# Progress record 1: Project Outline

## 1. Overall aims of project

A recent study (Stiller *et al.* (2024)) published in Nature reconstructed the phylogenetic history of avians using various locus types (different region). This study highlighted a strong effect of locus type on constructing gene trees and species trees, suggesting that intergenic regions yield more reliable results compared to exon regions, which produce substantially different tree outcomes.

However, the original study used Modeltest-NG v0.1.3 and RAXML-NG v0.9.0 for evolution model estimation, potentially leading to a oversimplified evolutionary model for gene tree estimation. Furthermore, the original study used nucleotide sequence alignment for phylogenetic analysis in exon region, while the amino acid alignemnt would be more stable for phylogenetic analysis. These might unfairly attribute the poor performance of exon regions solely to locus type.
As a consequence, in this study, we aim to use MixtureFinder (Ren *et al.*, 2024) to re-estimate the gene trees and species trees for both intergenic and exon regions to explore whether enhanced complex evolutionary models provide more accurate phylogenetic results.

In specific:

- Whether the estimated species trees of intergenic regions would change when using mixture models
- Whether the estimated speices trees of exon regions would be more similar to the one estimated from intergenic regions when using mixture model

## 2. Dataset used

We will use the dataset from the avian evolution paper by Stiller *et al.* (2024)

Reason: This dataset is the original data of the avain evolution paper, and the procedure of this project is to re-estimate the GTs and species trees using mixture models. So of course I should use the same dataset as the original paper.

## 3. A brief description of the dataset(s)

The dataset contains a comprehensive range of data(size of alignment in this data set is 50 times larger than the most complex study before), including

- Alignments of **363 species across 218 families** (covering 92% of total bird families in **37 orders and 11 evolution clades**)
- Tree files (1435 species trees in total)
- Polytomy test results...

The alignemnt file contains alignment in different regions using different loci numbers, data have been cleaned and filtered by the author, and are descripted as table below (data from Stiller *et al.* (2024) ):

| Data | Datatype | Description | Loci | Base pairs |
|---|---|---|---|---|
| 94K | Intergenic regions | All intergenic loci | 94,402 | 94,402,000 |

| Data | Datatype | Description | Loci | Base pairs |
|---|---|---|---|---|
| 80K | Intergenic regions | Excluding overlap with exons | 80,047 | 80,047,000 |
| 63K | Intergenic regions | Exucluding overlap with exon or introns | 63,430 | 63,430,000 |
| Intron | Introns | All introns | 44,846 | 136,940,000 |
| UCE | UCES | Ultraconserved Element loci | 4985 | 25,259,810 |
| Exon | Exons | All Exon loci | 14972 | 18,975,346 |

The data required for this project is

- Alignment file
- gene tree file

for both the exon and intergenic regions, which have been cleaned and filtered by the author. detailed description of required data as below:

**Exon:**

- 14972 Nucleotide Alignements of 14972 different exon locis (each of these alignments contain at least 4 taxa) in FASTA format (`all_alns.tar.gz`) (see `data/example_exon_data` for some examples)
- 14972 Gene trees built from alignment above in newick format with support as aLRT values (`exon_c12_atleast4taxa.gene.trees.tar.gz`) (see `data/example_exon_data/treefile` for some examples)

**Intergenic region:**

- Nucleotide Alignments of 63430 different loci from intergenic region in FASTA format for which gene trees were built (`63430.alns.tar.gz`) (see `data/example_intergenic_data` for some examples)
- 63430 gene trees built from alignment above (after collapsing branches with aLRT values below 0.95). These trees were used to construct the species tree in original paper (`63430.aLRT-0.95-collapsed.gene.trees.gz`)

# 4. A small subset of the data that you can test your code on

## Subsetting loci:

I would randomly use 10 loci as the first small attempt. For further analysis, maybe I would use the first x% (maybe 20% of the loci?). I think I can't use all, since 400 loci takes 3.5h in 128 threads, and 14972 and 63430 loci would take too long to finish.

## Subset of species (if required?):

The original data contains 218 families and 363 species in 37 orders.

There are 6 "large Orders" containing more than 10 species (for example, Passeriformes has 173 species and Charadriiformes has 29 species), for those order less than 10 species, just retain all of these;

For the species in 6 large orders, I might rank the occurance number of those species in selected loci, and select the top 10% species for each of these large orders (for example around 17 for passeriformes).

# 5. The plan for commands needed (e.g. IQ-TREE commands, ASTRAL commands)

**Subsetting orders:**

I would write a small python script (haven't done yet) to show the Order subsetting process, which creates a tabel file showing the occurance number of all species in 40 loci files, we could use R to further showing the result of first 10% species in the 6 “large orders”.

**Phylogenetic analysis process:**

## 1. Generate the tree in paper

### 1.1 Estimate gene trees

If not subsampling species, we could just subset the gene trees from the orginal data based on our selected loci.

If we subsample species, we should filter the selected loci alignment files with subeset species (this would need another python script), then use these alinments to estimating gene trees using the original model in paper through a command like this:

```
iqtree2 -S bird_subloci -m papermodel --prefix subloci -T 128
```

- `bird_subloci` is the name of directory containing alignment files of subsampled loci

> [!NOTE] Question1: For this part, should we use the same model (by Modeltest-NG) in original paper?

> Question2: We might use `-st NT2AA` option to perform the estimation based on translated AA sequences. However, it don’t works so far since the exon nucleotide alignment contain gaps, and the original data also excluded the third condon position, which makes the number of sites is not multiple of 3 and unable translate.

### 1.2 Estimate species trees

Using the `subloci.tree` file produced above (file containg gene trees of our selected loci), using ASTRAL command like this:

```
astral -i subloci.treefile -o astral_species.tree 2> astral_species.log
```

The the number of loci subsetted would differ between intergeneic regions, but the process to estimate gene tree and species tree should be similar.

> [!NOTE] Question3: The original paper used ASTRAL to construce species tree which takes a treefile containing a set of gene trees to estimate. While the IQTree used the alignment file (and also the

> partition file) to estimate species tree. In this study maybe I should use ASTRAL all the time?

## 2. Re-estimating new trees using mixturefinder:

### 2.1 Gene trees:

First we use MixtureFinder to find the best model

The -S option to specify a file directory with alignment files doesn't work in MixtureFinder, so we need to **concatenate** the alignments among different loci into a single alignment file first:

```
iqtree2 -p bird_subloci --out -aln bird_conloci --out-format NEXUS
```

- `bird_subloci` is the name of directory containing alignment files of subsampled loci

Then Perform the MixtureFinder on concatenated alignments:

```
iqtree -s bird_conloci -m MIX+MF -T AUTO
```

Then use the new mixture model to calculate the gene trees (here assume mixture model is `MIX{TIM2+FO,TPM3+FO,HKY+FO,TPM3+FO}+I+G`)

```
iqtree2 -S bird_subloci -m “MIX{TIM2+FO,TPM3+FO,HKY+FO,TPM3+FO}+I+G” --prefix
mixsubloci -T 128
```

> [!NOTE] Question 4: It seems like the `-m` option would use the same mixture model for all alignments in the `-S bird_subloci` directory, maybe I should write a scipt to use MixtureFinder for each loci to estimate different models for each loci alignment?

### 2.2 Species trees:

similar as above:

```
astral -i mixsubloci.treefile -o astral_mixspecies.tree 2> astral_mixspecies.log
```

## 3 Calculating several metrics

### 3.1 ROBINSON-Foulds distance

After calculating both(paper and MixtureFinder) species trees for both intergeneic and exon regions, we could calculate RF distances:

```
iqtree2 -rf astral_species astral_mixspecies
```

(we would also calculate the RF distance between Intergenic species trees from original paper and from using MixtureFinder)

**3.2 Concordance factor**

gCF:

```
iqtree2 -te astral_mixspecies.treefile --gcf mixsubloci.treefile --prefix concord
```

qCF:

```
astral -q astral_mixspecies.treefile -i mixsubloci.treefile -t 2 -o
astral_mixspecies_annotated.tree 2> astral_mixspecies_annotated.log
```

And also sCF...

```
# First approch
iqtree2 -te astral_mixspecies.treefile -s ALN_FILE --scfl 100 --prefix concord2

# Second approch to calculate gCF and sCF at the same time
iqtree2 -t astral_mixspecies.treefile --gcf mixsubloci.treefile -s ALN_FILE --scf
100
```

## 6. What I will measure and why

1: The bootstrap/UFBoot value to measure the statistical supports that our topology is supported by the data.

2: Compare the concordance factor between gene trees from the original paper and those new trees using MixtureFinder.

Concordance factor is calculated as the proportion of gene trees/ or quartets/or deceive sites that have same topologies as the species tree. It is a measurement of the similarity between the gene tree and species tree.

**If the concordance factor of new gene trees has greater value than that of the original paper. It indicates that the new model would improve the quality of gene tree estimation.**

3: Compare the Robinson-Foulds distance between the species trees of the exon and intergenic region.

If the RF distance between the 2 species trees of different region (using MixtureFinder) is less than that RF distance between the 2 species trees of intergenic and exon region in original paper, it would indicates that our mixture model helps to construct better phylogenic result for the exon region. Furthermore, Robinson-Fould distance could also used to check whether the intergenic species tree changed.

> [!NOTE] Question 5: If I subsampled the loci (or species), that means the RF distance in original paper and the Rf distance in this project would based on different data amount, is this acceptable?

# # Progress record 2: Bird-tree Progress update

## 1. Full datasets

Based on

- gene coverage (the number of genes that contain that species) (using `select_gene_species.py` in the script folder),
- selecting 2 species from either side of the deepest split for each clade,
- the remaining selected species are closest to the crown node of each clade

**24 species** ( 7 for Telluraves, 5 for Elementaves, 5 for Columbaves, 3 for Mirandornithes, 4 for Galloanseres ) were selected.

After that, **4255 exon loci** containing all of these 24 species were filtered using the `filterespecies_new.py` script in the script folder. Each of the 4255 filtered alignment files now contains only the 24 species instead of 363, and these files are placed in the `4255species_exondata_filtered` folder.

The analysis below was based on these 4255 exon loci. For the **full dataset**, I would change the filter logic to

- **include any loci that contain 4 or more of these 24 selected species**

, which would result in **14194 (in 14972)** exon loci.

## 2. Analysis Process (Exon)

Up to now, I have analysised exon loci.

### Model finder

For the ModelFinder, the following command was used (on the server) to calculate gene trees.

```
iqtree2 -S 4255exon -m MFP -pre
/data/changsen/4255exon_modelfinder_result/4255mf_combined -T 75
```

This command produces a `4255mf_combined.treefile`, which contains 4255 gene trees.

Then, I used ASTRAL (on my laptop) to generate the species tree:

```
astral -i 4255mf_combined.treefile -o astral_4255species_mf.tree 2>
astral_4255species_mf.log
```

It produces a species tree file called `astral_4255species_mf.treefile`.

After that, I calculated the **quartet concordance factor (qCF) and the posterior probability (pp)** using

```
astral -i 4255mf_combined.treefile -q astral_4255species_mf.tree -t 2 -o
astral_4255species_mf_qcf.tree  2> astral_4255species_mf_qcf.log
```

It produces a `astral_4255species_mf_qcf.tree` file, in which the qCF (q1,q2,q3) and the posterior probability (pp1) are listed in the annotation (`[]`).

After filtering out other information using a Python script, we finally produce a `4255_mf_qcf_pp_annotated.tree` file.

### MixtureFinder

First, I ran the `write_multiple_linuxcmd.R` script to write 4255 command into a text file (`iqtree_commands.txt` in the script folder).

A single line of that command is:

```
/data/changsen/bin/iqtree2 -s
/data/changsen/4255exon/GALGAL_R00006.fa.nt_ali.filtered.fasta -m MIX+MF -T 1 -
mset GTR -pre /data/changsen/4255exon_result/R00006
```

Then, I used the bash script `run_cmd_parallel_new.sh` to run the 4255 command parallelly one the server (maximum CPU was set to 75)

After that, MixTurefinder would produce 4255 gene trees, and R script `combine_genetree.R` was used to combine them into a single tree file.

Then I used ASTRAL (on my laptop) to generate the species tree

```
astral -i 4255combined.treefile -o astral_4255species_mix.tree 2>
astral_4255species_mix.log
```

It produces a species tree file called `astral_4255species_mix.treefile`

Then for the qCF and posterior probability (pp)

```
astral -i 4255combined.treefile -q astral_4255species_mix.tree -t 2 -o
astral_4255species_mix_qcf.tree  2> astral_4255species_mix_qcf.log
```

It produces a `astral_4255species_mix_qcf.tree` treefile, further cleaned up to `4255_qcfpp_annotated.tree`

### ROBINSON-Foulds distance

The Robinson-Foulds distance between the species trees calculated by MixtureFinder and ModelFinder was calculated by:

```
iqtree2 -rf astral_4255species_mix.tree astral_4255species_mf.tree
```

$$ \text{Standardized RF Distance} = \frac{\text{Original RF Distance}}{\text{Maximum RF Distance}} $$

Where:

- **Original RF Distance** is the value obtained from the `iqtree2 -rf` command.
- **Maximum RF Distance** is calculated as:

$$ \text{Maximum RF Distance} = 2 \times (n - 3) $$

Here, ( n ) is the number of species in the tree.

The analysis of the Intergenic region will be performed after finishing exon...

# 3 Exon results and Outline for Result Assessment

The assessment of the results, as well as the example results of the 4255 exon, were shown below:

## 3.1: The species tree with qCF and pp annotated:

The species tree of 4255 loci using ModelFinder, annoated with qCF and pp value：Modelfinder Image

The species tree of 4255 loci using ModelFinder, annoated with qCF and pp value：Mixturefinder Image

> [!NOTE] Problem1: Even when using 4255 exon loci (approximately 30% of overall exon loci), both ModelFinder and MixtureFinder could not correctly recover the main clades.
> Problem2: **The qCF of MixtureFinder is sometimes lower than the qCF in ModelFinder, which is unexpected.** There may be some issues in the species tree generating or the qCF generating process, but I haven't figured out why.

## 3.2: Table showing the qCF and pp for each Clade in each method

Since the Elementaves and Columbaves could not be recovered in both plots above, here I **only show the 3 clades that were recovered or basically recovered** (the 3 red circles in the tree above):

| Clade | qCF_MF | qCF_Mix | pp_MF | pp_Mix |
|---|---|---|---|---|
| Telluraves | 0.41 | 0.399 | 1.0 | 1.0 |
| Mirandornithes | 0.62 | 0.604 | 1.0 | 1.0 |
| Galloanseres | 0.753 | 0.742 | 1.0 | 1.0 |

From the table above, we see that in **nodes spliting 3 main clades, qCF of MF are always larger than qCF of Mixture Finder**

## 3.3: Boxplot & density plot to show the distribution of qCF between different method

boxplot

From the plot above, we can see that in the large qCF value range (qCF > 0.9), the qCFs of ModelFinder are usually larger than those of MixtureFinder. However, in most of the lower qCF value range, the qCFs of MixtureFinder are usually larger than those of ModelFinder

### 3.4: ROBINSON-Foulds distance (species vs species, gene trees vs species)

The ROBINSON-Foulds distance between the 2 species trees of MixtureFinder and ModelFinder is calculated based on the formula above, and the result is 8/42 = **0.1905**.

After finishing the intergenic region analysis, I would also calculate **the RF distance between the exon and intergenic species trees in 2 different method (MF and MixureFinder)**, the RF distance using MixtureFinder is expected be less than that calculated using ModelFinder.

Furthermore, I also calculated the RF distance between the **gene trees and the species tree** produced by those gene trees:

```
iqtree2 -rf astral_4255species_mix.tree 4255combined.treefile
```

```
iqtree2 -rf astral_4255species_mf.tree 4255mf_combined.treefile
```

The result could be visualized for different method (MF and MixureFinder). The RF distance using MixtureFinder is expected be less than that calculated using ModelFinder.

> [!NOTE] Problem3: However, my results so far seems strange: all gene trees have the same original RF distance (92) to the species tree in both MixtureFinder and ModelFinder...

## 4 Future Plan

- Solve quesition 1 to 3 above
- Perform the analysis for Intergenic region
- Try to also calculate sCF and gCF, and generate the concordance table and clade plot similar to that one in the concordance paper

# # Progress record 3: Third Progress update

## 1. Exon

For exon region, several things are done:

- Changed the MixtureFinder command to `-m MIX+MFP` and reran the analysis on 4255 loci.
- After modifying the species filter logic, applied ModelFinder to 14194 exon loci, and calculated qCF and pp values.
- Fixed the species tree in original paper in ASTRAL to get the correct clade grouping (2 kind of paper tree used), then calculate the qCF and pp result for both ModelFinder and Mixture Finder
- Calculate the BIC difference (BIC_MIX - BIC_MF) for 4255 exon loci and visulized it
- Calculated qCF/pp results for the filtered 36794 intergenic region loci using MixtureFinder.

### 1.1 ModelFinder in 14194 loci:

| Clade | qCF_MF_4255 | qCF_MF_14194 |
|---|---|---|
| Telluraves | 0.41 | 0.4 |
| Mirandornithes | 0.62 | 0.62 |
| Galloanseres | 0.753 | 0.85 |

In the tree and table above, even with 14194 (out of 14972) exon loci, the wrong clade grouping persists. Moreover, the qCF in 2 of 3 recognizable clades does not improve compared to the 4255 loci result.

Therefore, I may continue using the original filter logic (keeping 4255 for exon and 36794 for intergenic) since there is **no great improvement with a larger number of loci.**

### 1.2 Rerun the Mixture Finder and fixed the paper tree to calculate qCF:

As descripted above, `-m MIX+MFP` were used to rerun the analysis, then, 2 kind of orginal paper species trees:

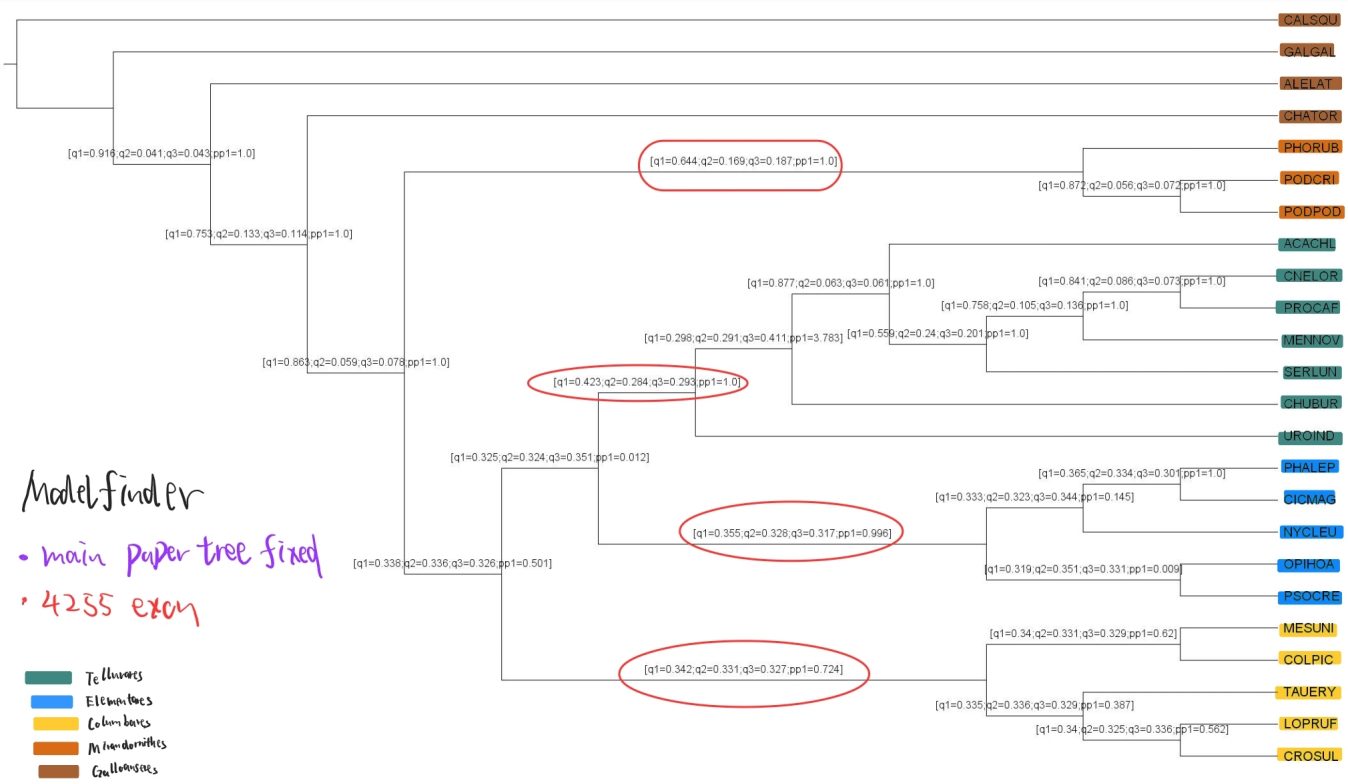
- The main species tree resulting from 63K intergenic regions analyzed with ASTRAL
- The species tree resulting from 63K intergenic regions analyzed with RAxML-NG **concatenation**

Both trees were trimmed to contain only the 24 selected species using `drop.tip`, and the species tree was fixed to ASTRAL to calculate qCF and pp values.

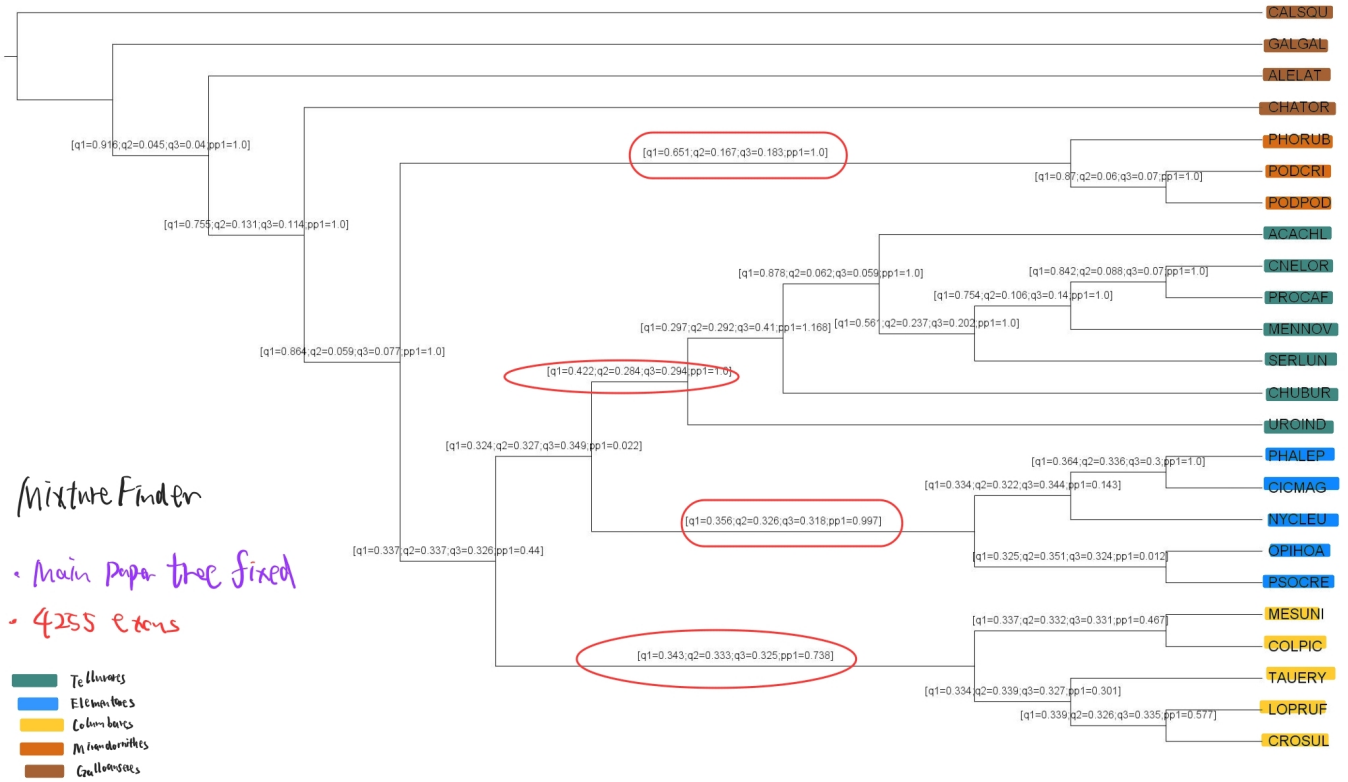
However, the main species tree appears to be loaded incorrectly in R (I tried several methods, including `read.tree()`, but the main species tree could not be loaded properly, with NaN branch lengths).

#### 1.2.1 Main species tree fixed:

**ModelFinder:**




**MixtureFinder:**



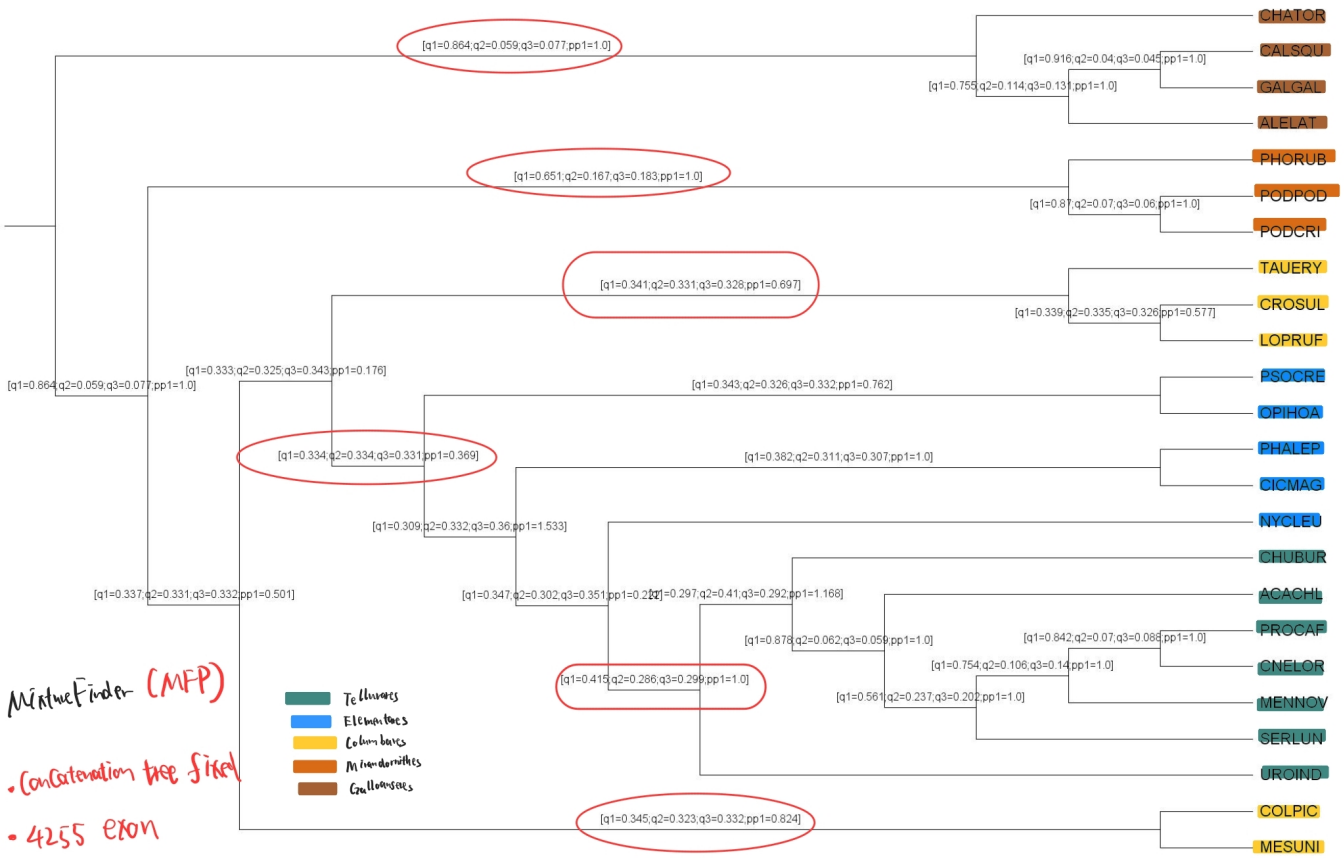

| Clade | qCF_MF | qCF_Mix | pp_MF | pp_Mix |
|---|---|---|---|---|
| Telluraves | 0.423 | 0.422 | 1.0 | 1.0 |
| Elementaves | 0.355 | 0.356 | 0.996 | 0.997 |
| Columbaves | 0.342 | 0.343 | 0.724 | 0.738 |
| Mirandornithes | 0.644 | 0.651 | 1.0 | 1.0 |

From the result fixing main tree, **the Mixture finder tends to have greater qCF and pp value in 3 of 4 clades** (only 4 in total, since the qCF of nodes defining Galloanseres is not shown in figure above)

### 1.2.2 Concatenation species tree fixed:

**ModelFinder:**

**MixtureFinder:**




| Clade | qCF_MF | qCF_Mix | pp_MF | pp_Mix |
|---|---|---|---|---|
| Telluraves | 0.417 | 0.415 | 1.0 | 1.0 |
| Elementaves | 0.335 | 0.334 | 0.395 | 0.369 |
| Columbaves | 0.343 & 0.347 | 0.341 & 0.345 | 0.766 & 0.904 | 0.697 & 0.824 |
| Mirandornithes | 0.644 | 0.651 | 1.0 | 1.0 |
| Galloanseres | 0.863 | 0.864 | 1.0 | 1.0 |

The concatenation tree could be loaded **without the problem of NaN branch length**. However, when fixed the concatenation tree, the Columbaves tends to diverge into 2 groups (I think it is because the clade grouping in ther original paper used the main tree rather than the concatenation tree)

## 1.3 BIC difference plot on 4255 Exon loci:

The below figure shows the MixtureFinder-ModelFinder BIC difference along with the 4255 loci, and the density distribution of the MixtureFinder-ModelFinder BIC difference:


Modelfinder Image

From the plot above, **only in few (24 in 4255) loci, MixtureFinder tends to have worse BIC value than ModelFinder**, and in actually 17 of these 24 loci, the amount of the BIC difference is less than 5. The loci that MixtureFinder have obvious worse results (difference greater than 5) are shown below

| Exon loci | Mixture BIC | MF BIC | BIC Difference |
|---|---|---|---|

| Exon loci | Mixture BIC | MF BIC | BIC Difference |
|---|---|---|---|
| R02147 | 11652.047 | 11322.546 | 329.5012 |
| R09991 | 7059.250 | 6989.755 | 69.4947 |
| R15264 | 3752.677 | 3684.638 | 68.0389 |
| R09188 | 2900.066 | 2837.831 | 62.2349 |
| R14193 | 6252.653 | 6193.582 | 59.0710 |
| R16467 | 5084.128 | 5032.581 | 51.5466 |
| R15019 | 7393.064 | 7371.869 | 21.1952 |

## 2. Intergenic region

Using the same filter logic of the 24 selected species, **I got 36794 (63430 in total) intergenic loci**

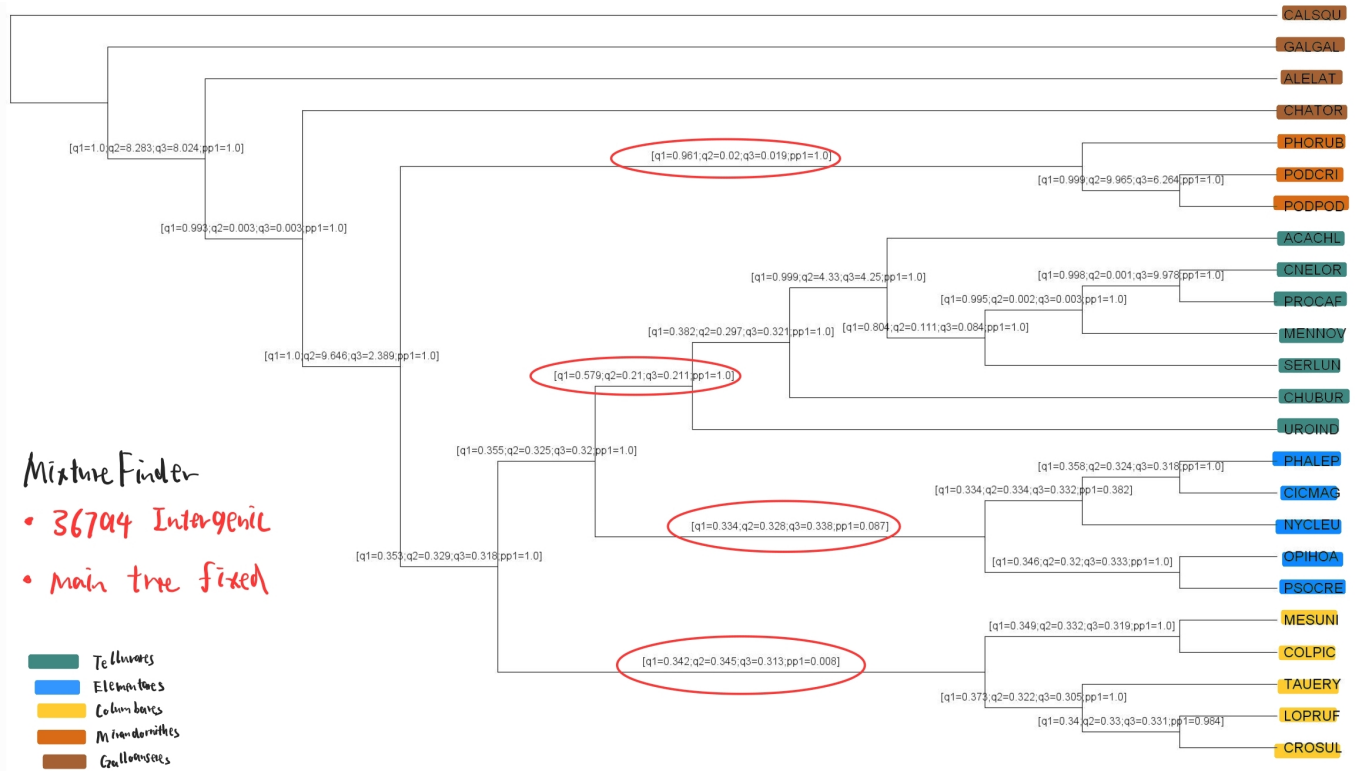
Then I run the same process as in exon: ModelFinder vs MixtureFinder

> [!NOTE] Problem1: However, ModelFinder failed to run on the server due to a resource error. I checked top, and both thread and memory resources were sufficient. I also reduced the number of threads used, but the following error still persists:

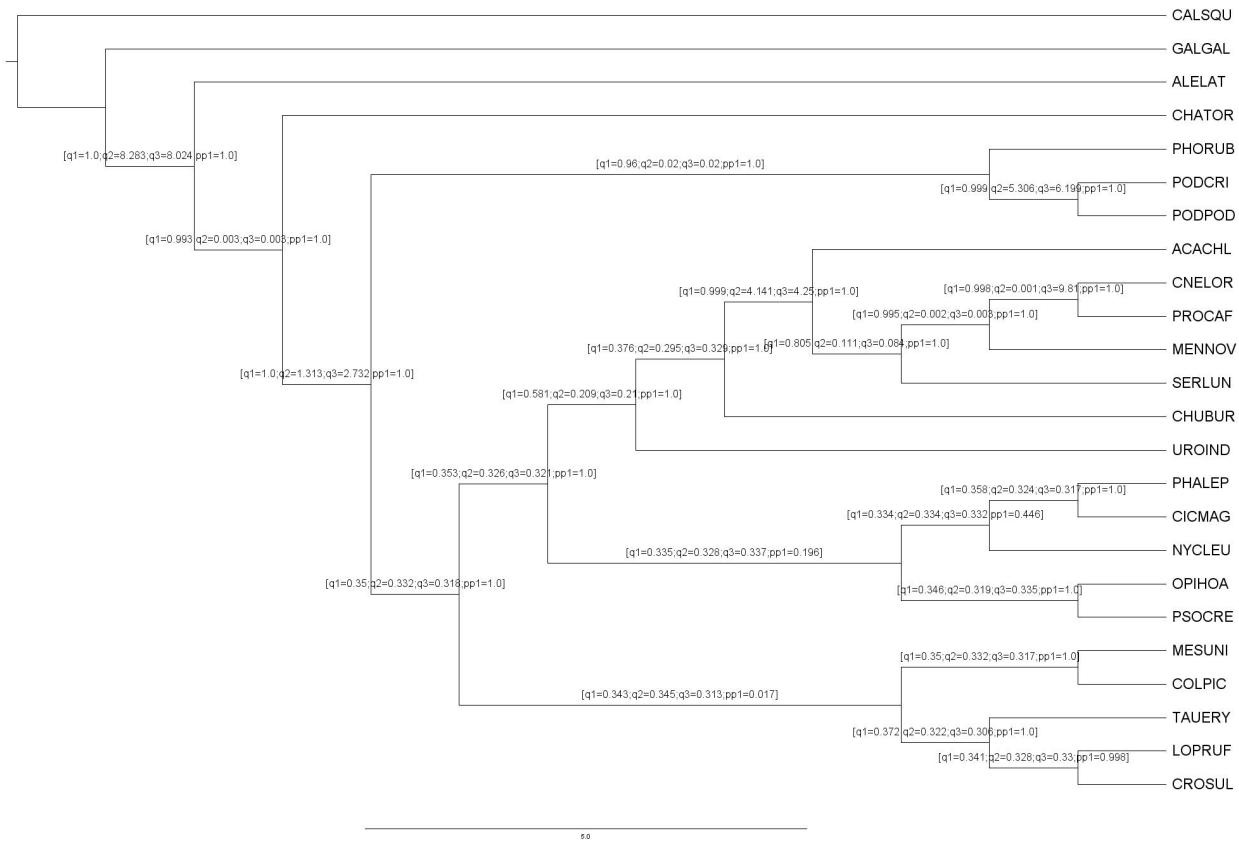
Error in ModelFinder:

```
libgomp: Thread creation failed: Resource temporarily unavailable
```

MixtureFinder Result of 36794 Intergenic region:



ModelFinder Result of 36794 Intergenic region:




| Clade | Intergenic_qCF_Mix | Intergenic_pp_Mix | Intergenic_qCF_Mf | Intergenic_pp_Mf |
|---|---|---|---|---|
| Telluraves | 0.579 | 1.0 | 0.581 | 1.0 |
| Elementaves | 0.334 | 0.087 | 0.335 | 0.196 |
| Columbaves | 0.342 | 0.008 | 0.343 | 0.017 |
| Mirandornithes | 0.961 | 1.0 | 0.96 | 1.0 |
| Galloanseres | 1.0 | 1.0 | 1.0 | 1.0 |

## 3. Try full species analysis

I have randomly selected 100 exon loci with full species and run with both MixtureFinder and ModelFinder

In ModelFinder, it takes **1h : 31m to run with 75 thread**.

However, in MixtureFinder, using **128 threads**, the analysis of just 100 loci took about **35 hours to finish**, so it is **extremely slow...**

## 4. Several things haven't done:

1: Using `treespace` to visulize the **4255 gene trees around the estimated species tree.** Since my tree results are unrooted, I used the "RF" method rather than the default

> [!NOTE] Problem2： I tried the following code to visulise, but too many tree points makes the plot noisy, and when I try to edit the plot, R studio always crashed...(Maybe due to too many trees)

```
species_tree <- read.tree("astral_4255species_mix_mfp.tree")
gene_trees <- read.tree("4255combined_newmix.treefile")
all_exon_trees <- c(species_tree,gene_trees)
treespace <- treespace(all_exon_trees, method = "RF")
plotGroves(treespace$pco, lab.show=TRUE, lab.cex=1.5)
```

2: The intergenic analysis using ModelFinder (as stated above)

3: Summarize the results of Model

## 5. Qcf & pp Result Table

| Clade | qCF_MF | qCF_Mix | pp_MF | pp_Mix |
|---|---|---|---|---|
| Telluraves | 0.423 | 0.422 | 1.0 | 1.0 |
| Elementaves | 0.355 | 0.356 | 0.996 | 0.997 |
| Columbaves | 0.342 | 0.343 | 0.724 | 0.738 |
| Mirandornithes | 0.644 | 0.651 | 1.0 | 1.0 |

| Locus Region | Method | Clade | qCF(q1) | q2 | q3 | pp1 | pp2 | pp3 |
|---|---|---|---|---|---|---|---|---|
| **Exon** | **ModelFinder** | Telluraves | 0.4231 | 0.2838 | 0.2930 | 1.0 | 0 | 0 |
| | | Elementaves | 0.3553 | 0.3276 | 0.3170 | 0.9955 | 0.0029 | 0.0015 |
| | | Columbaves | 0.3419 | 0.3311 | 0.3270 | 0.7244 | 0.1623 | 0.1133 |
| | | Mirandornithes | 0.6441 | 0.1690 | 0.1869 | 1.0 | 0 | 0 |
| | | Galloanseres | 0.8627 | 0.0592 | 0.0781 | 1.0 | 0 | 0 |
| | **MixtureFinder** | Telluraves | 0.422 | 0.284 | 0.294 | 1.0 | 0.0 | 0.0 |
| | | Elementaves | 0.356 | 0.326 | 0.318 | 0.997 | 0.002 | 0.001 |
| | | Columbaves | 0.343 | 0.333 | 0.325 | 0.738 | 0.176 | 0.086 |
| | | Mirandornithes | 0.651 | 0.167 | 0.183 | 1.0 | 0.0 | 0.0 |
| | | Galloanseres | 0.864 | 0.059 | 0.077 | 1.0 | 0.0 | 0.0 |
| **Intergenic** | **ModelFinder** | Telluraves | 0.581 | 0.209 | 0.21 | 1.0 | 0.0 | 0.0 |
| | | Elementaves | 0.3348 | 0.328 | 0.3372 | 0.196 | 0.036 | 0.768 |
| | | Columbaves | 0.343 | 0.345 | 0.313 | 0.017 | 0.983 | 0.0 |
| | | Mirandornithes | 0.96 | 0.02 | 0.02 | 1.0 | 0.0 | 0.0 |
| | | Galloanseres | 0.999 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 |
| | **MixtureFinder** | Telluraves | 0.5788 | 0.2103 | 0.2109 | 1.0 | 0.0 | 0.0 |

| Locus Region | Method | Clade | qCF(q1) | q2 | q3 | pp1 | pp2 | pp3 |
|---|---|---|---|---|---|---|---|---|
| | | Elementaves | 0.3341 | 0.3278 | 0.3382 | 0.0868 | 0.0205 | 0.8927 |
| | | Columbaves | 0.3422 | 0.345 | 0.3128 | 0.008 | 0.992 | 0.0 |
| | | Mirandornithes | 0.9609 | 0.0197 | 0.0194 | 1.0 | 0.0 | 0.0 |
| | | Galloanseres | 1 | 0 | 0 | 1.0 | 0.0 | 0.0 |

# Final Progress update

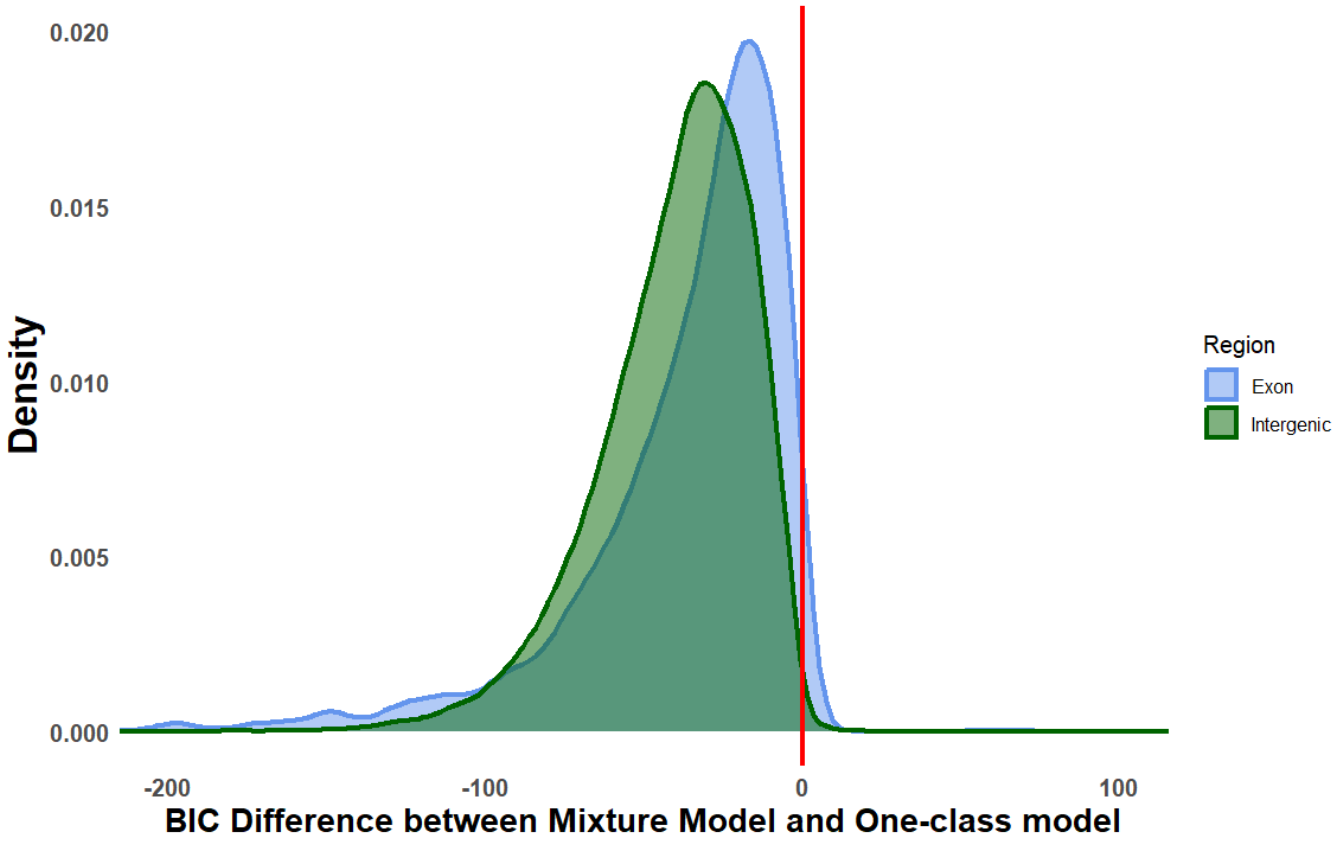
- **This progress file contains the results from the latter semester, which were used for the research report**

## 1. Comparison between the one-class model and mixture model

BIC (Bayesian Information Criterion) (Schwarz, 1978) is a criterion for model selection that measures the trade-off between model fit and complexity. In model selection, a lower BIC value indicates a better fit. The BIC values of the substitution models were extracted from the output files generated when estimating gene trees using IQ-TREE (version 2.3.5.1).

Specifically, the BIC value for the best model in each locus from both regions and both model types (single model from ModelFinder and mixture model from MixtureFinder) were extracted. We then calculate the difference between the BIC values of the mixture model and the Single Model in each locus (BIC mixture model - BIC single model) and plot the density distribution of the BIC differences.

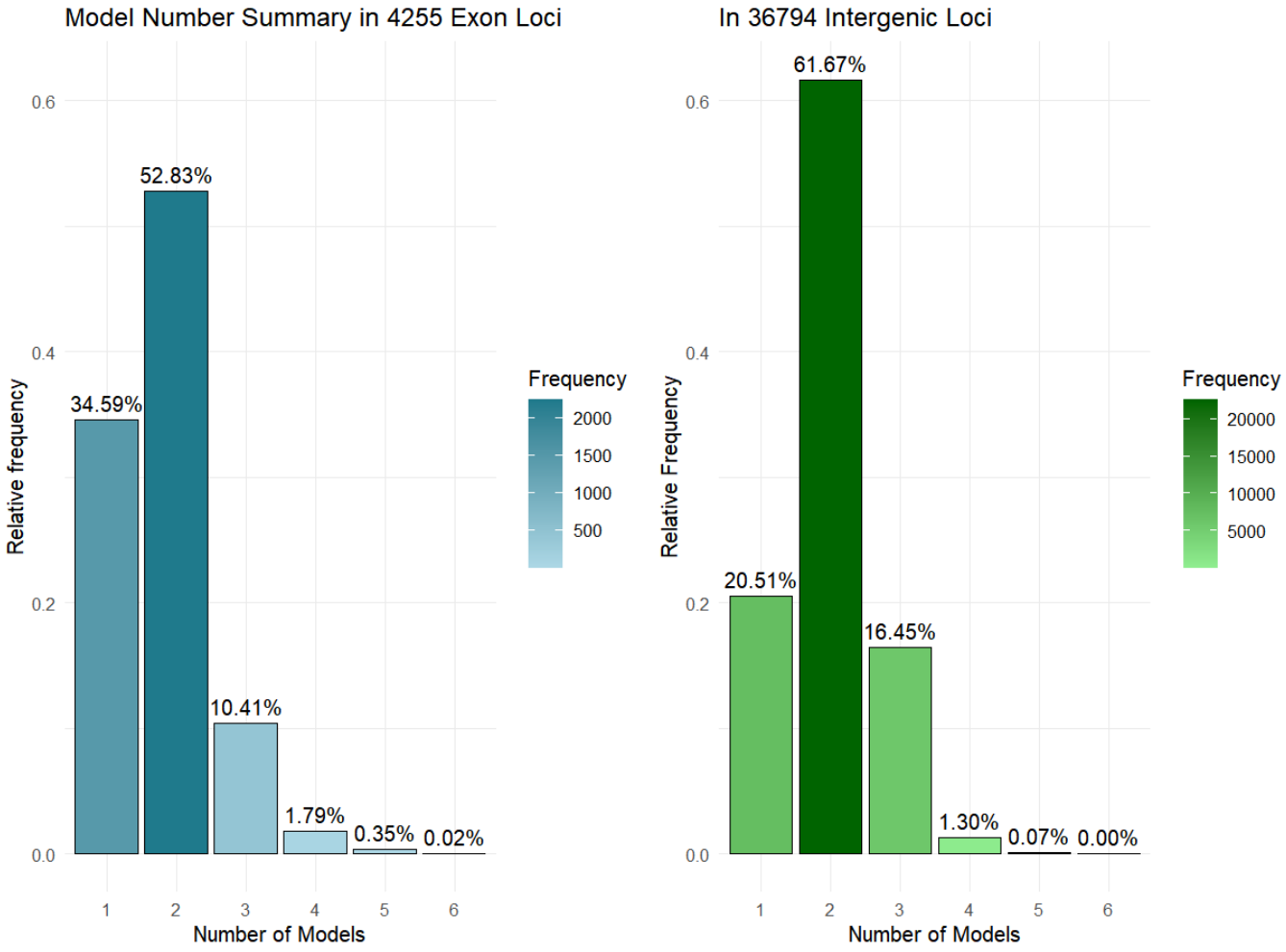
### Density distribution of the difference between BIC for the mixture model and the one-class model in the exon and intergenic regions




The classes number is the number of classes in the best-fit model for each locus. In a single model from ModelFinder, the model number would always be one, while in a mixture model selected by MixtureFinder, the number might be greater than one. Model numbers are also extracted from the output files generated when estimating gene trees using IQ-TREE (version 2.3.5.1), and the model number for each locus in both regions was extracted.

We use the number of models in the best-fit model for each locus and the distribution of BIC differences to compare the one-class model and the mixture model for each locus.

### Relative frequency of the number of models selected by MixtureFinder in (a) Exon and (b) Intergenic region



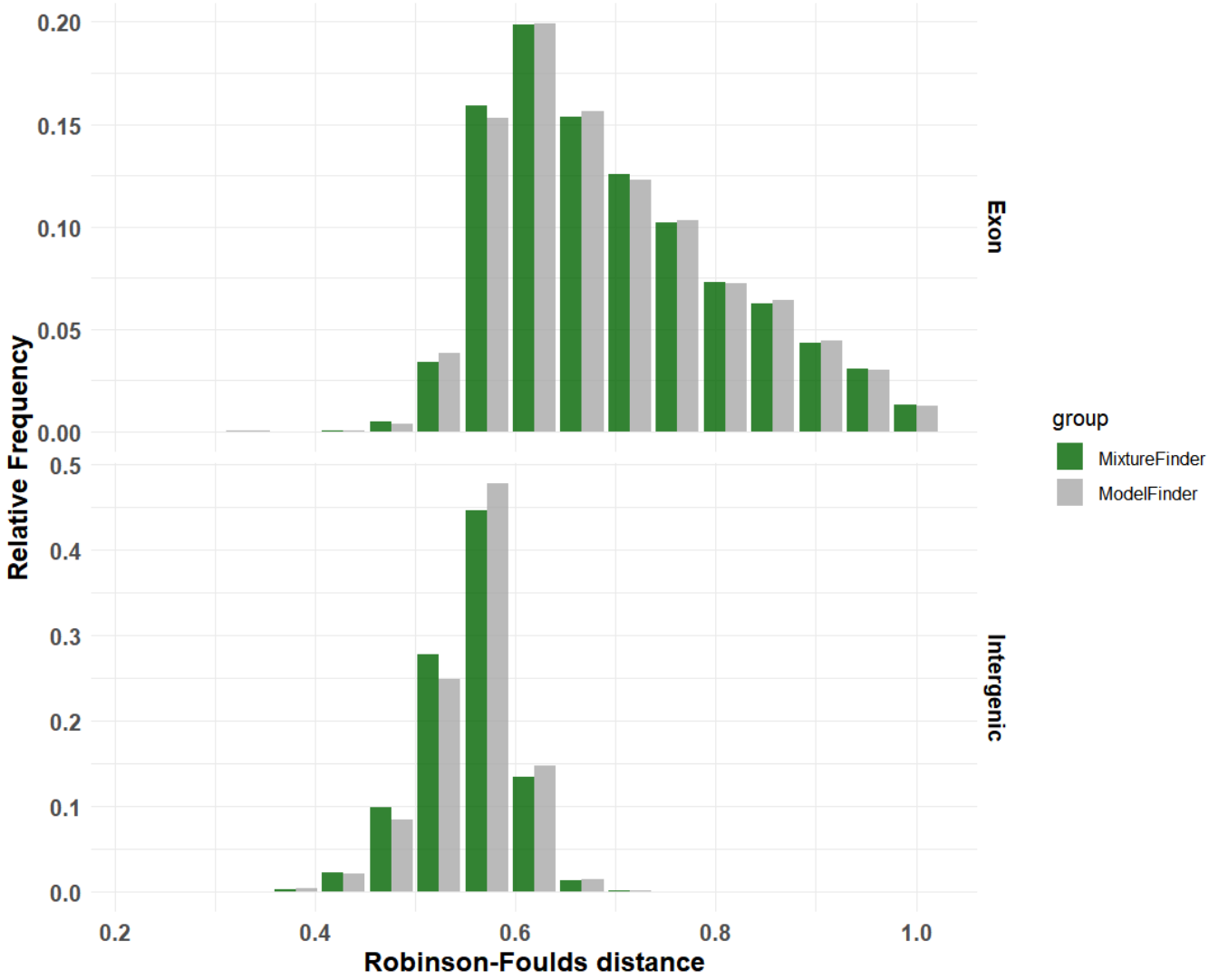

## 2. Evaluation of the gene trees

We use the normalized Robinson-Foulds distance and the tree length of gene trees to evaluate the gene trees produced using the One-class model and the Mixture model.

The Robinson-Foulds distance (Robinson & Foulds, 1981) is a measurement of the similarity between two trees. We calculate the normalized Robinson-Foulds distance (nRF, Robinson-Foulds distance divided by the maximum possible Robinson-Foulds distance for the pair of trees being compared) between each gene tree in a specific gene tree set and the corresponding ASTRAL species tree summarized by that gene tree set using IQ-TREE (version 2.3.5.1).

In general, a high nRF value indicates a significant difference between two trees, while a low nRF value indicates greater similarity (nRF value of 0 indicates that the two trees are identical). If the mixture model has a substantial impact on improving gene tree estimation, the nRF value should be lower than that using One-class model.

### Distribution of Normalised Robinson-Foulds (nRF) Distance




Tree length is simply defined as the sum of all branch lengths in a tree. The branch lengths for each tree in the four gene tree sets (as defined above) were extracted from the output files of IQ-TREE and summed into tree lengths.

### Tree length of Gene trees estimated by different method in exon and intergenic region

![alt text](d:\BIOL8706_project\data\final_progress_plot\gene tree length.png)

## Evaluation of the species trees of One-class model and mixture model for both regions

We first use IQ-TREE (version 2.3.5.1) to calculate the normalized RF distances between each of the two species trees of the exon region (one from the single model, and another from the Mixture model) and the two species tree estimated from the intergenic region (one from original study of in Stiller et al. (2024), another from ASTRAL), to directly show the similarity between exon and intergenic species trees.

| **nRF distance** | **Main tree from Stiller et al.** | **ASTRAL Intergenic species tree (One-class model)** |
|---|---|---|
| ASTRAL Exon treeOne-class model | 0.667 | 0.667 |

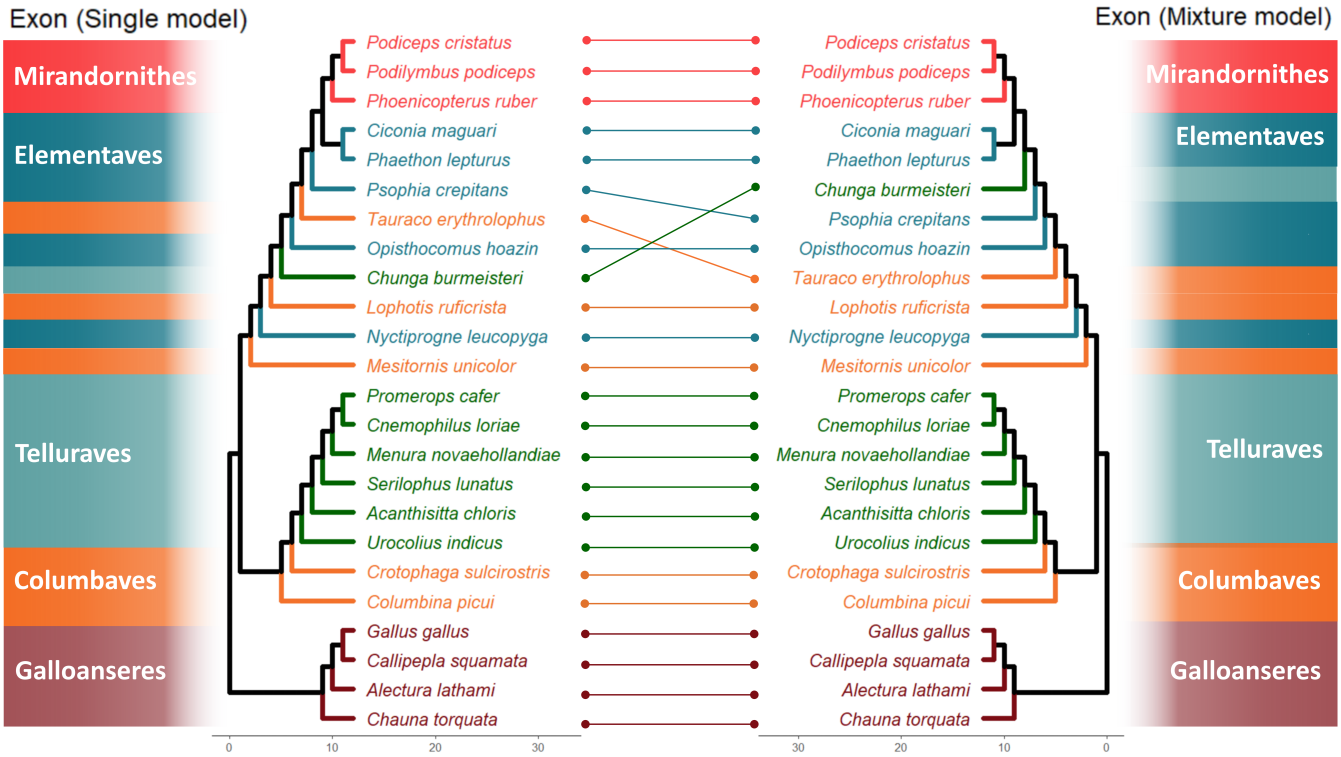
| nRF distance | Main tree from Stiller et al. | ASTRAL Intergenic species tree (One-class model) |
|---|---|---|
| ASTRAL Exon treemixture model | 0.667 | 0.667 |

We also use the qCF (quartet concordance factor) and posterior probability (pp) to implicitly evaluate the exon and intergenic species trees. The qCF (quartet concordance factor) is a measurement that counts the proportion of relevant quartets (subtrees of four taxa extracted from gene trees) associated with the reference topologies for the branch of interest in the species tree. Posterior probability is a measure assessing the confidence that the reference topology for the branch of interest is the most frequent among the gene trees (Lanfear & Hahn, 2024). We fix the subsampled main tree the original species tree estimated from the intergenic region in Stiller et al. as the assigned reference tree and calculate the four sets of qCF and pp values using each of the four gene tree sets (as defined above).

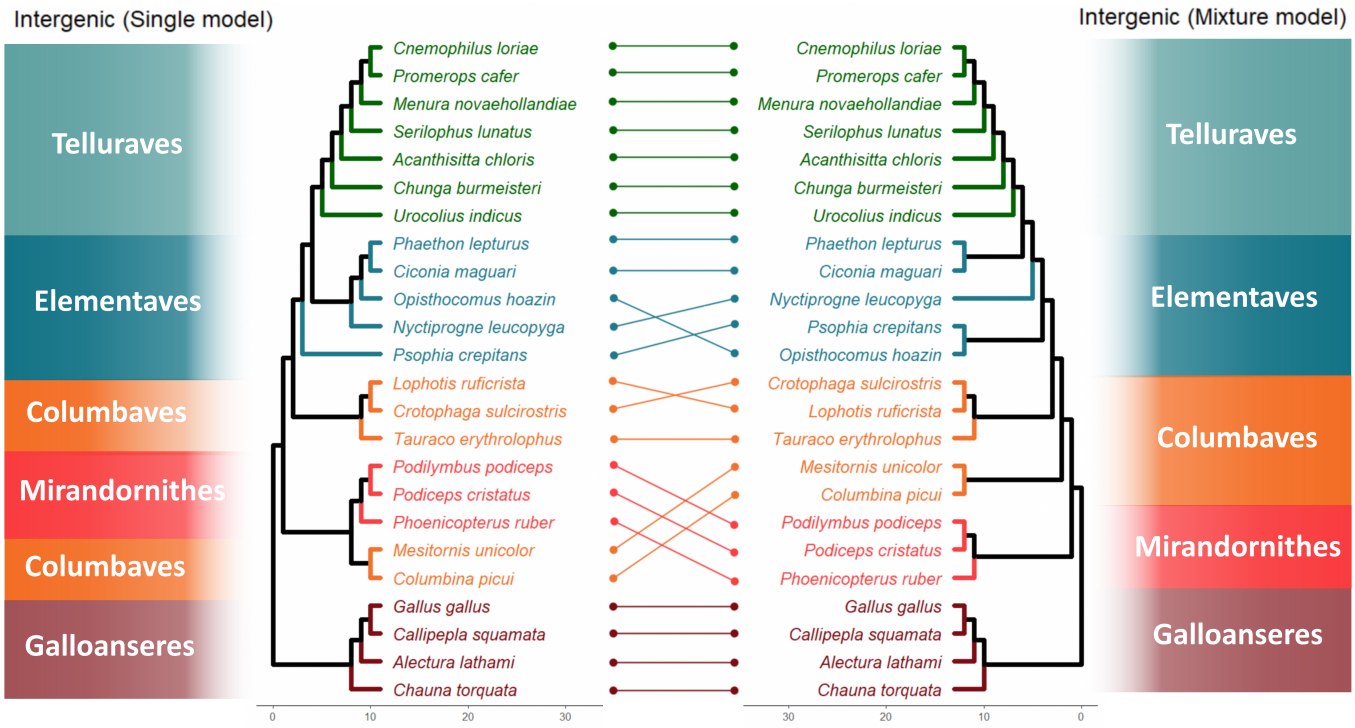
Since the exon species tree is summarized from the geen trees, greater similarity between the species trees of the exon and intergenic region would also be reflected in the gene trees. In oher words, if the mixture model has a substantial impact on making the exon species tree more similar to the intergenic tree, the qCF value using the mixture model should be higher than that using the one-class model. Supplementary Materials 2 show the tabel with full record of qCF and pp results.

## Tanglegrams of species trees:

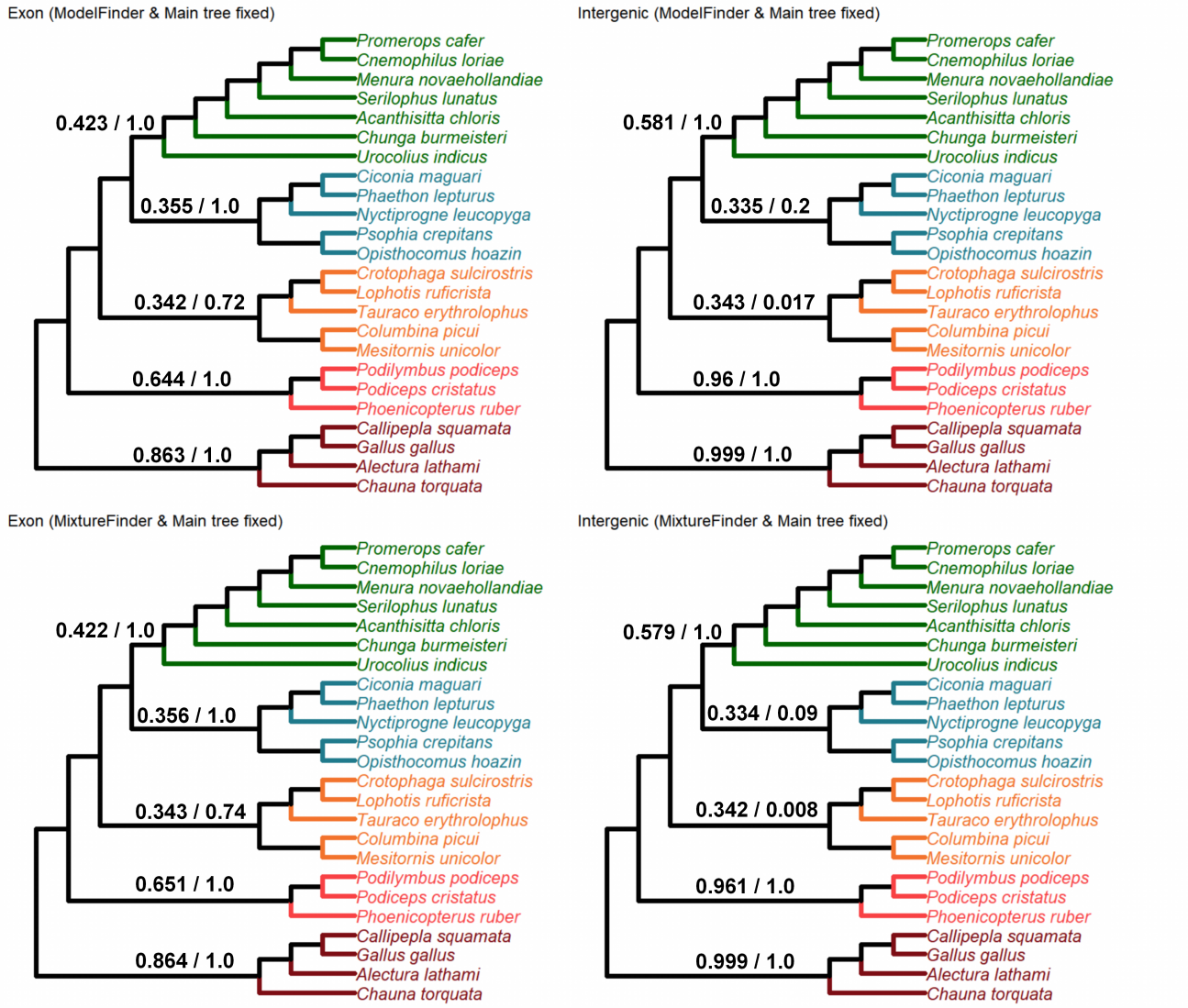
Exon region:




Intergenic region:

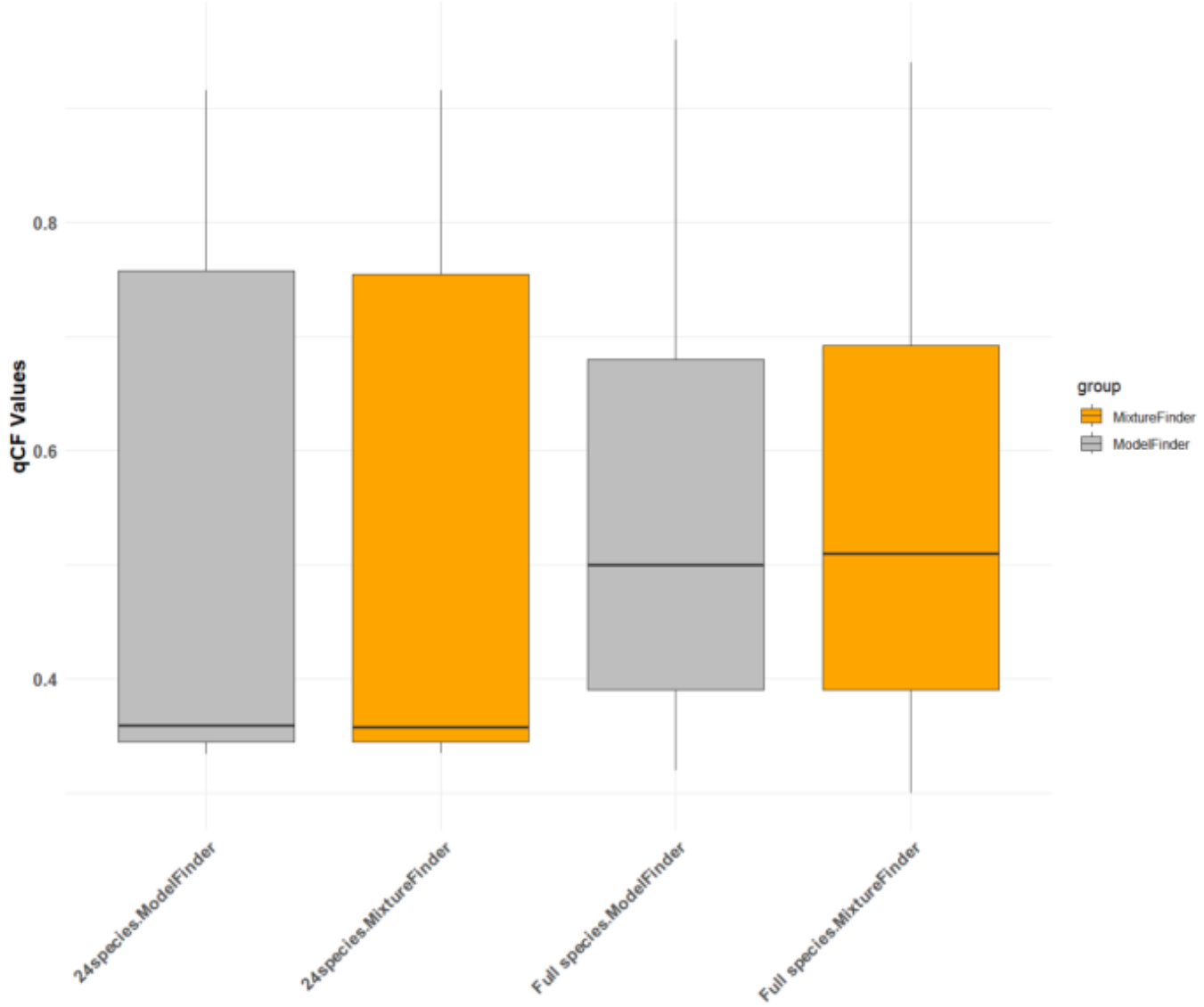



## Species tree with qcf/pp value:




## A small retry using full species

We randomly selected 100 exon loci that contained all species as a trial. The results showed that after using MixtureFinder, the qCF values for the exon region slightly increased. In contrast, the qCF distribution for 4255 exon loci containing only 24 species showed almost no change after using MixtureFinder. This supports our hypothesis that the limited species sampling in our methodology may have contributed to the less significant results.




This plot shows the distribution of qCF values in the tree estimated from 100 exon loci with the full species set (363 species) (two boxes on the right) and the qCF values in the tree estimated from 4255 exon loci with 24 selected species (two boxes on the left). Grey and orange colors represent the results from ModelFinder (One-class model) and MixtureFinder, respectively. From the figure, we can observe that when using 100 loci with the full species set, MixtureFinder increases the overall distribution of qCF values (two boxes on the right). However, when using 4255 loci with 24 species, there is no obvious improvement in qCF (two boxes on the left). qCF values were calculated using ASTRAL-5.7.1.