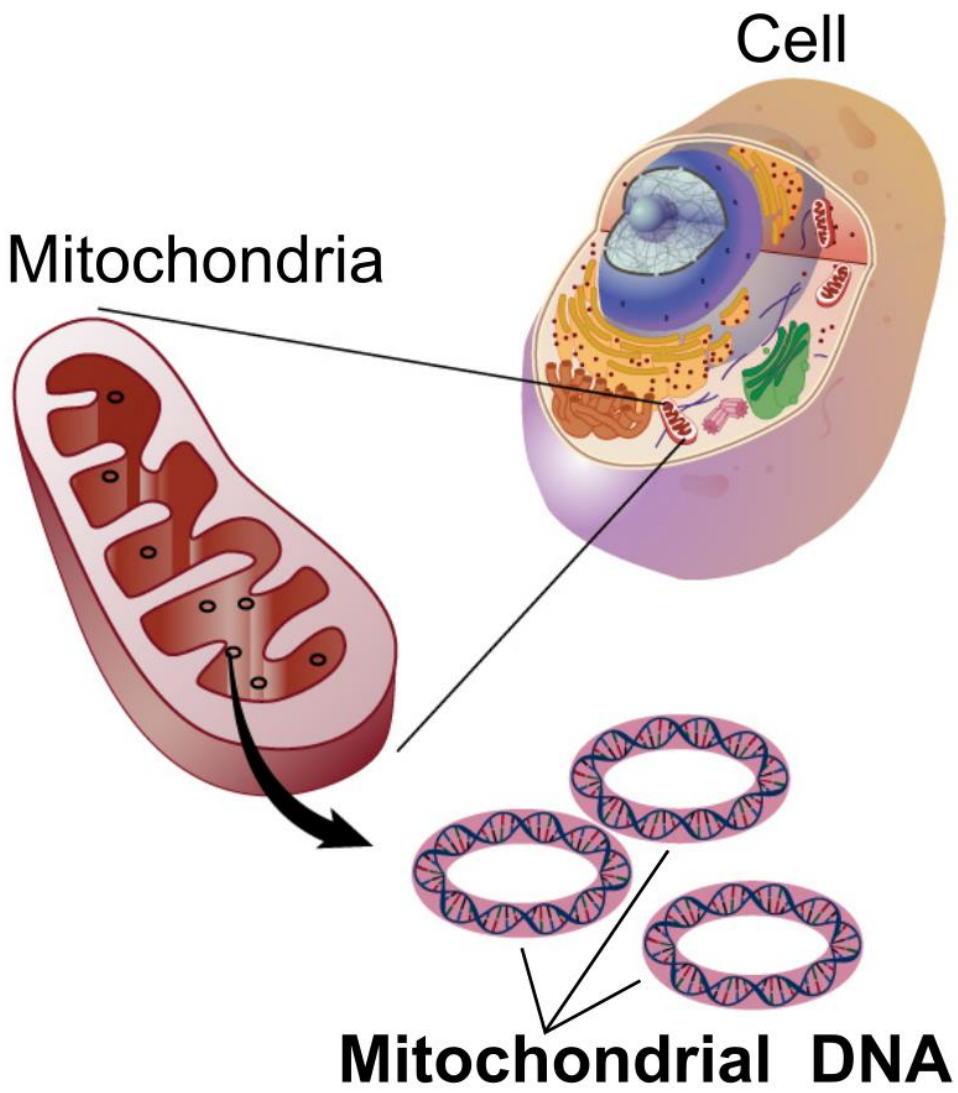
Exploring the differences in Clustering patterns between ND2 and Cyt-b genes using the Drosophilidae Family as a Case Study

Storyboard by Iroayo Toki

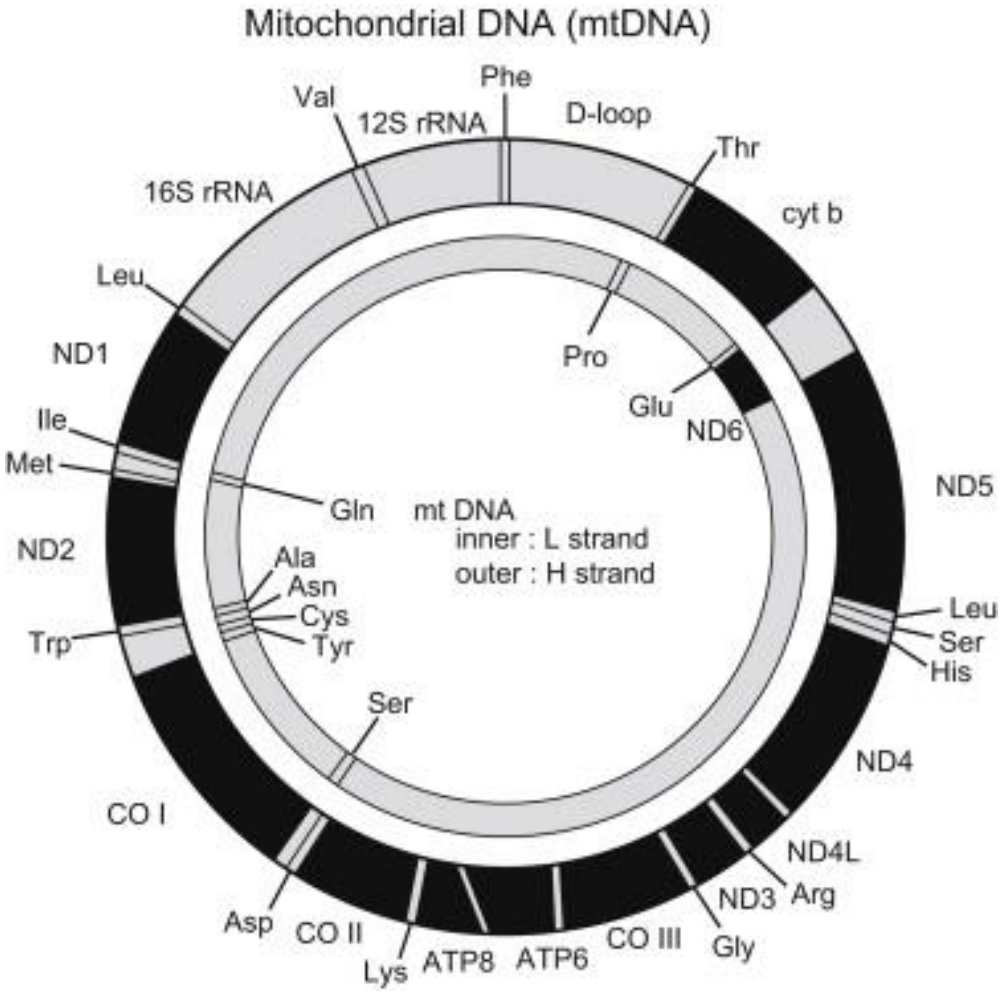


Introduction

- Mitochondrial genes are often used for phylogenic studies due to the relative ease of sequencing when compared with other genes.
- Another reason they are used is rapid evolution rate which can be used to find differences in closely related Taxa due to their sequence variability. (Finnegan et al., 2025).
- They are also passed down through maternal inheritance which means they generally do not undergo recombination and can make lineage tracing easier . (Manisha et al., 2023)

Introduction

- NADH Dehydrogenase Subunit 2(ND2) and Cytochrome b (Cyt-b) are two commonly used mitochondrial genes.
- The key differences between the two of them are:
- Difference in sequence length: While both sequence lengths are comparable, ND2 is generally longer than Cyt-B and this is also reflected in this study.
- Difference in **sequence variability**: Cyt-B is generally more conserved than ND2 and has more identical nucleotides.(Campillo et al., 2019)
- Due to this reason, we can compare the ability of these genes to function as classifiers of sequences and as a result classifiers of taxonomic groups.

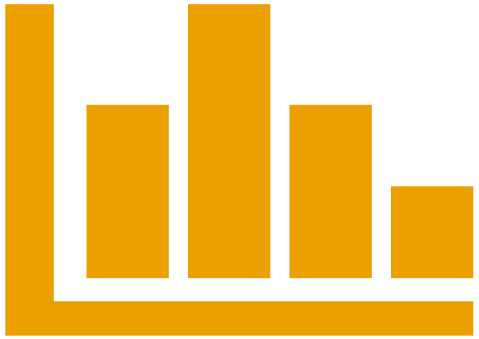


Hypotheses and Research Question

- Due to the reasons listed above we can then generate these interesting contrasting hypotheses.
- Highly variable genes act as great classifiers as their divergence shows evolutionary signals than can be used to spot differences between taxa.
- Better conserved genes with reduced variability act as better classifiers as they would keep ancestral relationships.
- These hypotheses can the help us to generate the following research question:
- Do fast-evolving mitochondrial genes and slow-evolving mitochondrial genes differ in their ability to resolve taxonomic relationships?

Objectives

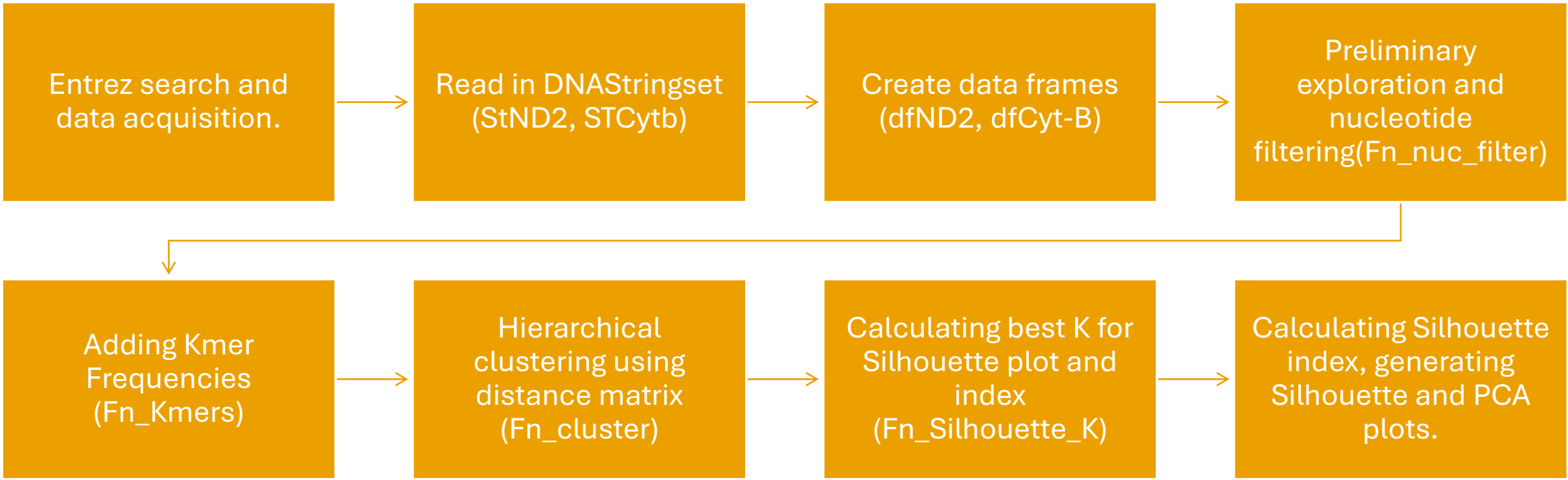
- The main Objective is to **compare hierarchical clustering patterns between the two genes and use internal measures of cluster strength to understand the reliability of these clustering patterns.**
- To achieve this, we must consider the key assumptions of hierarchical clustering and use them to develop our filtering and analysis objectives
- It assumes that the data quality includes comparable sequences- **Check for length variability across each gene and missing data or Ns.**
- It assumes that the distance metric used properly reflects the similarity of the sequences- **Use K-mer (Dinucleotide and Trinucleotide) frequencies to accurately depict the relationship between individual sequences.**
- The silhouette index assumes that number of clusters(k) chosen for silhouette plots are meaningful to the data- **Use silhouette width to estimate most meaningful k value to represent the data**



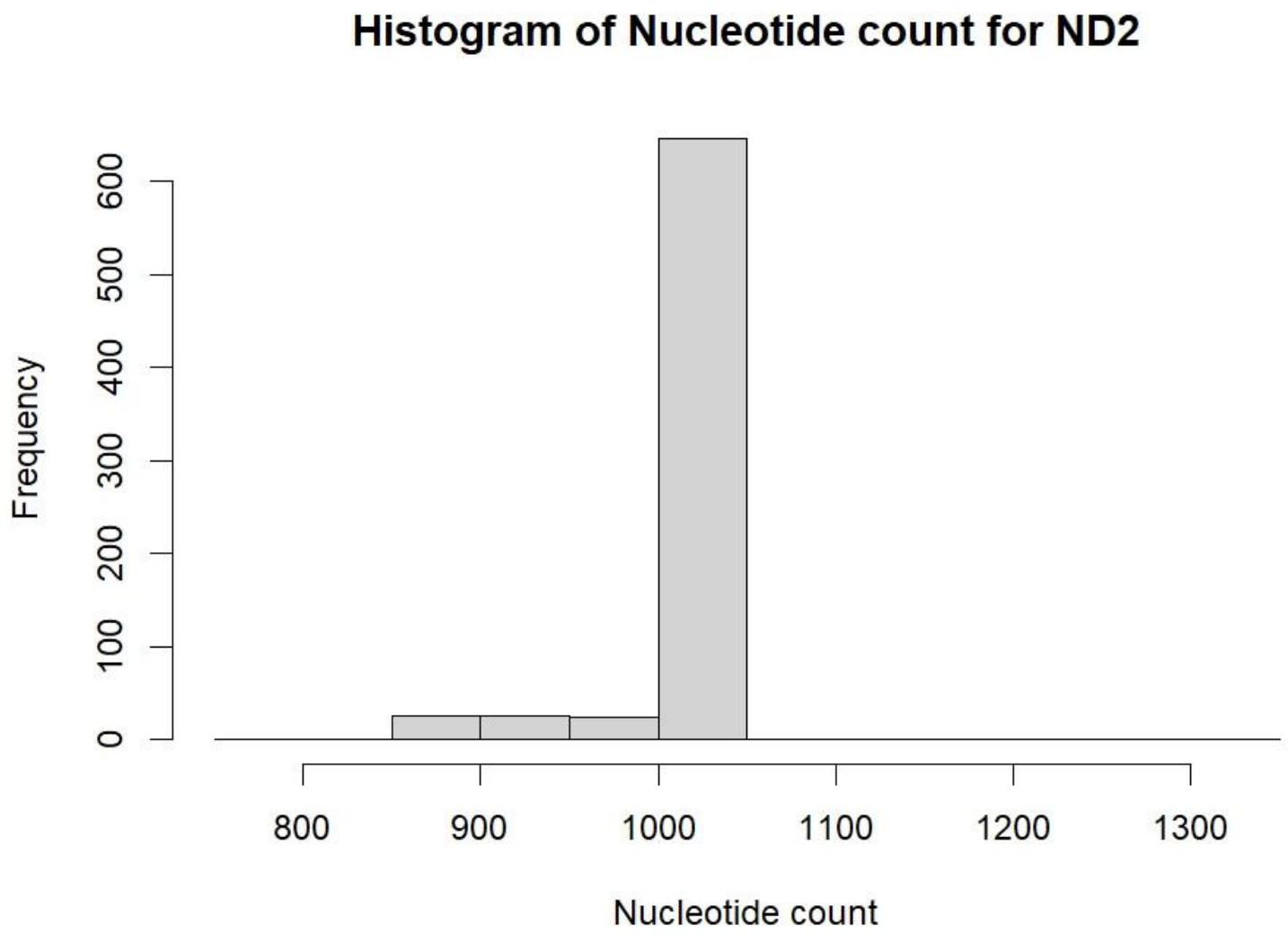
Description of Data set

- To study these objectives the 2 data sets that were chosen were for the insect family Drosophilidae that mainly comprises of fruitflies. Datasets were obtained for the ND2 and Cyt-B gene separately.
- This data was obtained on the 28th of November., 2025 from the nucleotide database of the **National Center for Biotechnology Information(NCBI)** website.
- The data was generated using the entrez search terms “drosophilidae [ORGN] AND ND2 [gene]” for ND2 and drosophilidae [ORGN] AND Cytb [gene] for cyt-B.
- They were then downloaded as .fasta files and read in into R-studio as DNAStringsets. This contained 1336 samples for ND2 and 1022 samples for Cyt-b.
- After filtering steps, they then contained 726 samples for ND2 and 541 samples for Cyt-b

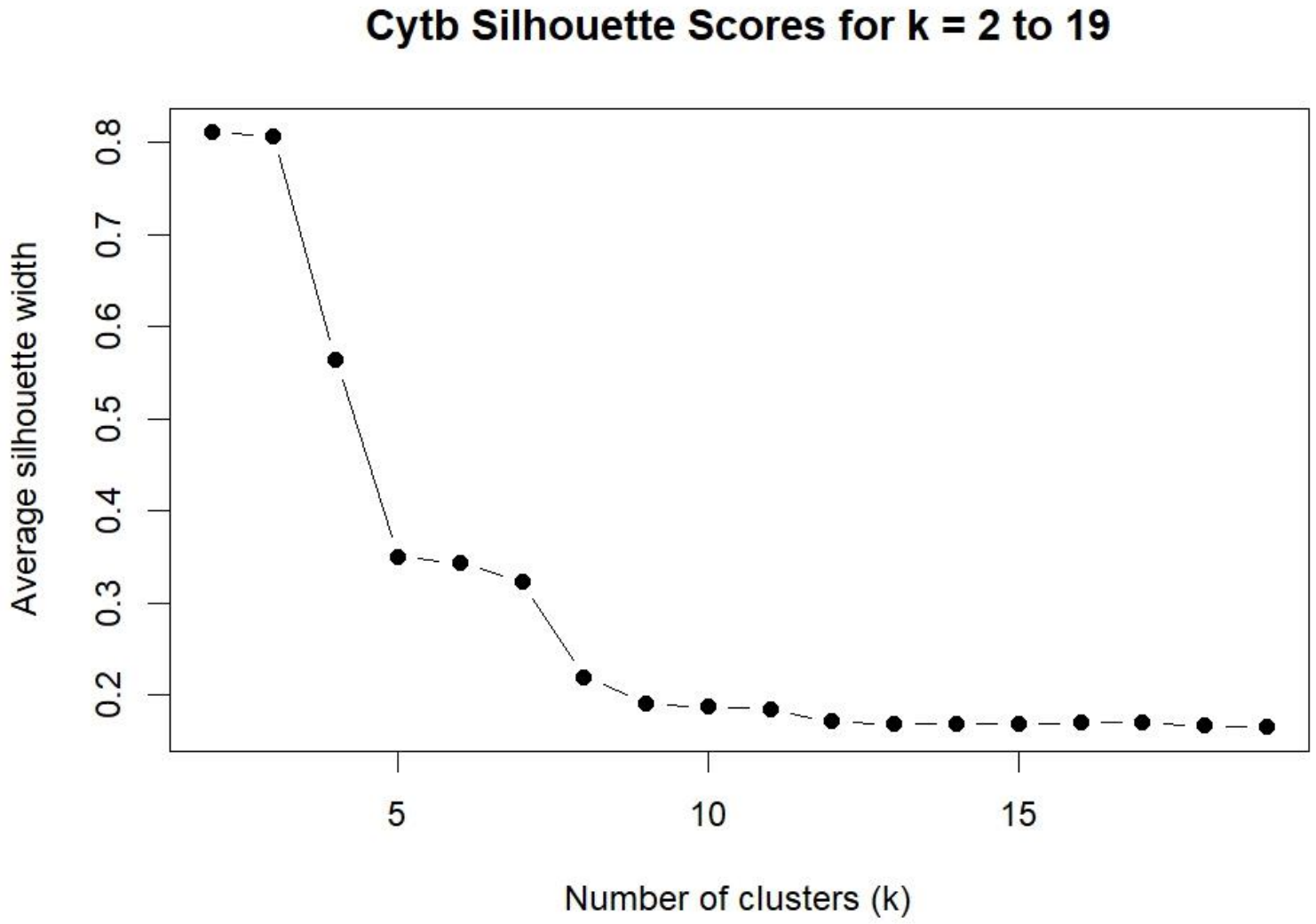
Methods


Entrez search and data acquisition.
Read in DNAStringset (StND2, STCytb)
Create data frames (dfND2, dfCyt-B)
Preliminary exploration and nucleotide filtering(Fn_nuc_filter)
Adding Kmer Frequencies (Fn_Kmers)
Hierarchical clustering using distance matrix (Fn_cluster)
Calculating best K for Silhouette plot and index (Fn_Silhouette_K)
Calculating Silhouette index, generating Silhouette and PCA plots.

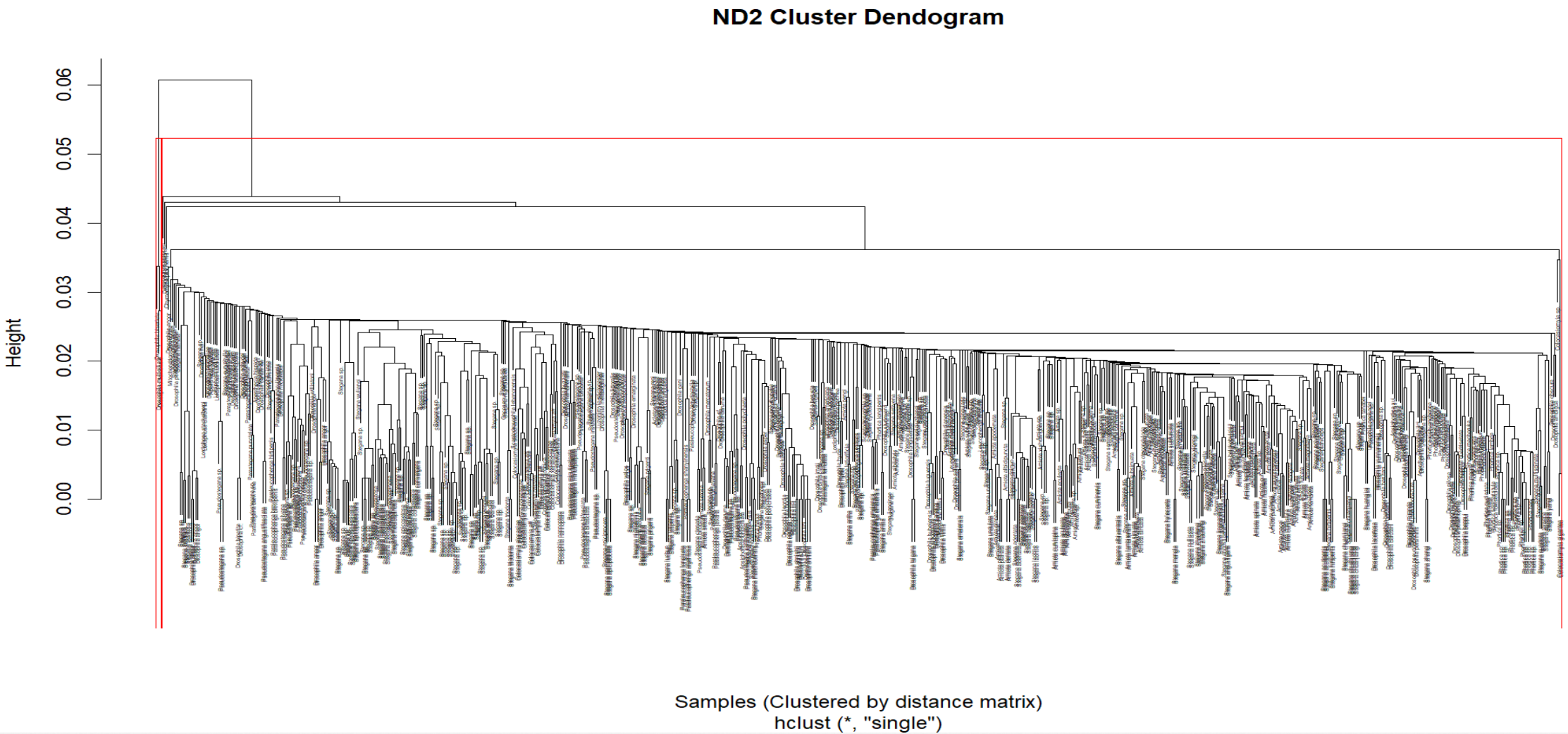
Nucleotide count aggregate near the median=1026 after filtering. This was true for both datasets



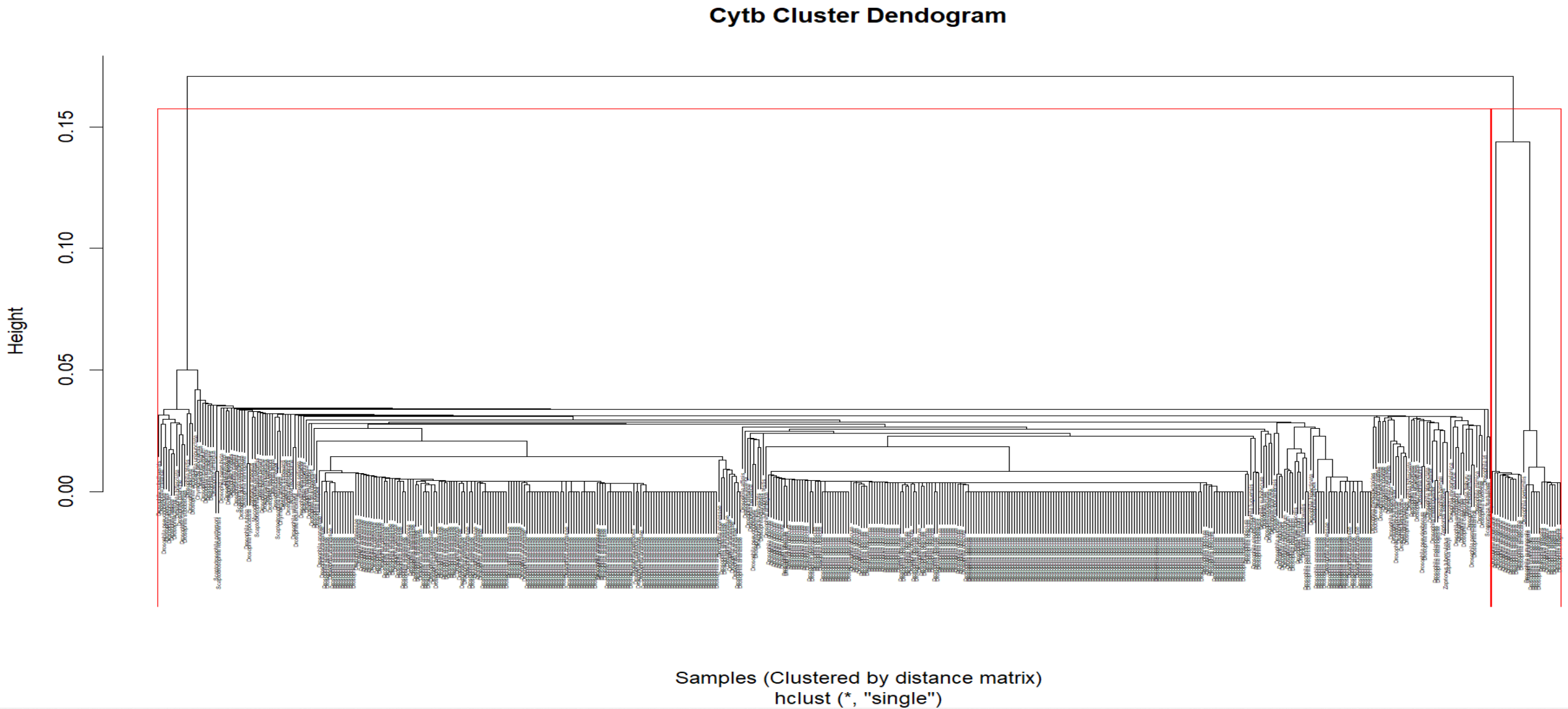
Silhouette scores are best at number of clusters(k) =2 for both datasets, after that there is a sharp decline



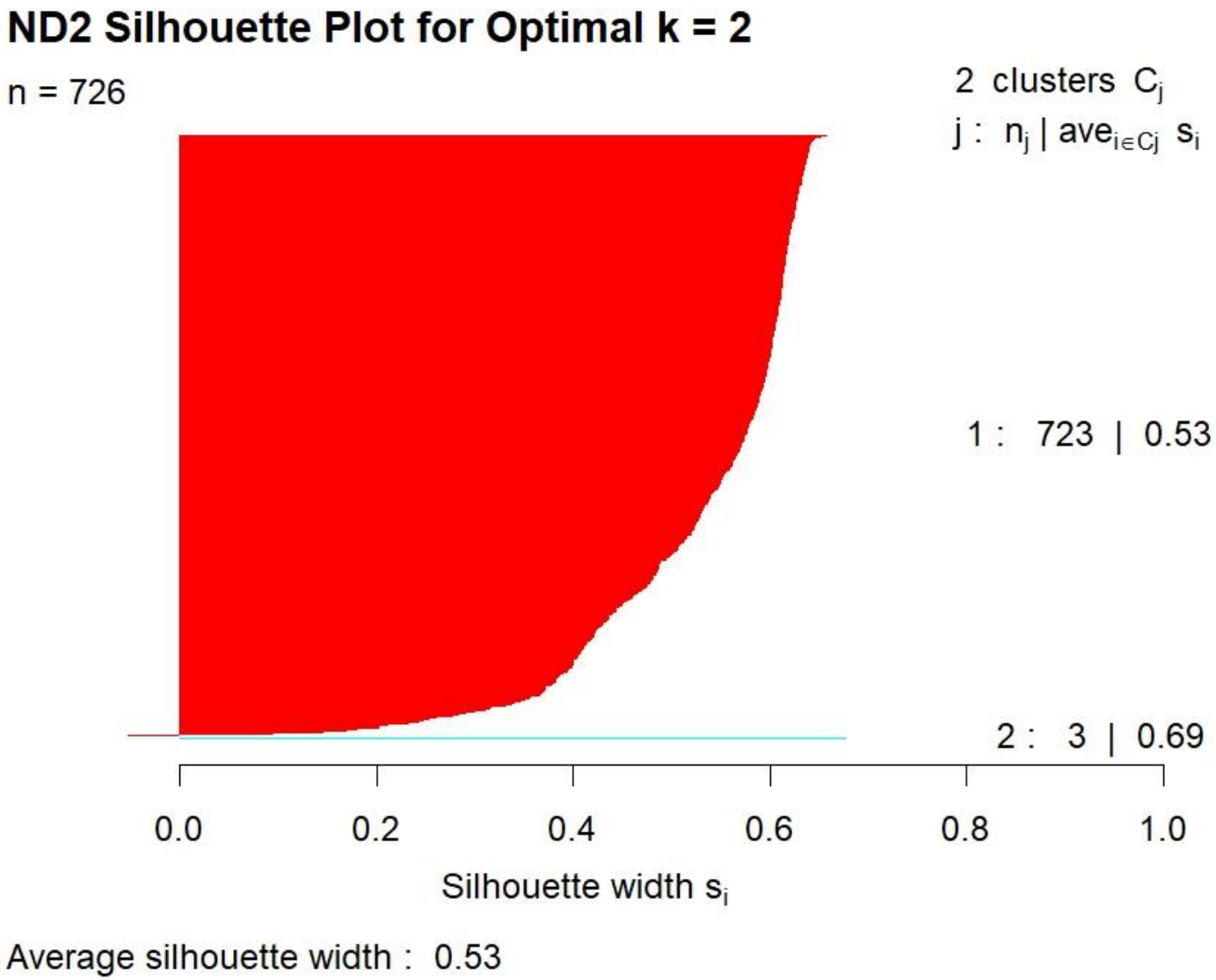
The dendogram has highly variable branch lengths and the minor cluster for ND2 is extremely small and not structured well.



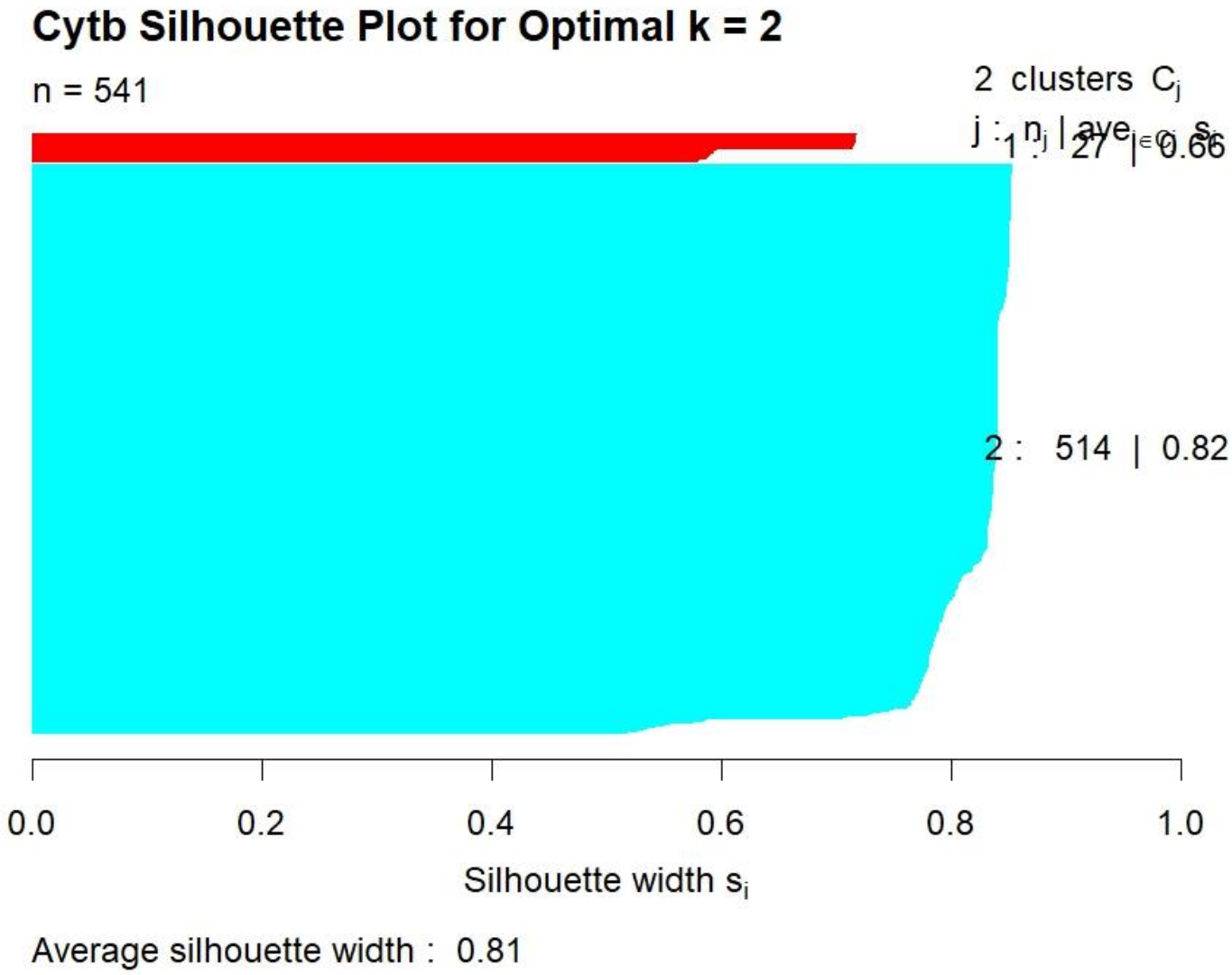
The dendogram has highly similar branch lengths and the minor cluster for Cyt-b is well structured .



Silhouette plot for ND2 shows a small average silhouette width and a poorly defined minor cluster with only 3 data points.



Silhouette plot for ND2 shows a large average silhouette width of 0.81 and a well-defined minor cluster with 27 data points.



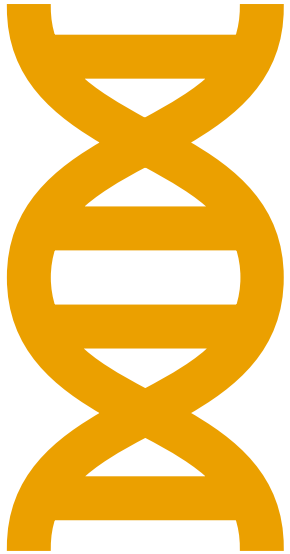
Results

- On average ND2 genes have longer sequences than Cyt-B genes mean = 1013 vs 748, this is supported by previous literature. (Sansook., 2009)
- The silhouette index describes both datasets can best be divided into 2 strong cluster groups with **Cyt-b** having more **well-defined clusters** and an index of **0.81** while **ND2** has an index of **0.53**
- The k-divided dendogram supports the above notion and provides more information as to how Cyt-b minor cluster is more well structured.
- These are further supported by the PCA plots (Present in the figs file)
- The dendograms also show ND2 branches having more variable length when compared to Cyt-b.

Discussion

- Our results show that fast evolving and slow evolving genes do behave differently in a number of ways.
- Firstly, we are shown that at the taxonomic level used for this research (Drosophilidae **Family**), Cyt-b the slower evolving gene acts as a better classifier probably due to its ability to remain conserved with small changes over longer periods of time thus maintaining its ability to show the ancestral relationship.
- This is supported by existing literature as a study done on mammals in 2010 shows Cyt-B having high classification ability at the Super order, order and family taxonomic levels. (Tobe et al., 2010)
- Secondly, the highly varying branch length of ND2 suggests that while it might not be a very accurate classifier at higher levels it could prove useful at lower levels like species and subspecies levels.
- This is again corroborated by a recent study that showed that combining Cyt-b and ND2 genes act as a stronger classifier than any singular mitochondrial gene. (Finnegan et al., 2025)
- Our final hypothesis can then be modified to state that:
- The classification ability of a mitochondrial gene is determined by both the nucleotide variability (fast vs slow evolving) and the taxonomic level being classified.

Limitations (Caveats)



- In publicly available databases like NCBI, there is a sampling bias towards some genes like COI that heavily represents most of the sequence data, this makes it hard to find large datasets to compare other genes.
- Most genes include different numbers of species, and the same pattern appears here. This can affect comparisons because randomly subsampling the larger dataset would throw away useful information, while limiting the analysis to only overlapping species might leave too few taxa to work with.
- Due to the differing structure of the genes across different taxonomic groups, the results of this study cannot be generalized unless further studies are conducted.

Next steps

- Comparing clustering differences of both genes in other taxonomic classes.
- Comparing clustering clustering differences of both genes at different taxonomic levels (order, class e.t.c.)
- Investigating the effect of combining Cyt-b and ND2 and other fast and slow evolving genes on the accuracy of a taxonomic classifier.

Reflection

- During the course of this project and the course at large I have learnt a few things.

i. Data acquisition should always be done with an good understanding of the research question and objectives to help select the right dataset.

ii. Data filtering should also be done contextually with full understanding of your data structure(for length variability of this dataset the numbers were changed from that in the example script as different genes have different acceptable length variability.

iii. From this project, the bioinformatics seminar and other classes, a pattern of combining methods and analysis is recurrent, this shows that to get better results one often has to combine different tools and functions properly even the conclusion of this projects suggests that combining genes might work as a better classifier.

Acknowledgements

- I would like to acknowledge Dr Karl Cottenie and the whole Maters of Bioinformatics cohort for the contributions during the various class activities and group activities where we discussed many issues ranging from correct coding syntax to reducing code redundancy and coming up with project ideas. I would also like to acknowledge my course mate and friend Sodiq Oluwaseun Dada for the continued correspondence on code correction, technical issues and project ideas.

References

- Campillo LC, Burns KJ, Moyle RG, Manthey JD. Mitochondrial genomes of the bird genus *Piranga*: rates of sequence evolution, and discordance between mitochondrial and nuclear markers. Mitochondrial DNA B Resour. 2019 Jul 16;4(2):2566-2569. doi: 10.1080/23802359.2019.1637286. PMID: 33365629; PMCID: PMC7687373.
- Finnegan N, Lima MGM, Lynch JW. Mitochondrial DNA for Phylogeny Building: Assessing Individual and Grouped mtGenes as Proxies for the mtGenome in Platyrrhines. Am J Primatol. 2025 Mar;87(3):e70017. doi: 10.1002/ajp.70017. PMID: 40059324; PMCID: PMC11891386.
- Maechler, M., Rousseeuw, P., Struyf, A.,Hubert, M., Hornik, K.(2025). cluster: Cluster Analysis Basics and Extension s. R package version 2.1.8.1.
- Manisha Munasinghe, J. Arvid Ågren. When and why are mitochondria paternally inherited?. Current Opinion in Genetics & Development, Volume 80, 2023, 102053. ISSN 0959-437X. https://doi.org/10.1016/j.gde.2023.102053.
- Sansook Boonseub, Shanan S. Tobe, Adrian M.T. Linacre. The use of mitochondrial DNA genes to identify closely related avian species. Forensic Science International: Genetics Supplement Series, Volume 2, Issue 1, 2009. 275-277, ISSN 1875-1768. https://doi.org/10.1016/j.fsigss.2009.08.050.
- Tobe SS, Kitchener AC, Linacre AM. Reconstructing mammalian phylogenies: a detailed comparison of the cytochrome B and cytochrome oxidase subunit I mitochondrial genes. PLoS One. 2010 Nov 30;5(11):e14156. doi: 10.1371/journal.pone.0014156. PMID: 21152400; PMCID: PMC2994770.