HP-SEE
DNA muligene approach on HPC using RAxML software

www.hp-see.eu

HP-SEE

High-Performance Computing Infrastructure for South East Europe's Research Communities

Luka Filipović
Danilo Mrdak
Božo Krstajić
University of Montenegro

The HP-SEE initiative is co-funded by the European Commission under the FP7 Research Infrastructures contract no. 261499

Phylogeny

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- Phylogenetic analysis
  - standard and essential tool in any molecular biologist's bioinformatics in the context of protein sequence analysis
  - enables us to study the evolutionary history and change of proteins and their function.
- Phylogenetic techniques have been used to explore the family tree of relationships between specific genes shared by many types of organisms
- Phylogenetic trees are mathematical structures that shows the evolutionary history of a group of organisms or genes.
- Computational phylogenetic
  - Consist of computational algorithms, methods and programs for phylogenetic analyses.
  - challenging even for the most powerful supercomputers.

Scope of our research

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- Phylogeny analysis of
  - fish species: salmons, trouts, grayling, ...
  - specially comparison of spices like
    - Mediterranean, Adriatic Danubian and Atlantic linage brown trouts,
    - Marbled, soft-muzzled, Ohrid lake trouts,
    - Atlantic salmon
    - River Huchen
    - Grayling

- Mitochondrial DNA D-loop, Cytochrom b gene analysis

About RAxML

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- RAxML
  - Random Axelerated Maximum Likelikhood,
  - Program for Maximum Likelihood based inference of large phylogenetic trees
- Developed by A. Stamatakis, The Exelixis Lab
- Built on base of FASTDNAML and DNAML program
- Sequential and parallel version
- Available for the Unix/Linux, Mac, and Windows operating systems

RAxML : how it works?

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- First step
  - RAxML generates large number of **starting trees** (defined by bootstrap) by adding the sequences one by one in random order, and identifying their optimal location on the tree under the parsimony optimality criterion.
- Second step
  - involves a method known as **lazy subtree rearrangement** (LSR) where all possible subtrees of a main tree are clipped and reinserted at all possible locations as long as the number of branches separating the clipped and insertion points is smaller than N branches. RAXML estimates the appropriate N value for a given data set automatically, but one can also run the program with any fixed value. The LSR method is first applied on the starting tree, and subsequently multiple times on the currently best tree as the search continues, until no better tree is found.

RAxML parallelization

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- Parallelization methods:
  - Coarse grained parallelization – using MPI
    - Division of bootstraps by CPU cores
  - Fine grained parallelization – using OpenMP and later Pthreads
    - Over number of patterns
  - Hybrid version, which combines coarse and fine grained parallelization.
- Experimental versions have been developed for Cell Broadband Engine and BlueGene/L

RAxML @ HP SEE

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

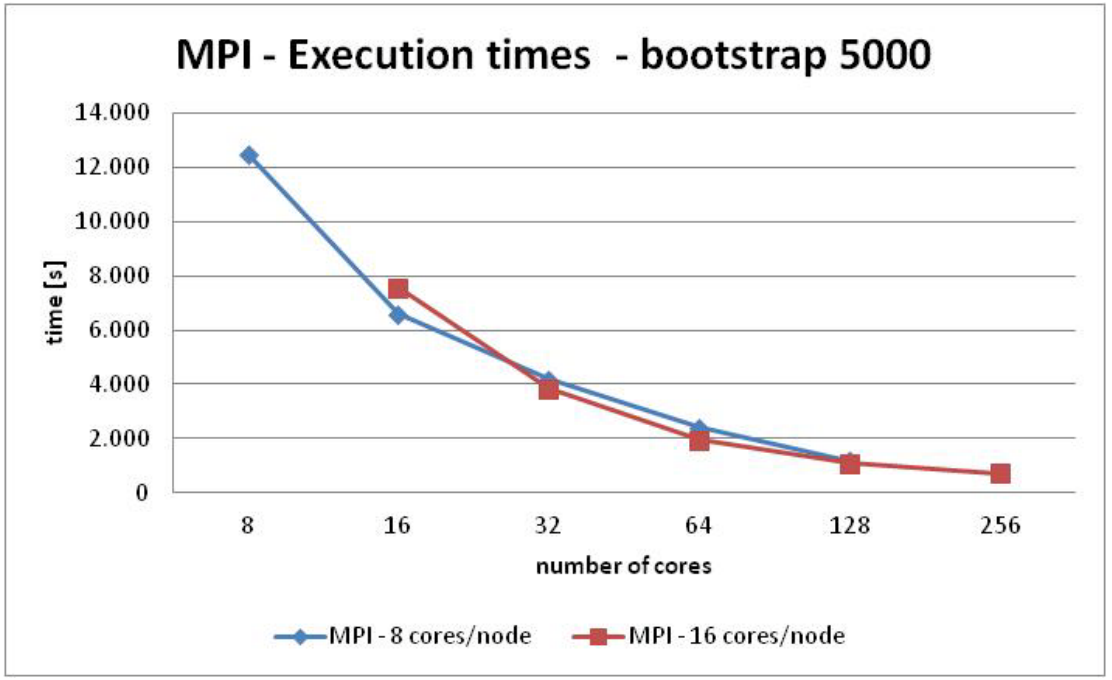
- Tested at
  - HPCG – Bulgaria
  - Debrecen & Pecs – Hungary
- Up to 256 cores
- Testing for one to five genes simultaneously

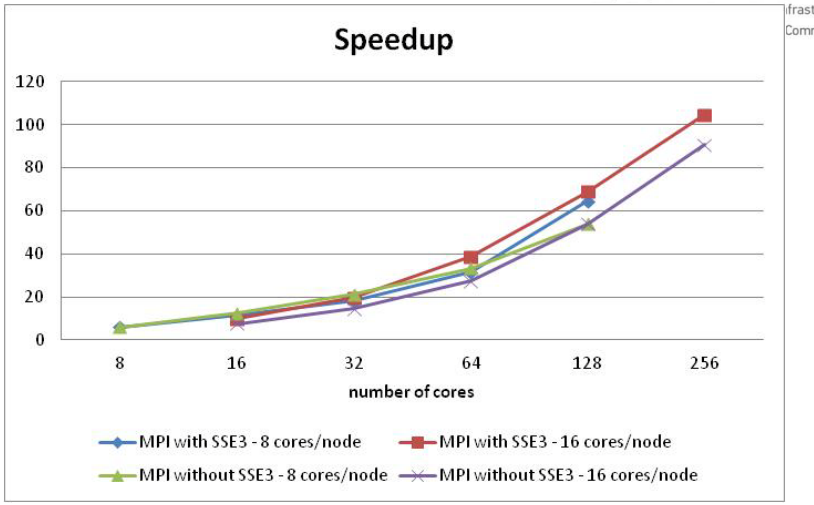
Analysis of one gene

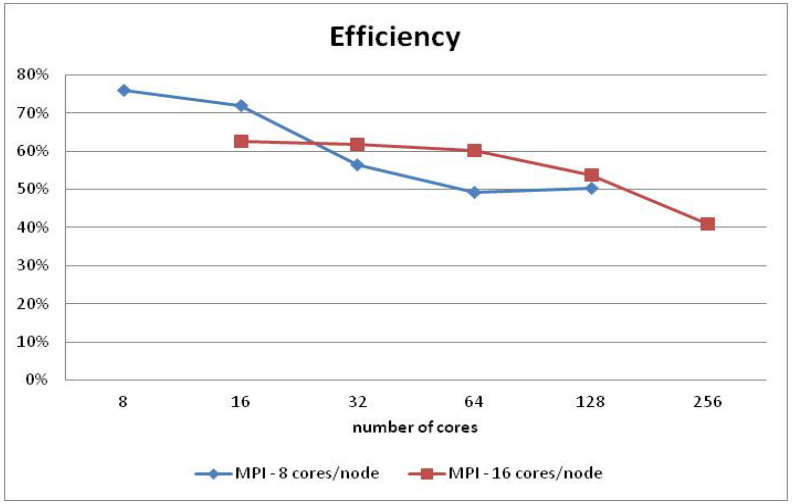
HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- 123 different DNA sequences of Salmo trutta (Linnaeus, 1758) from Eurasian geographical region,
- 552 base pairs per DNA sequence
- Source :
  - genebank;
  - UoM – Faculty of natural sciences, biology department
- Actions before RAxML analysis : multiple DNA alignment

Analysis of one gene Scalability results

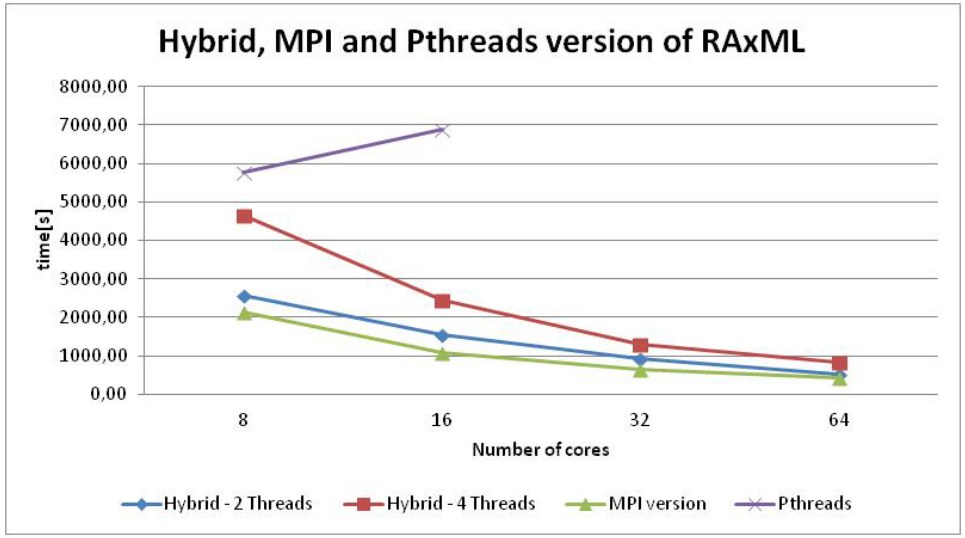


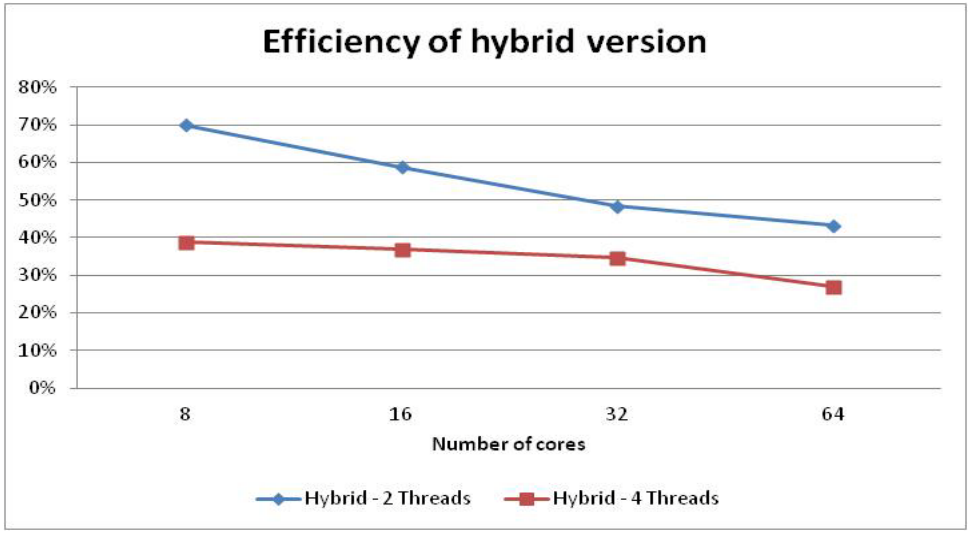




MPI results : execution time, speedup and efficiency for 5000 bootstraps

Analysis of one gene Scalability results

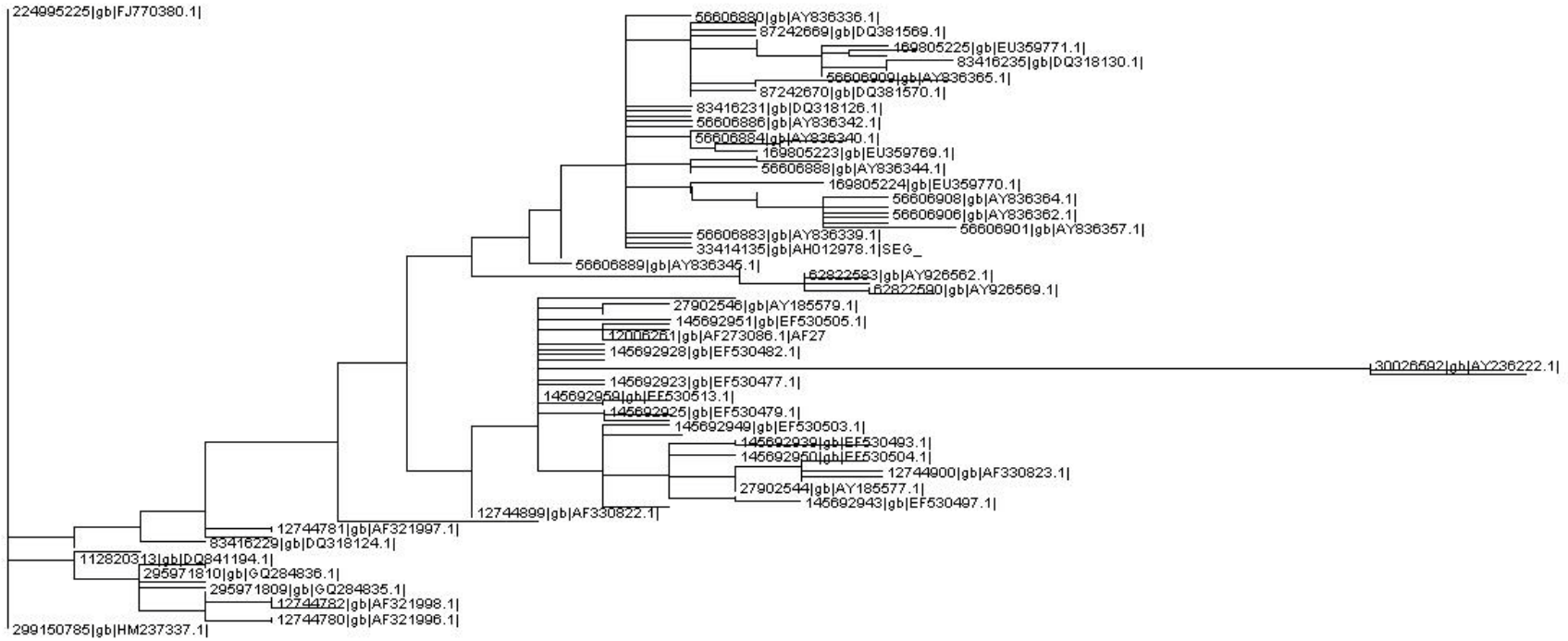




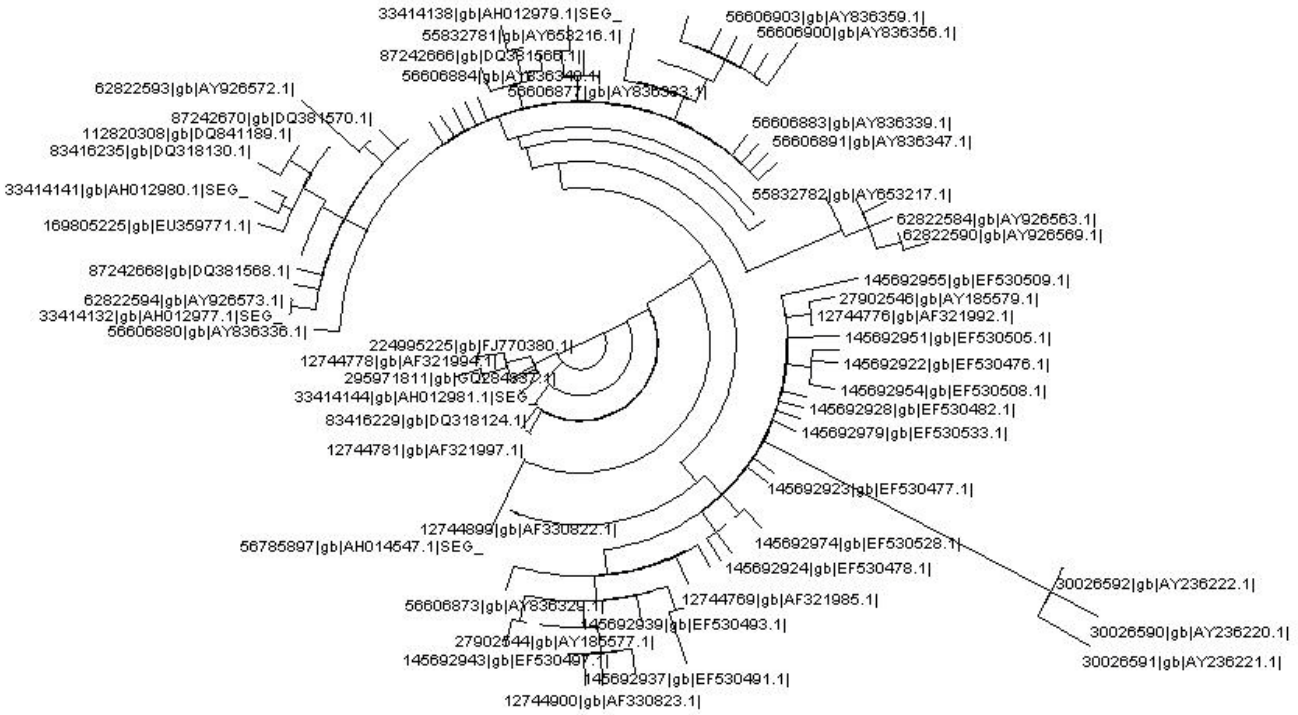
MPI, Pthreads and Hybrid execution time and efficiency

Tree for 123 DNA sequences

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities



Tree for 123 DNA sequences

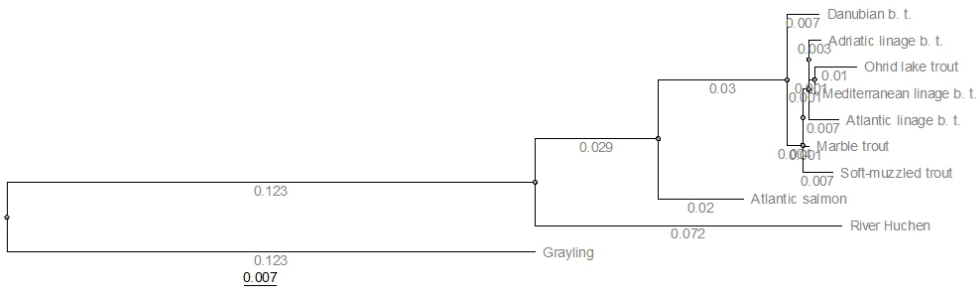


Multi-gene analysis

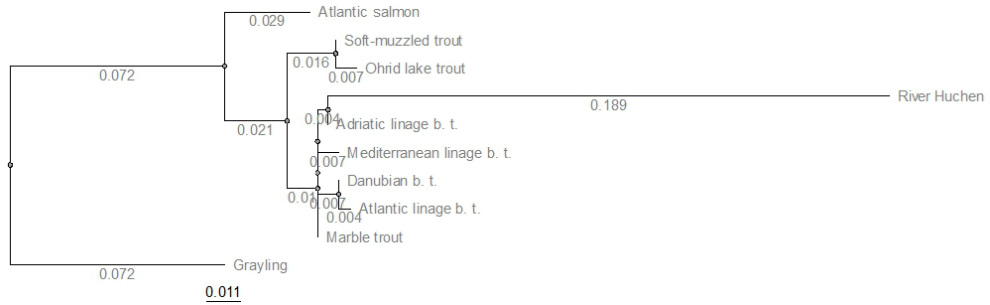
HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- Phylogeny analysis of
  - Mediterranean, Adriatic, Danubian and Atlantic linage brown trouts,
  - Marbled, soft-muzzled, Ohrid lake trouts,
  - Atlantic salmon
  - River Huchen
  - Grayling (outgroup)
- 5 genes
  1. Cytochrome b – 744 base pairs
  2. D-loop – 277 base pairs
  3. Modified Cytochrome b
  4. modified D-loop
  5. Hybrid (fake) gene -

Multi-gene analysis
Real genes - 1st and 2nd gene

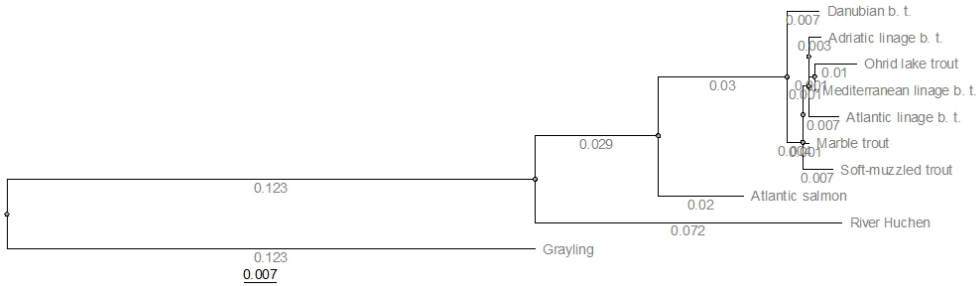


Cytochrome b

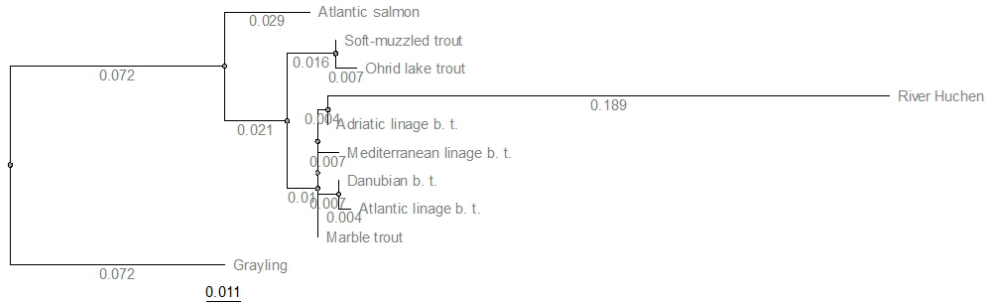


D-loop

Multi-gene analysis
Modified genes – 3rd & 4th gene

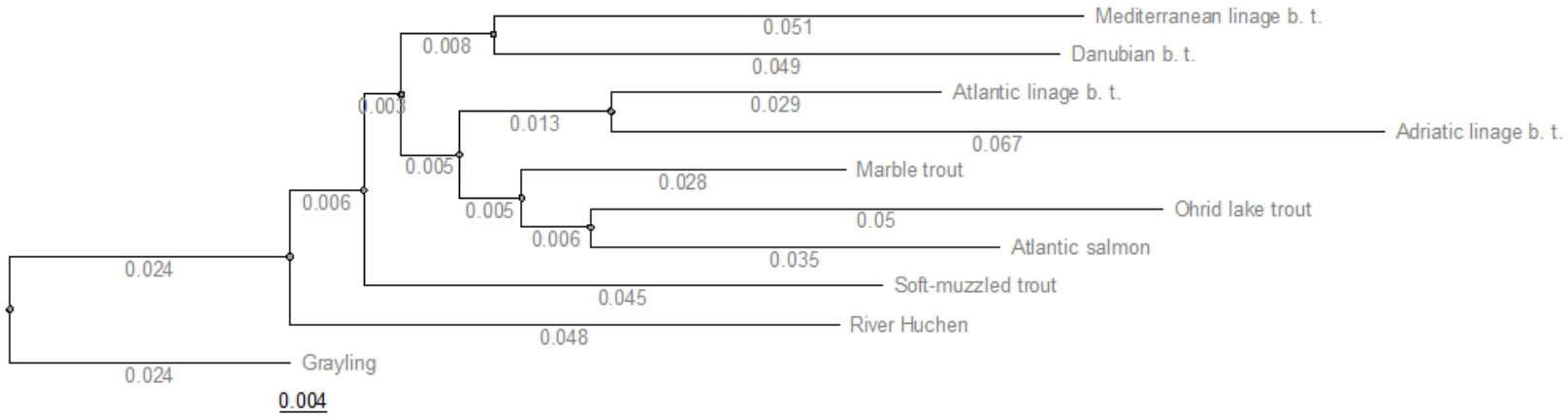


Modified Cytochrome b



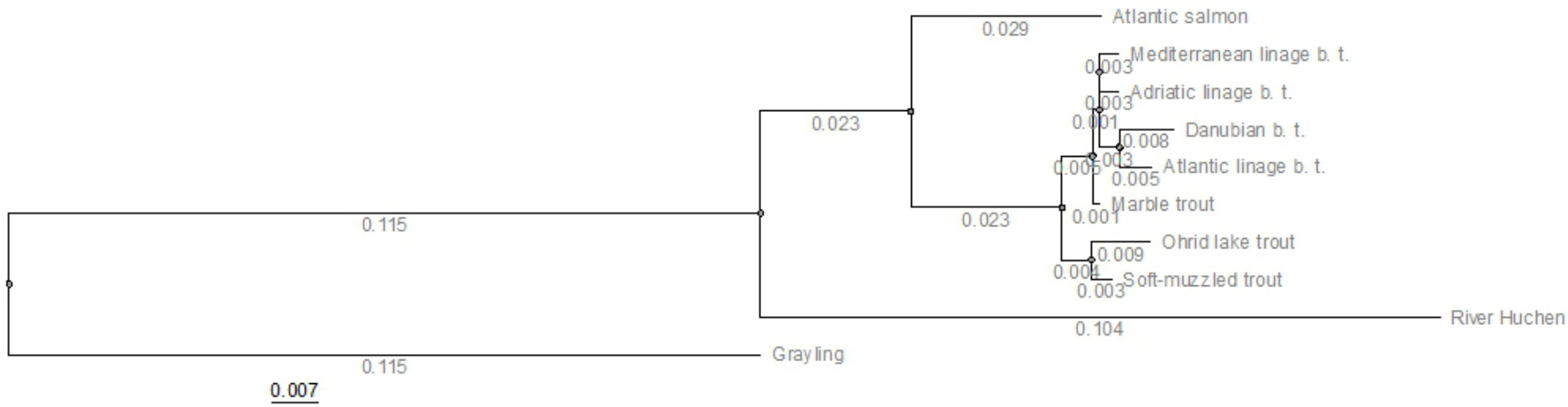
Modified D-loop

Multi-gene analysis 5th gene



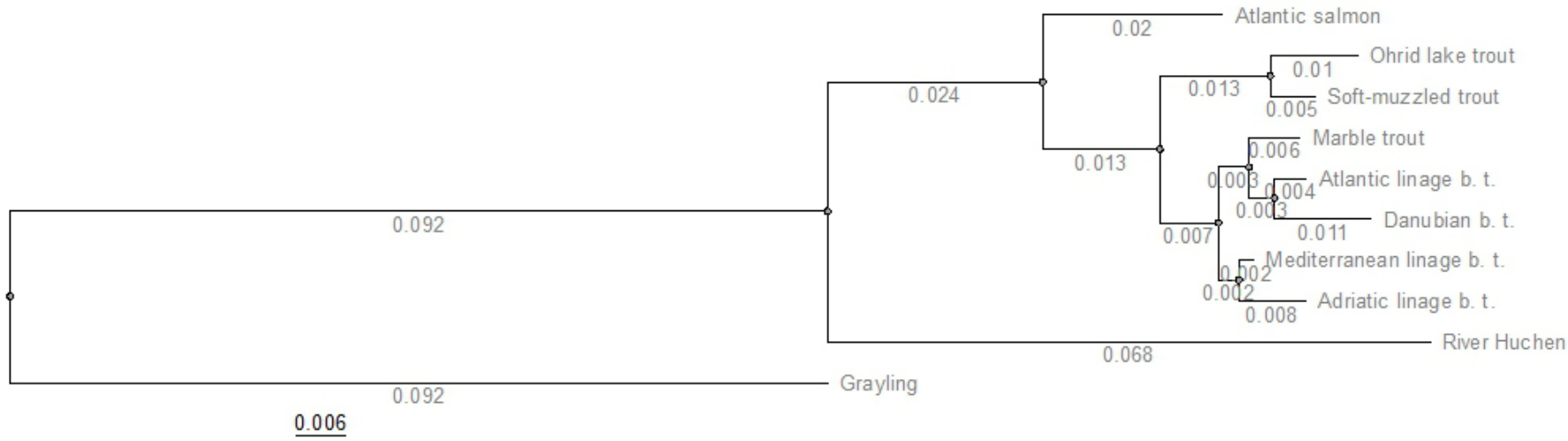
Hybrid (fake) gene
with bigger evolution distance between trouts

Multi-gene analysis



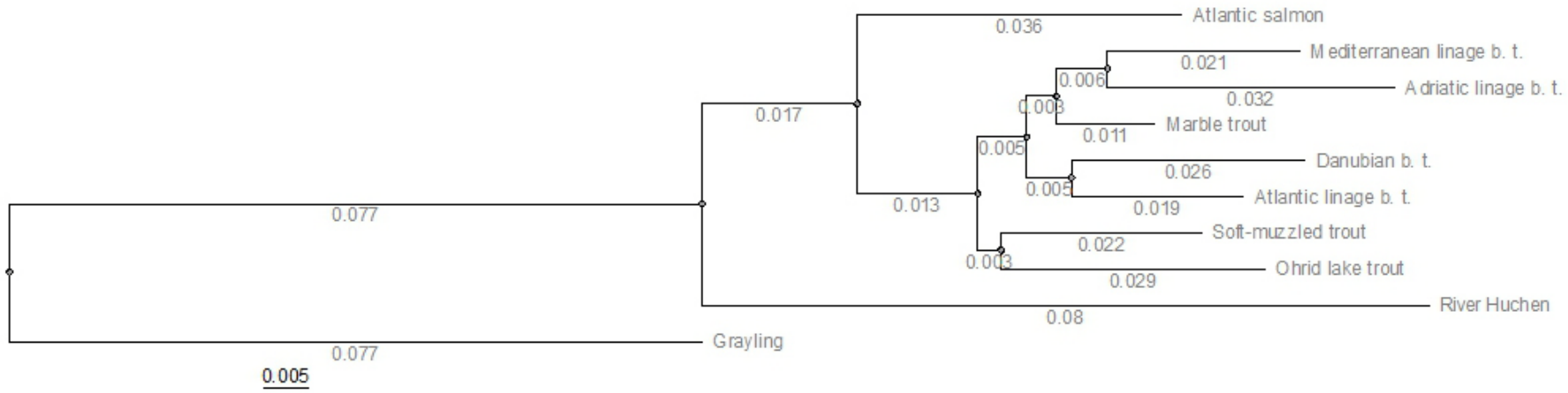
- Cytochrome b + D-loop

Multi-gene analysis



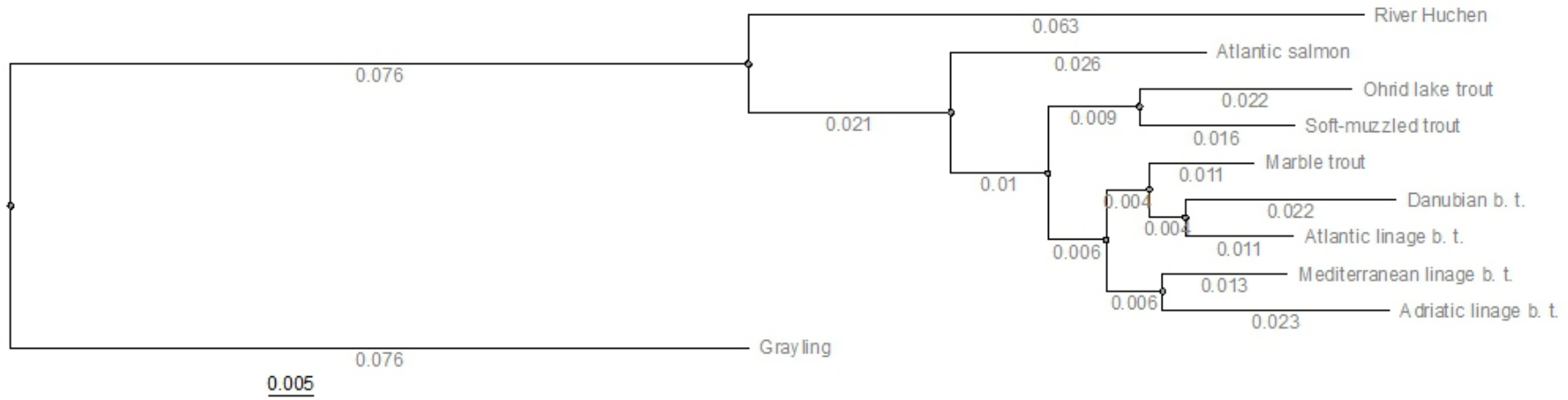
- Cytochrome b + D-loop + modified Cytochrome b + modified D-loop

Multi-gene analysis



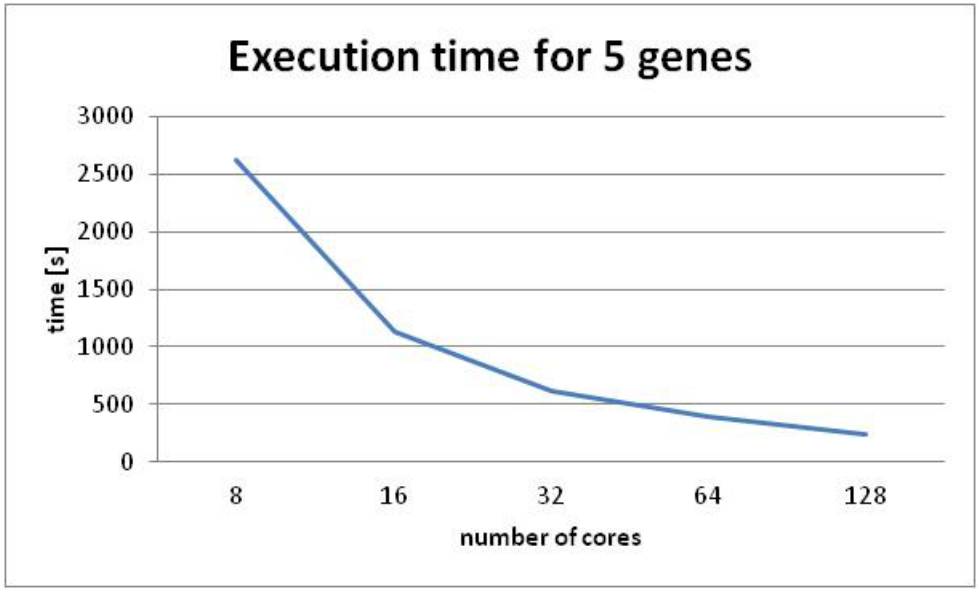
- Cytochrome b + D-loop + fake gene
- Phylogenetic information is correct, based on gene phylogeny 1 and 2 is not lost
- Gene 5 has influence, but no significant, on census tree

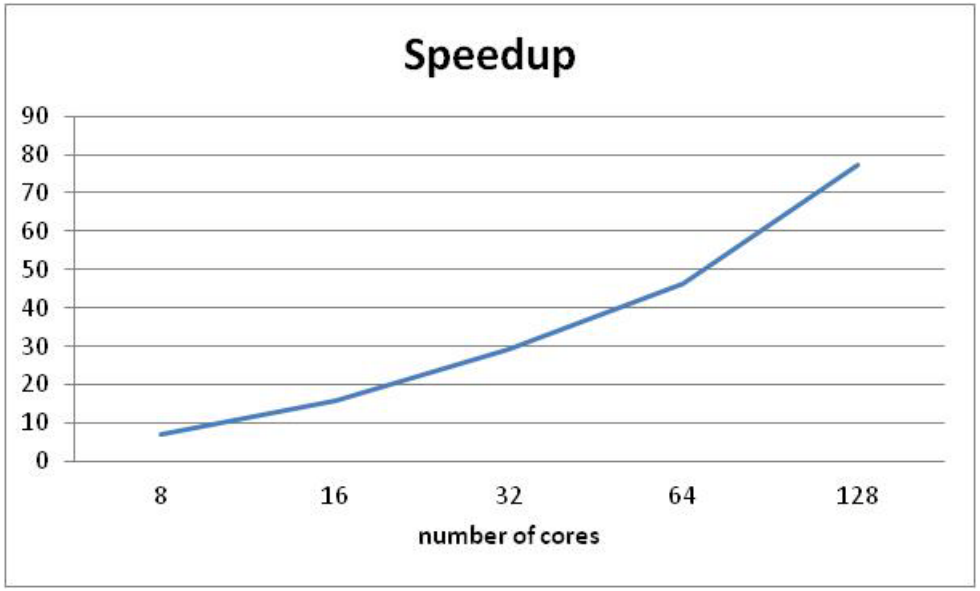
Multi-gene analysis



- All five genes
- 3025 base pairs divided in 5 genes
- Phylogenetic information is not lost, but the distance between the species have changed

Multi-gene analysis Scalability graphs





Execution time and speedup for bootstrap 10000

Conclusion

HP-SEE
High-Performance Computing Infrastructure for South East Europe's Research Communities

- We analyzed :
  - Scalability of RAxML
  - Single gene analysis on large dataset of spieces
  - Impact of muted and fake genes on correct phylogenetic information