

Treerecs with Seaview: gene tree inference from alignment to reconciliation, with a graphical interface

Eric Tannier,
INRIA, LBBE, University of Lyon

Phylogenetic software workshop, Montpellier
August 17, 2018



Phylogenetic reconciliation

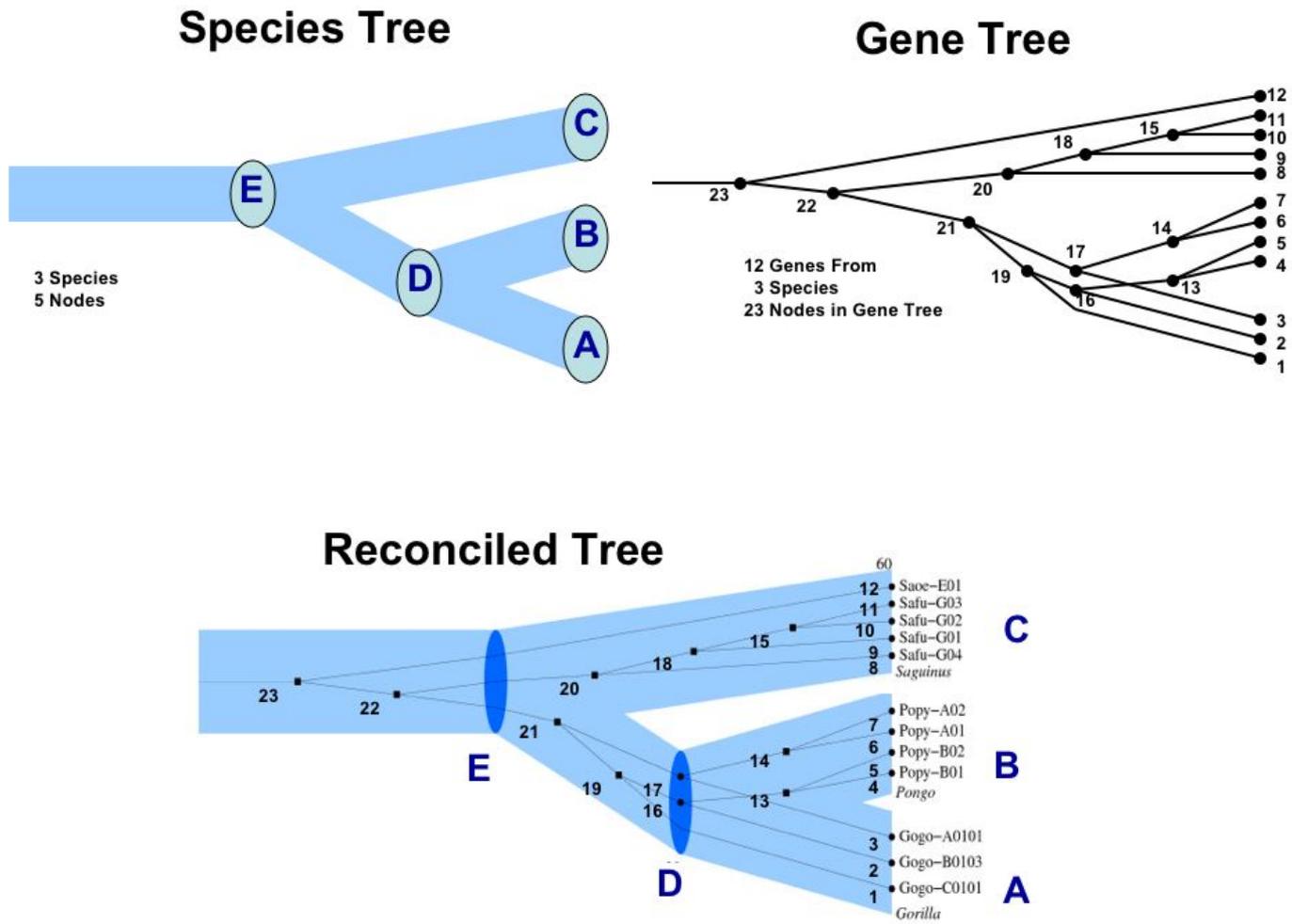
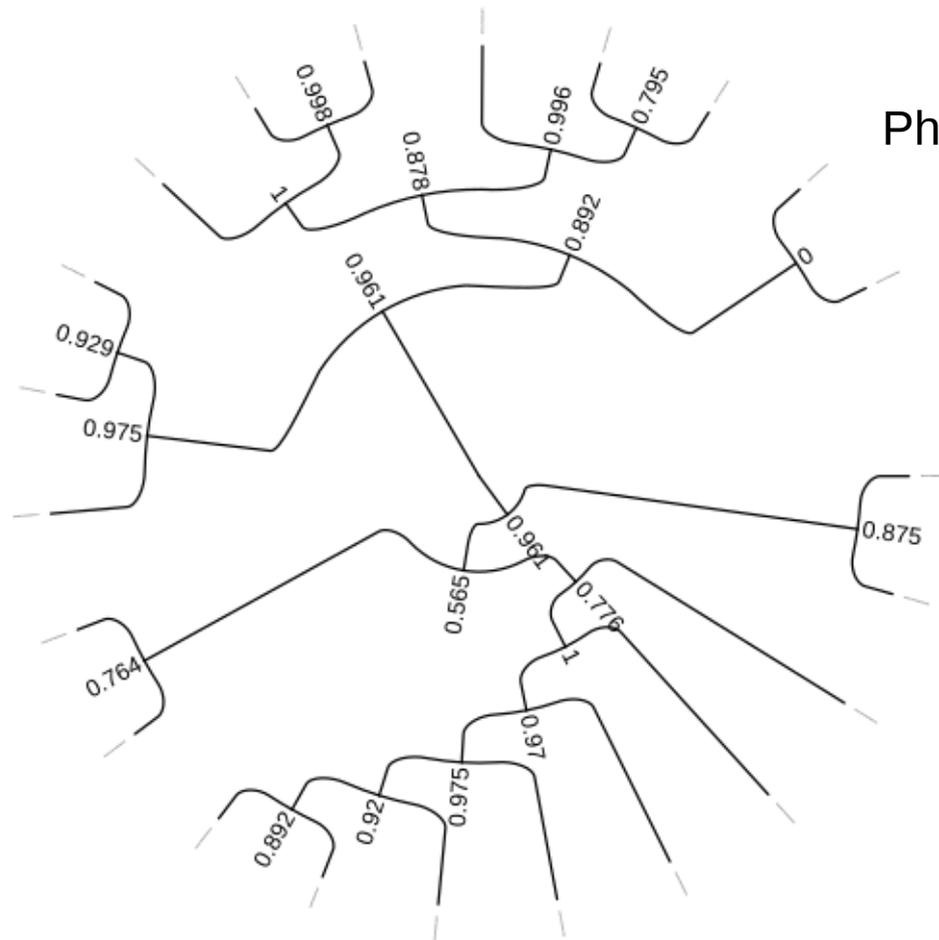


Fig from IpTOL, 2011

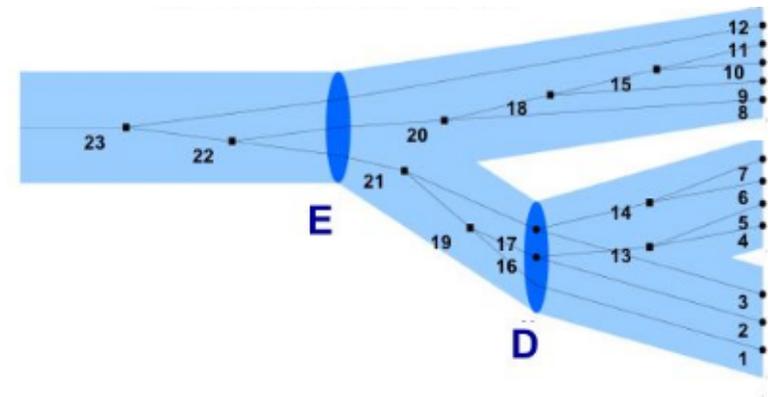
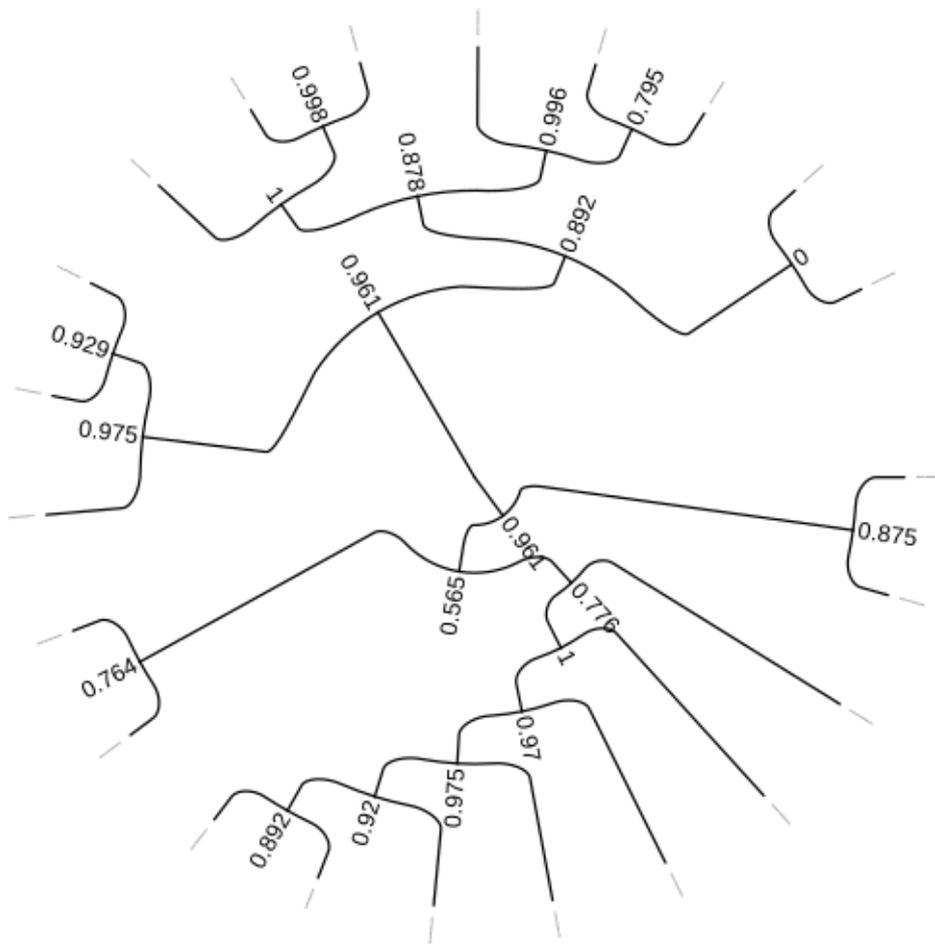
Typical output of a phylogenetic inference method from multiple sequence alignment



PhyML, RAxML, IQTree, FastTree,...

Unrooted, several uncertain branches

Different equivalent solutions (topology and root) according to the alignment might have a different reconciliation score (different D,T,L scenarios)



A good offer of phylogenetic reconciliation tools

ecceTera, Notung, Ranger-DTL, Prime-GSR, ALE, ...

- Usability is not optimal:

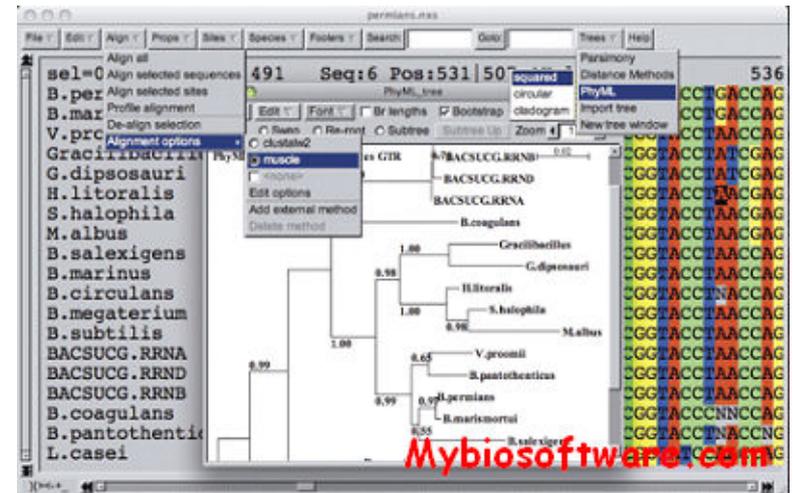
Notung has a nice graphical interface, but no integrative tool like **Seaview**

ecceTera is very efficient but dependent on several libraries

Formats are very strict (gene-species mapping)

- Basic functions are not always implemented

None of these software can easily and efficiently correct and root a gene tree





Basic functioning

Input:

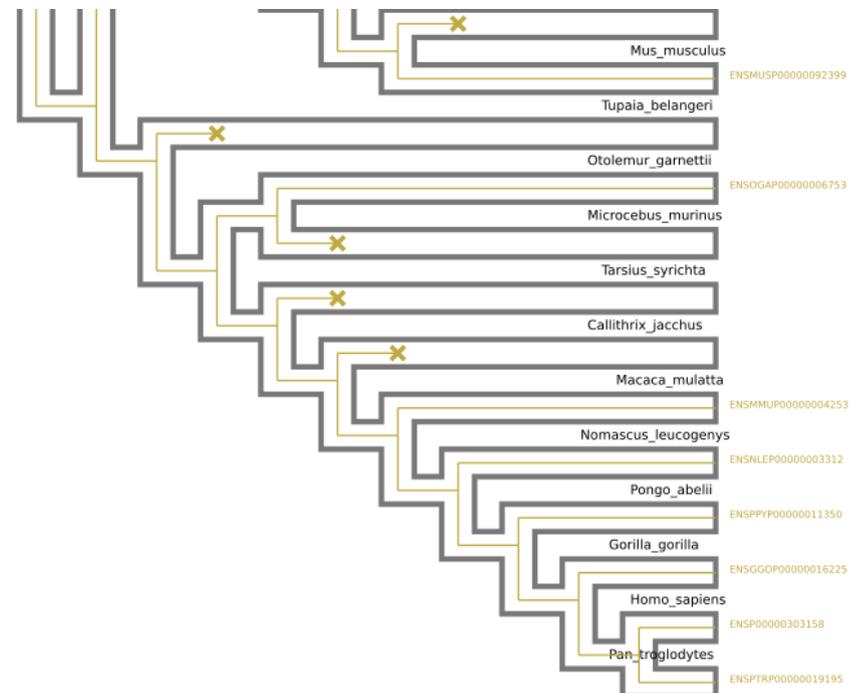
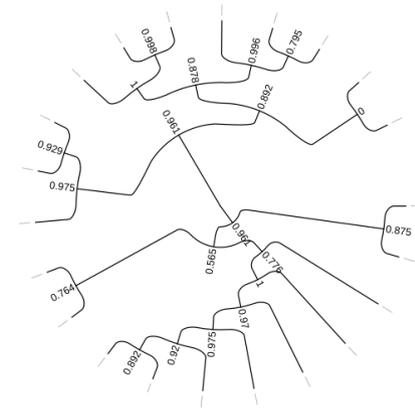
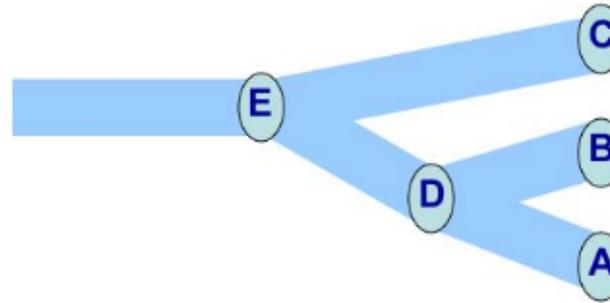
unrooted gene tree,
species tree,
threshold for supports
(can be estimated),
duplication and loss scores
(by default 2 and 1, can be estimated)

Output:

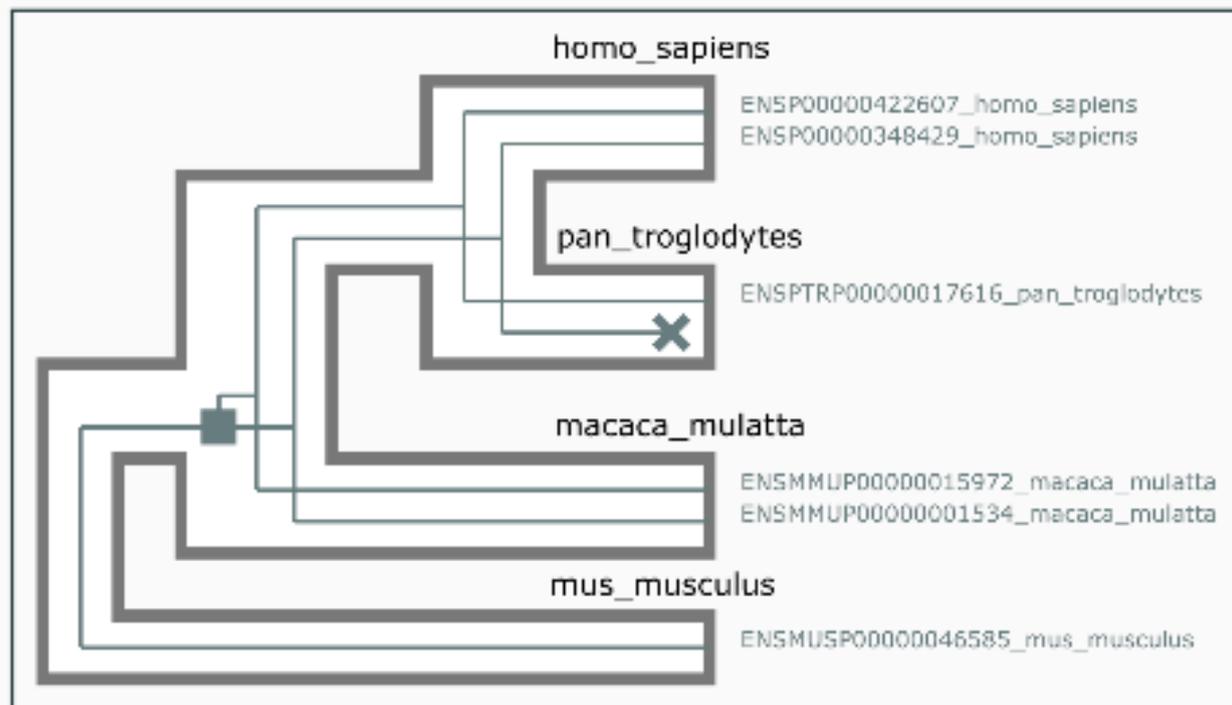
rooted gene tree and reconciliation,

keeping all supported branches (above threshold)

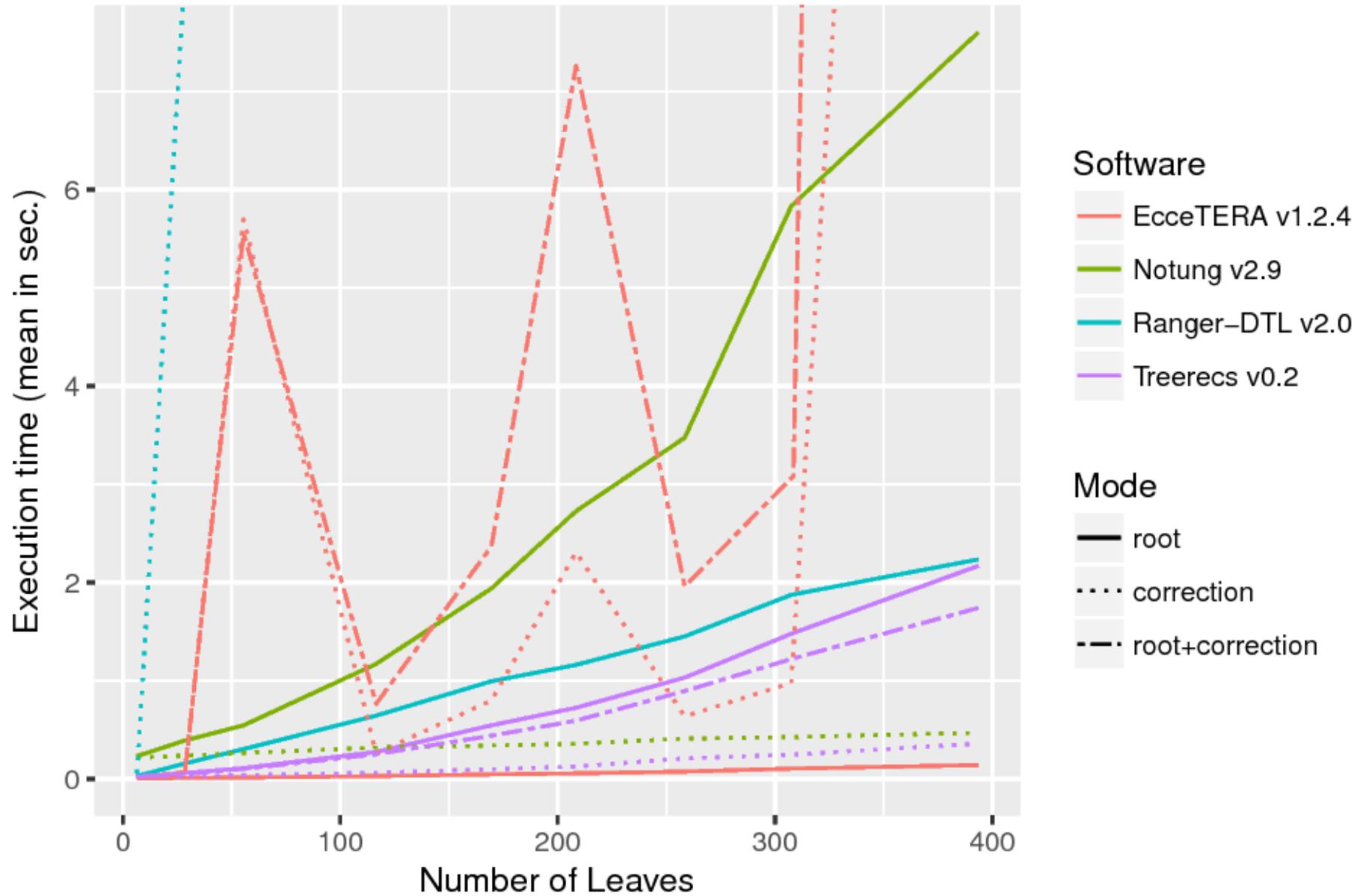
minimizing the duplication-loss score
(among all such rooted gene trees)



SVG output of a reconciled tree

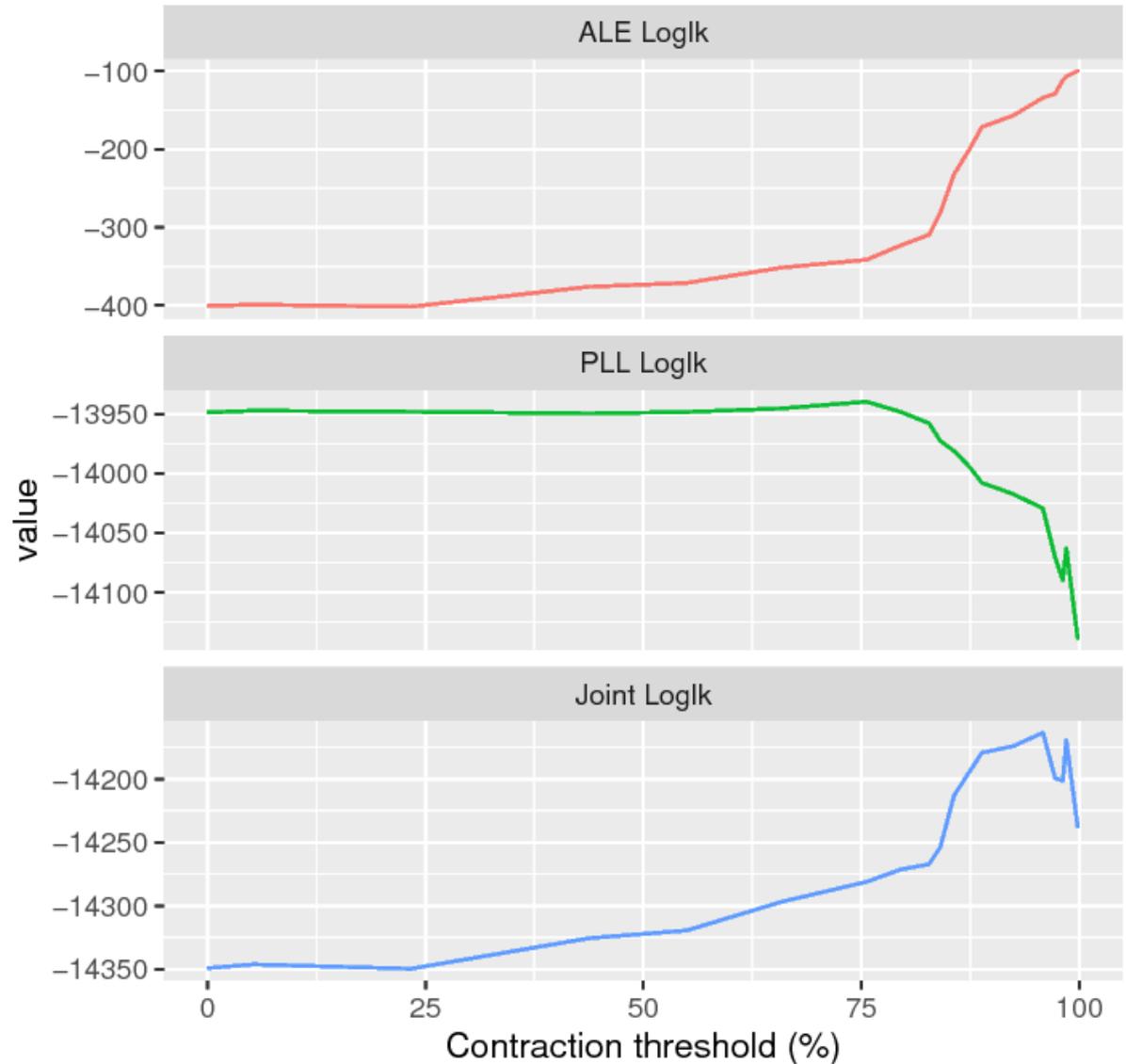


Efficiency



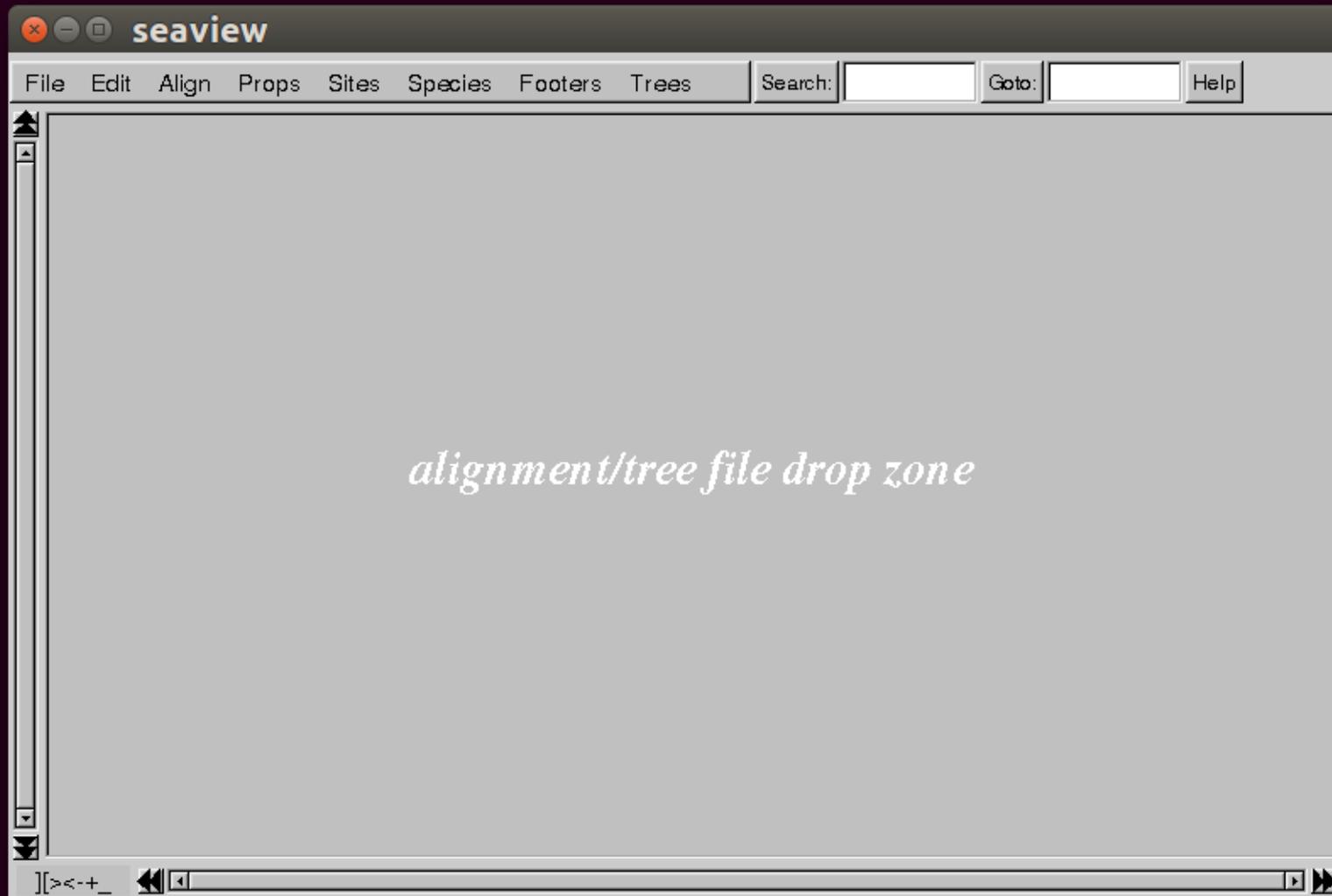
Integration with phylogeny based on multiple sequence alignment

Treerecs can also take a multiple sequence alignment of genes or proteins as input



Support threshold estimation with integrated likelihood

Integration with Seaview



Integration with Seaview

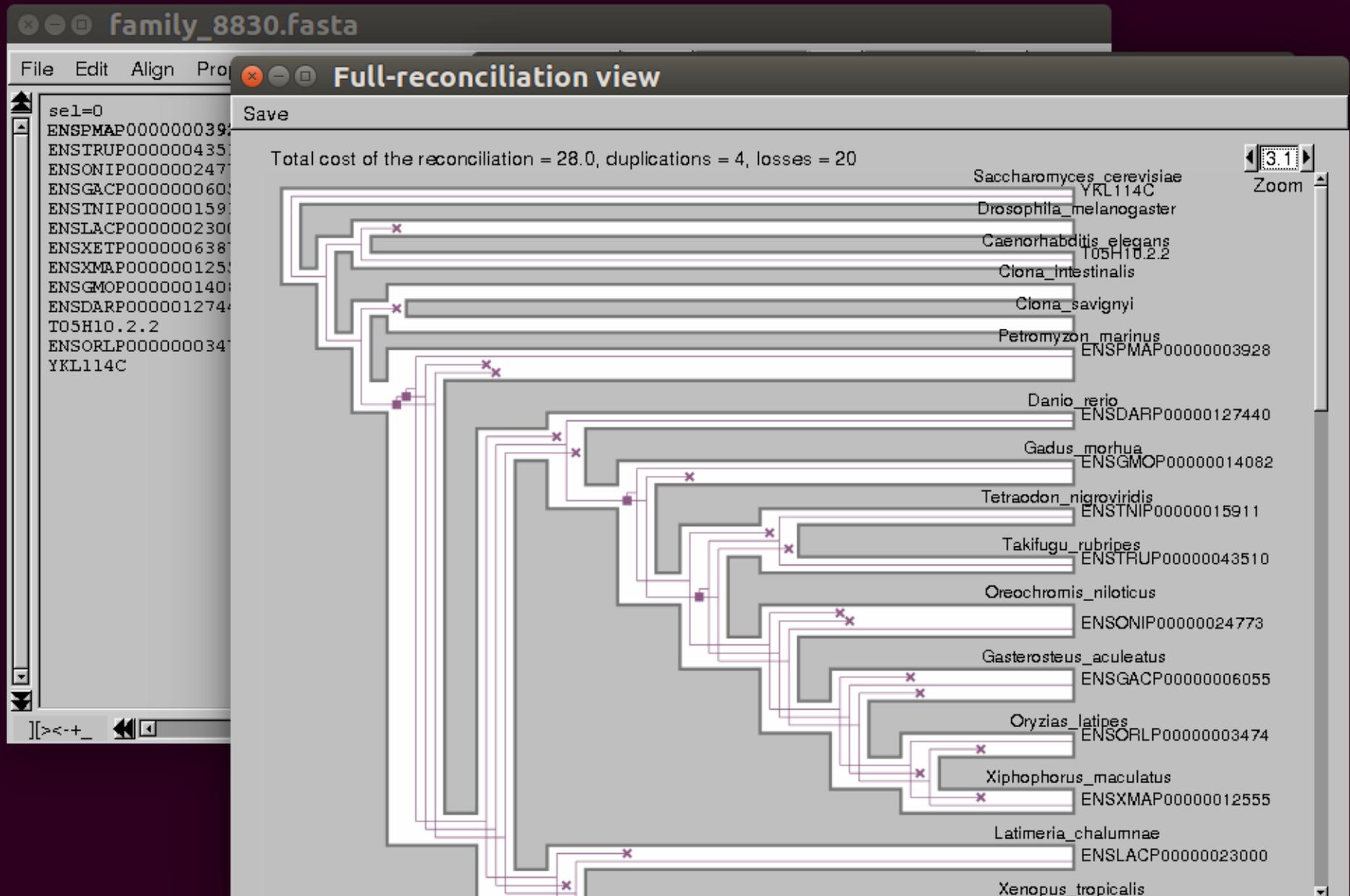
```
family_8830.fasta
File Edit Align Props Sites Species Footers Trees Search: Goto: Help
sel=0 271 344
ENSPMAP00000003928 VIENMCGGHTVGGQFEELRGIIDLVDQSRIGVCLDTCHAFAA-GYDVSSQEGVQHTLDEFHRIVGLKFLKAV
ENSTRUP00000043510 VLENMSGGGSIVGGRFSELRSIIDKVRDQSRVGVCLDTCHAFAA-GYDLAAEGGVKAMLDQFDQDVGLRYLRAI
ENSONIP00000024773 VLENMSGGGSIVGGKFCELSIIDRVRDQTRVGVCLDTCHAFAA-GYDLAAEGGVKAMLDQFDQEVGLQYLKAI
ENSGACP00000006055 VLENMSGGGSIVGGKFSELRSIIDKVRDQTRVGVCLDTCHAFAAEGYDVAAEGGVKAMLDQFEQEVGLQYLKAL
ENSTNIP00000015911 VLENMCGGATVGGRFSELRSIIDKVRDQSRVGVCLDTCHAFAA-GYDLAAEGGVKAMLDQFDQEVGLHLYKAV
ENSLACP00000023000 VIENMSCGNTVGGRFQELRGIIDGVEDKSRVGVCLDTCHAFAA-GYNLSTEDGLSQMLEEFSVVGLOYLKAV
ENSXETP00000063870 VLENMSCGGSIVGGRFSELRSIIDRVRDRSRVGVCLDTCHAFAA-GHDLSSKAGLEHMLDEFNKVVGLSFLKAI
ENSXMAP00000012555 VLENMSGGGSIVGGRFCELRRIIDRVRDRSRVGVCLDTCHAFAA-GHDLAADGGVAAMLQFDQEVGLRYLRAV
ENSGMOP00000014082 VLENMCGGHTVGGQFSELRSIIERVQDQSRVGVCLDTCHAFAA-GYDLAAVGGVSAMLDQFDTEVGLHLYRAV
ENSDARP000000127440 VLENMSGGGSIVGGQFSELKGIIDRVRDRSRVGVCLDTCHAFAA-GYDISPPGGVNNMLDEFDRVVGLHLYRAV
T05H10.2.2 VLEITMAGGNSIGGTFFELKFIIDKVKVKS RVGVCIDTCHIFAG-GYDIRTQKAYEEVMKNFGEVVGWNYLKAI
ENSORLP00000003474 VLENMSGGGSIVGGKFSELRSIIDRVRDQSRVGVCLDTCHAFSA-GYDLAAEGGVKAMLDQFDQEVGRHLYKAV
YKL114C VLENMAGTGNLVGSSSLVDLKEVIGMIEDKSRIGVCIDTCHTFAA-GYDISSTTETFNNFWKEFNDVIGFKYLSAV
```


Integration with Seaview

The image displays the Seaview software interface with three overlapping windows:

- family_8830.fasta**: A sequence alignment window showing a multiple sequence alignment of protein sequences. The alignment is color-coded by amino acid type. The first sequence is ENSPMAP00000003928 and the last is YKL114C. The alignment length is 271 residues.
- family_8830-PhyML_tree**: A window showing a phylogenetic tree. The tree is rooted and has a scale bar of 0.2. The tree is labeled with the same sequence IDs as in the alignment window. The tree is currently collapsed, showing a single branch with a support value of 0.98.
- Treerecs configuration**: A dialog box for configuring the Treerecs tool. It includes the following settings:
 - Treerecs global settings**:
 - Species tree: a.73.species.tree (Select)
 - Branch support threshold: -1.0
 - Duplication cost: 2.0
 - Loss cost: 1.0
 - Find best root:
 - Gene <> Species mapping method**:
 - Use gene names: Use file:
 - Auto Separator: Species before:
 - Map file: 73.protein.nhx.emf (Select)
 - Output tree**:
 - Gene tree: Full reconciliation:

Integration with Seaview



Integration with Seaview

The image displays the Seaview software interface with three overlapping windows:

- family_8830.fasta**: A protein alignment window showing a multiple sequence alignment of 16 sequences. The alignment is color-coded by amino acid type. The first sequence is ENSPMAP00000003928 and the last is YKL114C.
- family_8830-PhyML_tree**: A window showing a PhyML tree with a log-likelihood value of -5808.7483. The tree is partially visible, showing a bootstrap value of 0.98 at a node. The tree includes labels for sequences like T05H10.2.2 and YKL114C.
- Treerecs configuration**: A dialog box for configuring the Treerecs tool. It includes the following settings:
 - Species tree: a.73.species.tree
 - Branch support threshold: 0.8
 - Duplication cost: 2.0
 - Loss cost: 1.0
 - Find best root: checked
 - Gene <> Species mapping method: Use file (selected)
 - Auto Separator: checked
 - Species before: checked
 - Map file: 73.protein.nhx.emf
 - Output tree: Full reconciliation (selected)

Integration with Seaview



Ongoing developments

- For the moment we have only duplications and losses

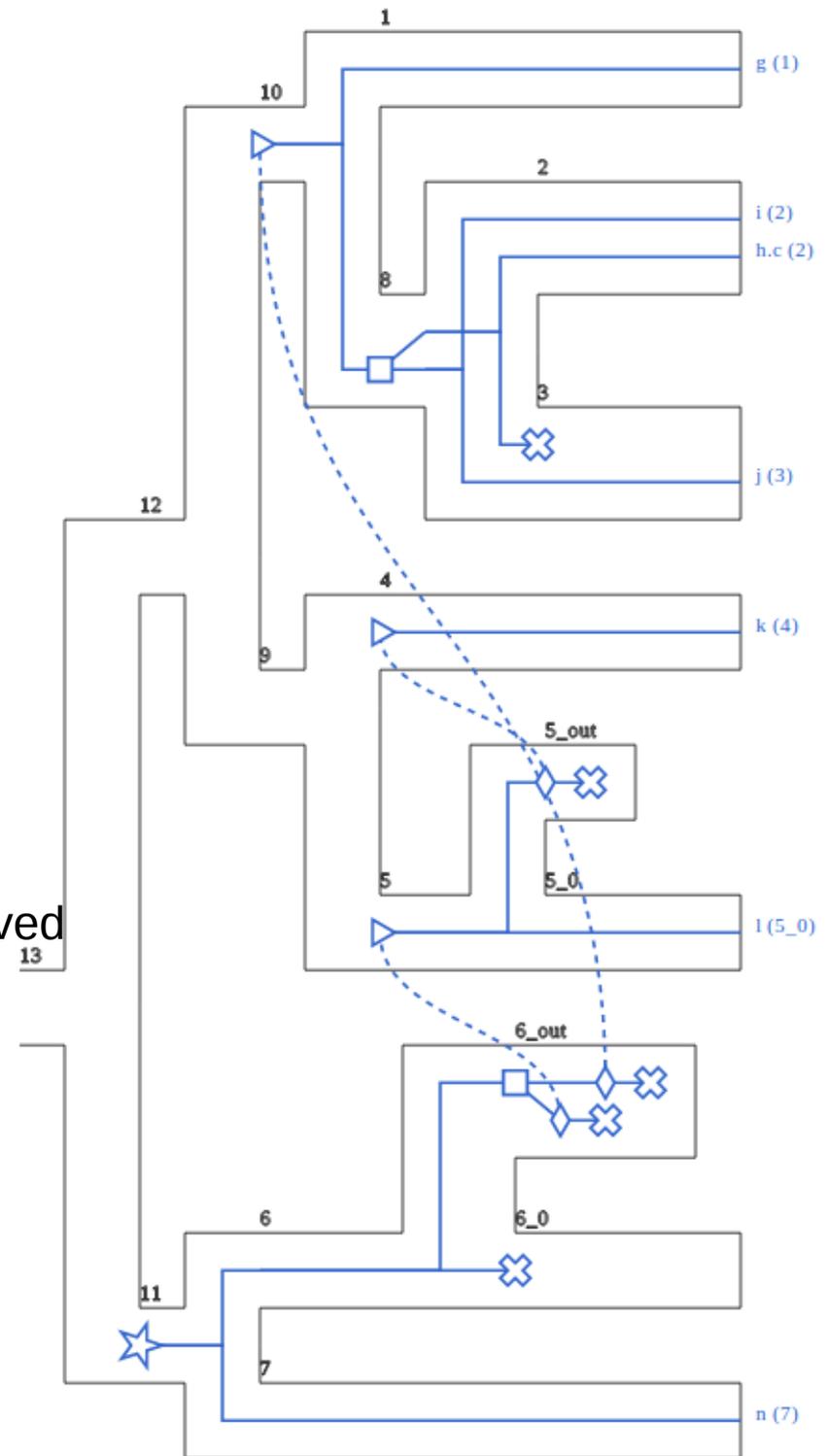
We will soon have transfers, based on ecceTera and ALE (short term)

We would like to have deep coalescence intergrated (long term)

- For the moment the species tree needs to be fully resolved

We will soon allow unresolved species tree (mid term)

We will soon allow no species tree at all, searching it in databases or reconstructing it from gene trees (long term)



With the participation of



Nicolas Comte
David Parsons



Benoît Morel
Alexandros Stamatakis



Vincent Daubin
Bastien Boussau
Celine Scornavacca
Manolo Gouy