

REVOLUTIONH-TL: A FAST AND ROBUST TOOL FOR DECODING EVOLUTIONARY GENE HISTORIES

Annachiara Korchmaros

joint work with K. A. Padilla, M. Hellmuth, A. J. Álvarez, M. H. Rosales, A. L. Sánchez, G. M. Medina, J. A. R. Rafael, P. F. Stadler (and other former collaborators).

Colloquium Combinatorics and Life Sciences

September 30, 2025



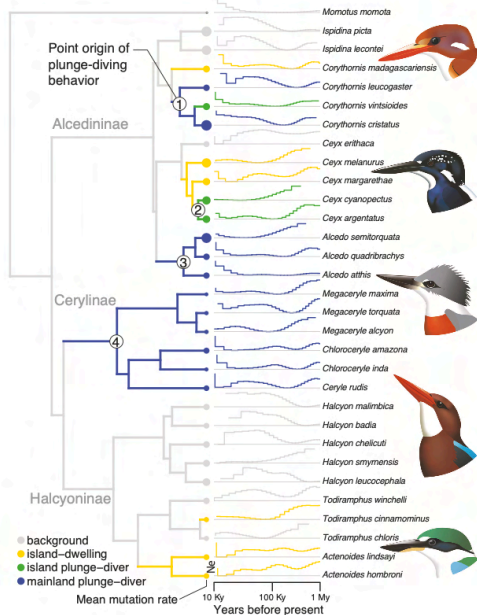
UNIVERSITÄT
LEIPZIG

BIO Leipzig
<INFORMATICS>

WHY DOES ESTIMATING TREE PHYLOGENIES MATTER?

WHY DOES ESTIMATING TREE PHYLOGENIES MATTER?

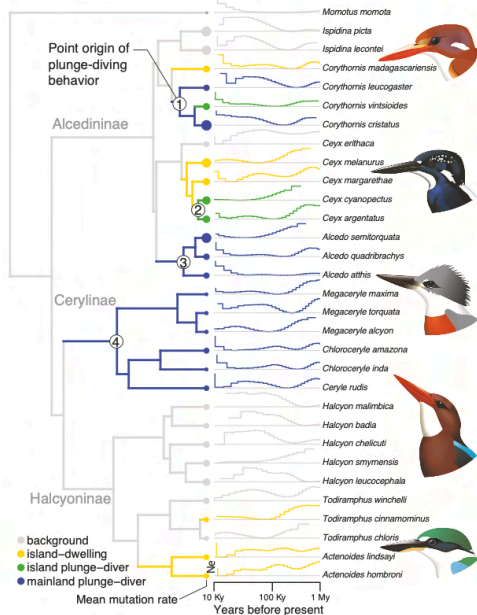
1. Localize trait/behavior change onto lineages.



(Eliason et al., 2023)

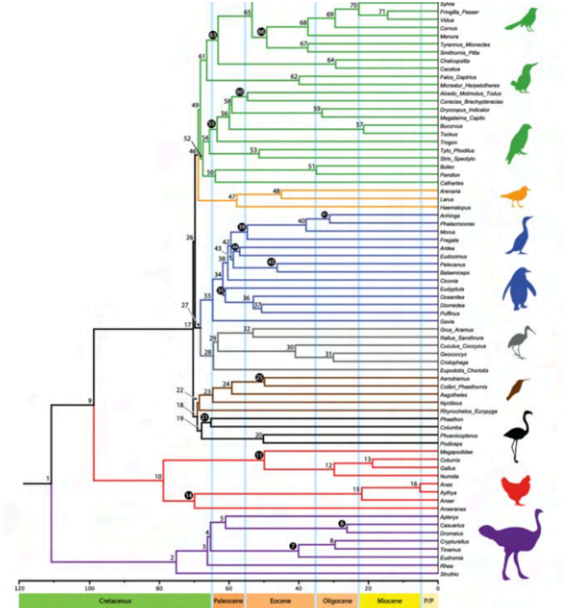
WHY DOES ESTIMATING TREE PHYLOGENIES MATTER?

1. Localize **trait/behavior change** onto lineages.



(Eliason et al., 2023)

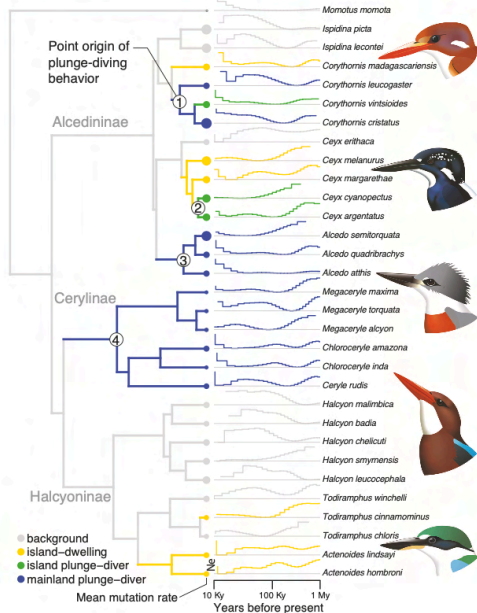
2. Map time divergence events.



(Ksepka et al, 2015)

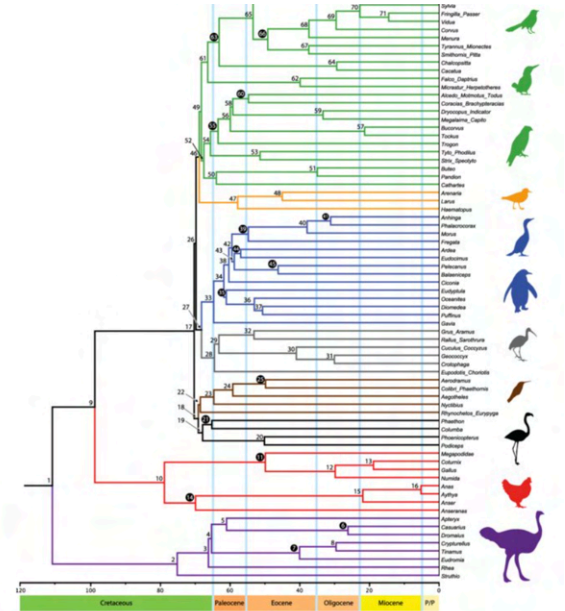
WHY DOES ESTIMATING TREE PHYLOGENIES MATTER?

1. Localize **trait/behavior change** onto lineages.



(Eliason et al., 2023)

2. Map **time divergence events**.



(Ksepka et al, 2015)

3. Help predict the **function** of newly discovered **genes**.



DL SCENARIOS & ORTHOLOGS

Gene ID

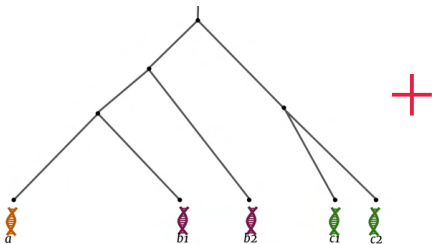
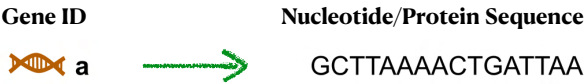
 a



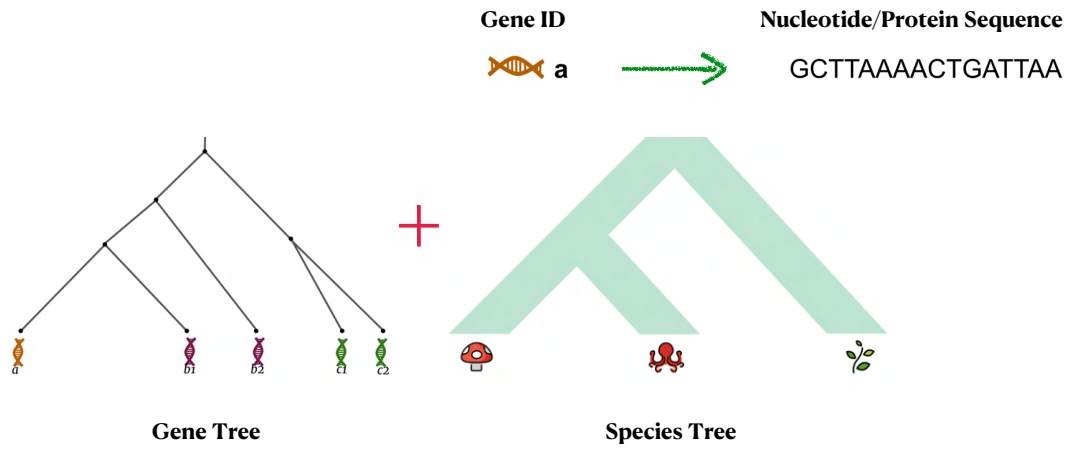
Nucleotide/Protein Sequence

GCTTAAACTGATTAA

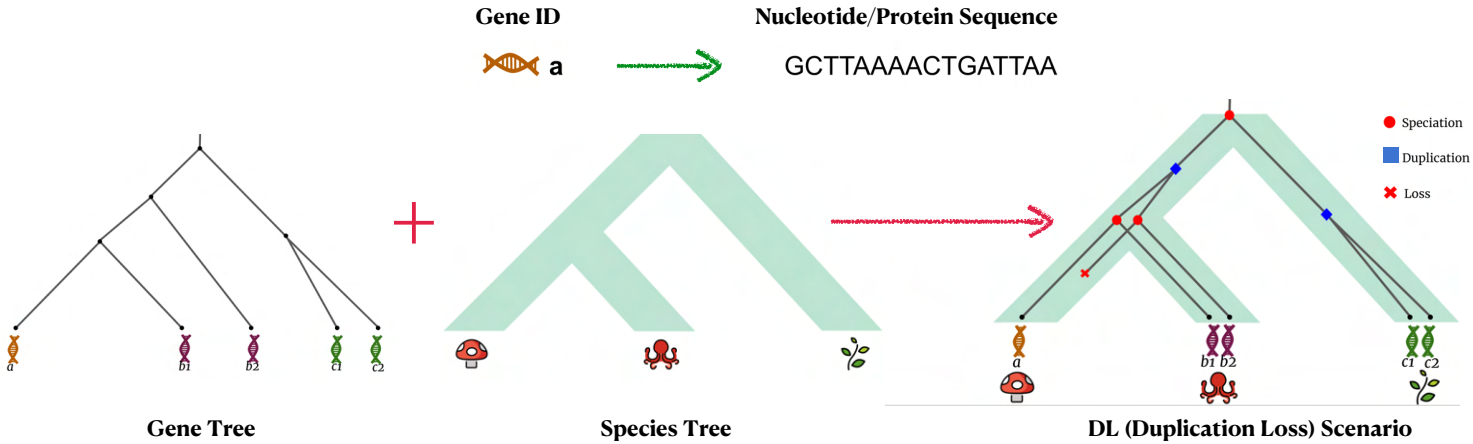
DL SCENARIOS & ORTHOLOGS



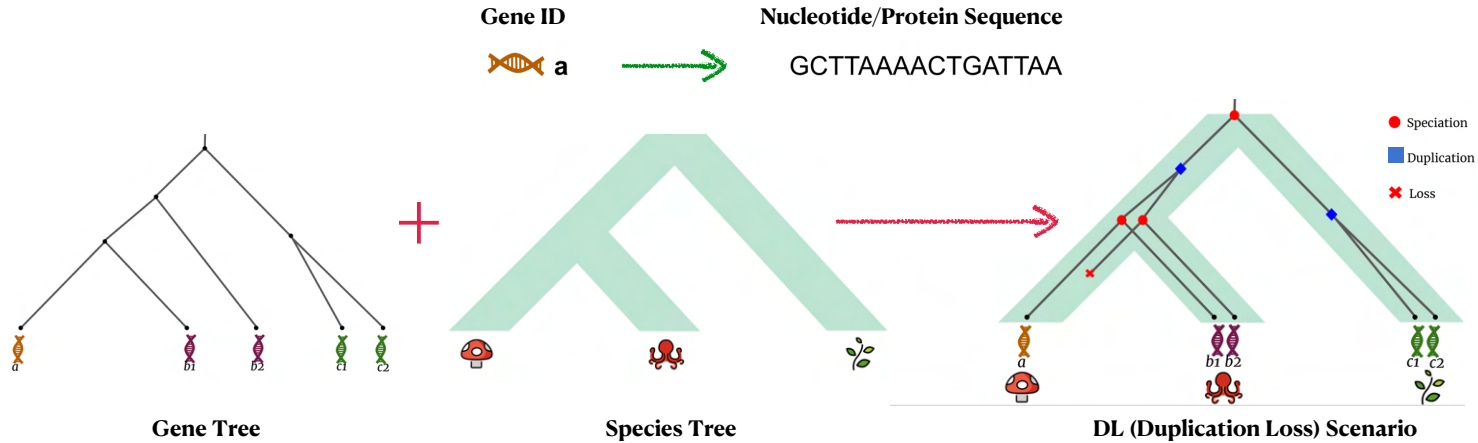
DL SCENARIOS & ORTHOLOGS



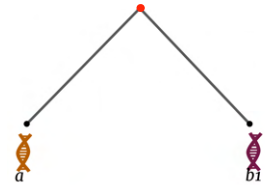
DL SCENARIOS & ORTHOLOGS



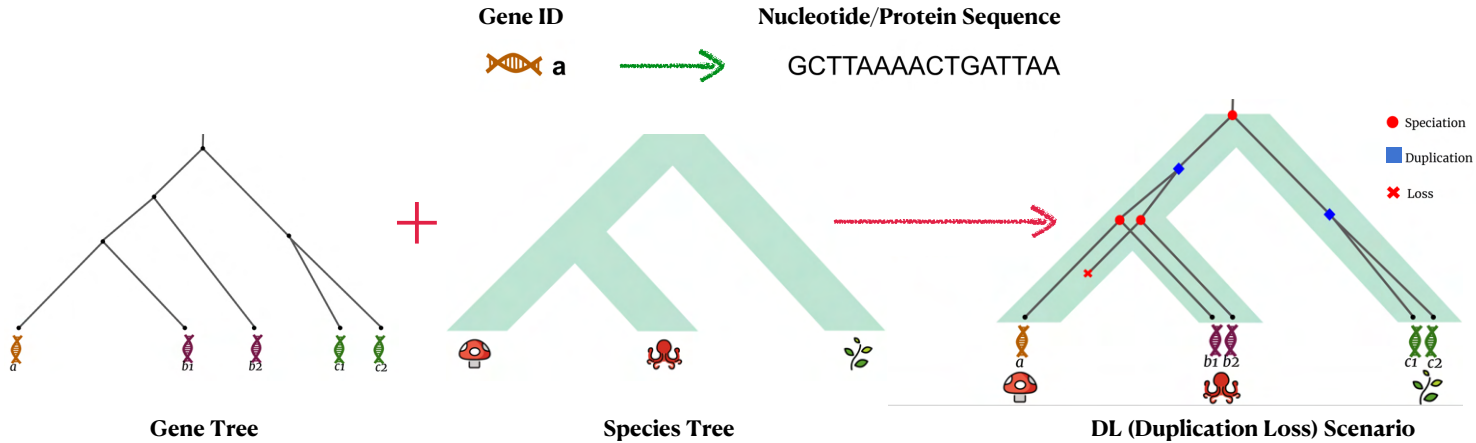
DL SCENARIOS & ORTHOLOGS



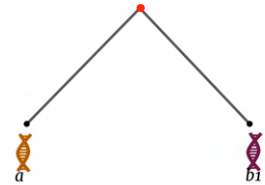
► **Orthologs:** genes from different species that split from one ancestral gene labelled as ●



DL SCENARIOS & ORTHOLOGS



- **Orthologs:** genes from different species that split from one ancestral gene labelled as ●
- **Orthogroup:** set of orthologs originating from a single last common ancestor gene.



THE PROBLEM

Species

Gene ID

Sequence



a

GCTTAAAACTGATTAACCTGCT



b1

GCTTCTGATTAACCTGCT



b2

GCTTCTGACTAGTCTGCT



c1

GCTTCTGATTTCTCCCTGCC



c2

GCTTCTGATTACTCCCTGCC

Sequences

?

Phylogenetic tree
& orthology
assignment

PHYLOGENY ESTIMATION METHODS

Maximum Likelihood Estimation

log likelihood:

$$\ell(\theta) = \ln f(D|\theta) = \ln \prod_{i=1}^N f(x_i|\theta) = \sum_{i=1}^N \ln f(x_i|\theta)$$

therefore, $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \ell(\theta)$

Bayesian Inference

posterior distribution

likelihood function

prior distribution

$$p(\theta|D) = \frac{p(D|\theta) p(\theta)}{\int p(D|\theta) p(\theta) d\theta}$$

PHYLOGENY ESTIMATION METHODS

Maximum Likelihood Estimation

log likelihood:

$$\ell(\theta) = \ln f(D|\theta) = \ln \prod_{i=1}^N f(x_i|\theta) = \sum_{i=1}^N \ln f(x_i|\theta)$$

therefore, $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \ell(\theta)$

computationally expensive
model-based



Bayesian Inference

posterior distribution

likelihood function

prior distribution

$$p(\theta|D) = \frac{p(D|\theta) p(\theta)}{\int p(D|\theta) p(\theta) d\theta}$$

widely used
branch length



PHYLOGENY ESTIMATION METHODS

Maximum Likelihood Estimation

log likelihood:

$$\ell(\theta) = \ln f(D|\theta) = \ln \prod_{i=1}^N f(x_i|\theta) = \sum_{i=1}^N \ln f(x_i|\theta)$$

therefore, $\hat{\theta} = \arg\max_{\theta} \ell(\theta)$

computationally expensive
model-based



Bayesian Inference

posterior distribution

likelihood function

prior distribution

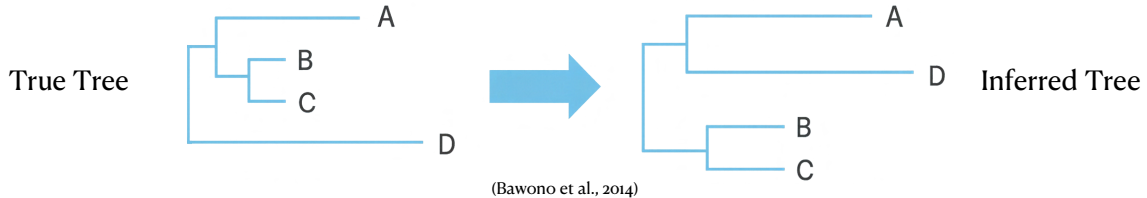
$$p(\theta|D) = \frac{p(D|\theta) p(\theta)}{\int p(D|\theta) p(\theta) d\theta}$$

(Ricketts, 2021)

widely used
branch length
not affected by LBA



Long Branch Attraction (LBA)



Maximum Parsimony



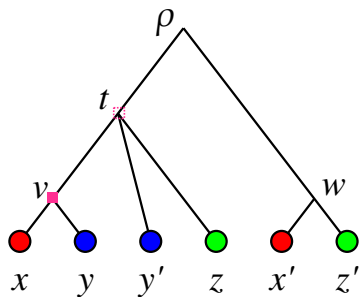
affected by LBA



not model-based (min. number of events)

BEST MATCH GRAPHS

- ▶ (T, σ) rooted gene tree, leaf coloring σ on leaf set $L(T)$
- ▶ $y \in L(T)$ is a **best match** of $x \in L(T)$ if
 1. $\sigma(x) \neq \sigma(y)$ and
 2. $\text{lca}(x, y) \preceq \text{lca}(x, z)$ for all $z \in L(T)$ with $\sigma(z) = \sigma(y)$

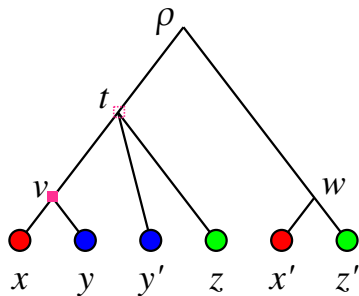


(T, σ)

v, t, w duplications/speciations

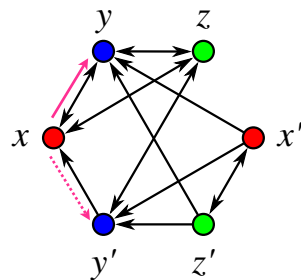
BEST MATCH GRAPHS

- ▶ (T, σ) rooted gene tree, leaf coloring σ on leaf set $L(T)$
- ▶ $y \in L(T)$ is a **best match** of $x \in L(T)$ if
 1. $\sigma(x) \neq \sigma(y)$ and
 2. $\text{lca}(x, y) \preceq \text{lca}(x, z)$ for all $z \in L(T)$ with $\sigma(z) = \sigma(y)$
- ▶ (G, σ) is **BMG** (T, σ) if vertices=leaves colored by σ and $x \rightarrow y$ iff y is a best match of x on (T, σ)



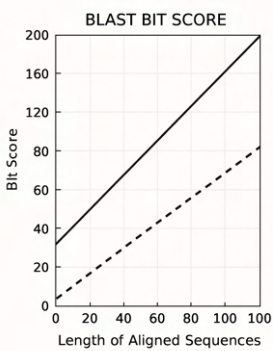
(T, σ)

v, t, w duplications/speciations



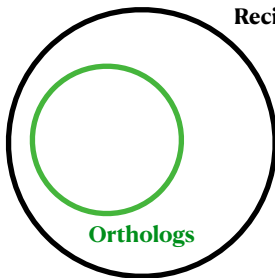
$\text{BMG}(T, \sigma)$

GRAPH-BASED ORTHOLOGY METHODS



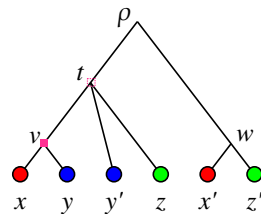
Orthologs Estimation

Bidirectional Best Hits

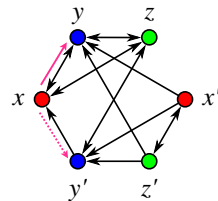


(Shaller et al., 2021)

Reciprocal Best Matches

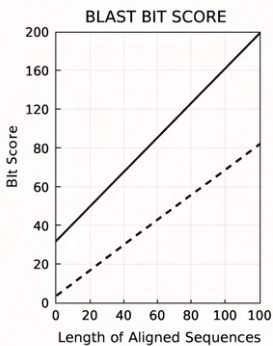


(T, σ)



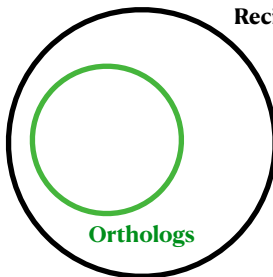
$\text{BMG}(T, \sigma)$

GRAPH-BASED ORTHOLOGY METHODS



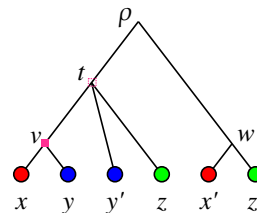
Orthologs Estimation

Bidirectional Best Hits

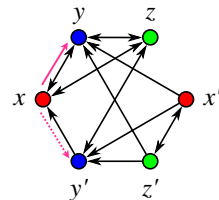


(Shaller et al., 2021)

Reciprocal Best Matches

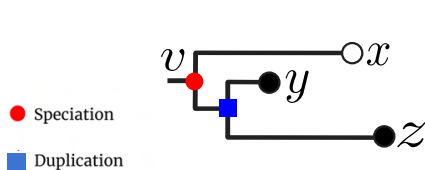


(T, σ)

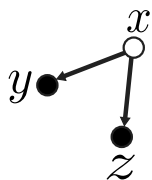


BMG(T, σ)

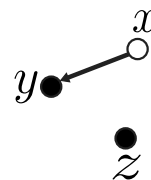
Problem: y, z co-orthologs of x with different evolutionary rates



(T, σ)

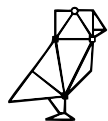
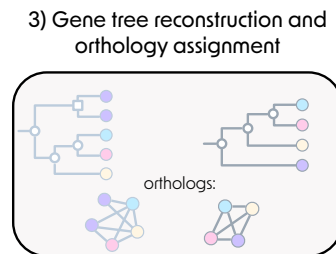
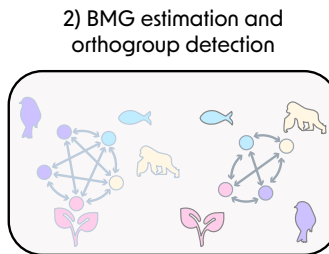
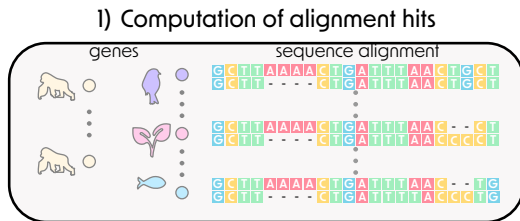
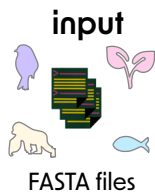


BMG(T, σ)



Best Hit Graph

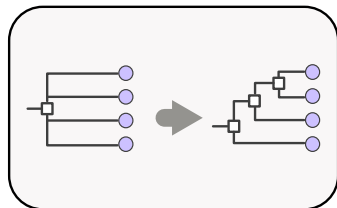
THE PIPELINE



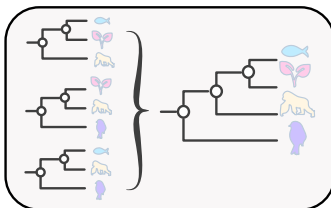
REvolutionH-tl

Reconstruction of Evolutionary Histories Tool

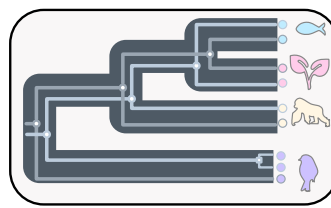
4) Resolution of duplication nodes



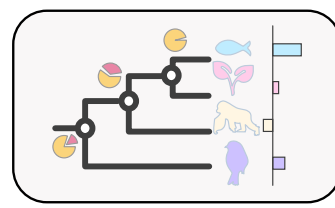
5) Species tree reconstruction



6) Tree reconciliation



7) Visualization



ORTHOGROUPS

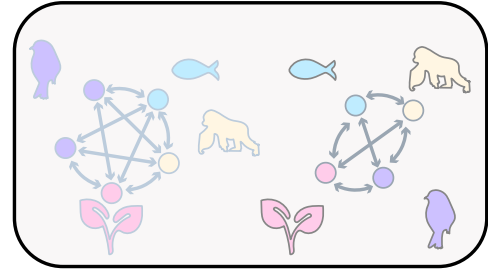
- ▶ **Best Hit graph:**

vertice=genes colored by σ

$x \rightarrow y$ if BLAST **bit score** is one of the **highest** wrt to $\sigma(y)$

- ▶ Its **connected components** estimate the **orthogroups**

Best Hit Graph



ORTHOGROUPS

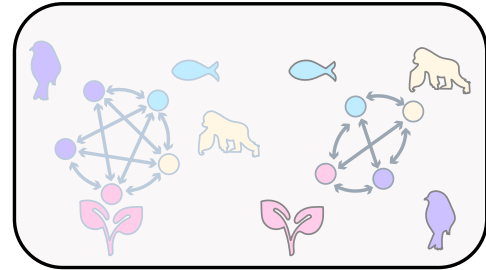
► Best Hit graph:

vertex=genes colored by σ

$x \rightarrow y$ if BLAST **bit score** is one of the **highest** wrt to $\sigma(y)$

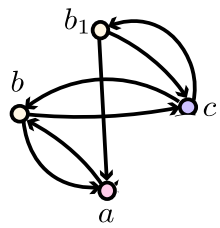
► Its **connected components** estimate the **orthogroups**

Best Hit Graph



RevolutionH-t1	Proteinortho	Orthofinder		True orthogroups	Predicted orthogroups
4070 (96.10 %)	3934 (92.89 %)	2427 (57.30 %)	Completely recovered		
8 (0.18 %)	179 (3.99 %)	841 (18.25 %)	Splitted		
148 (3.49 %)	113 (2.66 %)	928 (21.91 %)	Merged		
9 (0.21 %)	9 (0.21 %)	39 (2.52 %)	Missed		

GENE TREES

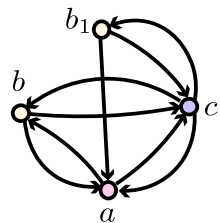


Best Hit Graph

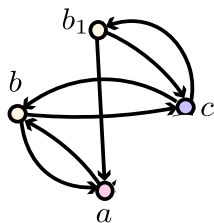
BMG editing
NP-complete (Schaller et al., 2021)



Best Match Graph



GENE TREES

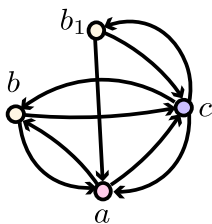


Best Hit Graph

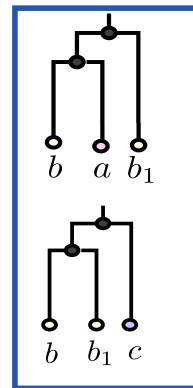
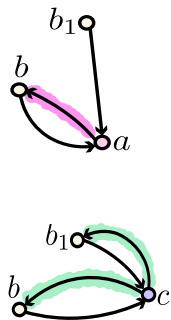
BMG editing
NP-complete (Schaller et al., 2021)



Best Match Graph



Polynomial Heuristic

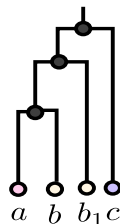


rooted triples

BUILD
+
Louvain



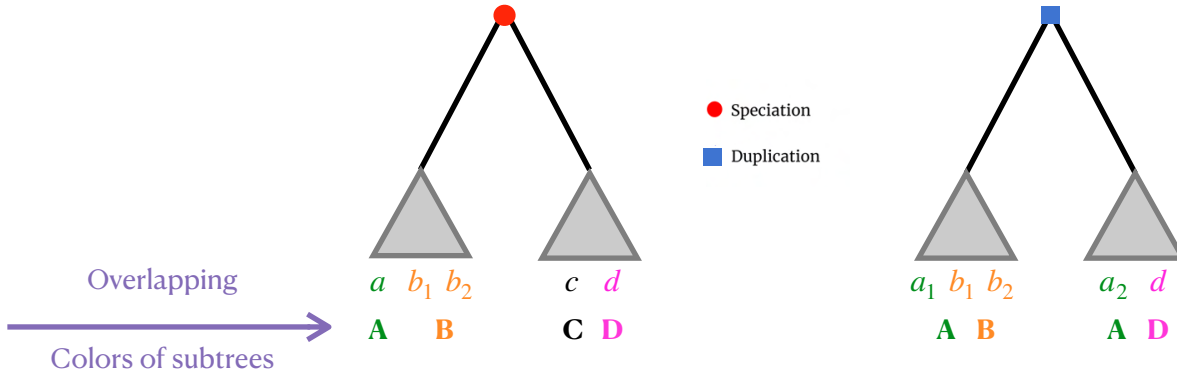
(Schaller et al., 2021)



Most of the
rooted triples are displayed

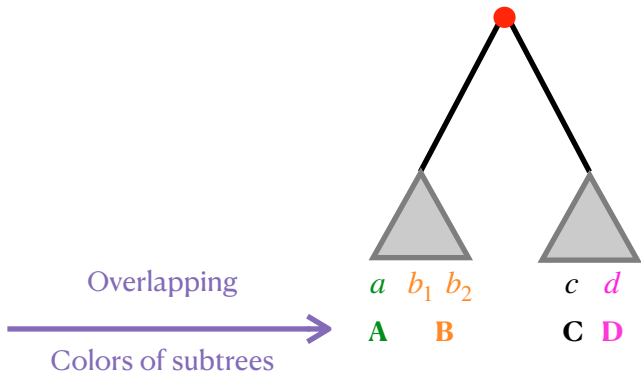
ORTHOLOGS

Species Overlap Method

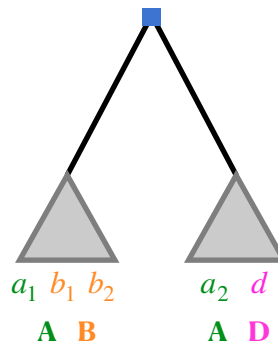


ORTHOLOGS

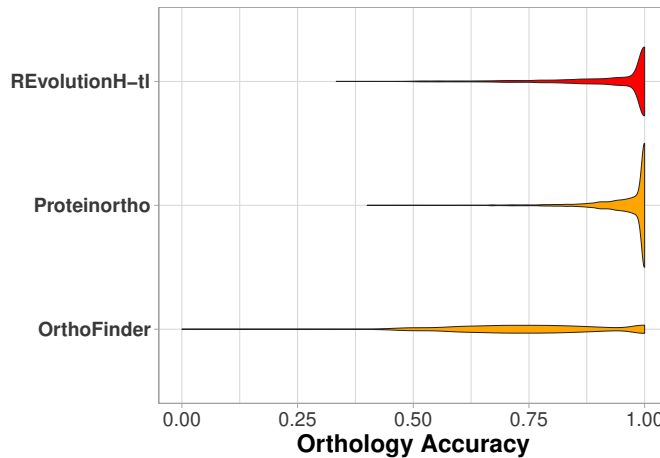
Species Overlap Method



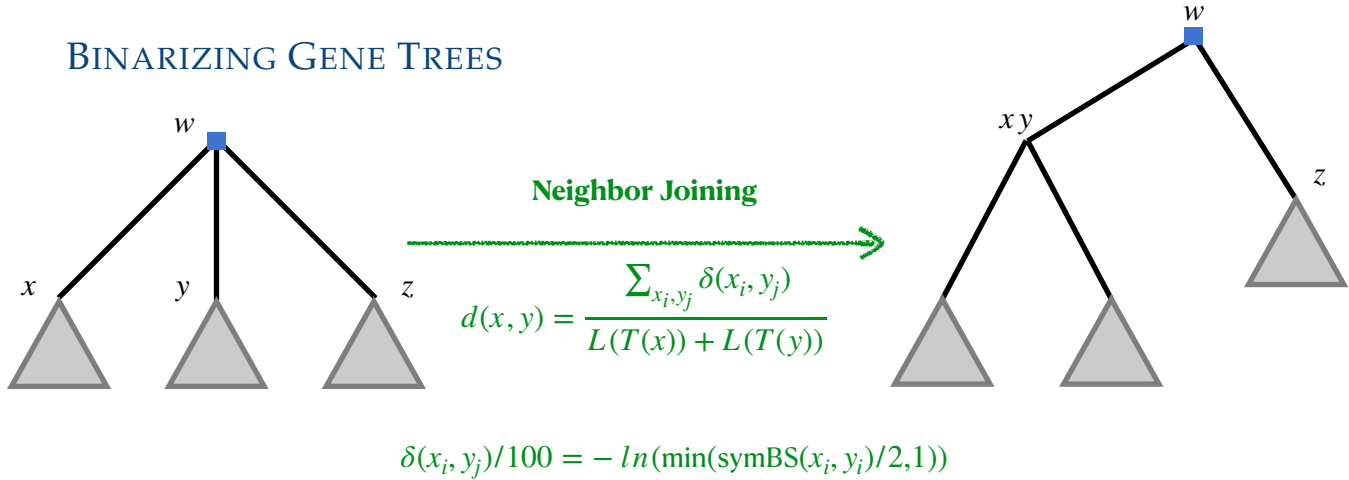
● Speciation
■ Duplication



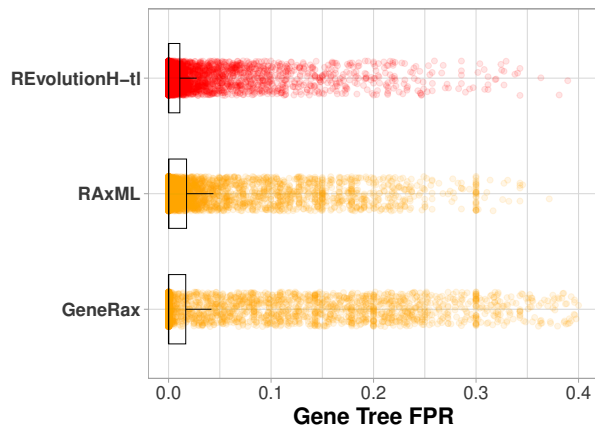
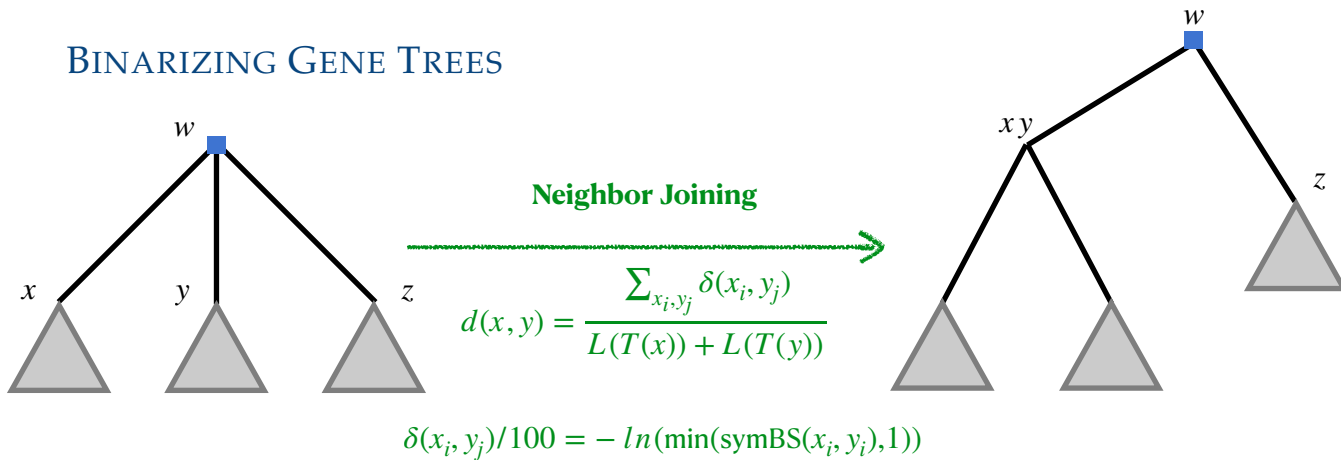
Synthetic Dataset:
~6k gene trees
~1.2k species trees
5-50 species
~1.2k loss & duplication rates



BINARIZING GENE TREES

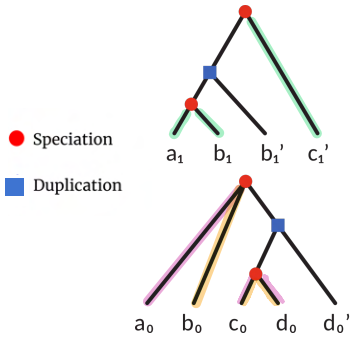


BINARIZING GENE TREES



SPECIES TREES

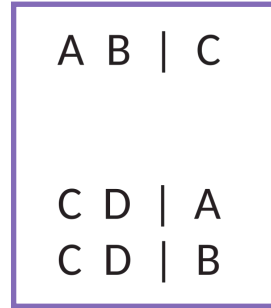
Gene Trees



rooted triples

LCA node is ●

Species Triples



BMG editing

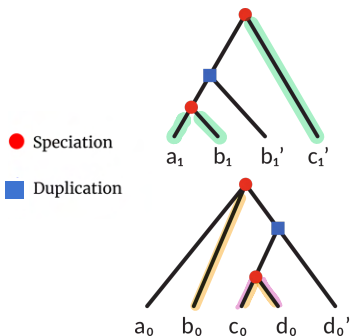
Heuristic

Species Tree



SPECIES TREES

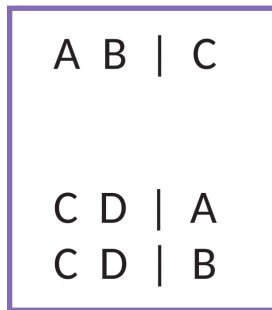
Gene Trees



rooted triples

LCA node is ●

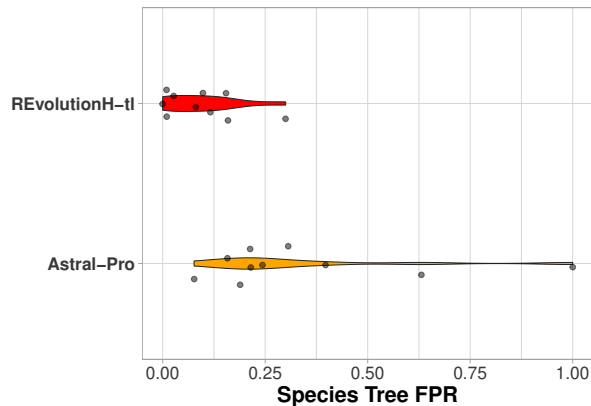
Species Triples



BMG editing

Heuristic

Species Tree

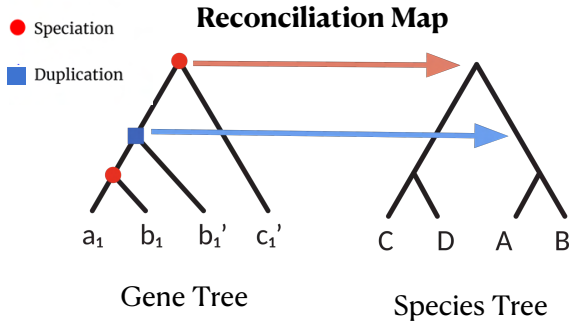


Synthetic Dataset:

~1.2k species trees

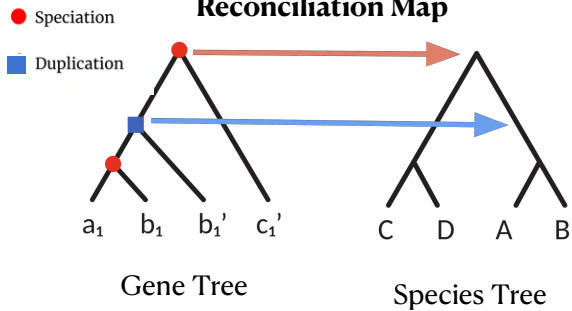
5-50 species

TREE RECONCILIATION

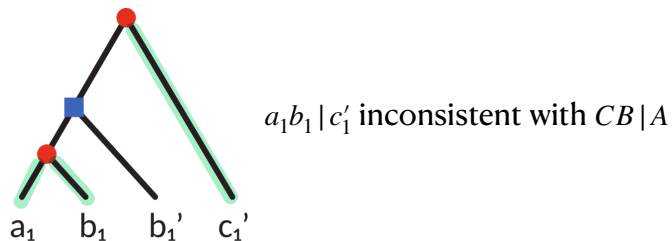


TREE RECONCILIATION

Reconciliation Map

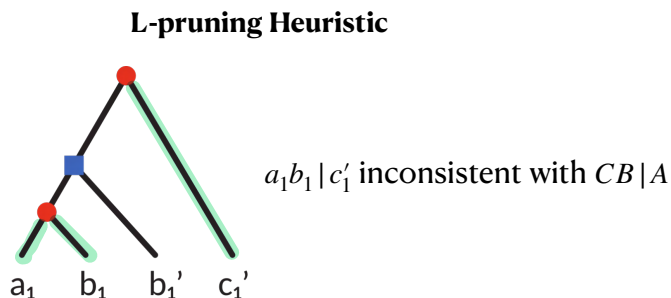
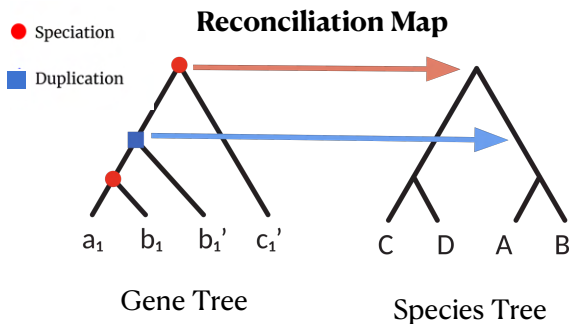


L-pruning Heuristic



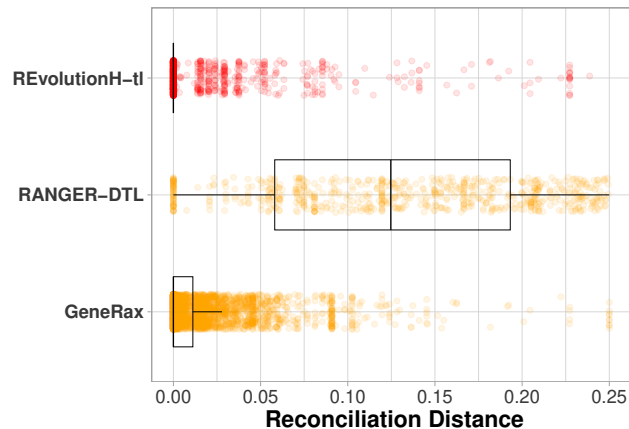
Genes in most of the inconsistent triples are removed

TREE RECONCILIATION



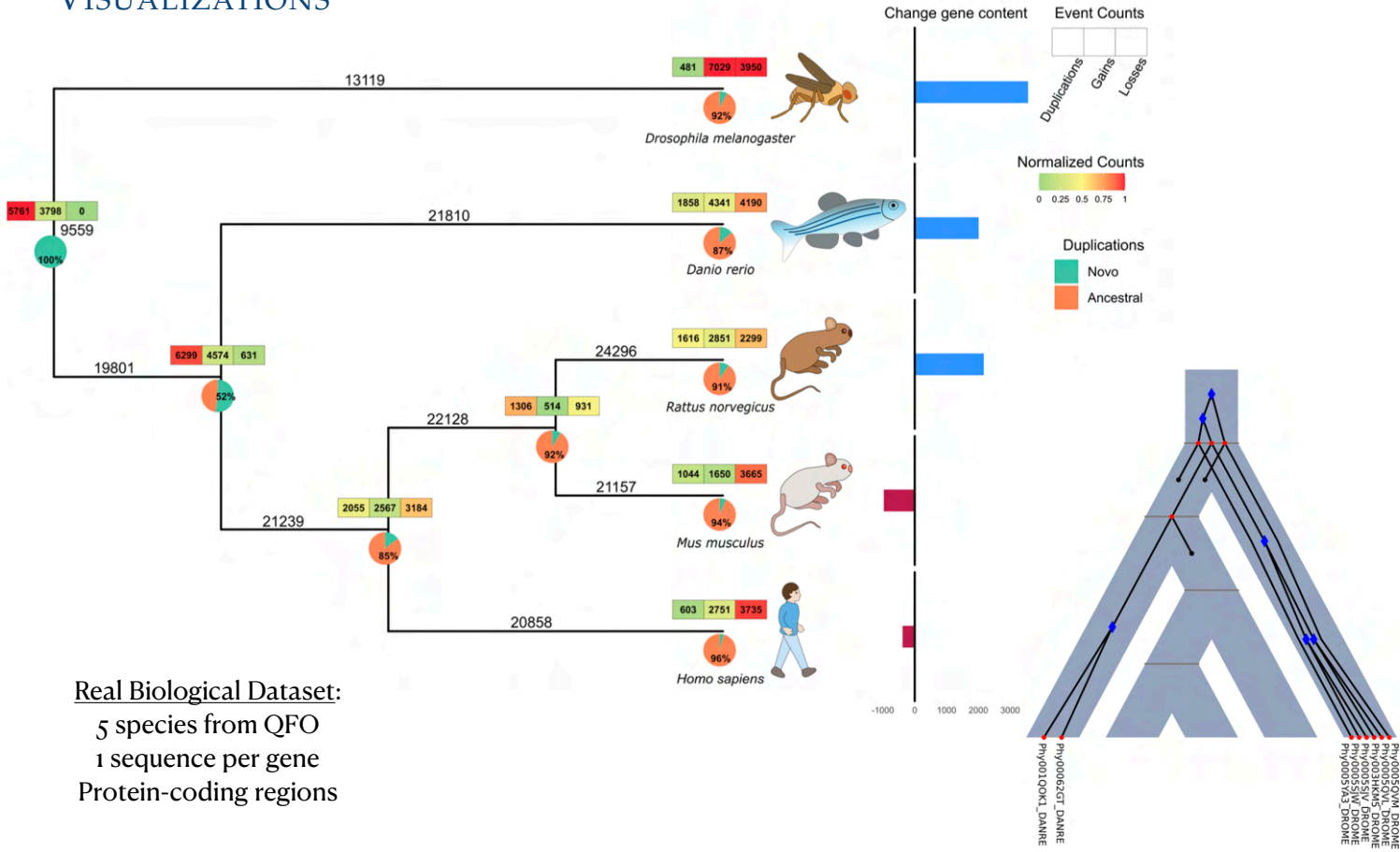
Genes in most of the inconsistent triples are removed

Benchmarking:
Same species trees for all tools
Same gene trees for REvolutionH-tl and GeneRax



PLR dissimilarity measure (Sánchez et al., 2025)

VISUALIZATIONS



RECENT APPLICATIONS

Open Access Article

Analysis of Stress Response Genes in Microtuberization of Potato *Solanum tuberosum* L.: Contributions to Osmotic and Combined Abiotic Stress Tolerance

by Lisset Herrera-Isidron ^{1,†} , Braulio Uribe-Lopez ^{1,†} , Aaron Barraza ² , José Luis Cabrera-Ponce ^{3,*}  and Eliana Valencia-Lozano ^{3,*} 

The chromosome-level genome assembly of the slug *Deroceras laeve* facilitates its use as a comparative model of regeneration

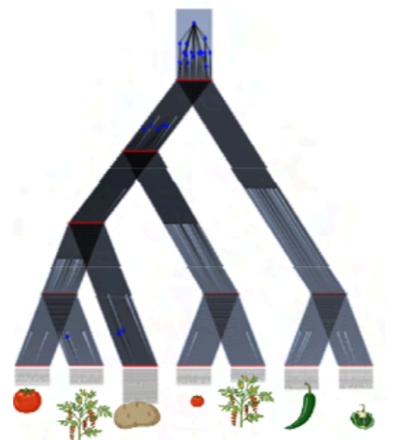
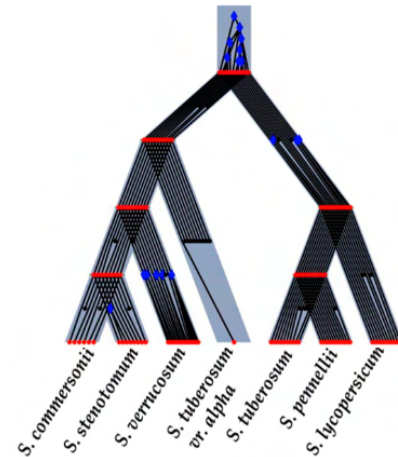
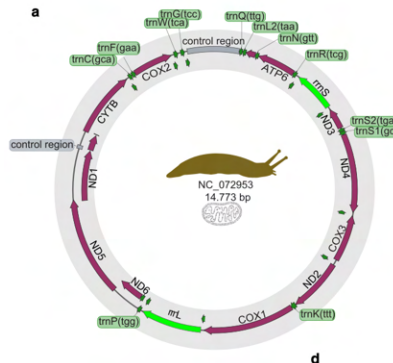
Jerónimo R Miranda-Rodríguez, Javan Okendo, Wilbert Gutiérrez-Sarmiento, Tobías Ber, Josef Pasulka, Kateryna Nemesh, Aranza Torrado-Tapias, Emilio Ortiz-Ávila, Obed Ramírez-Sánchez, Carlos Lozano-Flores ... Show more

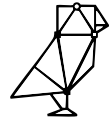
Author Notes

G3 Genes|Genomes|Genetics, jkaf164, <https://doi.org/10.1093/g3journal/jkaf164>

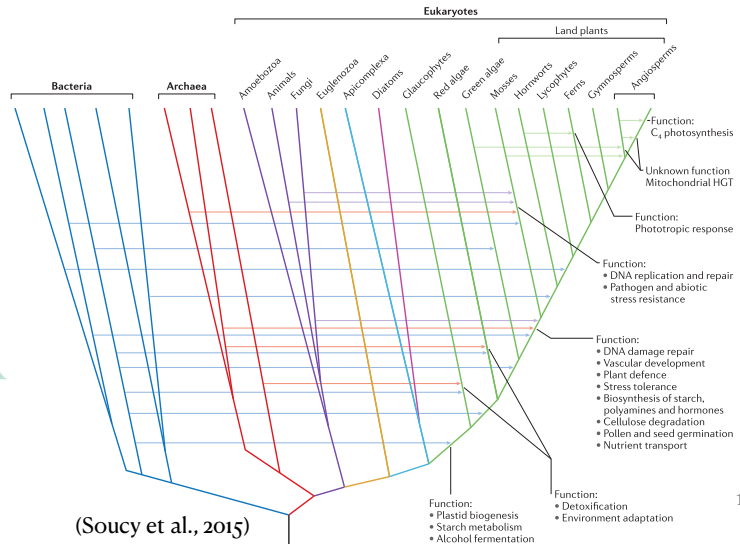
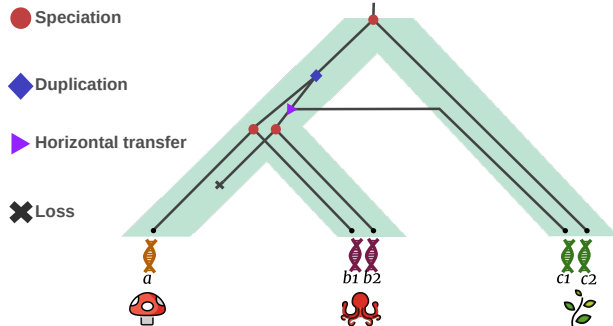
Insights into the Transcriptional Reprogramming in Tomato Response to PSTVd Variants Using Network Approaches

by Katia Aviña-Padilla ^{1,2,*} , Octavio Zambada-Moreno ¹ , Gabriel Emilio Herrera-Oropeza ³ , Marco A. Jimenez-Limas ⁴ , Peter Abrahamian ⁵ , Rosemarie W. Hammond ^{5,*}  and Maribel Hernández-Rosales ^{1,*} 





- **Expanding** orthology and orthogroup **benchmarking** for real biological data.
- Expanding to **DTL** (Duplication Transfer Loss) scenarios, where transfers are horizontal gene transfers.



Extra Slides

