UNIBE

Evolutionary Genomics – HS 2020

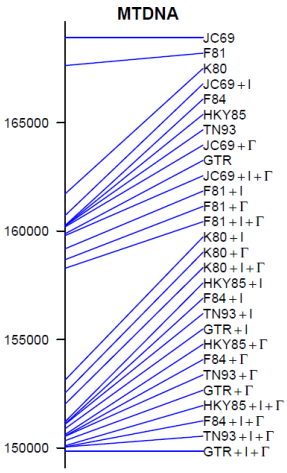
Phylogenetics

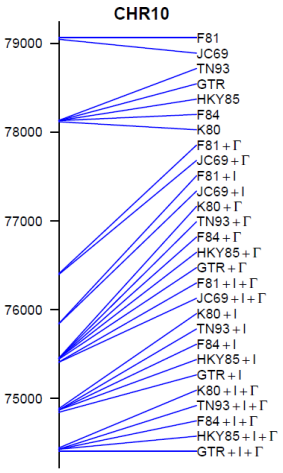
Thibault Schowing
17/04/2020

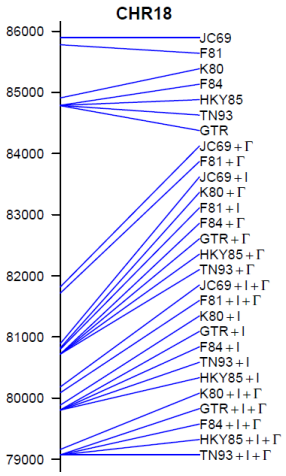
Maximum Likelihood with PhyML

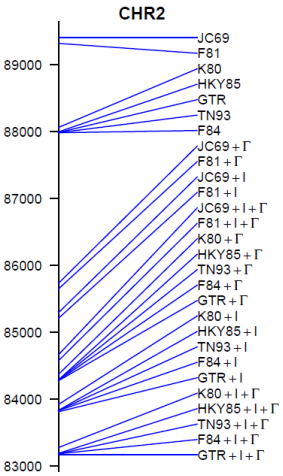
Find out which are the best models for each of the DNA regions available for the clownfishes and reconstruct a phylogenetic tree for each DNA region.

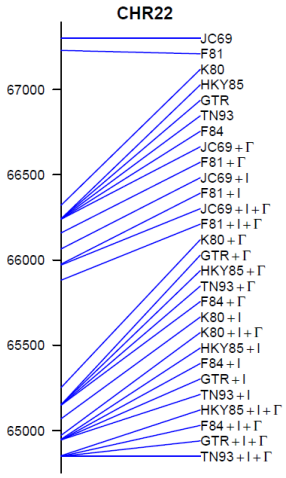
For each region, GTR (General-time reversible) is the best or is among the best models, including with the concatenated genome.

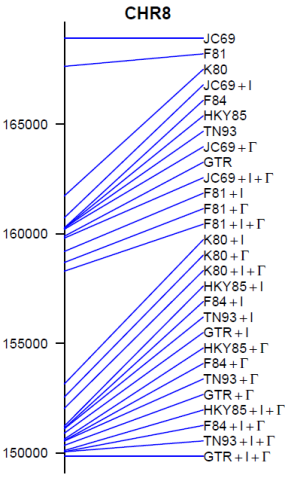


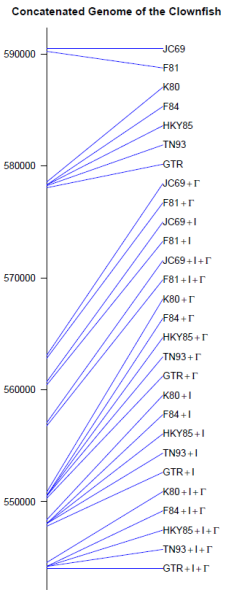






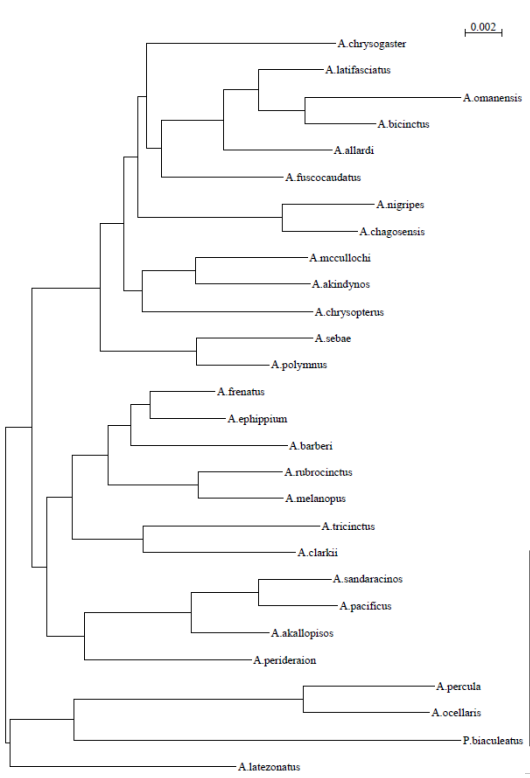


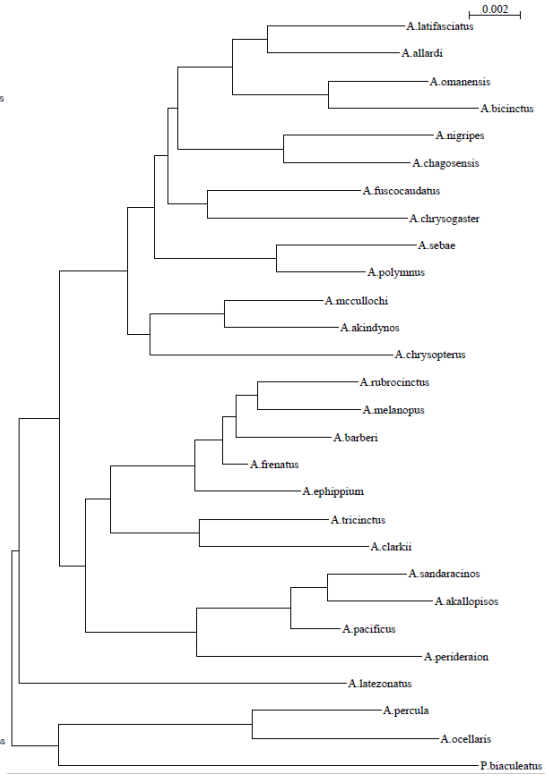




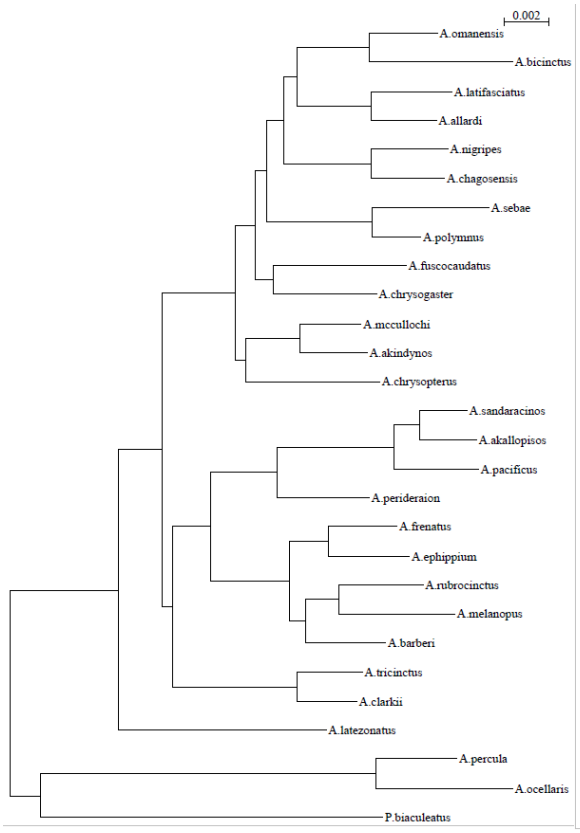
Trees generated using GTR for each DNA region:

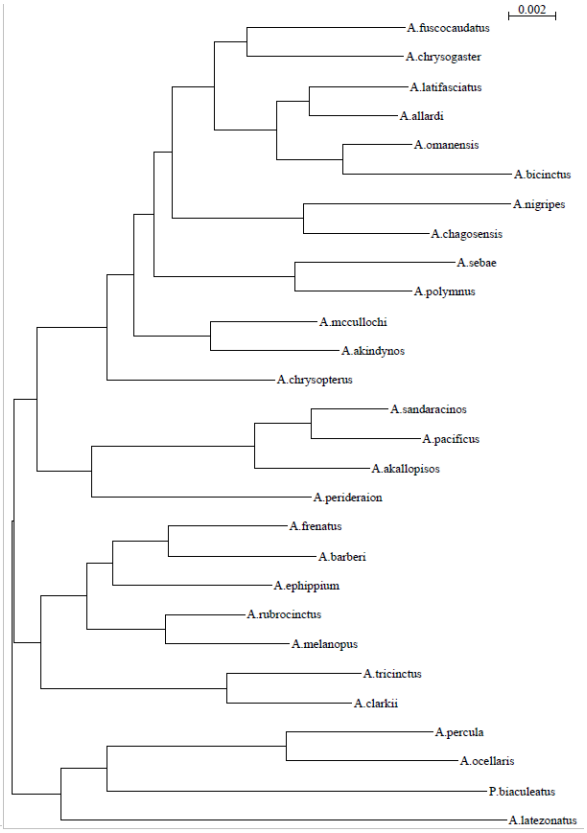
Chr02 and Chr03:



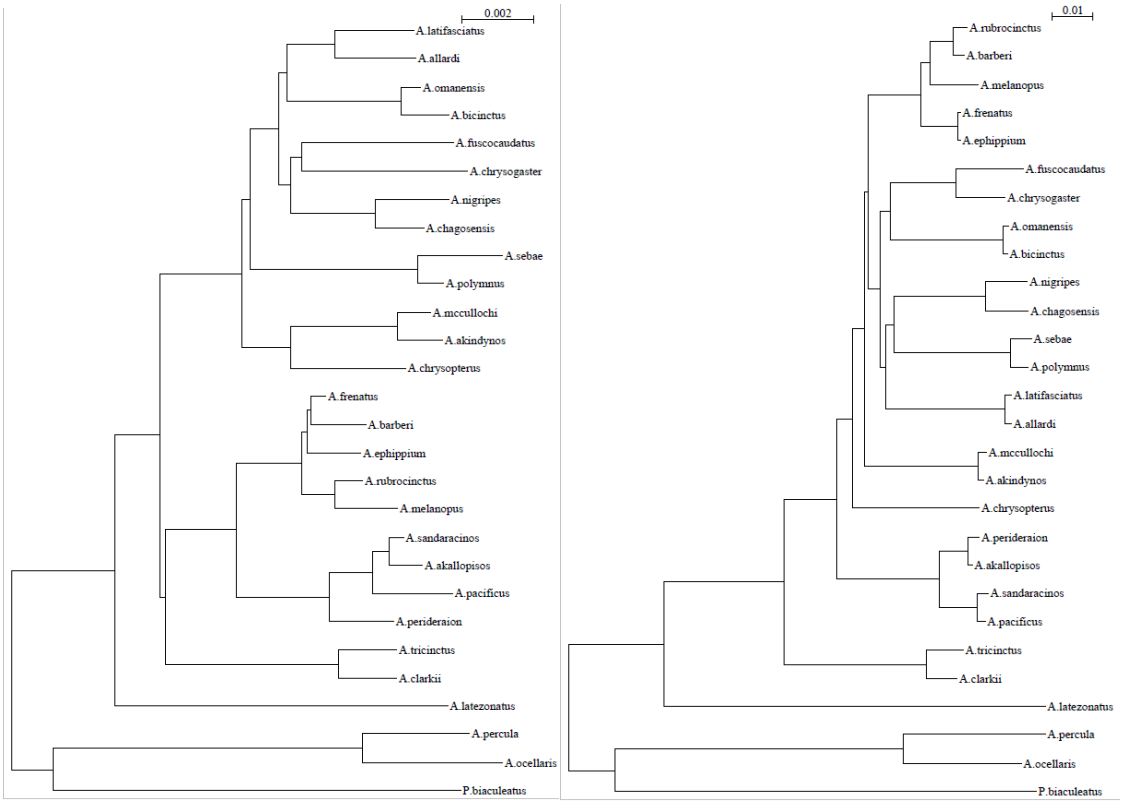


Chr10 and Chr18:



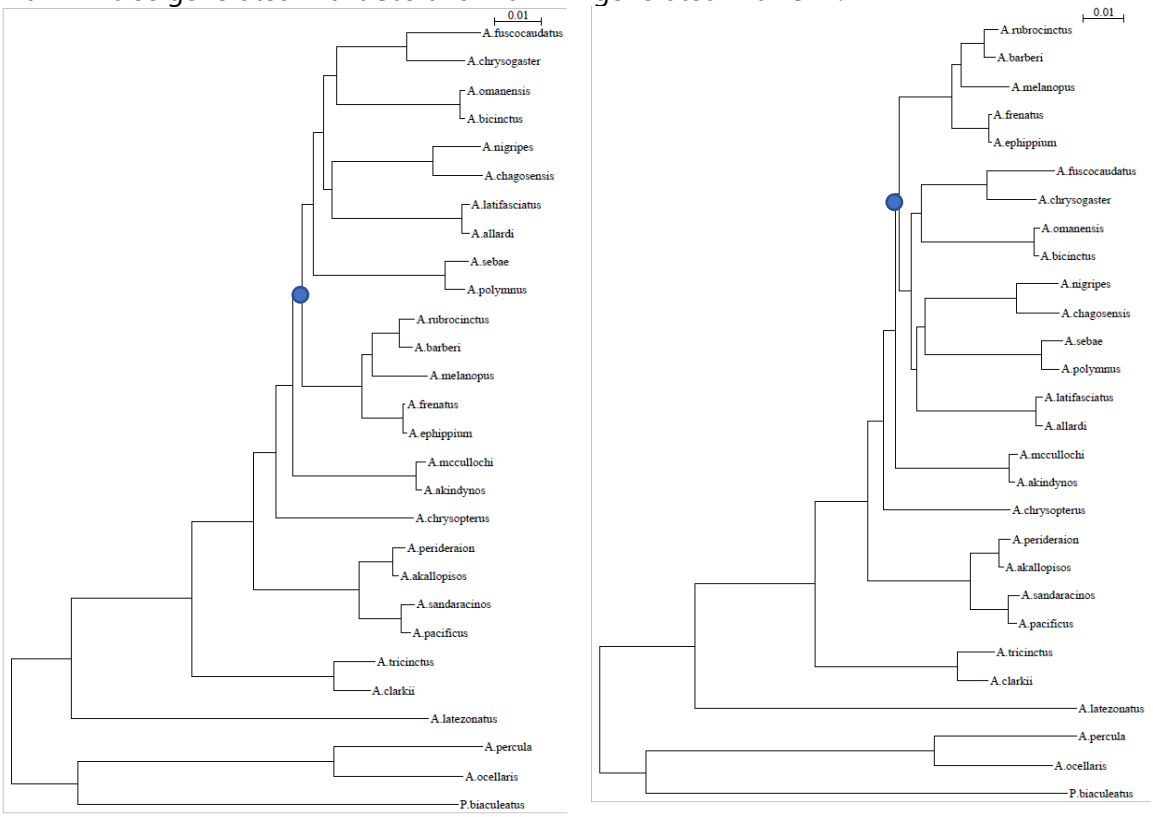


Chr22 and Mt-DNA:



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?

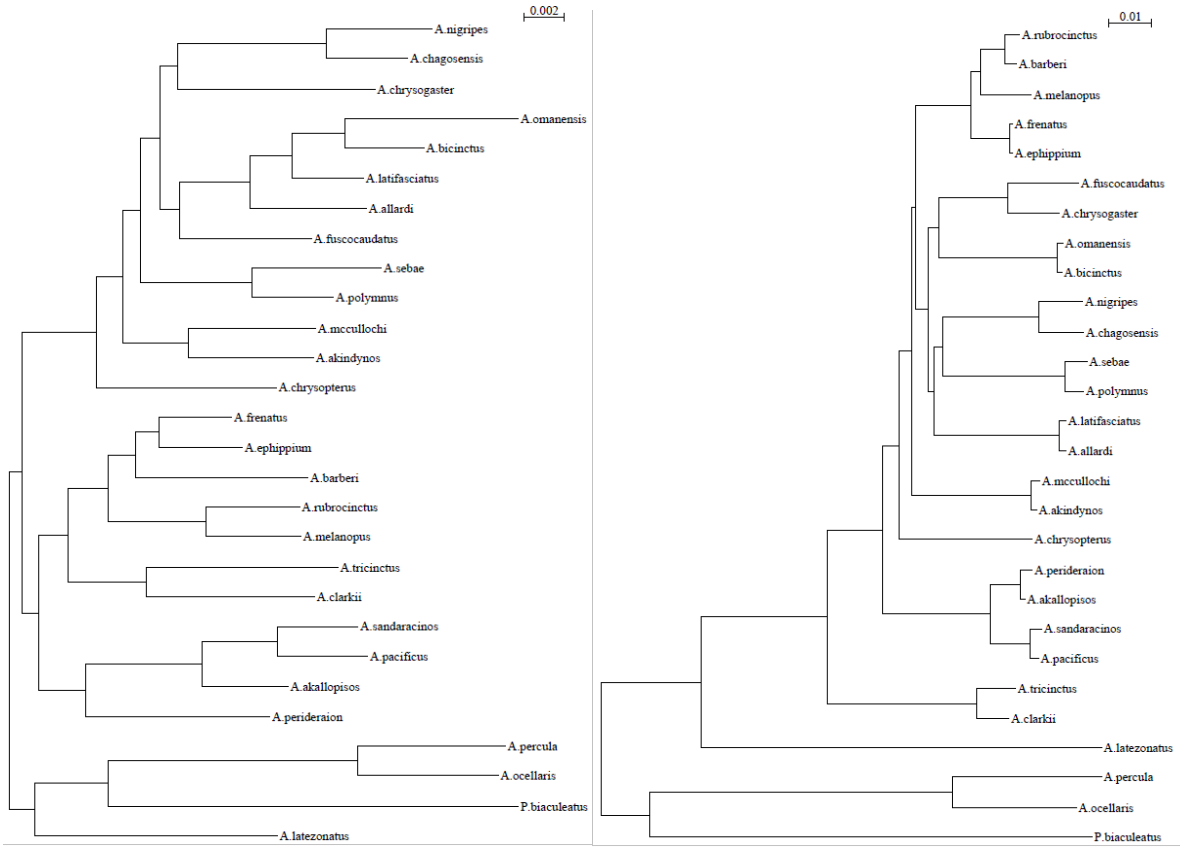
Mt-DNA tree generated with JC69 and Mt-DNA generated with GTR:



We can see that the oldest organisms are identically placed on the tree up to *A. mccullochi*. After, the trees are rotated where the blue dot is placed. For JC69, all base frequencies and probabilities of mutation are the same which is not the case with GTR. For GTR, the parameters are different for each transition (mutation rate, base transition).

For one of the DNA regions (Mt-DNA), 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.

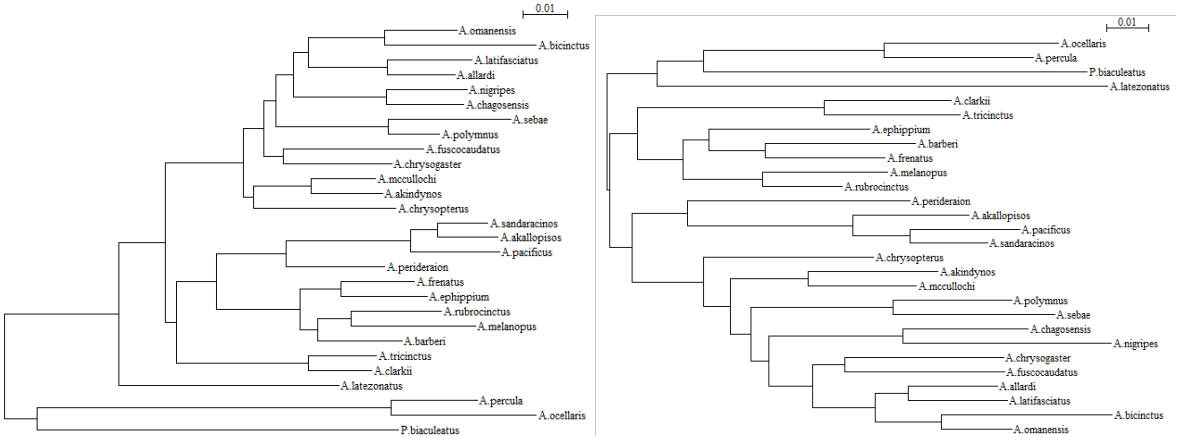
Mt-DNA tree with SPR swapping and Mt-DNA tree with NNI respectively:



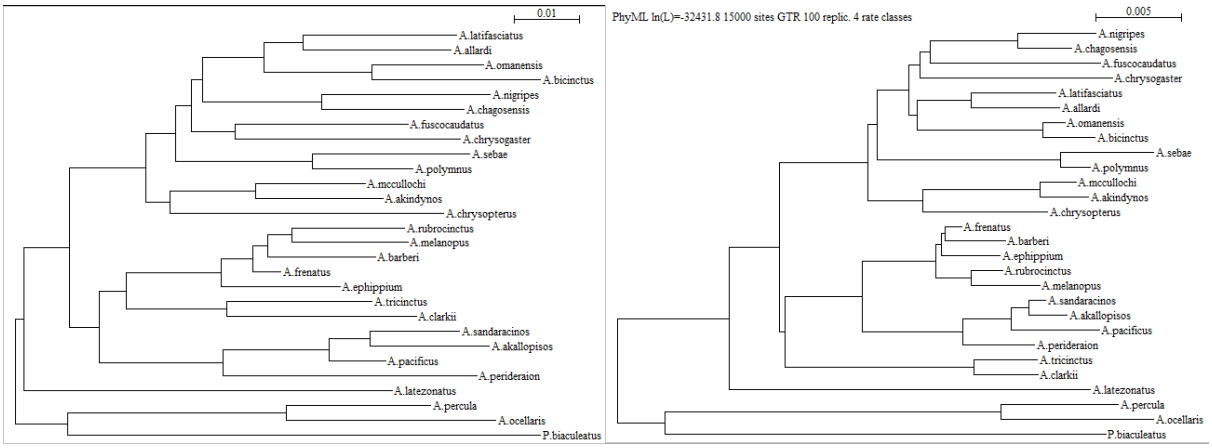
The two methods used above (SPR and NNI) are used to swap branches. SPR searches a bigger number of neighbors while NNI search for a local optimum only. We can observe in the two trees above, that SPR produce much longer branches. We can observe many more differences between these two trees than between the two previous models (JC69 and GTR) even if we can still see similar sub-trees grouping the same species each time (high confidence in some of the clades).

Assess the support for the different nodes of the phylogenetic trees built with each DNA region by running 100 bootstrap replicates. Use the best approach for the other options available in PhyML. Trees generated with GTR, NNI and 100 bootstrap, for each chromosome:

Chr10 and 18:

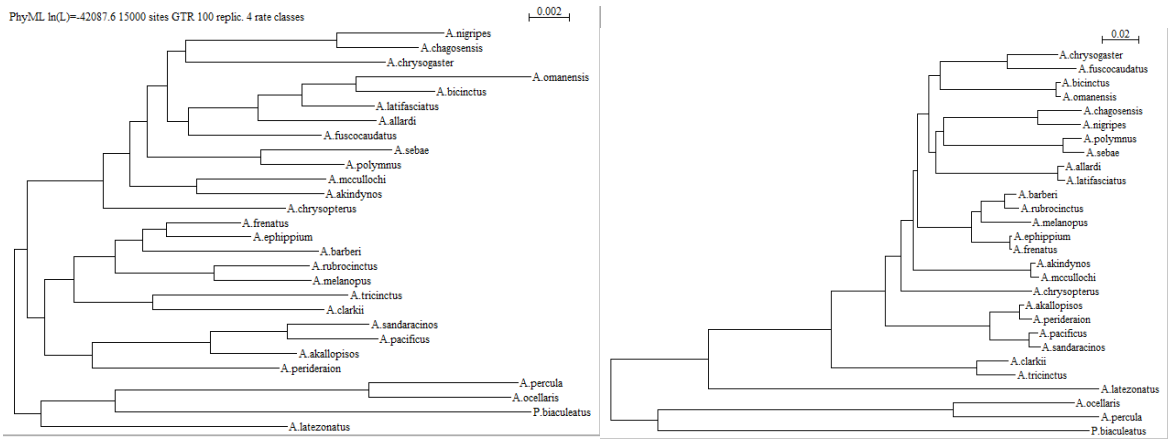


Chr08 and Chr22



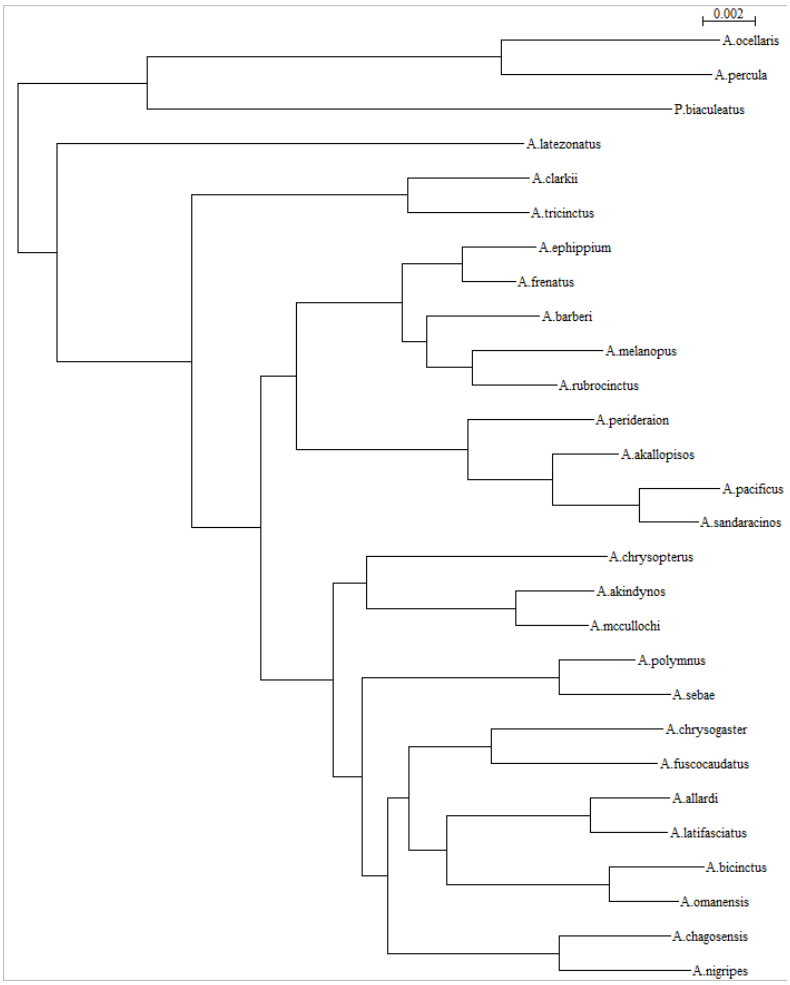
Chr02 and mtDna

Note: the mtDna tree with 100 bootstrap took more than 3 hours to compute and the tree-building completion took an endless amount of time (seaview crashed). The tree here is from Jacqueline Wyss.



Tree from concatenated genome

Here is the Phyml tree created with GTR and NNI for the concatenated genome:



Except a few changes in blocks order, the trees keep the same general order between species. Distances vary but stay in the same range between the separated DNA regions and the concatenated genome.

Topology tests

Shimodaira-Hasegawa on 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?

SH test done with chromosomes: 2, 8, 10, 18, 22 and mtDna respectively:

Trees	ln L Diff	ln L	p-value
1	-32559.64	156.60601	0.0002
2	-32486.92	83.88396	0.0190
3	-32442.97	39.93029	0.2509
4	-32543.70	140.6648	0.0002
5	-32403.04	0.00000	0.8400
6	-32772.55	369.51787	0.0000

Chromosomes 22 and 10 have a high p-value while mtDna, chromosome 2 and 18 are significant, meaning here that the tree topologies are significantly different given a tested DNA sequence but for instance chromosome 10 and 22 (trees #3 and #5) are not.

Gene trees species tree estimation

Questions

- is the species obtained by *BEAST2 the same as the PhyML trees? Why?
- which gene trees are incongruent with the species tree?

They are all more or less incongruent with the species tree and, also with each other. There is always consistency within small clusters for instance *A. ocellaris* is always with the ancient *A percula* but there are always inversions and switches between groups.

Genes can be duplicated and evolve apart from each other, be deleted or even being transferred horizontally. For this reason, the phylogenetic tree of a gene can be very different than the tree of the species. The species tree is a consensus of the gene trees that allow to have the most coherent tree in the end.

Positive selection

For each gene, modify/correct the .phy files as requested and reconstruct a phylogenetic tree as in the first part with the commands:

./PhyML_3.0_win32.exe -i Clownfish_slc9a6.phy -m GTR -b 0
./PhyML_3.0_win32.exe -i Clownfish_snai2.phy -m GTR -b 0
./PhyML_3.0_win32.exe -i Clownfish_tbx2.phy -m GTR -b 0
./PhyML_3.0_win32.exe -i Clownfish_rh1.phy -m GTR -b 0

Reroot the tree and add "#1" at the ancestral node.

Note: the names in the fasta file do not correspond to the fish names, it is thus impossible to know with which organism to reroot the tree. This has only been done for the *rh1* gene but the entire procedure stays the same.

Gene rh1

Log likelihood of the different models. *Ntime: number of branch length. Np: total number of parameters, including branch length.*

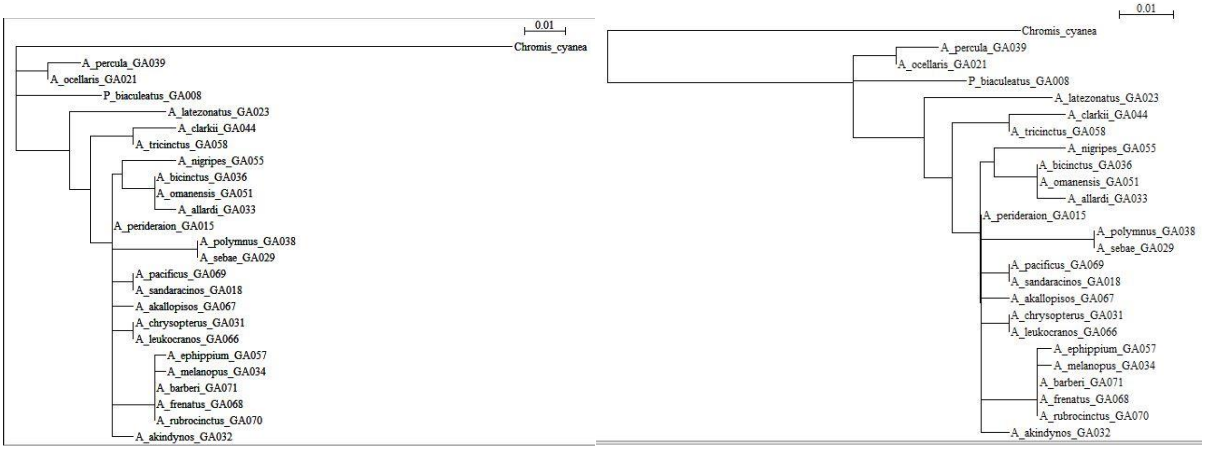
- A1alt: lnL(ntime: 48 np: 53): -2370.155507 +0.000000
- A1null: lnL(ntime: 48 np: 52): -2371.340911 +0.000000
- M1a: lnL(ntime: 48 np: 51): -2371.340901 +0.000000
- M2a: lnL(ntime: 48 np: 53): -2370.569786 +0.000000

The likelihood ratio test between M1a and M2a gives us a p-value of **0.46** and thus we reject M2a (alternative model) that is not significantly different from M1a even if it has a better likelihood.

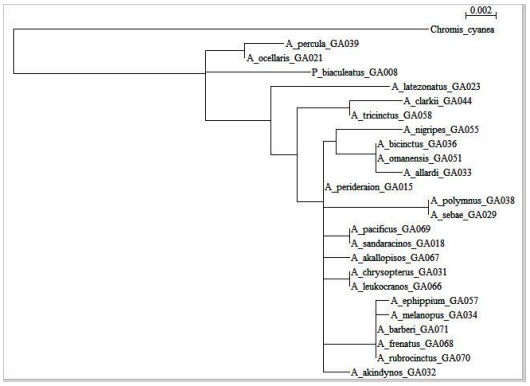
Comparing the A1null and A1alt models, we get a p-value of **0.12** with a degree of freedom of 1, so we also reject the alternative model but with a less obvious margin.

We can only compare A models with themselves and similarly with M1 and 2 because the models have to be nested to be comparable.

Tree length comparison **A1null** (left) and **M1null** (right):



And length estimation using HKY85 model:



With HKY85 the tree looks close to the one with M1null but, looking at the scale, we see that the branches are estimated to be much shorter.

The next steps with the other genes (**slc9a6, snai2 and tbx2)** could not be repeated, due to lack of information in the files (organisms name). Doing this for different genes can help to find evidence of past selective pressure and adaptation.