Report - Phylogenetics

Evolutionary Genomics HS 2020

Lionel Rohner

April 2020

1 Maximum Likelihood Methods

Introduction: During this practical, we used five 15Kb DNA regions from the nuclear genome (including Chromosomes 2, 8, 10, 18, and 22) and the full mitochondrial genome from Clownfishes. The aim was to construct phylogenetic trees of the Clownfishes using SeaView and to compare the different models we applied Akaike Information Criterion (AIC) estimation in R with the PhyML package.[2]
AIC is used to identify the most appropriate model from a selection of models. Unlike the Likelihood ratio test, AIC also allows comparing non-nested models (e.g. JC69+Γ and GTR). The AIC is composed of the maximum log-likelihood and the compensation term, that penalizes unnecessarily complex models (i.e. models with a lot of parameters to estimate).

$$AIC_i = -2ln(L_i) + 2p_i$$

Since SeaView and PhyML cannot read fasta files the first step was to convert the DNA-sequences into the extended phylip format (phy). This step was done using the "Analyses of Phylogenetics and Evolution" (ape) package in R.[4] Below you can find the R code that was used to convert fasta files into phy files as well as the calculation of the AIC for all models.

```
# Convert .fasta to .phy
d <- read.dna("clownfish_mtdna.fasta", format="fasta")
write.dna(d,"clownfish_mtdna.phy", format="sequential",
          nbcol=10000, colsep="", colw=100000)


# run phyml sequentially on each DNA model
#change the execname based on what you have on your computer
phyml.test <- phymltest(seqfile="clownfish_mtdna.phy",
                        execname="phyml_3.0_win32.exe -b 0 -o lr",
                        append = FALSE, "sequential")
#plot the results
plot(phyml.test)

# Save object to file
saveRDS(phyml.test, file = "model_compare_mtdna.rds")
```

Question 1.1: For the mitochondrial region, did the topology change between the JC69 model and the best model selected? Try to explain why. What else is different between the two trees built with different models?

The topology is different between the two models (Figure 1). Especially the initial branches towards the root of the tree differ in length. Two of the main sub-trees seem to be mirrored between JC69 and GTR+I+Γ, but the branch lengths are very similar (green and blue rectangles in Figure 1). The discrepancies could be due to the differences in the nucleotide frequency assumptions between the two models. In JC69 all the nucleotide frequencies are assumed to be identical (i.e. $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$). In comparison, the nucleotide sequences in the GTR model are assumed to be influenced by individual rates of changes from one nucleotide to another. Another striking difference is that the GTR model generates an overall bigger tree and has an overall better maximum log-likelihood (Figure 1) and AIC (Figure 3F) compared to JC69.

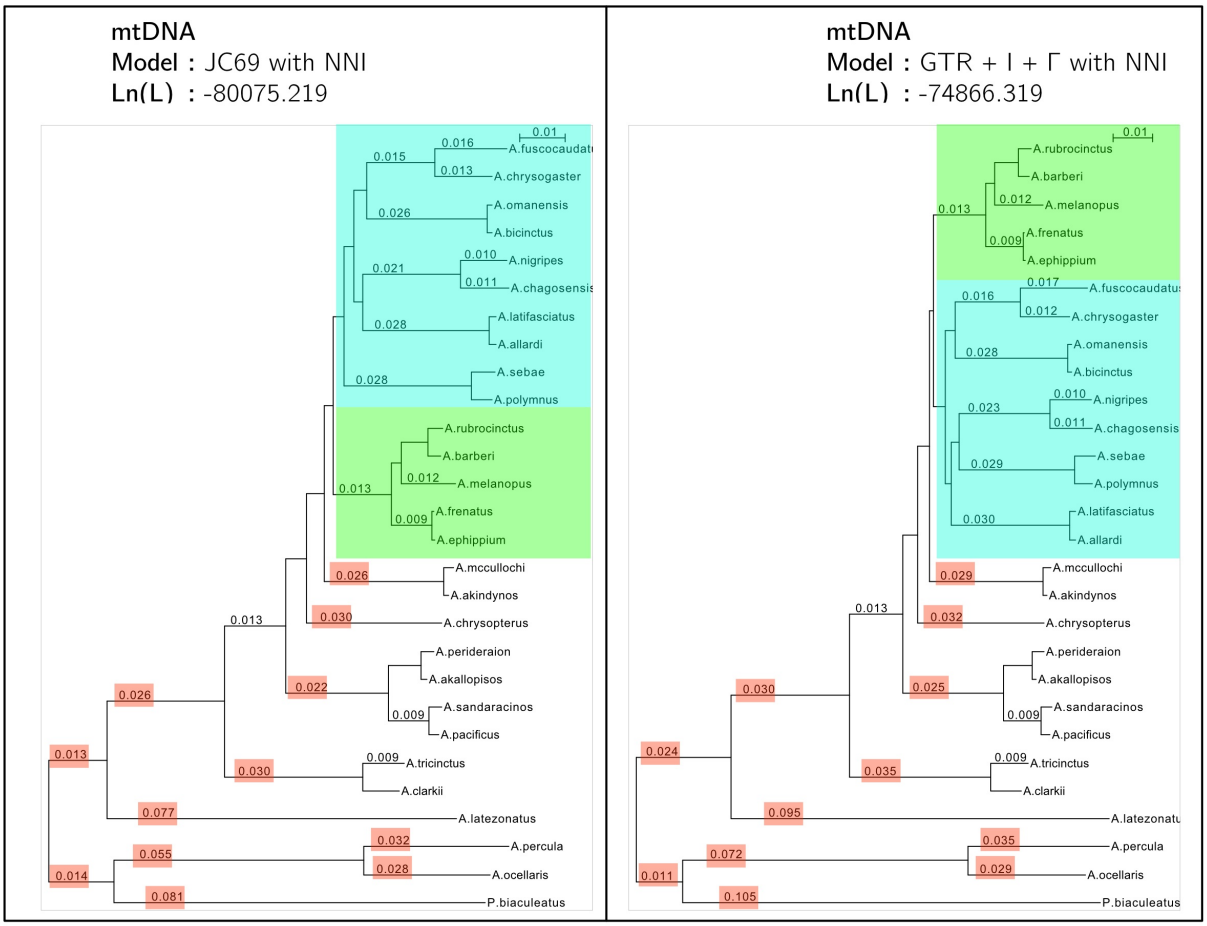

Figure 1: Difference between JC69 and GTR+I+Γ. Red numbers indicate different branch lengths between the two tested models. Green and blue blocks represent sub-trees that are identical in the compared models.Phylogenetic trees have been generated with SeaView. [1]

Question 1.2: For one of the DNA region (your choice), build a new phylogenetic tree by using the SPR branch swapping option instead of the NNI, which is selected by default. What are the consequences of setting this option? Compare the trees obtained and explain the differences if you see any.

First of all, the generation of a tree using the SPR option is much more time intensive compared to NNI. Moreover, the likelihood is only minimally better compared to the model with the NNI tree-rearrangement algorithm (Figure 2). This is because SPR compares much more neighbors. While NNI searches a local optimum in $2(n-3)$ neighbors, SPR searches through $4(n-3)(n-2)$ neighbors. Another difference is that two of the main sub-trees appear to be mirrored in the model with SPR (green and blue rectangles in Figure 2). Interestingly, the branch lengths in this mirrored sub-trees do not differ between the two tree-rearrangement algorithm. The same was already observed when comparing the JC69 and GTR model (Figure 1), suggesting that there is a high confidence for the these clades.

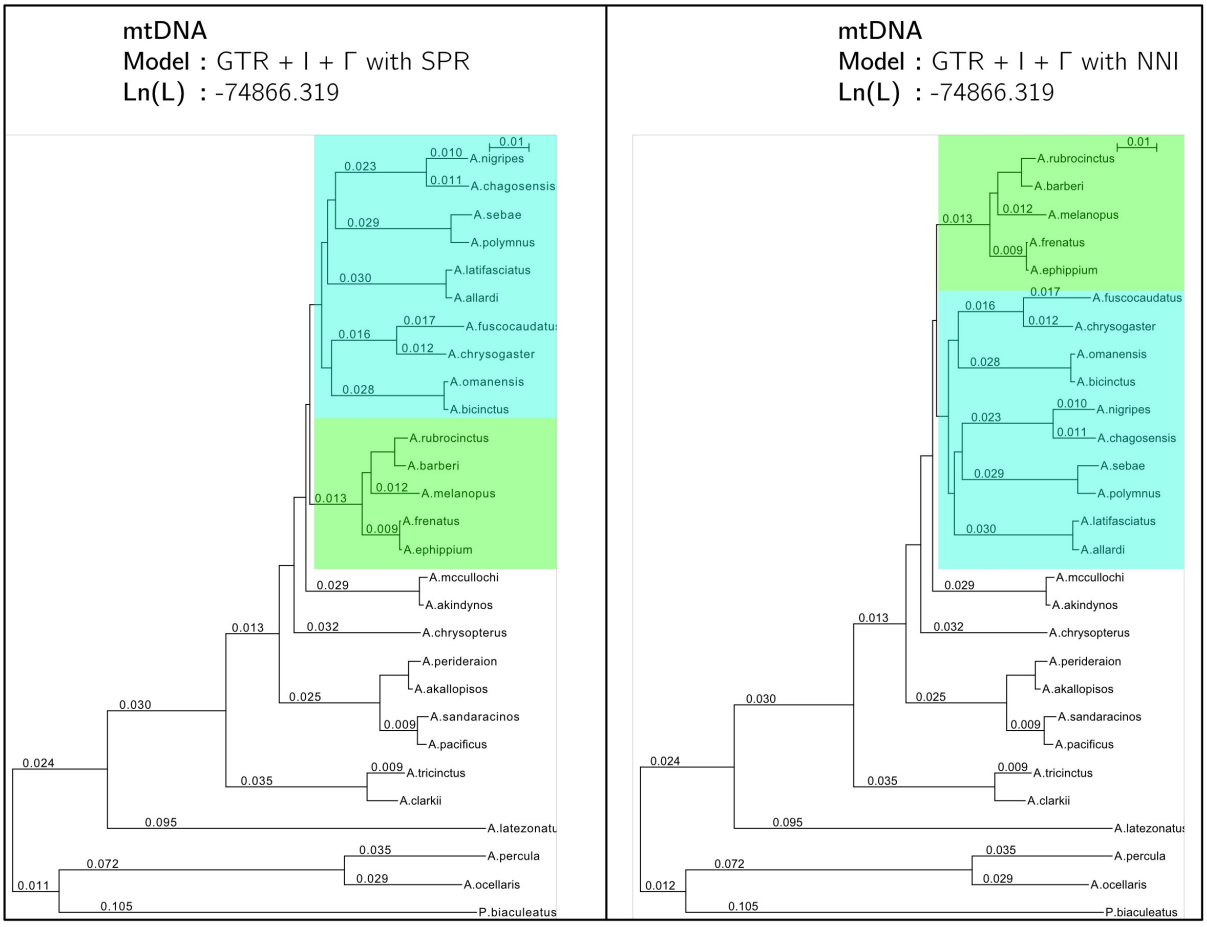

Figure 2: SeaView - Difference between Tree-Rearrangement Algorithm in mtDNA

Question 1.3: Assess the support for the different nodes of the phylogenetic trees built with each DNA region by running 100 bootstrap replicates.

The node support values indicate the plausibility of a particular branch estimated by bootstrapping. The values represent the number of times a given node appeared during the re-sampling technique compared to the total number of re-sampling replicates. For 100 bootstrap replicates we can assume that high support values (i.e. above 70) indicate that the branch is reliable. Interestingly, we can see that most values that are below 70 appear in the same sub-tree in all different chromosomes with the exception of chromosome 18 (Figure 3D). This could indicate that the genes in the chromosomes of these species are under selection or that adaptive radiation has recently occurred, which might explain why many species in this clade have emerged. Furthermore, whether the branch support is estimated by the approximate likelihood ratio test approach (aLRT - SH-like) or by bootstrap has no impact on the structure of the phylogenetic tree (compare Figure 2 and Figure 3F).

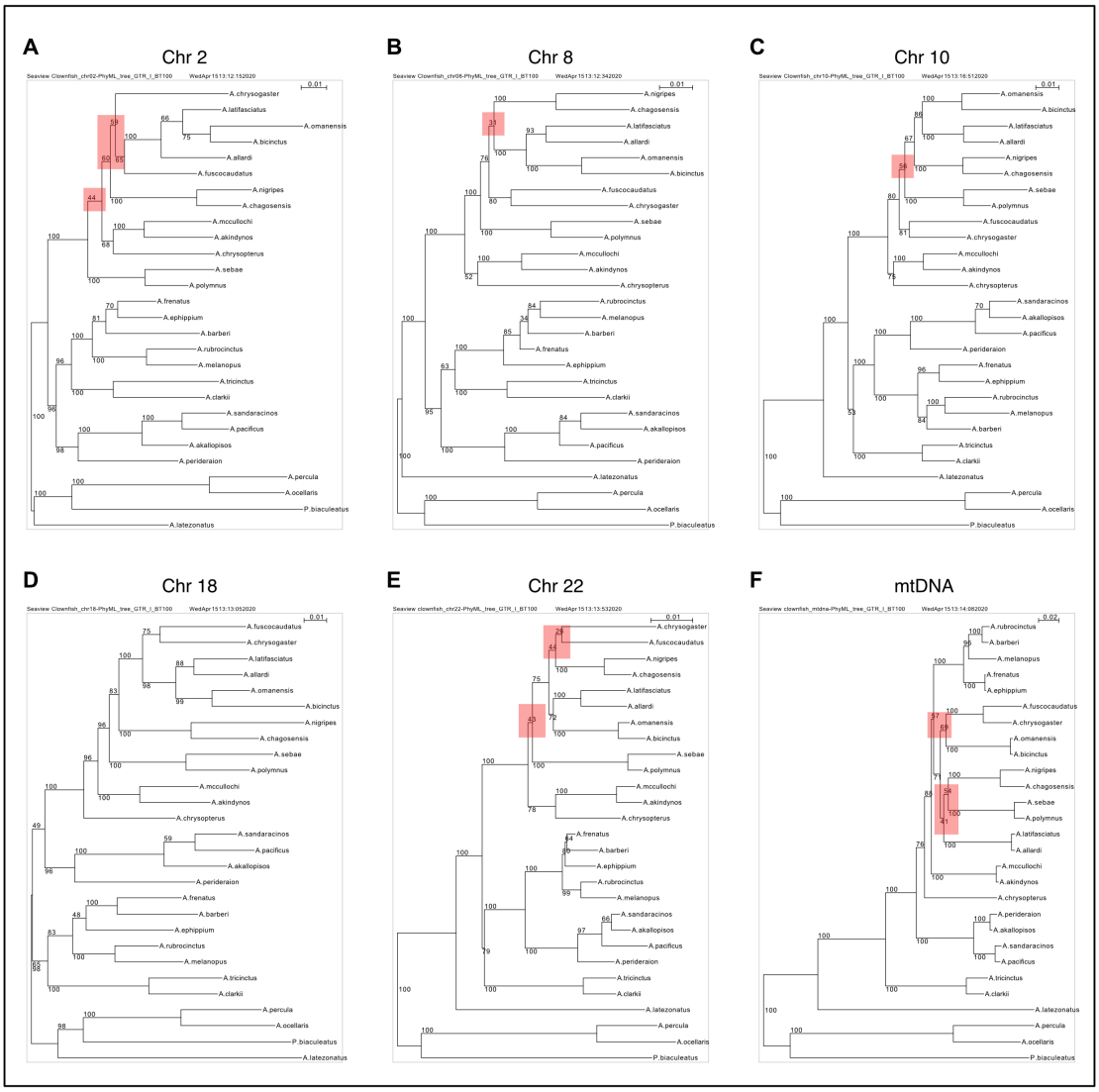

Figure 3: SeaView - Phylogenetic trees whose branch support were computed by bootstrapping. The phylogenetic trees of chromosome 2 (A), 8 (B), 10 (C), 18 (D), 22 (E), and mitochondrial DNA (F) are depicted. Red rectangles highlight low branch support values found in the same sub-tree.

Question 1.4: Concatenate the six DNA regions. Repeat the analyses for each DNA region (i.e. use each region as the DNA matrix and test if the trees explain their evolution equally well). Are the trees significantly different or not? Try to explain the results obtained?

We tested the AIC of the all available models, namely JC69, K80, F81, F83, HKY85, TN93, and GTR, for all individual chromosomes as well as the concatenated data containing all the Clownfish chromosomes (Figure 3). The best model for the concatenated data appeared to be the GTR+I+Γ model (Figure 3G), which is in line with the observations we have made when computing the AIC of the individual chromosomes (Figure 3A-F). In two chromosomes, namely chromosomes 18 and 22, the best model was TN93+I+Γ (Figure 3 D-E). It is important to point out that the AIC difference is minimal compared to the GTR+I+Γ. Since the different models differ in their assumptions of nucleotide frequencies and mutation rates, we decided to choose GTR+I+Γ for all chromosomes for later steps, in order to make the phylogenetic trees as comparable as possible.
Interestingly, we found that the parameters additional parameters proportion of invariable sites (I) and the shape parameter gamma, which implies that the degree of rate heterogeneity follows a gamma distribution, have a substantial effect on the AIC. It seems that these parameters are almost more important than the choice of the model. The models GTR, TN93, F84, and HKY85 have quite comparable AIC with the additional parameters (Figure 3A-F). Although JC69 performs much better with I and Γ, it still cannot compete with the other newer models. Similar trends were also observed with the models F81 and to some extent K80, which overall yield much higher AIC.

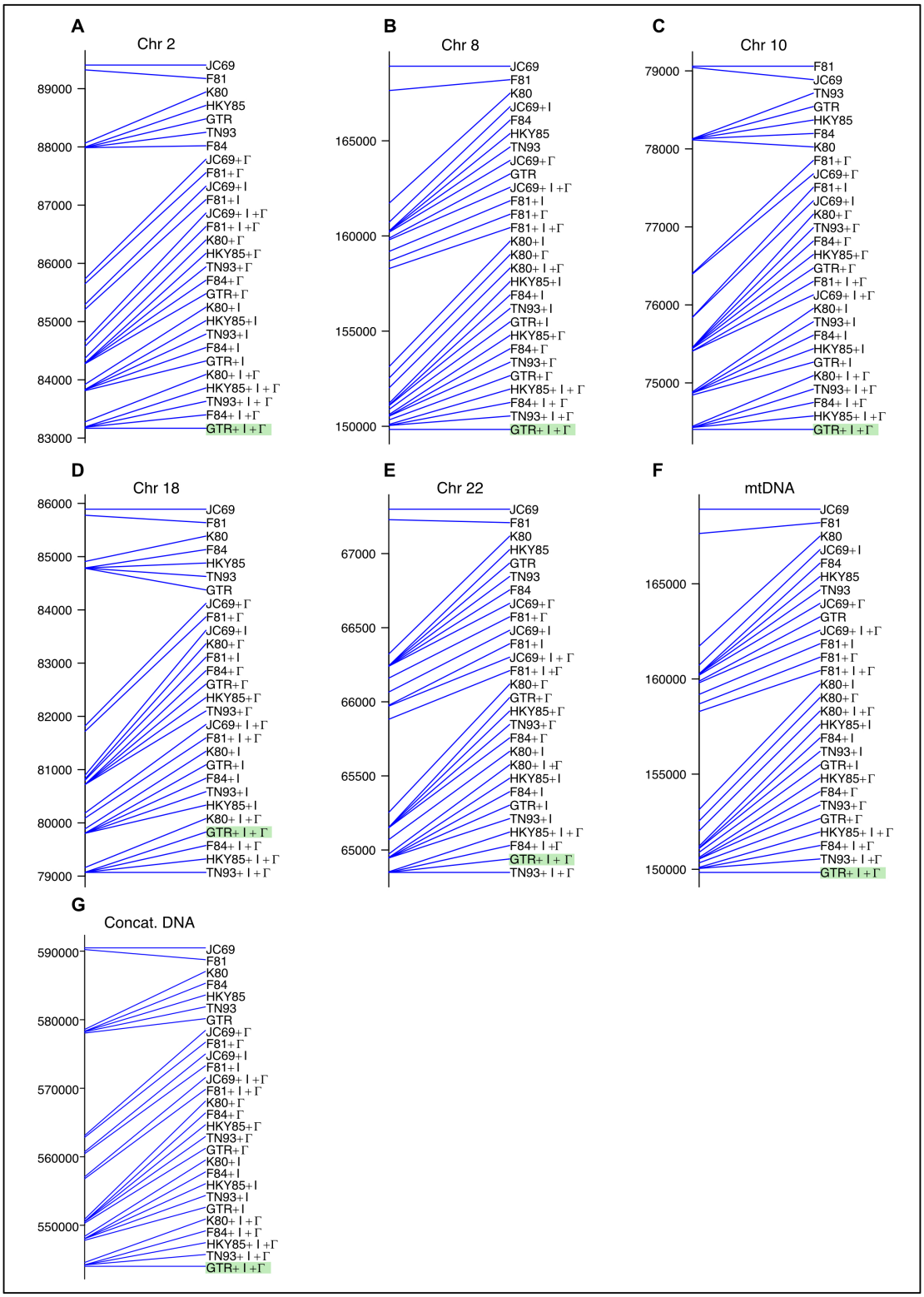

Figure 4: AIC of all tested models from every DNA sequence. AIC plot were generated using the APE package [4].

1.1 Topology Test

In the first part we used the MLE analysis to determine the different topologies for the Clownfish data set. The next task is to assess is to test the topologies of the constructed trees and to detect potential random errors in the tree reconstruction or real biological differences between the DNA regions that were tested. To assess the difference of the topology, we perform the Shimodaira-Hasegawa test.

Question 1.5: Repeat the analyses for each DNA region (i.e. use each region as the DNA matrix and test if the trees explain their evolution equally well). Are the trees significantly different or not? Try to explain the results obtained?

The Shimodaira-Hasegawa test (SH-test) is a bootstrap-based method that allows to evaluate the topological similarity of multiple candidate trees given a data set. This test uses the likelihood ratio test statistic to generate a null distribution using the bootstrapping re-sampling. The SH-test corrects for multiple testing and takes into account selection bias. The null hypothesis of the SH test states that all the tested candidate trees have the same expected score. For a given chromosome its respective tree yields a log-likelihood difference of 0 and consequently a high p-value, thus in this situation we cannot reject the null hypothesis. The bigger the difference of the log-likelihood, the smaller the p-value, which indicates that the tree topologies are significantly different given the tested DNA sequence. The amount of information that is contained in table 1 is quite vast, therefore only few striking findings will be discussed here. Overall, we can see that the tree generated from chromosome 2, chromosome 8 and mtDNA seems to differ significantly from any other tree in any non-redundant SH-test (i.e. tests that do not compare a particular tree with the DNA sequence it was generated from). It is important to note that the p-value of the mtDNA tree comparisons equals 0 in all non-redundant tests. This might be due to the fact that the SH-test assumes that all tested trees equal in likelihood, which does not apply for mtDNA, which has much lower log-likelihood values compared. Nonetheless, it is plausible that mtDNA has a substantially different topology compared to autosomal chromosomes as it is maternally inherited and is not subject to bi-parental recombination in the germ line. In contrast, chromosome 22 is only significantly different from chromosome 2, but is topologically similar to any other phylogenetic tree. Similar results were observed for chromosome 10.

Tested Data	Trees	ln$\mathcal{L}$	Diff ln$\mathcal{L}$	p-value
Chr 2	Chr 2	−41 542.7896	0	0.8063
	Chr 8	−41 710.8593	168.069 739	0.0019
	Chr 10	−41 715.0244	172.234 817	0.0018
	Chr 18	−41 586.1148	43.325 186	0.3188
	Chr 22	−41 698.1435	155.353 897	0.0059
	mtDNA	−42 149.9324	607.142 806	0
Chr 8	Chr 2	−39 495.6131	92.746 690 7	0.0291
	Chr 8	−39 402.8664	0	0.8897
	Chr 10	−39 433.2348	30.368 422	0.3981
	Chr 18	−39 458.0259	55.159 458 7	0.1267
	Chr 22	−39 427.347	24.480 568 1	0.4803
	mtDNA	−39 754.6835	351.817 107	0
Chr 10	Chr 2	−37 379.3145	189.368 756	0
	Chr 8	−37 274.8016	84.855 869	0.0412
	Chr 10	−37 189.9457	0	0.8734
	Chr 18	−37 314.7884	124.842 679	0.004
	Chr 22	−37 251.8893	61.943 555 1	0.1352
	mtDNA	−37 625.0677	435.122 015	0
Chr 18	Chr 2	−39 612.471 36	88.860 655 28	0.0493
	Chr 8	−39 601.428 85	77.818 138 86	0.0621
	Chr 10	−39 619.022 83	95.412 116 76	0.0236
	Chr 18	−39 523.610 71	0	0.8597
	Chr 22	−39 562.762 75	39.152 037 19	0.2904
	mtDNA	−39 990.788 85	467.178 145 8	0
Chr 22	Chr 2	−32 559.642 39	156.606 008	2.00×10^{-4}
	Chr 8	−32 486.920 34	83.883 958 14	0.0184
	Chr 10	−32 442.966 68	39.930 289 52	0.2466
	Chr 18	−32 543.701 27	140.664 881 9	3.00×10^{-4}
	Chr 22	−32 403.036 39	0	0.84
	mtDNA	−32 772.554 25	369.517 868 7	0
mtDNA	Chr 2	−76 515.3258	1663.169 97	0
	Chr 8	−76 278.8007	1426.644 82	0
	Chr 10	−76 048.6397	1196.4839	0
	Chr 18	−75 958.552	1106.396 15	0
	Chr 22	−76 645.2367	1793.080 88	0
	mtDNA	−74 852.1558	0	0.7645

Table 1: SH-test Output from R. SH-tests were performed using the Phangorn Library.[3]

2 Gene Trees

In this part of the practical, the aim was to infer species tree of Clownfish species given the DNA sequencing data that was used previously.
Question 2.1: Is the species obtained by *BEAST2 the same as the PhyML trees? Why?

Maybe, maybe not. **Question 2.2:** Which gene trees are incongruent with the species tree?

3 Positive Selection

Question 3.1:

References

[1] M. Gouy, S. Guindon, and O. Gascuel. "SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building". In: *Mol. Biol. Evol.* 27.2 (Feb. 2010), pp. 221–224.

[2] S. Guindon et al. "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0". In: *Syst. Biol.* 59.3 (May 2010), pp. 307–321.

[3] K.P. Schliep. "phangorn: phylogenetic analysis in R". In: *Bioinformatics* 27.4 (2011), pp. 592–593. URL: `https://doi.org/10.1093/bioinformatics/btq706`.

[4] E. Paradis and K. Schliep. "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". In: *Bioinformatics* 35 (2018), pp. 526–528.