distanceMatrix.COI))
```

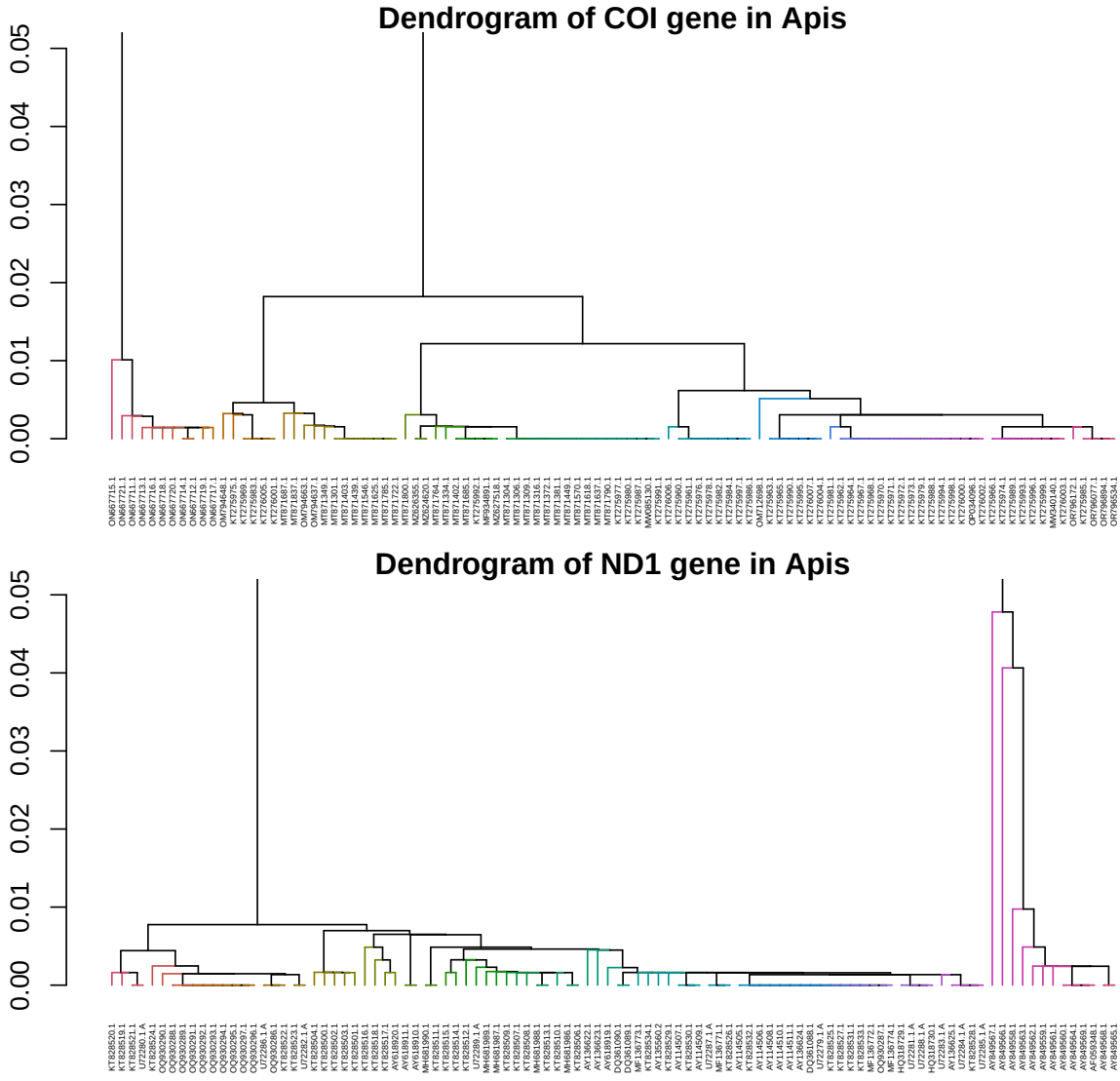

Figure 1: Dendrogram Results for Cluster Analysis of COI and ND1 genes in Apis species.

```
ND.cut.2 <- cutree(clusters.ND, 2)
ND.cut.3 <- cutree(clusters.ND, 3) #cut clustering data into 2 and 3 groupings, respectively
ND.sil.2 <- silhouette(ND.cut.2, dist(distanceMatrix.ND))
ND.sil.3 <- silhouette(ND.cut.3, dist(distanceMatrix.ND)) #run sil analysis on each set

COI_sil_2_plot <- fviz_silhouette(COI.sil.2) +
     ggtitle("COI Clustering with 2 Groupings") +
     labs(y = "Silhouette Width", x = "Cluster Sequence") +
     theme_minimal() +
     theme(axis.text.x = element_text(size = 2))
COI_sil_3_plot <- fviz_silhouette(COI.sil.3) +
    ggtitle("COI Clustering with 3 Groupings") +
    labs(y = "Silhouette Width", x = "Cluster Sequence") +
    theme_minimal() +
    theme(axis.text.x = element_text(size = 2))
COI_combined_plots <- plot_grid(COI_sil_2_plot, COI_sil_3_plot, ncol = 2, align = "h")
print(COI_combined_plots)

ND_sil_2_plot <- fviz_silhouette(ND.sil.2) +
    ggtitle("ND1 Clustering with 2 Groupings") +
    labs(y = "Silhouette Width", x = "Cluster Sequence") +
    theme_minimal() +
    theme(axis.text.x = element_text(size = 3))
ND_sil_3_plot <- fviz_silhouette(ND.sil.3) +
    ggtitle("ND1 Clustering with 3 Groupings") +
    labs(y = "Silhouette Width", x = "Cluster Sequence") +
    theme_minimal() +
    theme(axis.text.x = element_text(size = 3))
ND_combined_plots <- plot_grid(ND_sil_2_plot, ND_sil_3_plot, ncol = 2, align = "h")
print(ND_combined_plots)
#visualize silhouette analysis for each of the groupings, divided by starting gene.
```

The silhouette plots (Figure 2 and 3) show consistently that the clustering into 2 groupings has a higher strength than the 3 groupings - both sit close to exactly 1.0 cluster strength with 2 groupings while the 3 groupings vary much more. With COI, the largest cluster group drops in strength from 0.98 to an average of 0.62. In the ND1 analysis, the 2 grouping cluster shows similar trends to the COI, with the majority of samples sitting close to 1.0 (0.95 and 0.91). Where ND1 varies is the 3 grouping cluster. The largest group is still close to 1.0 (0.95) but the second group shows a significant difference with an average of 0.51 but some sequences falling below 0 into the negatives. In addition, the third grouping is incredibly small, consisting of a single sequence, and sits at a cluster score of 0.

```
COI_Dunn2 <- dunn(distanceMatrix.COI, COI.cut.2)
COI_Dunn3 <- dunn(distanceMatrix.COI, COI.cut.3)
ND_Dunn2 <- dunn(distanceMatrix.ND, ND.cut.2)
ND_Dunn3 <- dunn(distanceMatrix.ND, ND.cut.3)
#run Dunn analysis on usable data that was cut in previous set for sil analysis
Dunn_Results <- data.frame(
  Dataset = c("COI", "COI", "ND1", "ND1"),
  k = c(2, 3, 2, 3),
  DunnIndex = c(COI_Dunn2, COI_Dunn3, ND_Dunn2, ND_Dunn3)
) #create dataframe for visualizing results
ggplot(Dunn_Results, aes(x = factor(k), y = DunnIndex, fill = Dataset)) +
  geom_bar(stat = "identity", position = "dodge") +
```

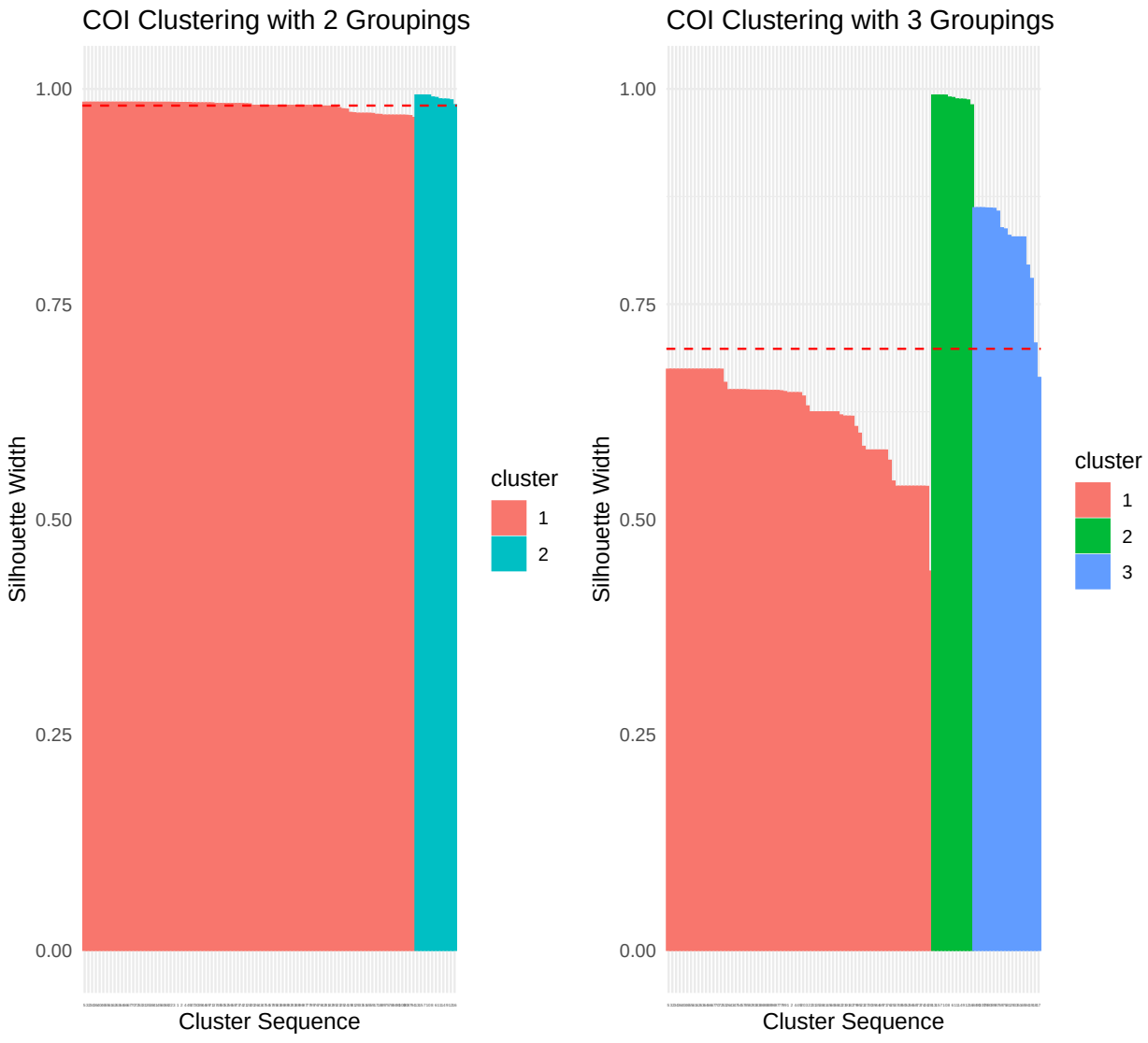

Figure 2: Silhouette plots of clustering data for COI gene using 2 and 3 cluster groupings for sample analysis.

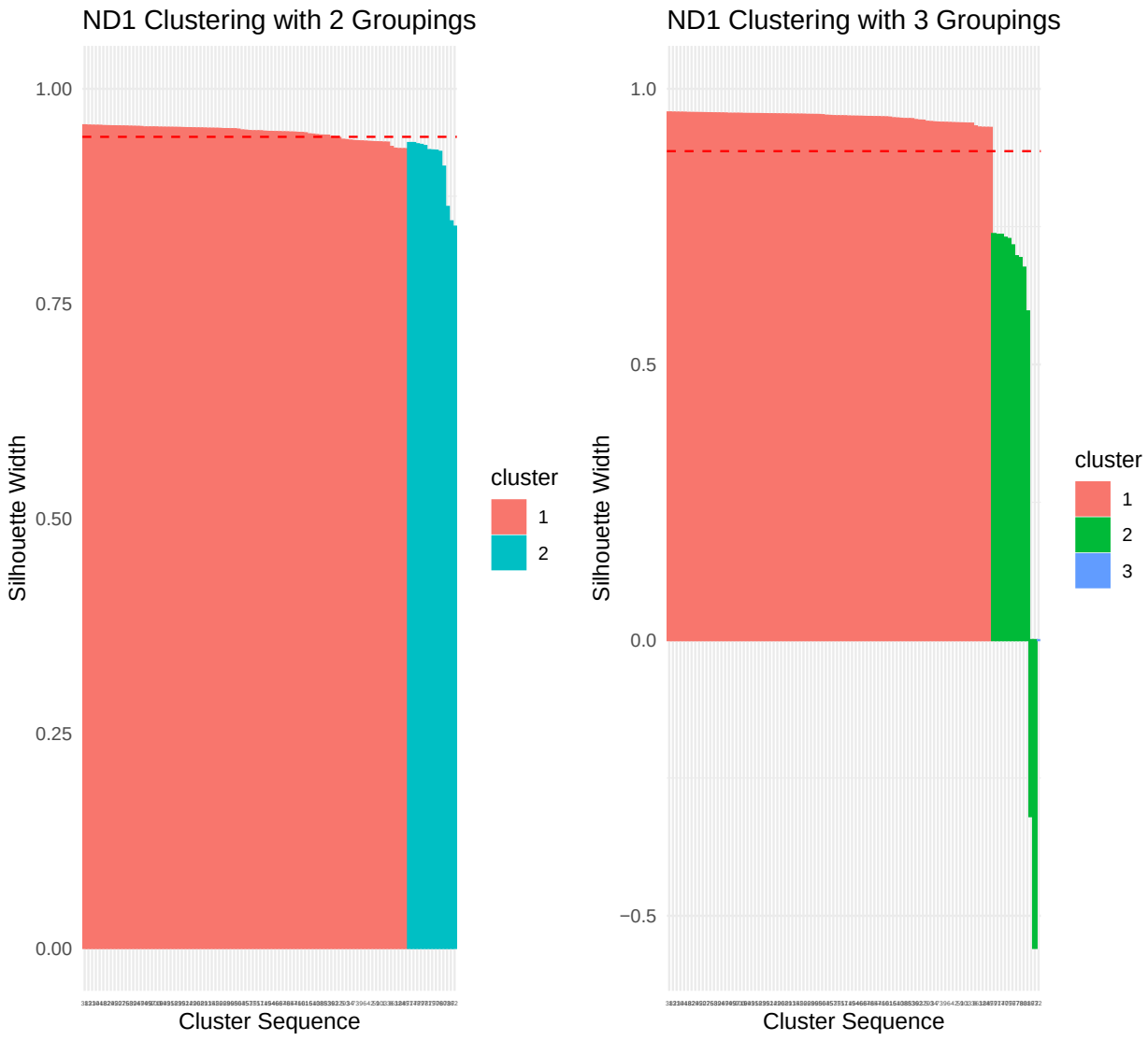

Figure 3: Silhouette plots of clustering data for ND1 gene using 2 and 3 cluster groupings for sample analysis.

```
labs(title = "Dunn Index Results for COI and ND1 Clustering in Apis Species",
     x = "Number of Groupings",
     y = "Dunn Index Result") +
theme_minimal() #comparison visualization of the results in a bar chart
```

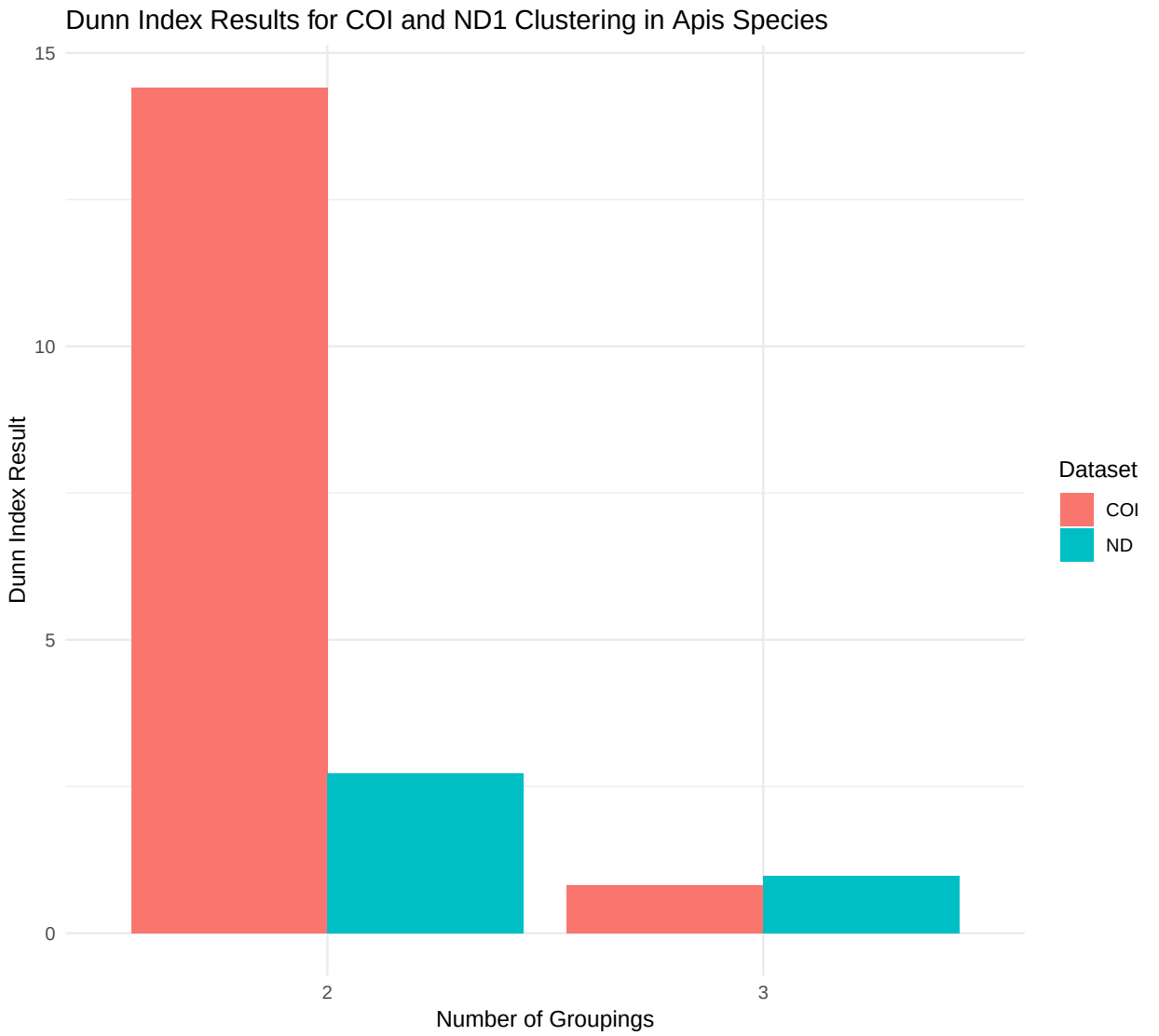

Figure 4: Bar graph of Dunn Index results for COI and ND1 clustering analysis at 2 and 3 cluster groupings.

The bar graph (Figure 4) demonstrates that the 2 grouping clusters for COI and ND1 both have a higher Dunn Index (14.4 and 2.72, respectively) than the 3 grouping clusters (0.816 and 0.975, respectively). These show a much stronger clustering distinction and relation between sequences at the 2 group stage than the 3.

Discussion and Conclusion

The Silhouette and Dunn Index analysis both showed a higher strength of sequence clustering and similarity at the 2 grouping cluster than the three (Fig 2-4). The COI results were consistently high for the 2 grouping data, with a Dunn Index of 14, indicating that it is both strongly associated and genetically distinct from

the other groups. The ND1 gene showed a higher cluster strength at 2 than 3 groupings but not to the same degree as COI. The Silhouette Index for COI was very strong and consistent at 2 groupings and had a big drop for two of the clusters when moving to three groups. The ND1 gene showed great strength at 2 groupings and very poor performance at 3 clusters. The single gene present in one group as well as scores varying so widely in cluster 2 indicate that the clustering is not effectively done at a grouping of 3.

The COI gene shows a high amount of strength in its clustering across both levels of groupings. The ND1 gene, in contrast, does not have the same level of strength as the COI analysis. This indicates that COI clusters well at 2 groupings that are both distinct and separate. The use of COI is fairly common in species identification and is known to be consistent for species identification (Ratnasingham & Hebert, 2007), so it is expected to show strong clustering patterns that are useful for species identification. The ND1 gene shows some clustering strength at the 2 grouping level but lowers greatly at 3 groupings. This can indicate that ND1 is more sensitive to species separation and could require further separation to accurately represent the differences in sequences between species. Through this analysis, I have demonstrated the effectiveness of clustering analysis using COI genes as they maintain their strength in different groupings and when analysed in multiple forms of clustering evaluation. The ND1 gene may require further study with either a more robust dataset, different forms of clustering, or more levels of grouping, to capture the full diversity of the gene. In future research, the species diversity and spread of the *Apis* genus is crucial for the continued survival of both the bee and human populations (Arpaia et al., 2021; Bartomeus et al., 2011, Ilyasov et al., 2021; Requier et al., 2023). The more data and avenues of species biodiversity that can be studied can allow for a greater understanding of these crucial pollinators around the world.

Acknowledgements

A continuous thank you to Dr. Karl Cottenie and Brittany MacIntyre for their regular instruction and advice throughout the process of this course. Additional thank you goes to Noah Zeidenberg, Pheonix Armstrong, and Dhruv Mishra for continuous soundboarding and emotional support throughout the process. Of course, the largest thank you must go to my coding partner, Aqua of Tofana, (Supplementary Figure 1) who patiently slept at my feet while I typed and retyped this assignment.

Supplementary Figures

Supplementary Figure 1: The goodest girl, Ms. Tofana.

References

Arpaia, S., Smagghe, G., & Sweet, J. B. (2021). Biosafety of bee pollinators in genetically modified agro-ecosystems: Current approach and further development in the EU. In *Pest Management Science* (Vol. 77, Issue 6, pp. 2659–2666). John Wiley and Sons Ltd. https://doi.org/10.1002/ps.6287

Bartomeus, I., Ascher, J. S., Wagner, D., Danforth, B. N., Colla, S., Kornbluth, S., & Winfree, R. (2011). Climate-associated phenological advances in bee pollinators and bee-pollinated plants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), 20645–20649. https://doi.org/10.1073/pnas.1115559108

Chuang, Y. L., & Hsu, C. Y. (2013). Changes in mitochondrial energy utilization in young and old worker honeybees (Apis mellifera). *Age*, *35*(5), 1867–1879. https://doi.org/10.1007/s11357-012-9490-y

Ilyasov, R. A., Han, G. Y., Lee, M. L., Kim, K. W., Park, J. H., Takahashi, J. I., Kwon, H. W., & Nikolenko, A. G. (2021). Phylogenetic Relationships among Honey Bee Subspecies Apis mellifera caucasia and Apis mellifera carpathica Based on the Sequences of the Mitochondrial Genome. *Russian Journal of Genetics*, *57*(6), 711–723. https://doi.org/10.1134/S1022795421060041

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System: Barcoding. *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Requier, F., Pérez-Méndez, N., Andersson, G. K. S., Blareau, E., Merle, I., & Garibaldi, L. A. (2023). Bee and non-bee pollinator importance for local food security. In *Trends in Ecology and Evolution* (Vol. 38, Issue 2, pp. 196–205). Elsevier Ltd. https://doi.org/10.1016/j.tree.2022.10.006