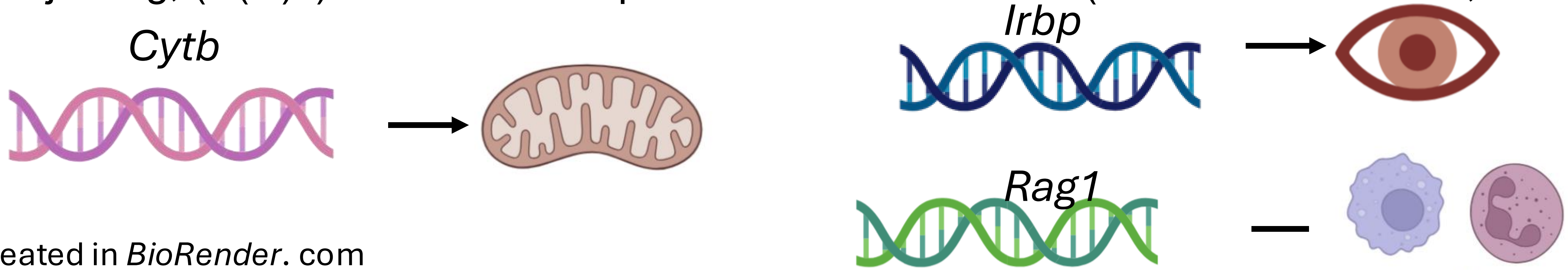
Evaluation of Different Machine Learning Classifiers in the *Muridae* Taxon Group

Assignment 04
BINF 6210
By: Liona Vu

The *Muridae* taxon and its importance in biomedical research

The *Muridae* group is a scientifically important taxonomic class as it contains known research model organisms such as both the *Mus musculus* (Standley et *al.*, 2024) and the *Rattus norvegicus* (Modlinska & Pisula, 2020) species. In fact, decades of important discoveries have been done extensively in *Mus musculus* (Standley et *al.*, 2024). Since mice and rats are important to biomedical research, having the ability to distinguish *Muridae* species can help with the reproducibility and validation of biological results.

Various genes have been well-studied in the *Muridae* group. Thus, its roles are well established in their respective biological processes. In fact, these genes are well annotated in *Muridae* making them ideal for comparative genomic studies. These genes include, *Cytb,* a mitochondrial gene that codes for a protein that is part of complex III in the electron transport chain, *Irbp*, a gene known for its function in eye vision, and *Rag1*, a gene involved in the variable, diversity, and joining, (V(D)J) recombination process of immune cells (UniProt Consortium, 2025).



Created in *BioRender*. com

Supervised machine learning can be used for classification

Many supervised machine learning models algorithms have been developed over the past decades. In fact, machine learning can be applied to distinguish species and validate previous work on species classification (Salles & Domingos., 2025). In addition, machine learning is useful to detect subtle differences between species that previous methods may have overlooked.

One feature that is commonly used for machine learning sequence analysis is k-mers, combinations of nucleotides A, T, C, and G of varying lengths such as dinucleotide k-mers (for a total of 16 combinations). This alignment-free approach is useful and computationally efficient because the number of sequencing data increases with each passing day.

The current gap in knowledge is that with larger genomic datasets, more automated analysis for species identification is needed while preserving accuracy and reliability. Thus, supervised machine learning models can address this.

Dinucleotide combos

AA
AT
AC
AG
TA
TT
TC
TG

CA
CT
CC
CG
GA
GT
GC
GG

Why use several machine learning classification algorithms?

Each machine learning model has its own inherit strengths and weaknesses. For example, random forests (RF) can control for overfitting and has high predictive accuracy (Sipper & Moore., 2021) but can be slow if too many trees are introduced. Partial least squares (PLS) use dimension reduction techniques to make data easier to interpret visually when there are lot of features but can easily overfit the data (Brereton & Lloyd., 2014).

By using multiple machine learning algorithms, it can cover the weaknesses of another model while strengthening the robustness of classification analysis. In addition, it can allow for direct comparisons and see whether each result of each machine learning algorithm is consistent with each other.

Defining the Hypothesis and Objectives

Hypothesis

K-mer dinucleotide frequencies have enough succinct features such that supervised machine learning algorithms can accurately train and detect different gene sequences and *Muridae* species

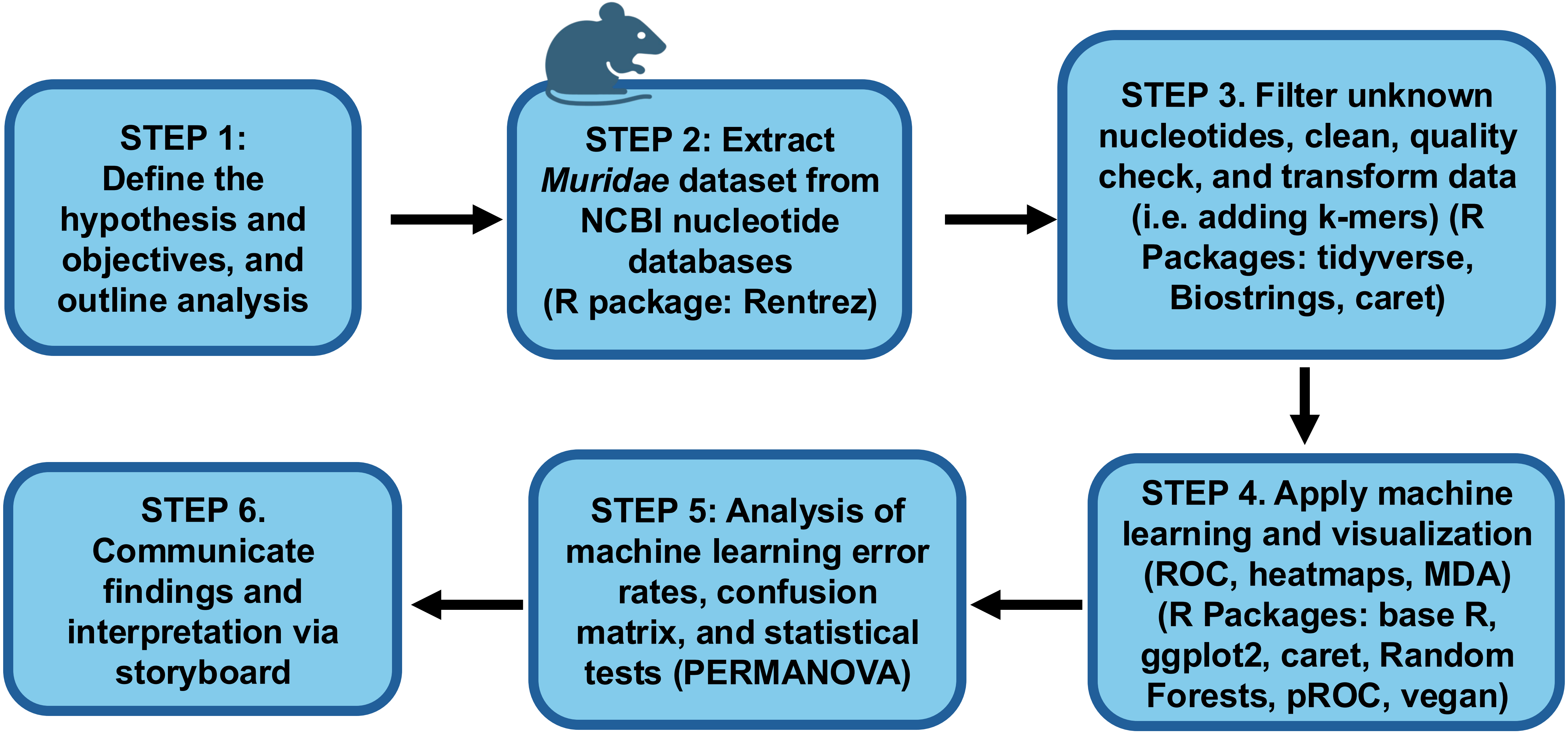
Null hypothesis

K-mer dinucleotide frequencies do not differ amongst different gene sequences and *Muridae* species (and is due to random chance)

Objectives

1. Apply multiple supervised machine learning classifiers such as RF and PLS to classify gene sequences and *Muridae* species
2. Compare and contrast how each machine learning model evaluates, trains, and classifies *Cytb*, *Irbp*, and *Rag1* gene sequences, and *Muridae* species

Methodological framework to investigate *Muridae* machine learning classification



STEP 1:
Define the hypothesis and objectives, and outline analysis

STEP 2: Extract *Muridae* dataset from NCBI nucleotide databases
(R package: Rentrez)

STEP 3. Filter unknown nucleotides, clean, quality check, and transform data (i.e. adding k-mers) (R Packages: tidyverse, Biostrings, caret)

STEP 4. Apply machine learning and visualization (ROC, heatmaps, MDA) (R Packages: base R, ggplot2, caret, Random Forests, pROC, vegan)

STEP 5: Analysis of machine learning error rates, confusion matrix, and statistical tests (PERMANOVA)

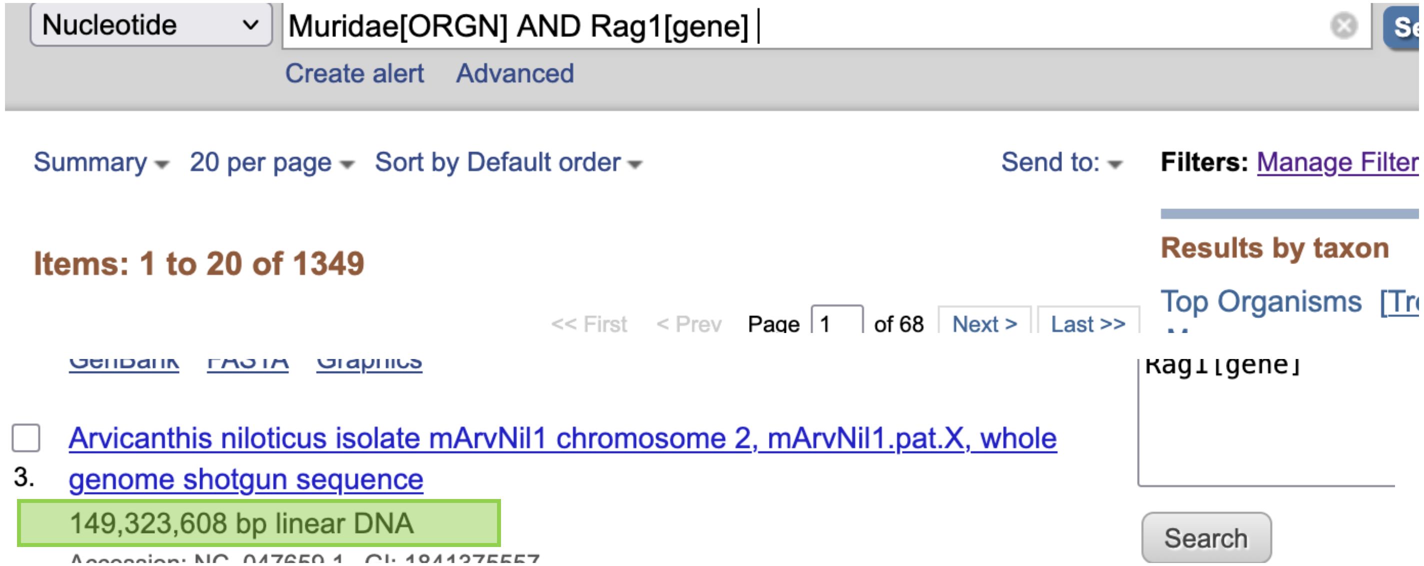
STEP 6.
Communicate findings and interpretation via storyboard

STEP 2: NCBI web search and Rentrez can be used to obtain relevant datasets

The National Center for Biotechnology Information (NCBI) is a well-established public resource that contains relevant bioinformatics databases for this analysis. Therefore, nucleotide sequences for *Cytb*, *Irbp*, and *Rag1* genes were obtained from NCBI. These genes were chosen because they have good number of sequencing records and are well-studied.

These genes were filtered for 1000 – 1500 base pairs because this was their average gene sequence size. In fact, some entries were over 100,000,000 base pairs long (shotgun sequences, see below for an example) and some sequences were low quality and are error prone such as short read sequences. Therefore, both needed to be filtered out. In addition, the Rentrez R Package was used to interact with the NCBI database to also download nucleotide sequences (Winter, 2017).

rentrez 1.2.4



Nucleotide

Muridae[ORGN] AND Rag1[gene]

Items: 1 to 20 of 1349

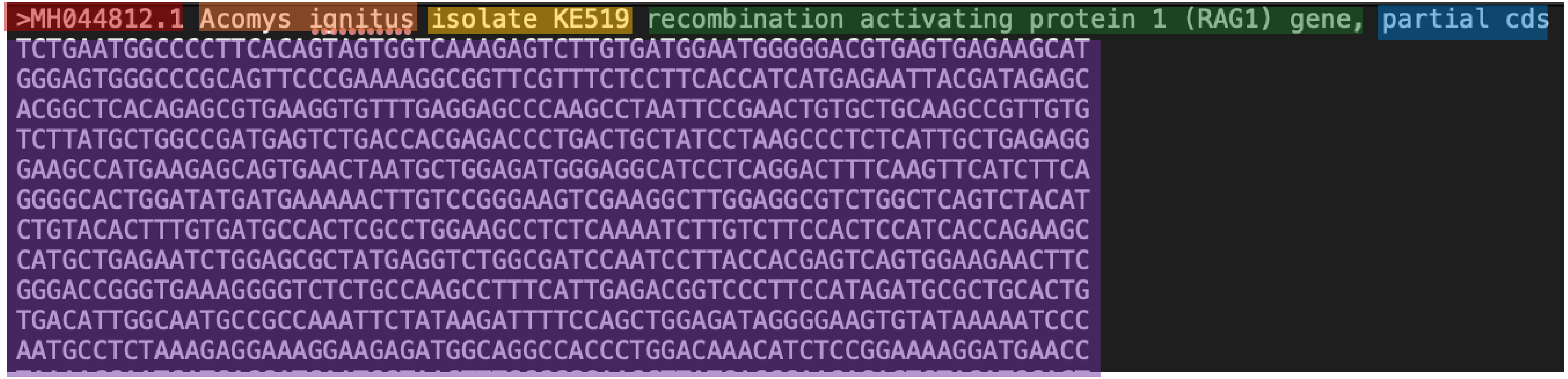
□ 3. Arvicanthis niloticus isolate mArvNil1 chromosome 2, mArvNil1.pat.X, whole genome shotgun sequence

149,323,608 bp linear DNA

Accession: NC_047659.1 GI: 1841375557

STEP 2: The dataset includes key properties and sequencing information

Properties of the data from NCBI include metadata such as the accession number, the species name (useful for grabbing species name for classifier), the isolate number, the gene/protein name and its full name, its coding DNA sequence (CDS) status and its accompanying nucleotide sequence in a FASTA file format. See below for an example:

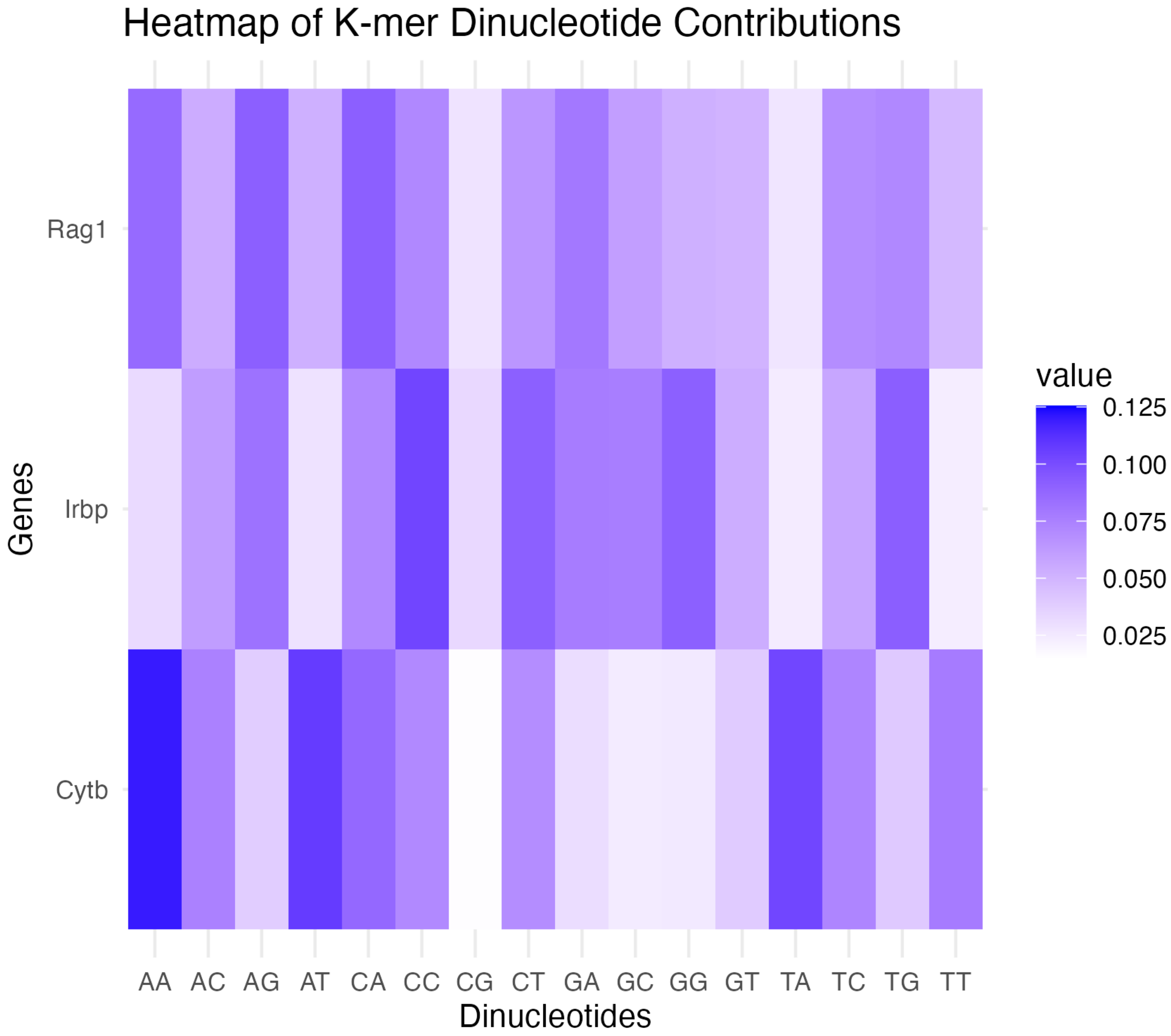


```
>MH044812.1 Acomys ignitus isolate KE519 recombination activating protein 1 (RAG1) gene, partial cds
TCTGAATGGCCCCTTCACAGTAGTGGTCAAAGAGTCTTGTGATGGAATGGGGGACGTGAGTGAGAAGCAT
GGGAGTGGGCCCGCAGTTCCCGAAAAGGCGGTTCGTTTCTCCTTCACCATCATGAGAATTACGATAGAGC
ACGGCTCACAGAGCGTGAAGGTGTTTGAGGAGCCCAAGCCTAATTCCGAACTGTGCTGCAAGCCGTTGTG
TCTTATGCTGGCCGATGAGTCTGACCACGAGACCCTGACTGCTATCCTAAGCCCTCTCATTGCTGAGAGG
GAAGCCATGAAGAGCAGTGAACTAATGCTGGAGATGGGAGGCATCCTCAGGACTTTCAAGTTCATCTTCA
GGGGCACTGGATATGATGAAAAACTTGTCCGGGAAGTCGAAGGCTTGGAGGCGTCTGGCTCAGTCTACAT
CTGTACACTTTGTGATGCCACTCGCCTGGAAGCCTCTCAAAATCTTGTCTTCCACTCCATCACCAGAAGC
CATGCTGAGAATCTGGAGCGCTATGAGGTCTGGCGATCCAATCCTTACCACGAGTCAGTGGAAGAACTTC
GGGACCGGGTGAAAGGGGTCTCTGCCAAGCCTTTCATTGAGACGGTCCCTTCCATAGATGCGCTGCACTG
TGACATTGGCAATGCCGCCAAATTCTATAAGATTTTCCAGCTGGAGATAGGGGAAGTGTATAAAAATCCC
AATGCCTCTAAAGAGGAAAGGAAGAGATGGCAGGCCACCCTGGACAAACATCTCCGGAAAAGGATGAACC
```

These were the number of hits (sample size) for each gene search result with the following search features in both NCBI and Rentrez searches:
Muridae[ORGN] AND *Gene*[gene] AND 1000:1500[SLEN]

Gene	Number of hits
Cytb	740
Irbp	1804
Rag1	854

Kmer contributions for RF classifier differs for all 3 genes



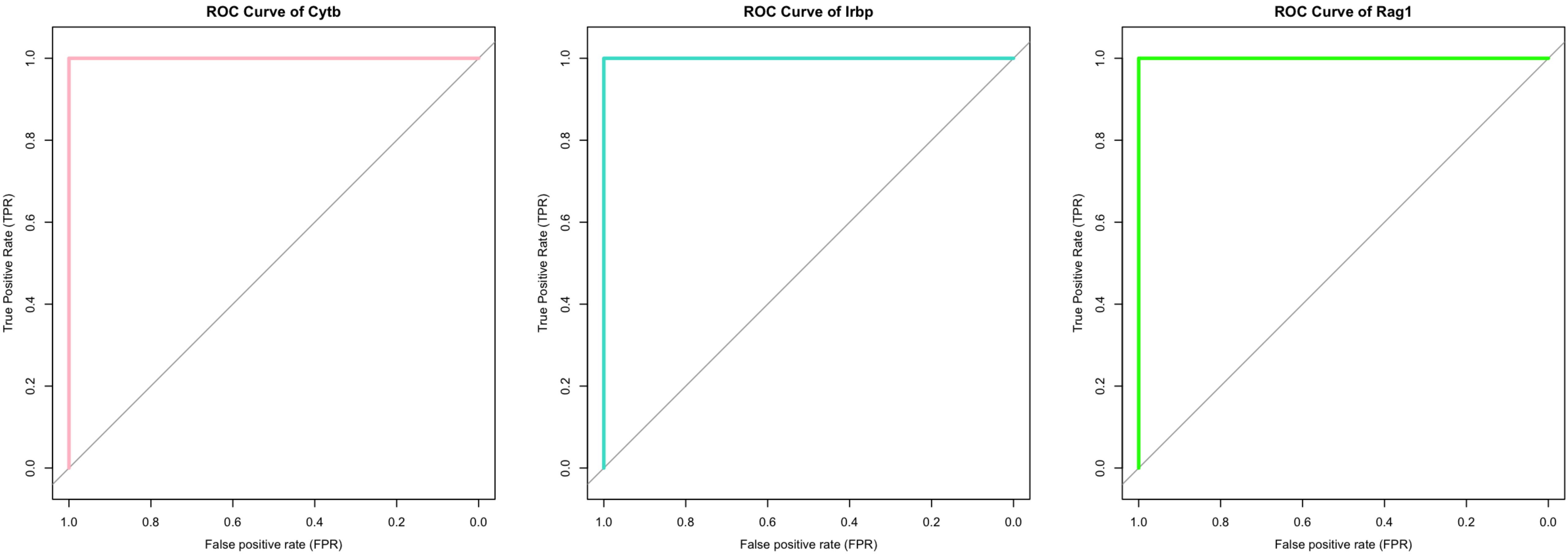
Kmer contributions for RF classifier differs for all 3 genes

From the heatmap before, the proportions of k-mer contributions during the training classification for each gene differs. Specifically, some notable observations include:

- *Rag1* has all k-mers of varying frequency values contribute to the training except for **CG** and **TA** where has low expression
- *Irbp* has a greater frequency of **CC**, **CT**, **GG** , and **TG** k-mers than *Rag1* and *Cytb*.
- *Cytb* gene has a higher proportion of **AA**, **AT**, and **TA** kmer contributions than the proportions in *Rag1* and *Irbp*.
- **CG** k-mer contributed very little to all 3 genes for the classifier. This suggests that this may be a key feature that distinguishes the *Muridae* taxon from other taxons.

This result suggest that each gene has different k-mer contributions and are different enough from each other for the RF machine learning model to classify each gene. This figure accomplished the first objective where machine learning (RF) was successfully applied for gene classifiers. The results were to be expected since each gene differs in their sequence composition.

ROC curves reveals perfect RF classifier for all 3 genes



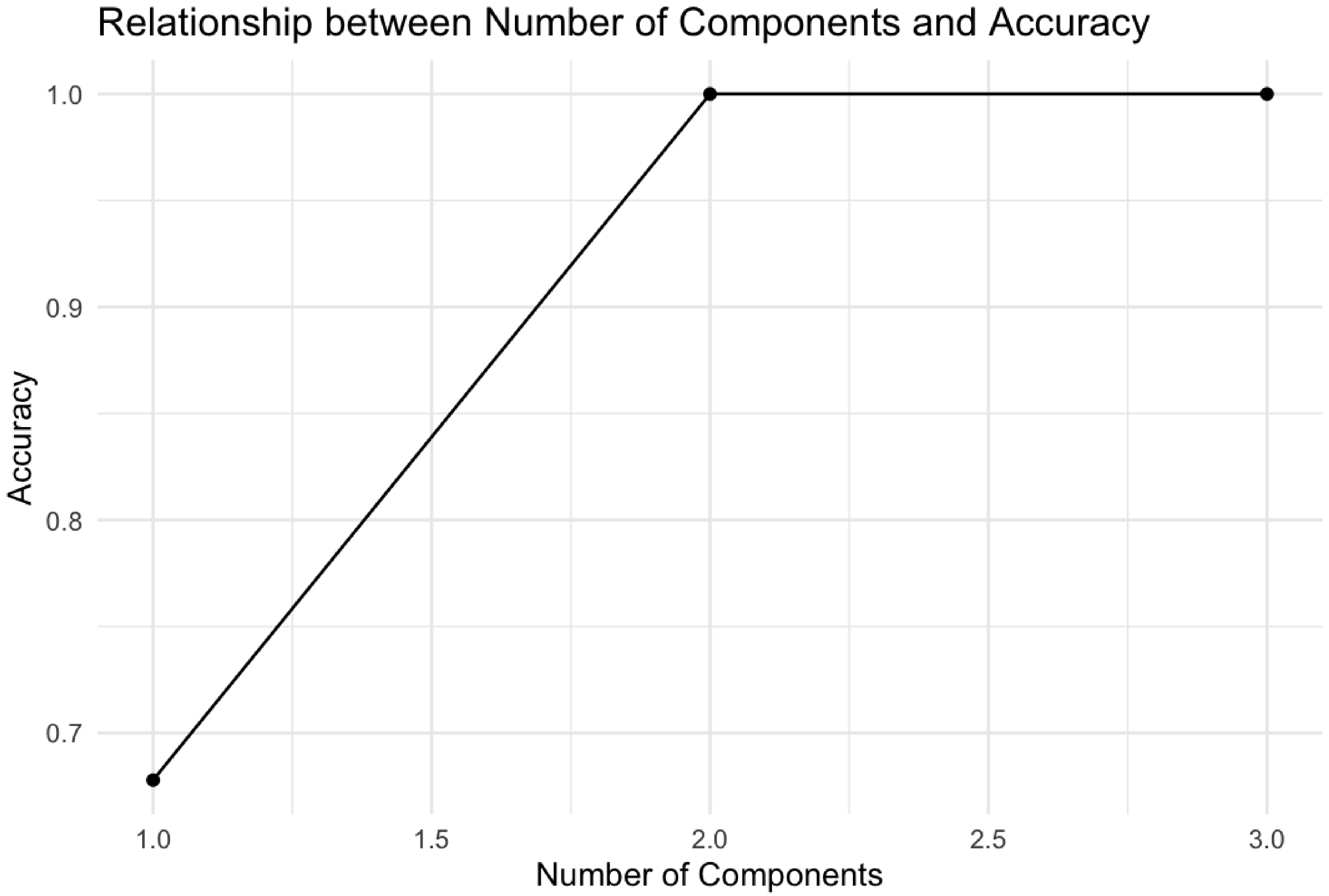
ROC curves reveals perfect RF classifier for all 3 genes

The receiving operator characteristics (ROC) curve can indicate performance on a machine learning model. The steep rise on the left side for each plot indicate that the model rarely makes false positives and indicate a perfect classifier for all 3 genes: *Cytb*, *Irbp*, and *Rag1*. Confusion matrix was also calculated on Rstudio and indicated an accuracy of 1 (not shown here).

This suggests that the classifier can classify gene classes with extremely high specificity, even on unseen data. In fact, the k-mer gene classification for *Muridae* was able to separate each gene, indicating that the RF is a strong classifier for the gene classification task.

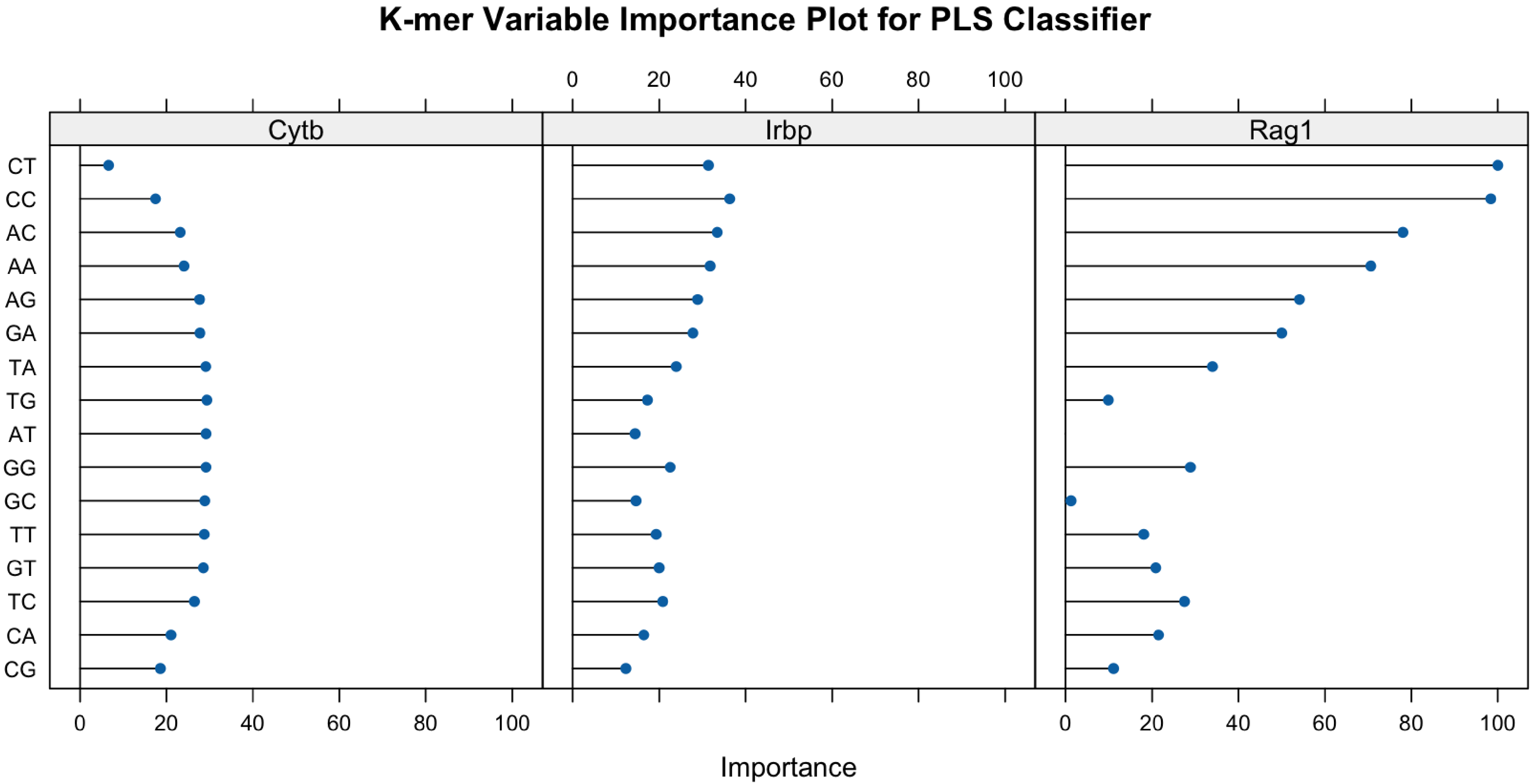
This figure supports the main hypothesis where features such as dinucleotide k-mers in machine learning algorithms can accurately train and detect different gene sequences and *Muridae* species.

Accuracy increases as number of components increases in PLS



Relationship between Number of Components and Accuracy

K-mers contributions differs for each gene of the PLS classifier during training



PLS and its classification on *Muridae* gene

From the previous figures, the accuracy of gene classifiers increased as the number of components increased, up to a maximum accuracy of 1.0 at 2 components and plateaus.

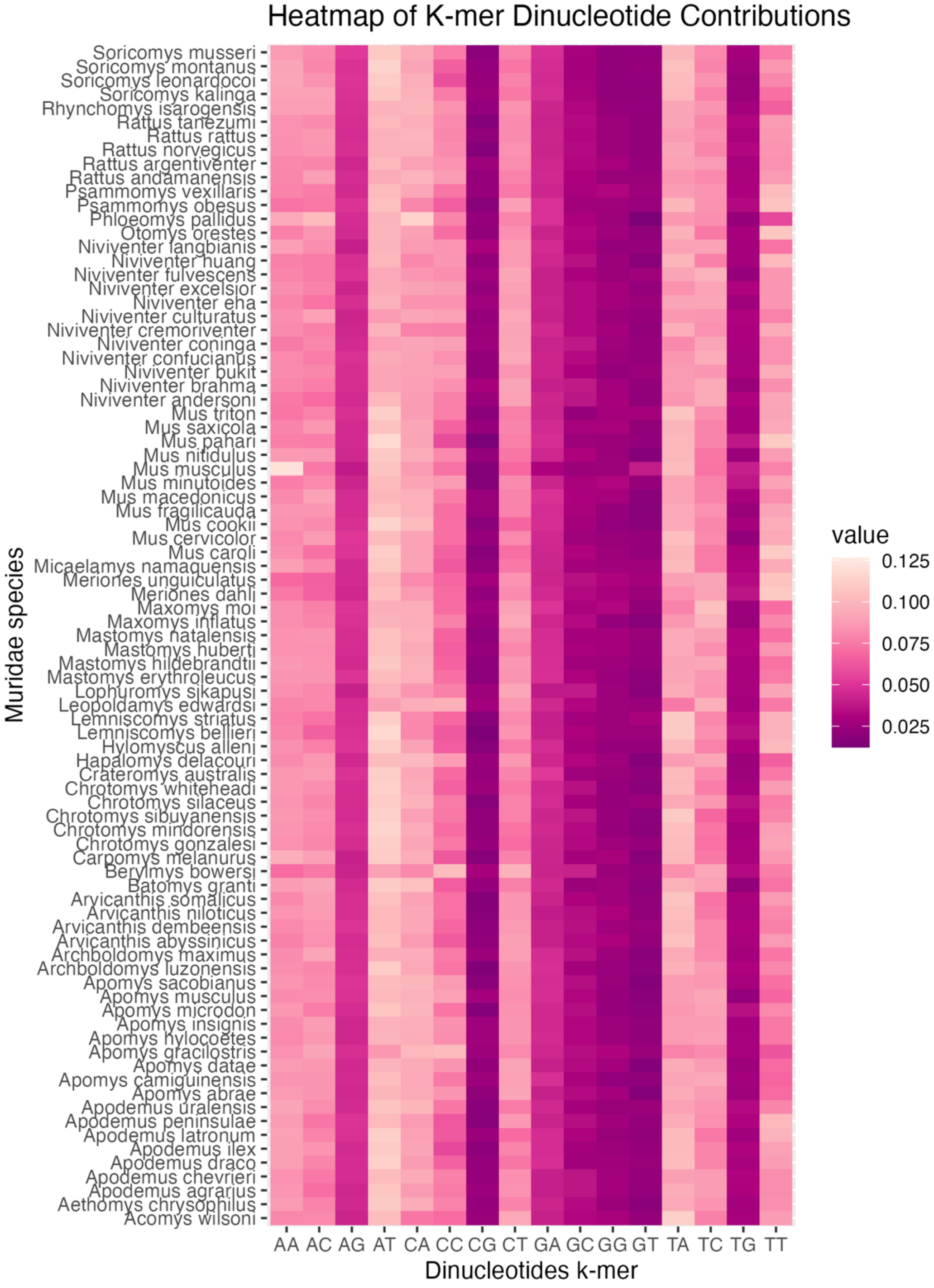
In addition, the proportions of k-mer contributions during the training classification for each gene differs. Specifically, some observations include:

- *Cytb* gene has equal importance of k-mer contribution but is low for **CT**, results are not consistent RF model
- *Irbp* has a wide variety of k-mer contributing to the training model, with the **CC** k-mer contributing the most(**CC** k-mer consistent with the RF model)
- *Rag1* shows more distinct contributions, with **CT**, **CC**, **AC**, **AA** contributing the most while **AT** and **GC** k-mer contributing the least (**GC** k-mer consistent with the RF model)

This suggests that each gene has different k-mer contributions and are different enough for the PLS algorithm learning to classify each gene. Some PLS k-mer contributions are consistent with the RF model.

This figure accomplished the first and second objectives where machine learning (RF) was successfully applied for gene classifiers and comparisons were made.

The RF model uses different kmers combinations during training to classify each *Muridae* species based on the *Cytb* gene



Dinucleotide combos

AA
AT
AC
AG
TA
TT
TC
TG

CA
CT
CC
CG
GA
GT
GC
GG

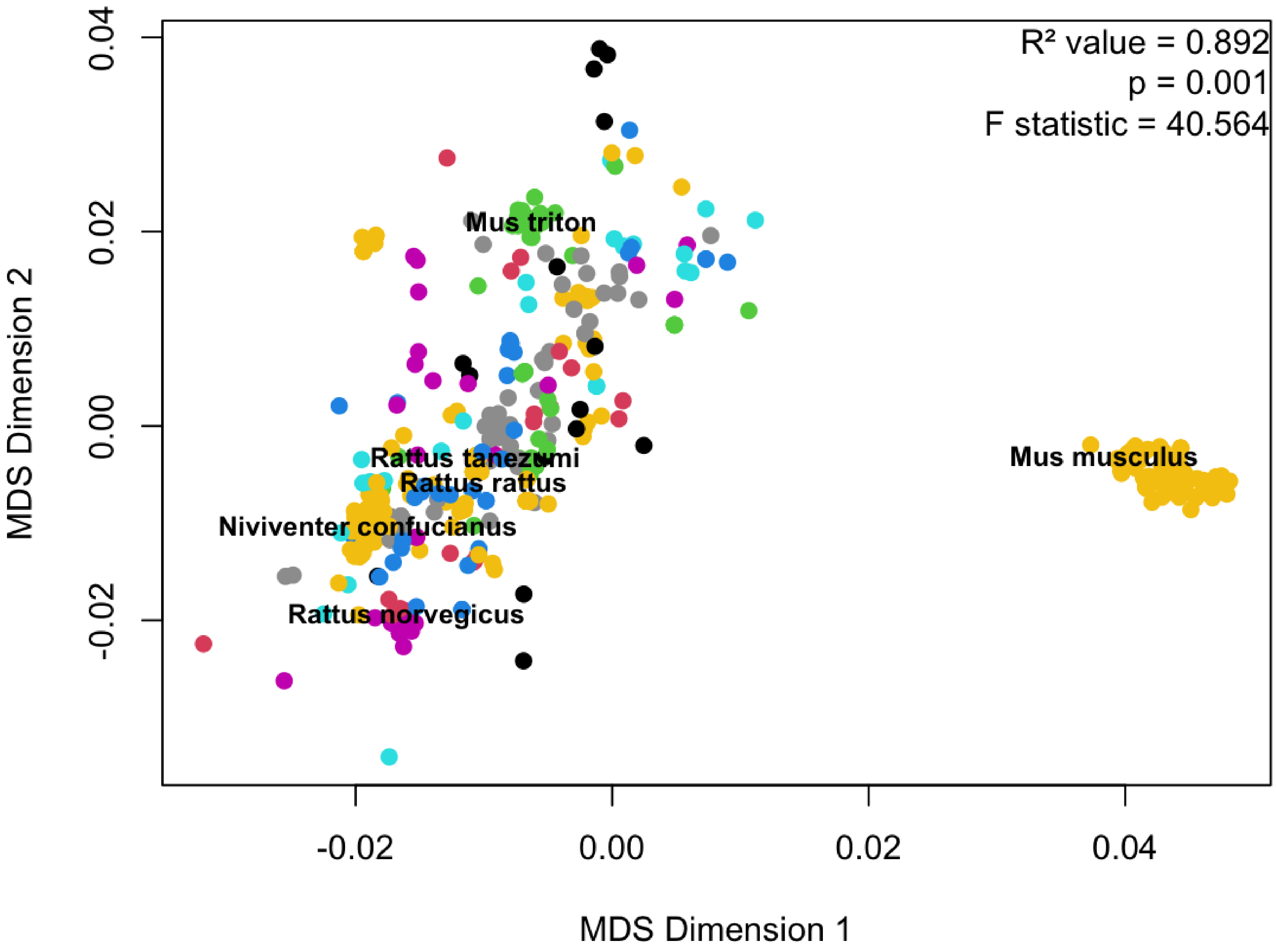
RF model uses different kmers combinations to classify each *Muridae* species based on the *Cytb* gene

From the heatmap, it shows the proportions of K-mers contributions for each species and how it differs for each species. For example, *Mus musculus* has a higher frequency of **AA** as compared to other species. This is consistent with the first heatmap, where the **AA** k-mer contributes to the most and was probably biased by the *Mus musculus* species.

AG, **CG**, **GA**, **GC**, **GG**, **GT** and **TG** k-mers has low frequency values across all species indicating a potentially conserved feature in the *Muridae Cytb* gene. To really confirm this observation, an outgroup from another taxon (i.e. *Rodentia* which is the higher classification or a sister taxon, *Cricetidae*) should be included to perform comparisons.

This figure supports the main hypothesis where k-mer dinucleotide frequencies contribute differently to each *Muridae* species.

The RF clusters *Muridae* species during training



Discussion: RF model training on *Muridae* species

From the previous figure, we can observe that the *Muridae* species cluster in species. Specifically, this can be seen with the *Mus musculus* species as it makes its own distinct cluster on the middle right (yellow cluster). The *Rattus norvegicus (*magenta cluster*)* and the *Mus triton (*green cluster*)* species also form their own distinct cluster.

The PERMANOVA test can also tell a lot of information:

- significant (p = 0.001)
- R^2 value is 0.892 and is close to 1, indicating a good fit
- F statistic value of 40.564 which is a high value and indicates that there is a greater variation between sample means relative to the variations within samples

Potential bias - An explanation to the one cluster formed for the *Mus musculus* group is that it is such as well-studied model organism as compared to other *Muridae* species and thus they have vast amounts of information available (Standley & Xie., 2024). If more sampling was done to other *Muridae* species, we can anticipate they can form their own distinct group as opposed to one giant mixed cluster.

Next steps and follow-ups in investigating *Muridae* species

Some of the *Muridae* species were misclassified in the RF model (such as the *Apodemus draco* misclassified as *Apodemus ilex*). To generate a more robust method of classification, other features and models can be used. Some include:

- GC content and different k-mer sizes in *Muridae* species (Li et al., 2020)
- Multi-gene classifier that combines the multiple genes at once instead of each gene separately (Klimov et al., 2019). For this analysis, it would be to use all 3 genes at once. This will let us see if using all 3 genes increases the accuracy of species classification.

Due to the limit of time, some next steps is to graph the validation data to see whether it lines up with the training data. This way, we can see if the training model is either overfitted or underfitted. In addition, the PLS model can be used to apply to the species classifier. In addition, we can repeat the analysis with another gene such as *Irbp* and *Rag1* for the species classifier.

In the broader context of classification, if there are unknown *Muridae* species with unknown IDs, analysis on machine learning models can identify these new sequences and identify them to be either current species or newly discovered species. In addition, population structure within species can be further investigated to see whether new species have emerged.

Bonus Reflection slide

I learned a lot from this assignment. Specifically, I learned that many different types of machine learning model functions are finicky and only like specific data formats. Most of the time, I was troubleshooting on how to get the data to the right format. Else it will throw errors to the console. I now realized that bioinformatics is 90% data wrangling and 10% visualizations.

I want to work on my understanding on more machine learning pipelines and why some are used more often than others and the context behind it. I believe that with more opportunities to practice, the better I will become at it. Some opportunities include replicating published pipelines on GitHub and in publications. This would give me an opportunity to learn more about what is standard in the field, the rationale as to why it was done that way, and the advantages and disadvantage of each pipeline. Moving forward, I learned a lot from plotting heatmaps and MDS graph since it has applications to gene expression analysis which I am interested in analyzing. This also may become relevant my future BINF6999 summer project where I am interested in looking at gene expression in the mouse brain.

Acknowledgement Slide

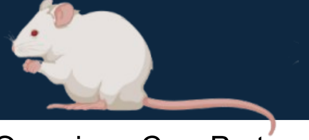
I would like to thank Farah Sadoon for helping me narrow down on some genes for my analysis and suggesting good ones. The *Shh* gene would have been interesting!

I would also like to thank the teaching assistant, Saira Asif, for helping me choose some candidate genes for my analysis.

I also want to thank Britany MacIntyre as her seminar, "Urinary metabolites as non-invasive biomarkers of cardiometabolic health: identification and validation" helped me to ultimately decide on which machine learning model algorithm to choose. I noticed that she used partial least square which inspired me to try to do the same with my assignment.

I want to thank my friends, Kazeera Aliar and Rebekah Hest, for helping me run the code and ensuring everything ran smoothly and providing me the suggestion to declutter my environment. I had a lot of data frames!

References and Resources Consulted



Standley A, Xie J, Lau AW, Grote L, Gifford AJ. Working with Miraculous Mice: Mus musculus as a Model Organism. Curr Protoc. 2024 Oct;4(10):e70021. doi: 10.1002/cpz1.70021. PMID: 39435766.

Modlinska, K., & Pisula, W. (2020). The Norway rat, from an obnoxious pest to a laboratory pet. *ELife*, *9*, e50651. https://doi.org/10.7554/eLife.50651

Standley, A., Xie, J., Lau, A. W., Grote, L., & Gifford, A. J. (2024). Working with miraculous mice: *Mus musculus* as a model organism. *Current Protocols*, 4, e70021. doi: 10.1002/cpz1.70021

The UniProt Consortium , UniProt: the Universal Protein Knowledgebase in 2025, *Nucleic Acids Research*, Volume 53, Issue D1, 6 January 2025, Pages D609–D617, https://doi.org/10.1093/nar/gkae1010. Accession numbers B1P8W3, A0A0A7NWN0, and P15919 were used to review gene and protein function

Sipper, M., Moore, J.H. Conservation machine learning: a case study of random forests. *Sci Rep* **11**, 3629 (2021). https://doi.org/10.1038/s41598-021-83247-4

Brereton, R.G. and Lloyd, G.R. (2014), Partial least squares discriminant analysis: taking the magic away. J. Chemometrics, 28: 213-225. https://doi.org/10.1002/cem.2609

Salles, M. M. A., & Domingos, F. M. C. B. (2025). Towards the next generation of species delimitation methods: an overview of machine learning applications. *Molecular Phylogenetics and Evolution*, *210*, 108368. https://doi.org/https://doi.org/10.1016/j.ympev.2025.108368

Winter DJ. 2017. rentrez: An R package for the NCBI eUtils API. PeerJ Preprints 5:e3179v2 https://doi.org/10.7287/peerj.preprints.3179v2

Li, X., Li, H., Yang, Z. *et al.* Exploring objective feature sets in constructing the evolution relationship of animal genome sequences. *BMC Genomics* **24**, 634 (2023). https://doi.org/10.1186/s12864-023-09747-x

Klimov, P.B., Skoracki, M. & Bochkov, A.V. *Cox*1 barcoding *versus* multilocus species delimitation: validation of two mite species with contrasting effective population sizes. *Parasites Vectors* **12**, 8 (2019). https://doi.org/10.1186/s13071-018-3242-5

Figures were taken from *BioRender*.com

Tutorials

https://cran.r-project.org/web/packages/caret/vignettes/caret.html
https://topepo.github.io/caret/using-your-own-model-in-train.html#illustrative-example-1-svms-with-laplacian-kernels
https://cran.r-project.org/web/packages/vip/vignettes/vip.html
https://r-graph-gallery.com/79-levelplot-with-ggplot2.html
https://www.geeksforgeeks.org/r-language/multidimensional-scaling-using-r/
https://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization
https://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/122-multidimensional-scaling-essentials-algorithms-and-r-code/