UNIL | Université de Lausanne

A fast pipeline for species tree inference using placement in Hierarchical Orthologous Groups

Sina Majidian[1,2], Adrian M Altenhoff[2,3], Christophe Dessimoz[1,2,4]

1 Department of Computational Biology, University of Lausanne.
2 SIB Swiss Institute of Bioinformatics.
3 Department of Computer Science, ETH Zurich.
4 Department of Computer Science, University College London.

@DessimozLab
@SinaMajidian

Inferring species trees: a fundamental problem

Orthologous Groups (marker genes):

Group of genes that emerged from a common ancestral gene through speciation.

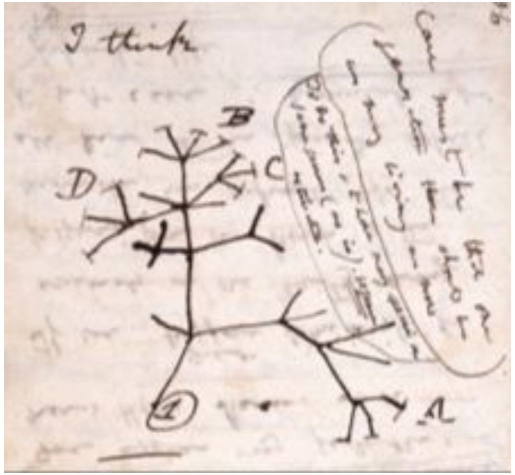
1) conventional orthology pipelines

❌ computationally intensive (not scalable)

2) Precomputed markers e.g. **BUSCO**

❌ not available for all clades

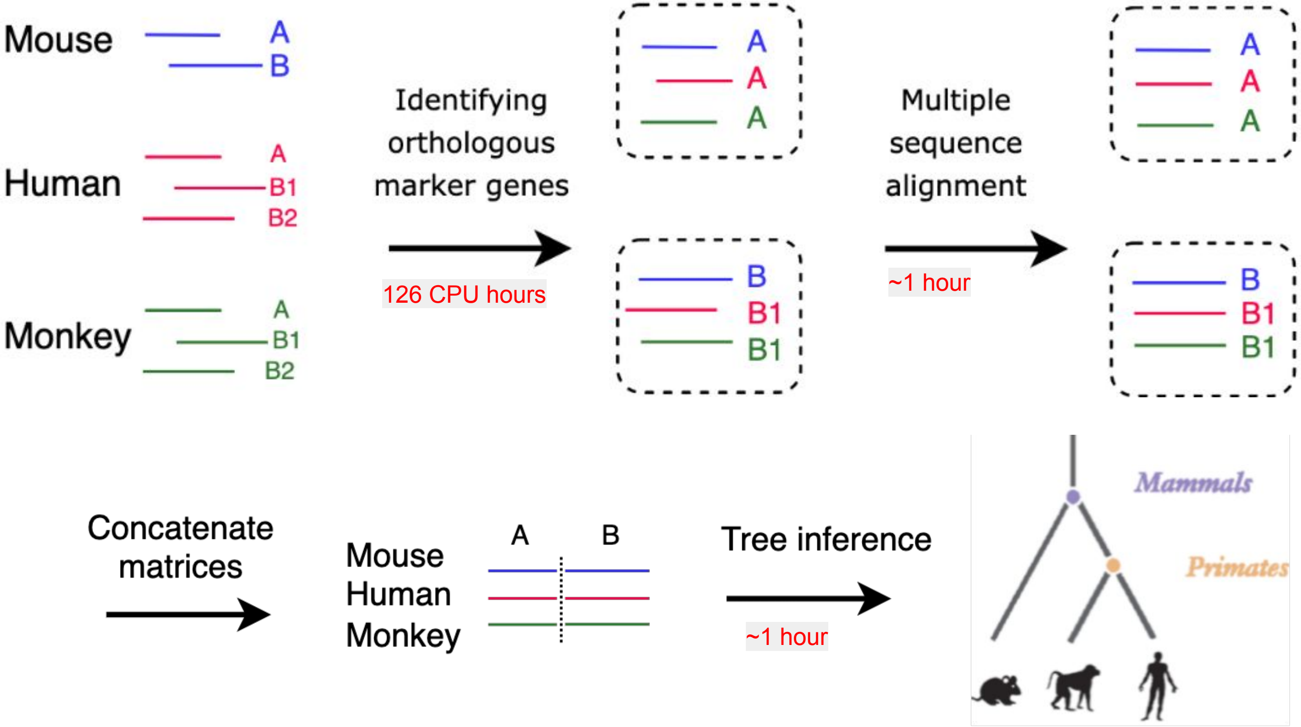
❌ ↑species → **U**niversal and **S**ingle-**C**opy **O**rthologs ↓



Darwin, 1837.

Sudhindra R., et al. "Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree." *Journal of Experimental Zoology: Molecular and Developmental Evolution* 2005.

Inferring species trees: the standard pipeline



Identifying orthologous marker genes

126 CPU hours

Multiple sequence alignment

~1 hour

Concatenate matrices

Tree inference

~1 hour

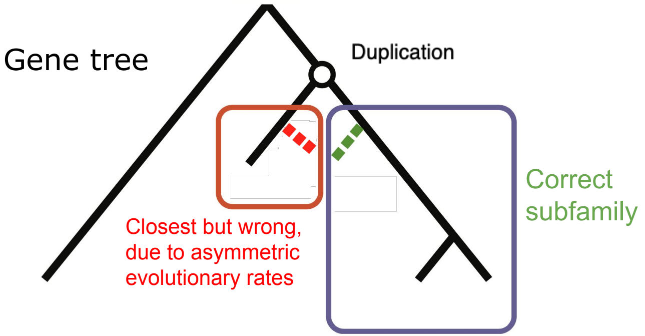
Dylus, David, et al. "How to build phylogenetic species trees with OMA." *F1000Research*. 2020.

Basis of the proposed pipeline (FastOMA)

- Map proteins to the database of subfamily of genes
- Traditionally achieved by finding the closest sequence (by BLAST or DIAMOND).

❌ the closest sequence might belong to a different subfamily (not ortholog)

Our solution is OMAmer:
A subfamily-level classifier
using subfamily-informed k-mers.



Gene tree

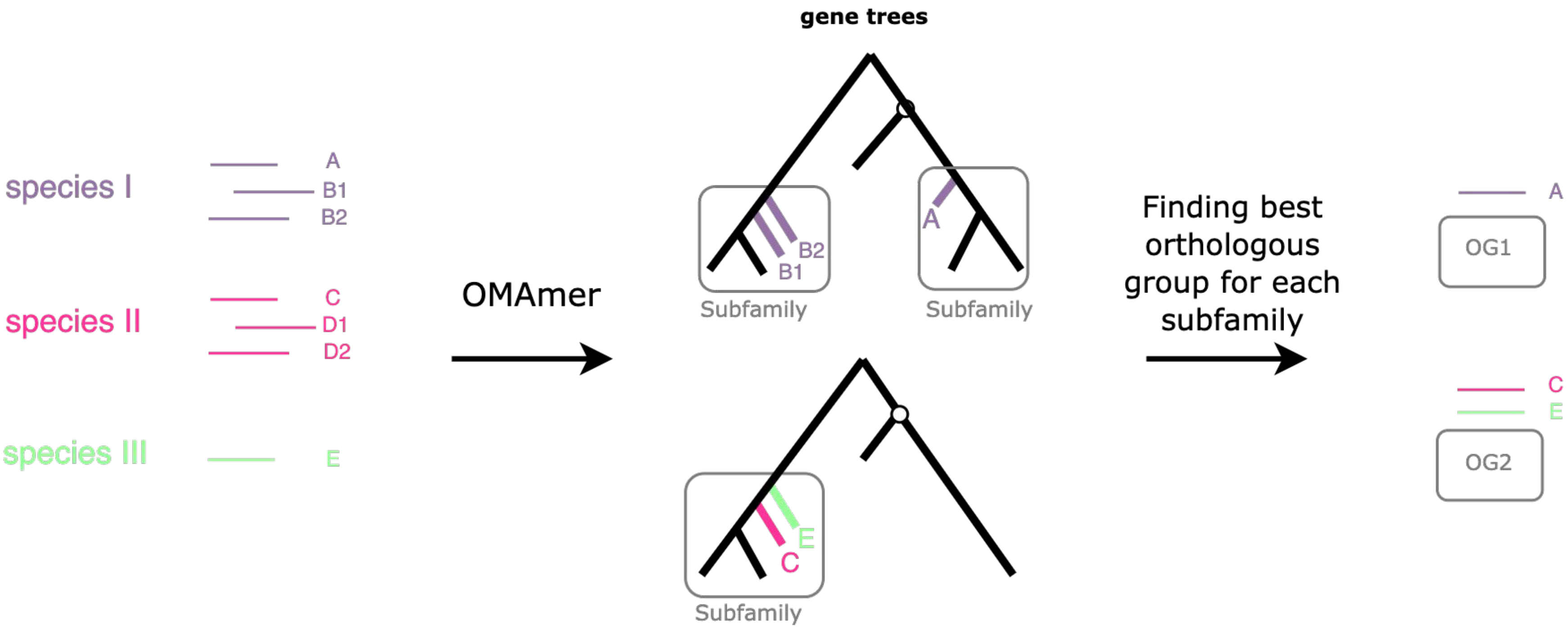
Duplication

Closest but wrong,
due to asymmetric
evolutionary rates

Correct
subfamily

Rossier, V., et al. "OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches." Bioinformatics 2021.

FastOMA: the accelerated pipeline



OMAmer

Finding best orthologous group for each subfamily

Evaluation on B10k dataset

- phase II: 363 birds

Run time in CPU hours

	Identifying orthologous groups (OGs)	Multiple sequence alignment (100 OGs)	Tree inference
Standard	1936	~1	56
FastOMA	49	~1	79

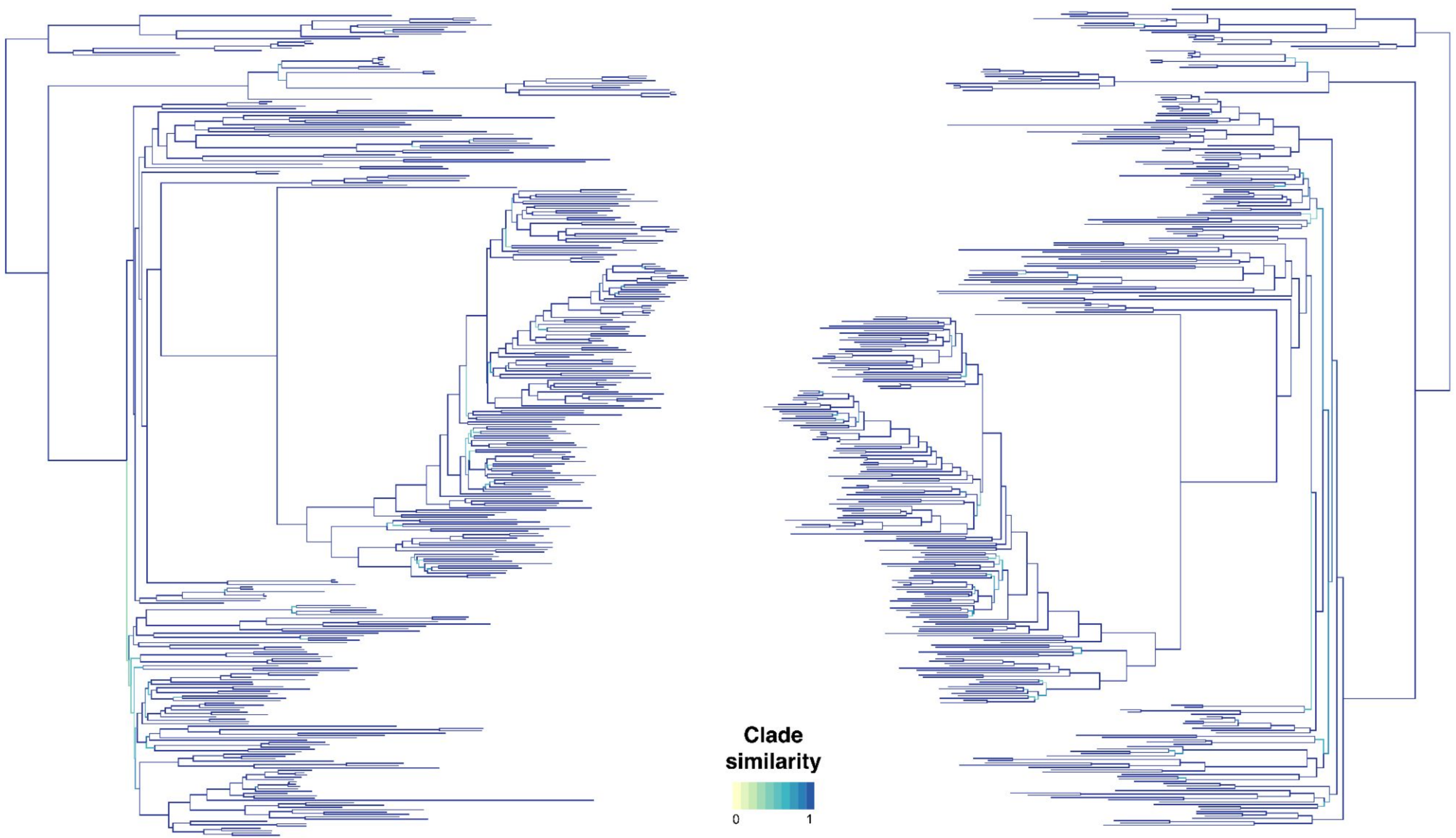
FastOMA is

15 times faster.

Shaohong F., et al. "Dense sampling of bird diversity increases power of comparative genomics." Nature, 2020.

Standard pipeline

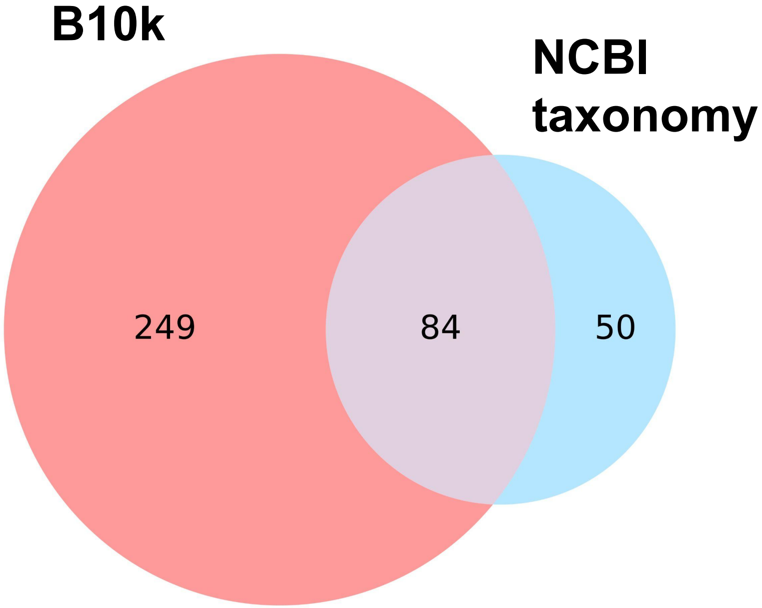
FastOMA



RF partitions that exist only in standard, not found in FastOMA **= 69 out of 704.**

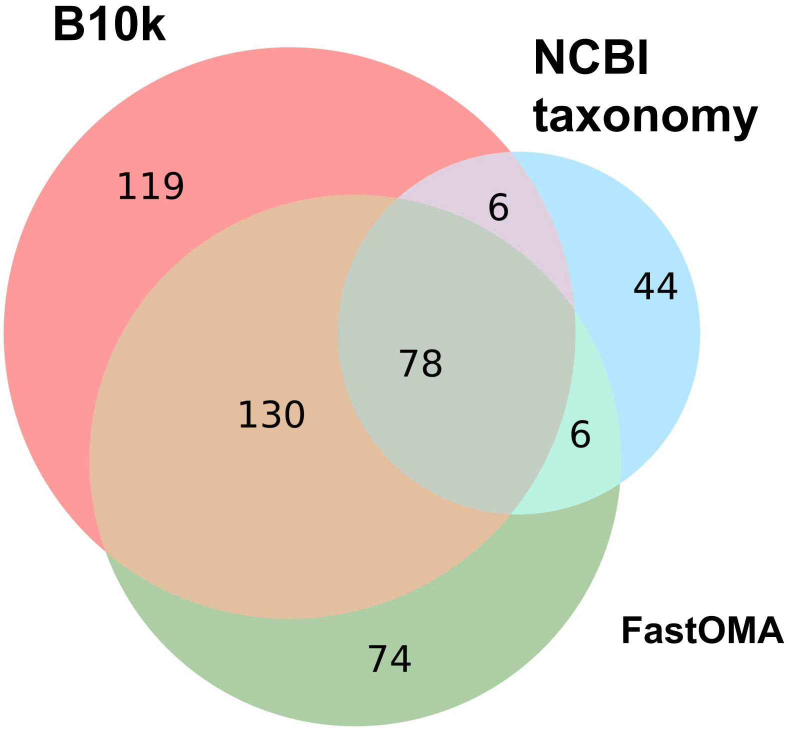
Bird phylogeny is challenging!

RF values

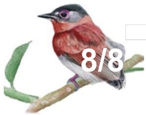


Bird phylogeny is challenging!

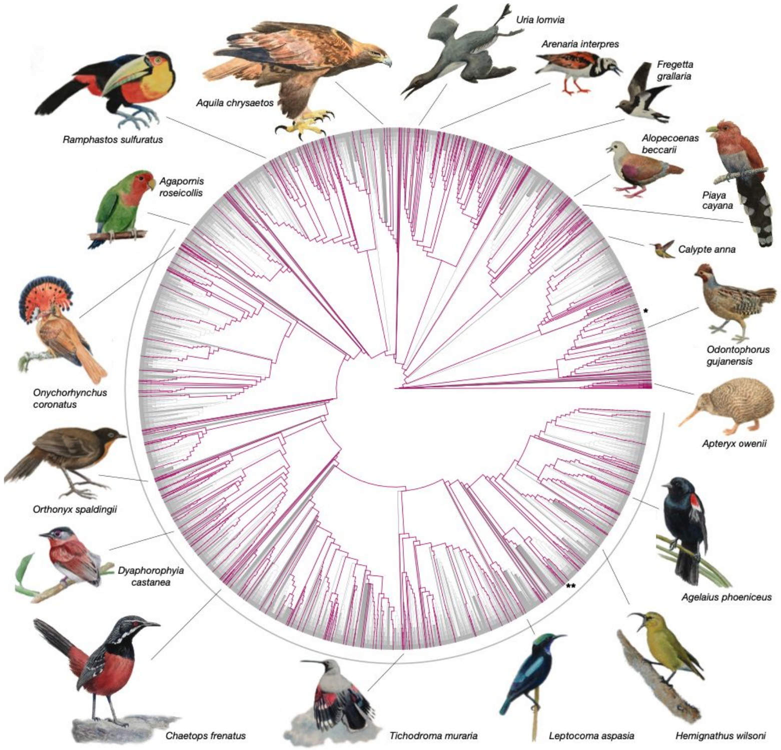
RF values



FastOMA is a reliable and scalable solution for the tree inference pipeline.



Thank you!



Shaohong F., et al. Nature, 2020.

Evaluation on B10k dataset

- phase II: 363 birds

Run time in CPU hours

	Identifying orthologous groups (OGs)	Multiple sequence alignment (100 OGs)	Tree inference
Standard	1936	~1	56
FastOMA	49	~1	79

FastOMA is

20 times faster.

Shaohong F., et al. "Dense sampling of bird diversity increases power of comparative genomics." Nature, 2020.