Computational comparative genomics in the era of BioGenome projects

(fast orthology and phylogeny tools)

Sina Majidian

Department of Computational Biology, University of Lausanne. SIB Swiss Institute of Bioinformatics.

November 6, 2023

@SinaMajidian
sina.majidian@unil.ch

Speaker biography

2021- 2024 Postdoctoral Fellow, Switzerland.

Comparative Genomics Lab

(Christophe Dessimoz & Natasha Glover)

2018-2019 Guest researcher, Netherlands.

Bioinformatics group (Dick de Ridder)

2015-2020 PhD, Signal processing, Iran.

Computational comparative genomics in the era of BioGenome projects

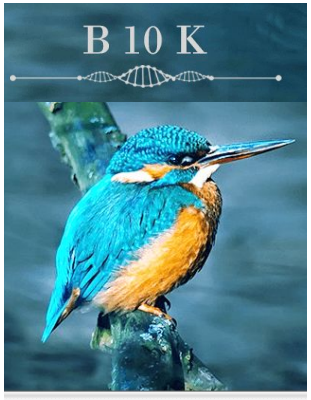
(fast orthology and phylogeny tools)

Outline

- Introduction
- Orthology inference with FastOMA
- Phylogeny inference from raw sequencing reads with read2tree

BioGenome projects

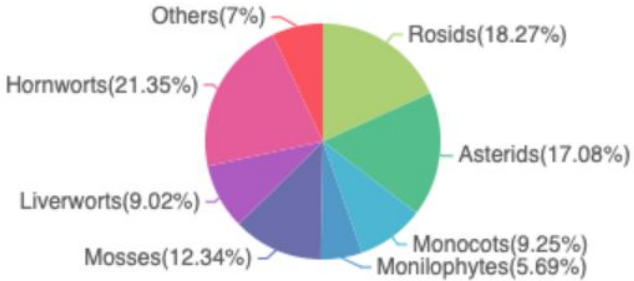
Sequence and characterize the genomes of all of Earth's eukaryotic biodiversity over ten years.



AQUATIC SYMBIOSIS GENOMICS

African BioGenome Project

10,000 Plants (10KP) aims to sequence a member of every genus!

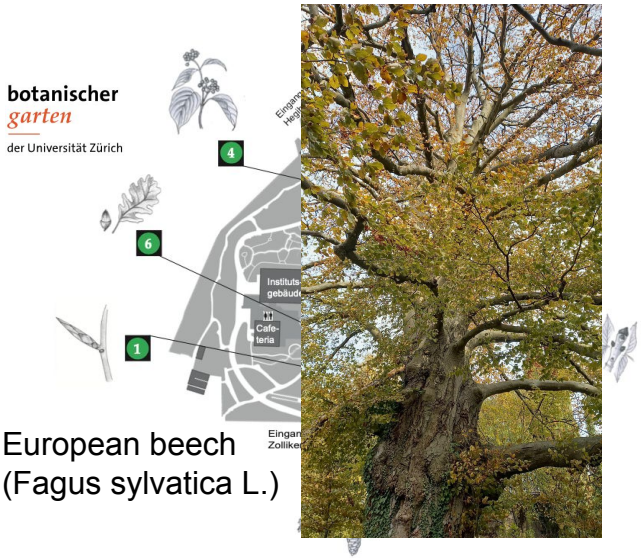


Darwin TREE of LIFE

Biodiversity Genomics Europe

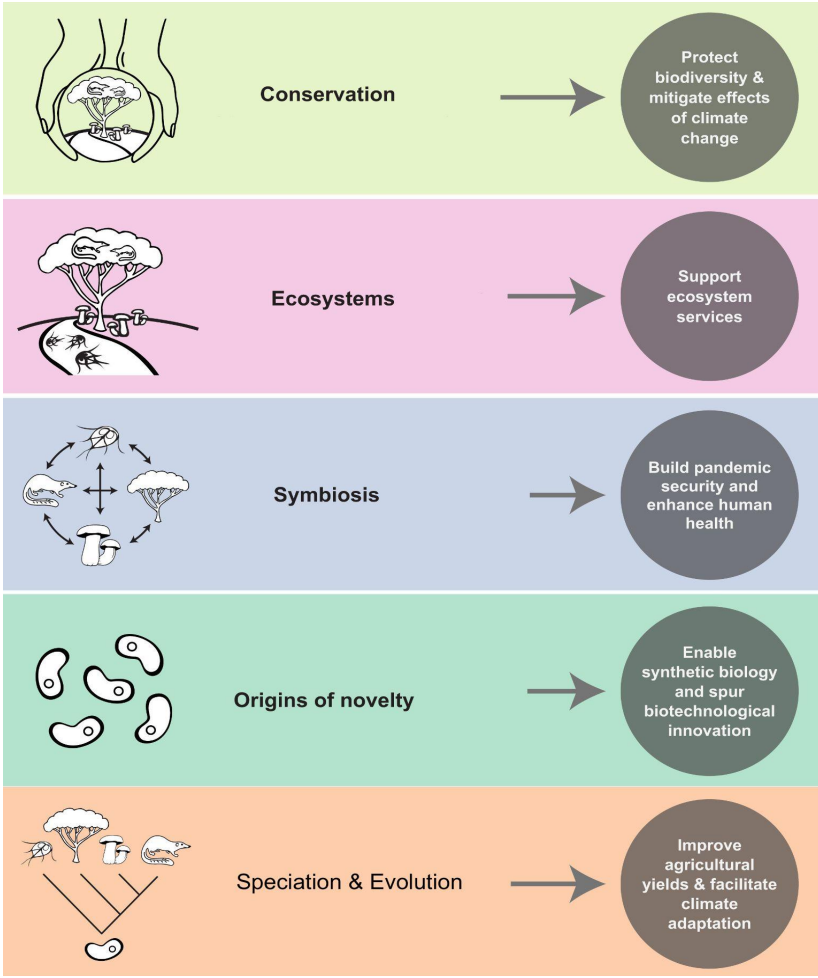
ERGA
EUROPEAN REFERENCE GENOME ATLAS

Genome project benefits



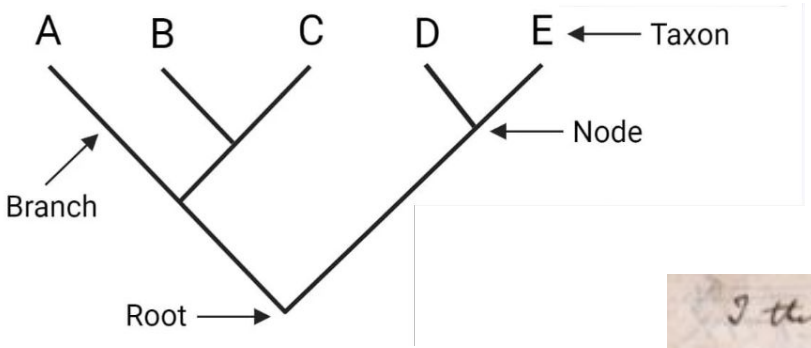
European beech
(Fagus sylvatica L.)

Genomic basis for drought resistance in European beech forests threatened by climate change. Elife 2022.



The Earth BioGenome project, The plant journal, 2019.vvvv Why sequence all eukaryotes? PNAS. 2022.

Evolutionary tree



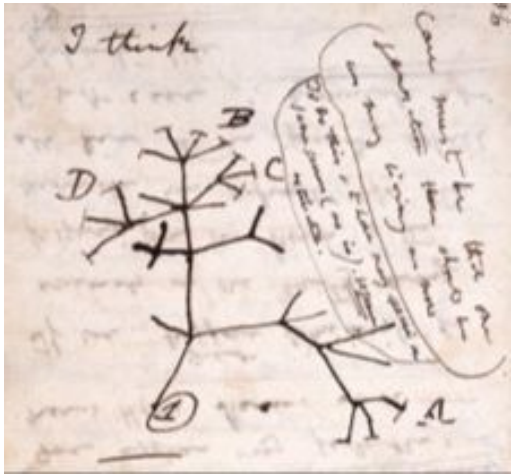
Inferring species trees

1) 16s rRNA

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

❌ not available for all clades

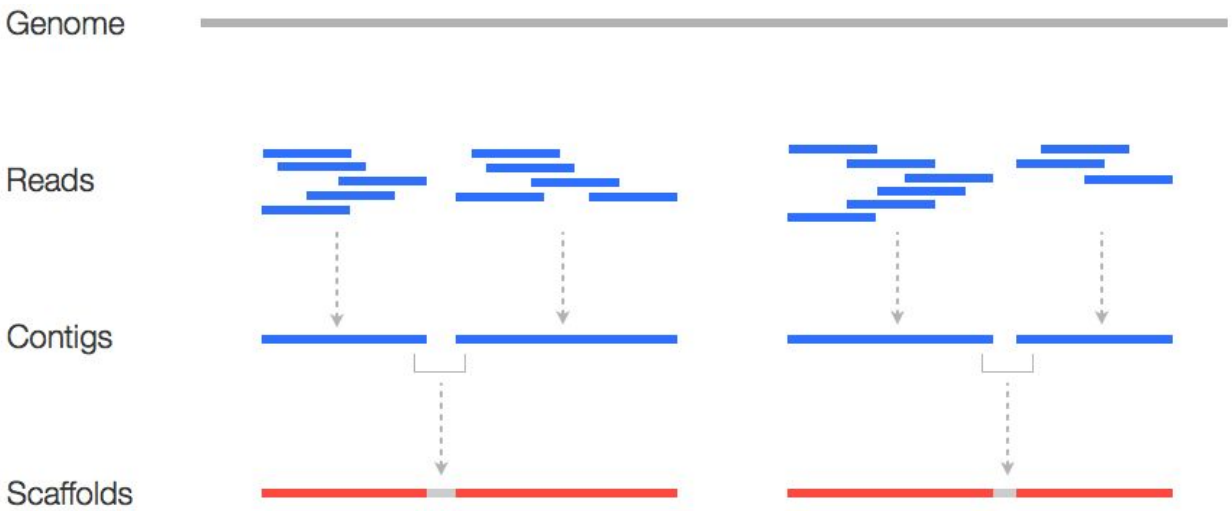
3) Gene markers



Darwin, 1837

Inferring species trees with concatenation

Assembly	Gene annotation	Finding gene markers	Multiple sequence alignment	Concatenation	Tree inference



EBP recommendation

ARIMA GENOMICS
OR
Dovetail Genomics
HiC/OmniC

PACBIO
HiFi >30X
OR
Oxford NANOPORE Technologies
ONT data >60X

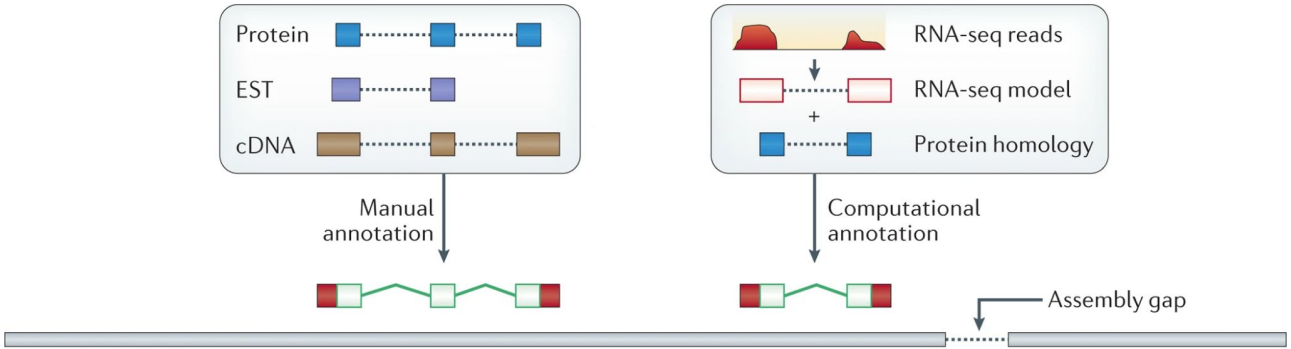
Inferring species trees with concatenation

Assembly	**Gene annotation**	Finding gene markers	Multiple sequence alignment	Concatenation	Tree inference

Finding genomic coordinates of coding regions

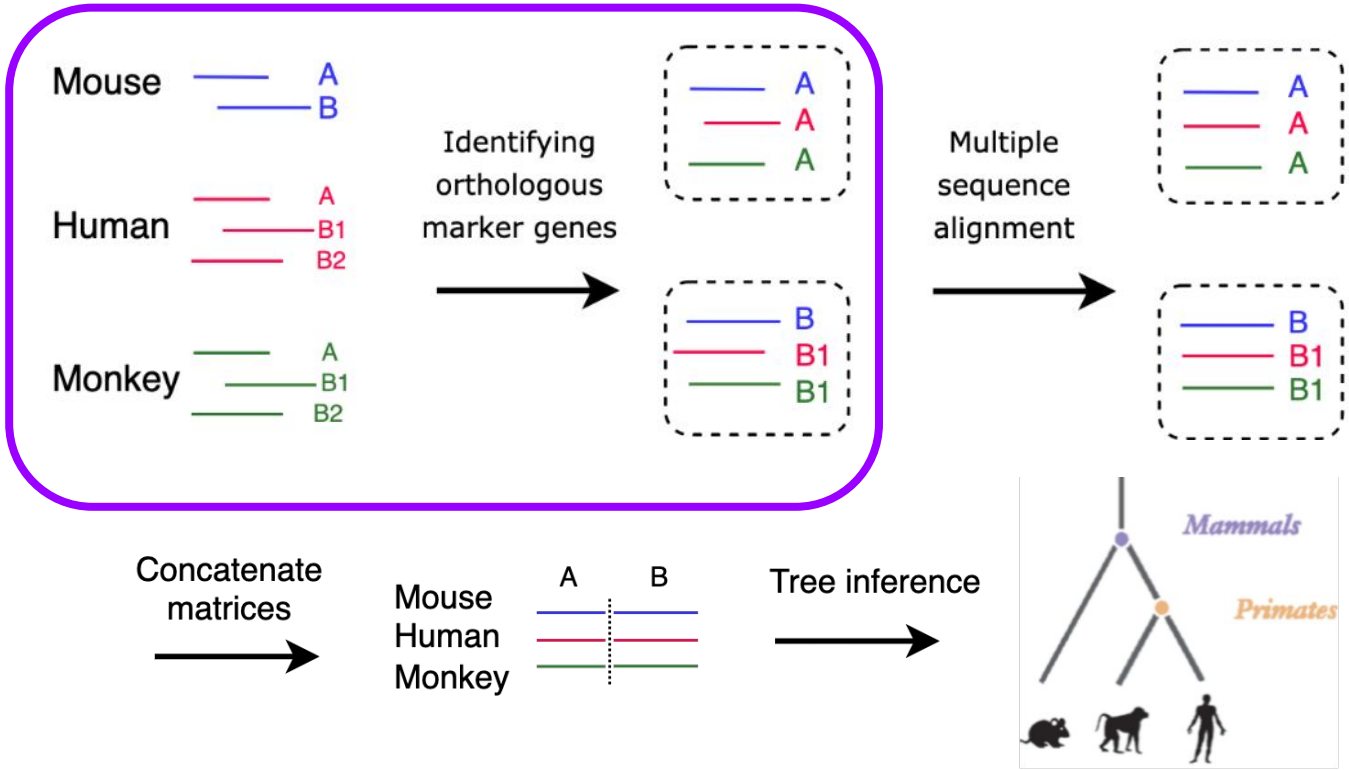
MAKER
Annotate this!

illumina
RNAseq
>5 tissue types
AND/OR
PACBIO
ISOseq



→ **proteome**

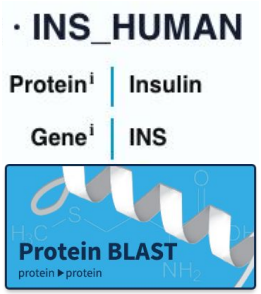
Assembly	Gene annotation	Finding gene markers	Multiple sequence alignment	Concatenation	Tree inference



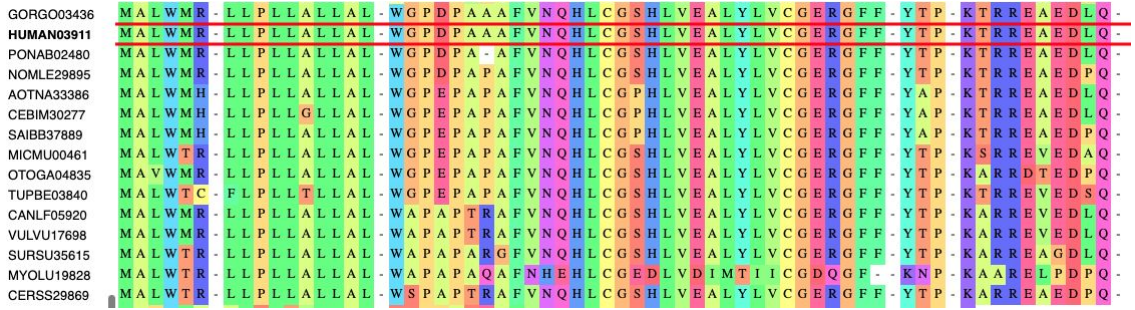
Dylus, et al. "How to build phylogenetic species trees with OMA." *F1000Research*. 2020.

Homology

homologous regions or regions of common ancestry.



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	insulin isoform UB [Homo sapiens]	Homo sapiens	226	226	100%	7e-74	100.00%	153	QMS45324.1
☑	insulin [Gorilla gorilla gorilla]	Gorilla gorilla go…	226	226	100%	8e-74	100.00%	153	XP_004050475.2
☑	Homo sapiens insulin [synthetic construct]	synthetic construct	224	224	100%	2e-73	100.00%	111	AAP36446.1
☑	insulin preproprotein [Homo sapiens]	Homo sapiens	223	223	100%	2e-73	100.00%	110	NP_000198.1
☑	insulin isoform X2 [Pongo abelii]	Pongo abelii	222	222	100%	1e-72	99.09%	110	XP_024110665.1
☑	INS [synthetic construct]	synthetic construct	221	221	100%	1e-72	99.09%	110	AKI70564.1
☑	insulin isoform U1 [Homo sapiens]	Homo sapiens	225	225	100%	1e-72	100.00%	204	QMS45321.1
☑	insulin preproprotein [Pan troglodytes]	Pan troglodytes	221	221	100%	2e-72	98.18%	110	NP_001008996.1



Similarity implies common ancestry.

Orthology vs. paralogy

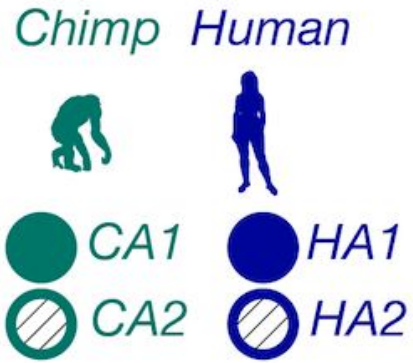
Two classes of homologous genes:

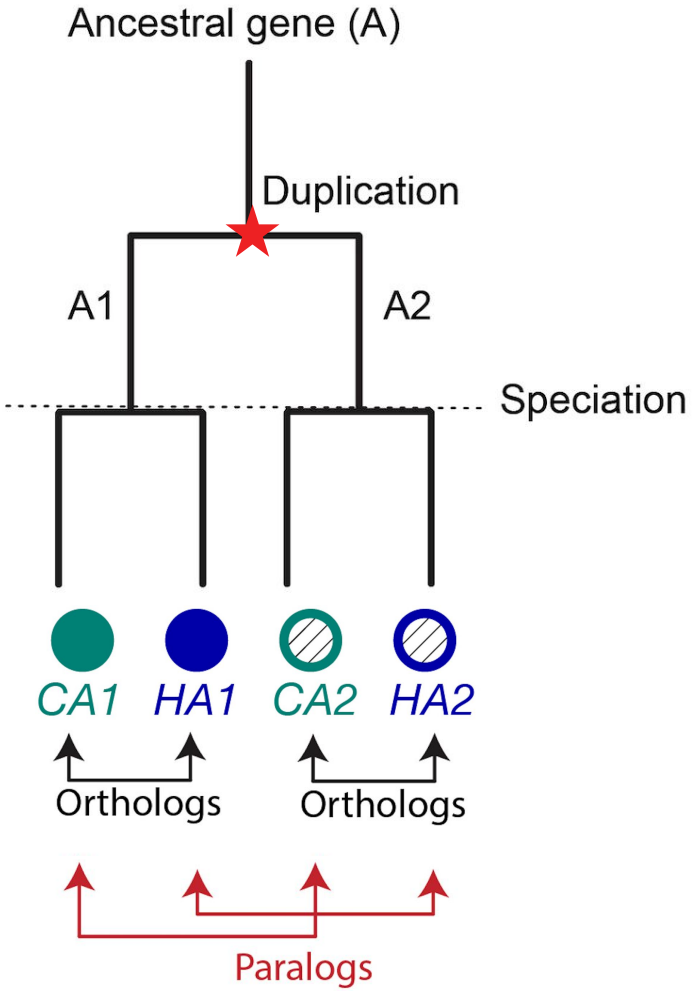
- **Orthology:** A relation between pairs of genes that started diverging via evolutionary speciation.
    - useful for species tree inference.
    - tend to have conserved functions.

- **Paralogy:** A relation between pairs of genes that started diverging via gene duplication.

Gabaldón, Koonin. Functional and evolutionary implications of gene orthology. Nat. Rev. Genet. 2013.

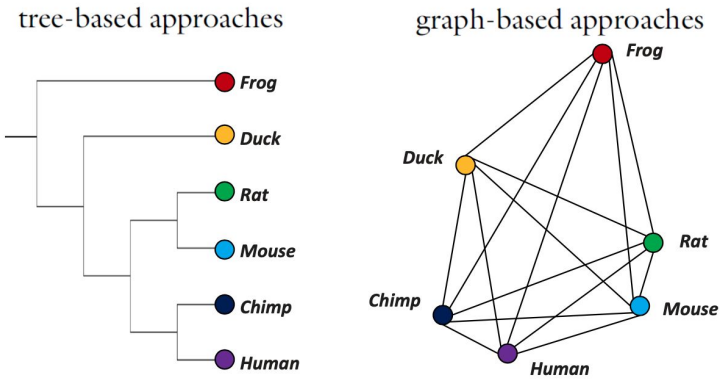
Orthology vs. paralogy





Orthology: promises and challenges, Evolutionary Biology. 2020.

Orthology inference methods

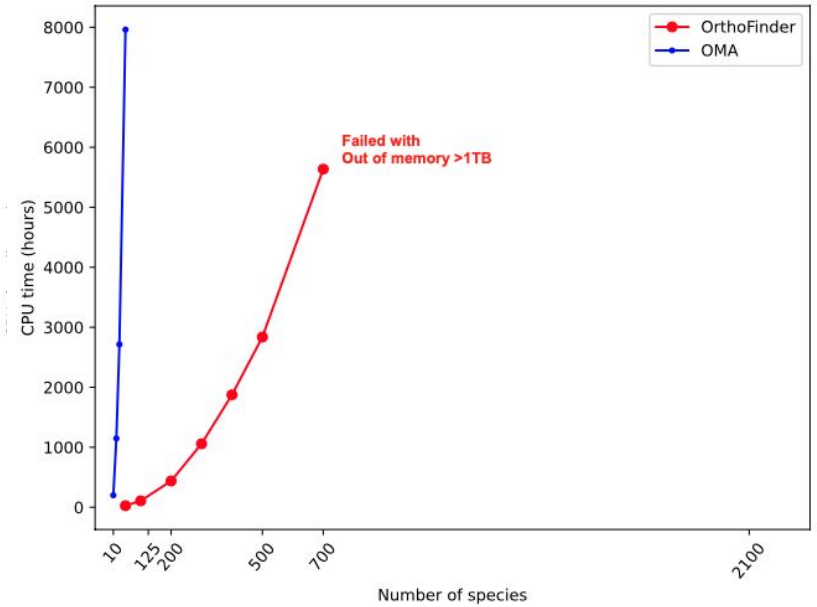


Method/resource	# species	
Oma orthologous matrix browser	2,851	Graph based
OrthoDB	20,110	Graph based
DB phylome	6,000	Tree based
NOG EggNOG 6.0.0	12,535	Graph + tree
OrthoFinder 2	-	Graph + tree

❌ All-vs-all comparison

Is orthology inference possible for >1000 species ?

- UniProt reference eukaryotic proteomes



We need new concepts and methods to handle data from BioGenome projects!

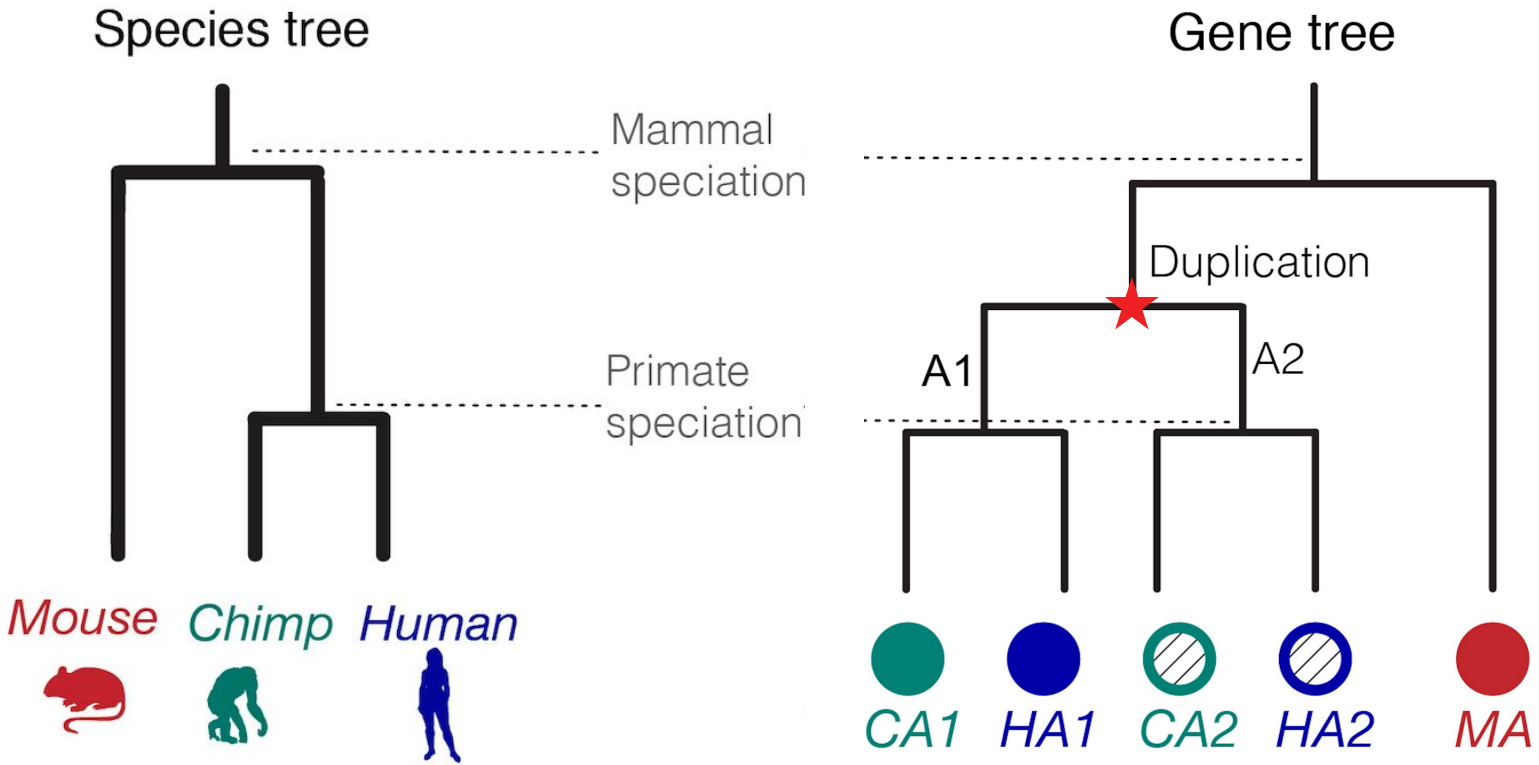
Orthologous groups

- **Orthologous groups:** Groups of genes which are all orthologous to each other.

- **Hierarchical Orthologous Group (HOG):** Groups of genes descended from a common ancestral gene at a specific taxonomic level

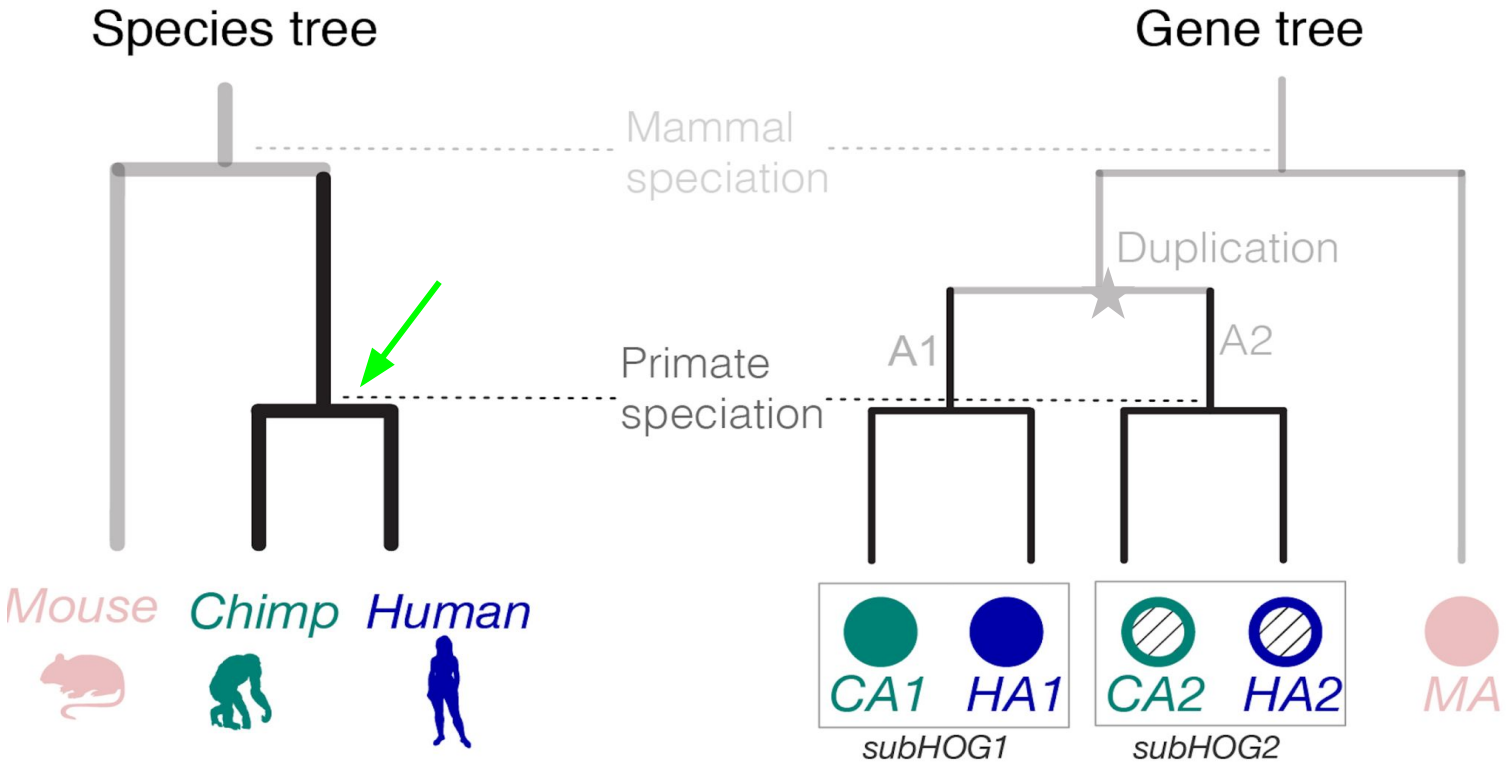
HOG Hierarchical Orthologous Group

- Genes descended from a common ancestral gene at a specific taxonomic level



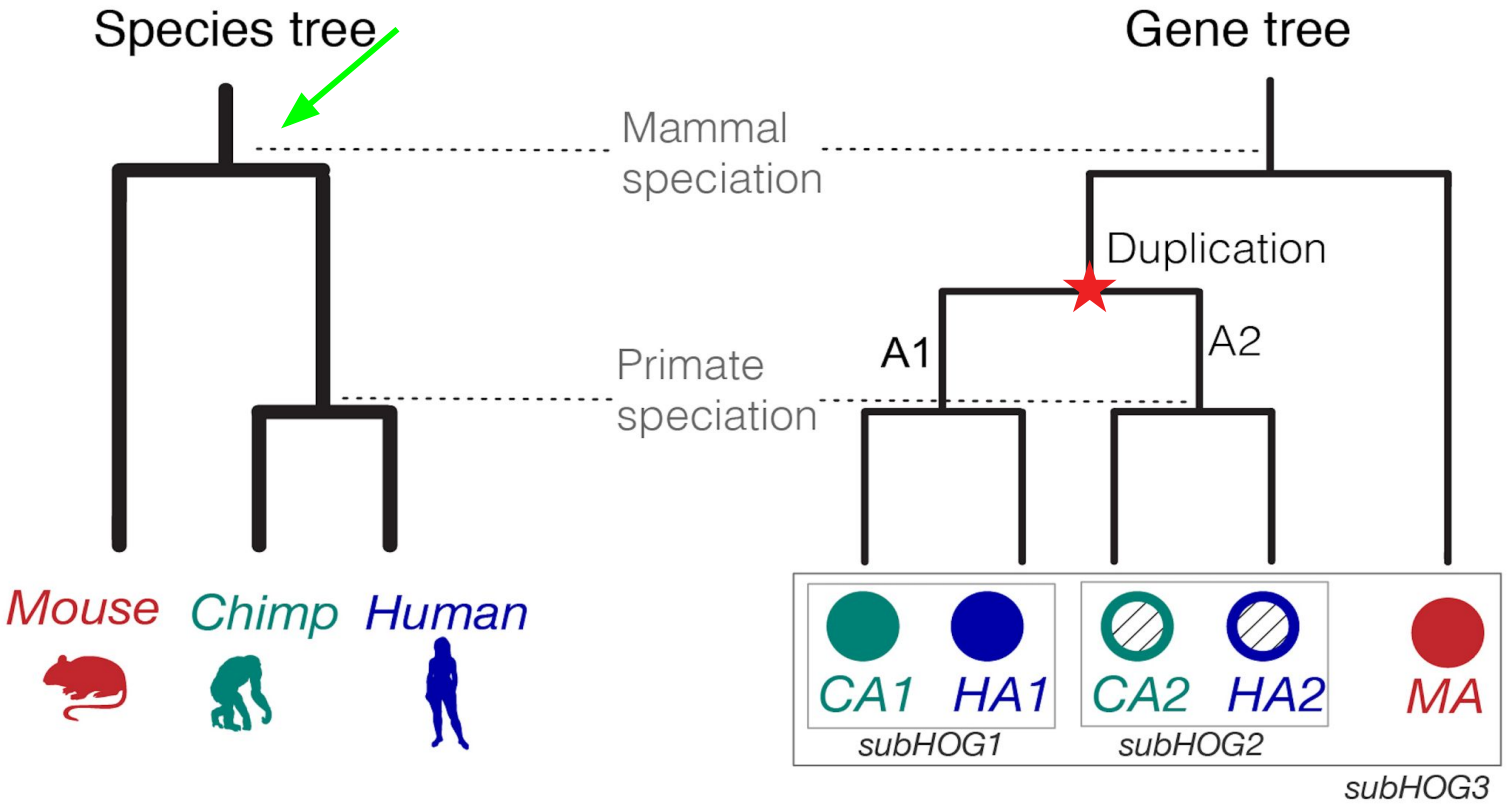
HOG Hierarchical Orthologous Group

- Genes descended from a common ancestral gene at a specific taxonomic level



HOG Hierarchical Orthologous Group

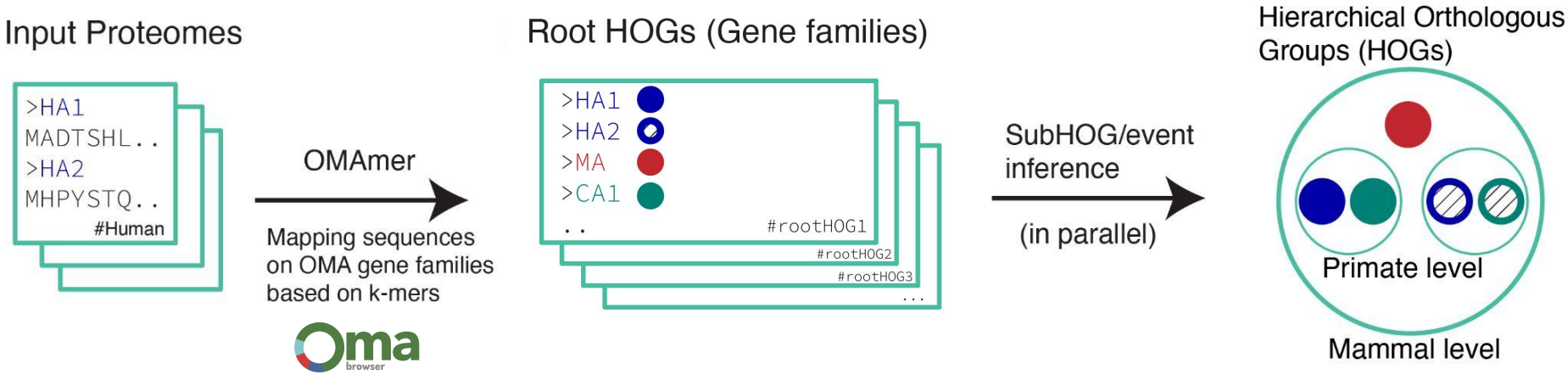
- Genes descended from a common ancestral gene at a specific taxonomic level



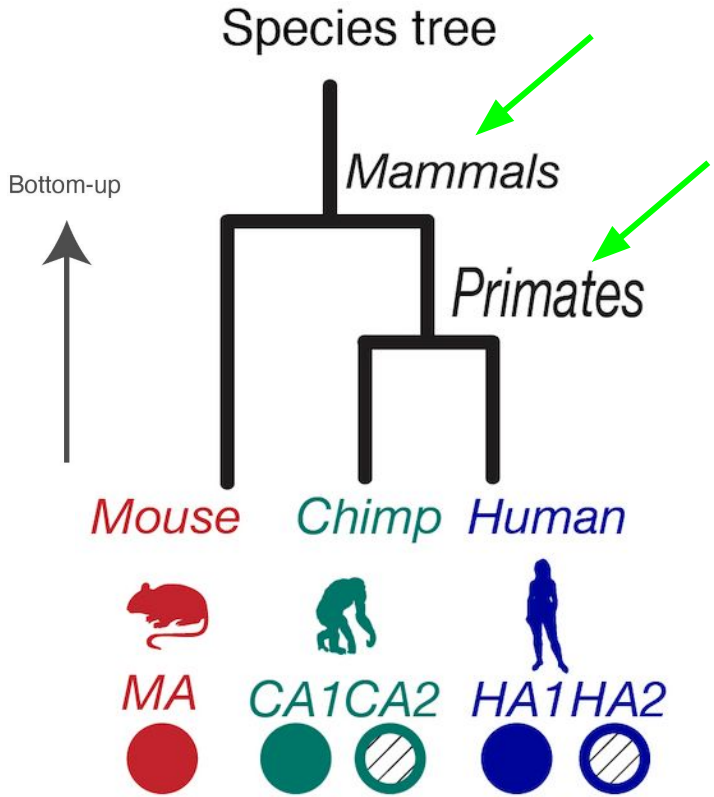
Outline

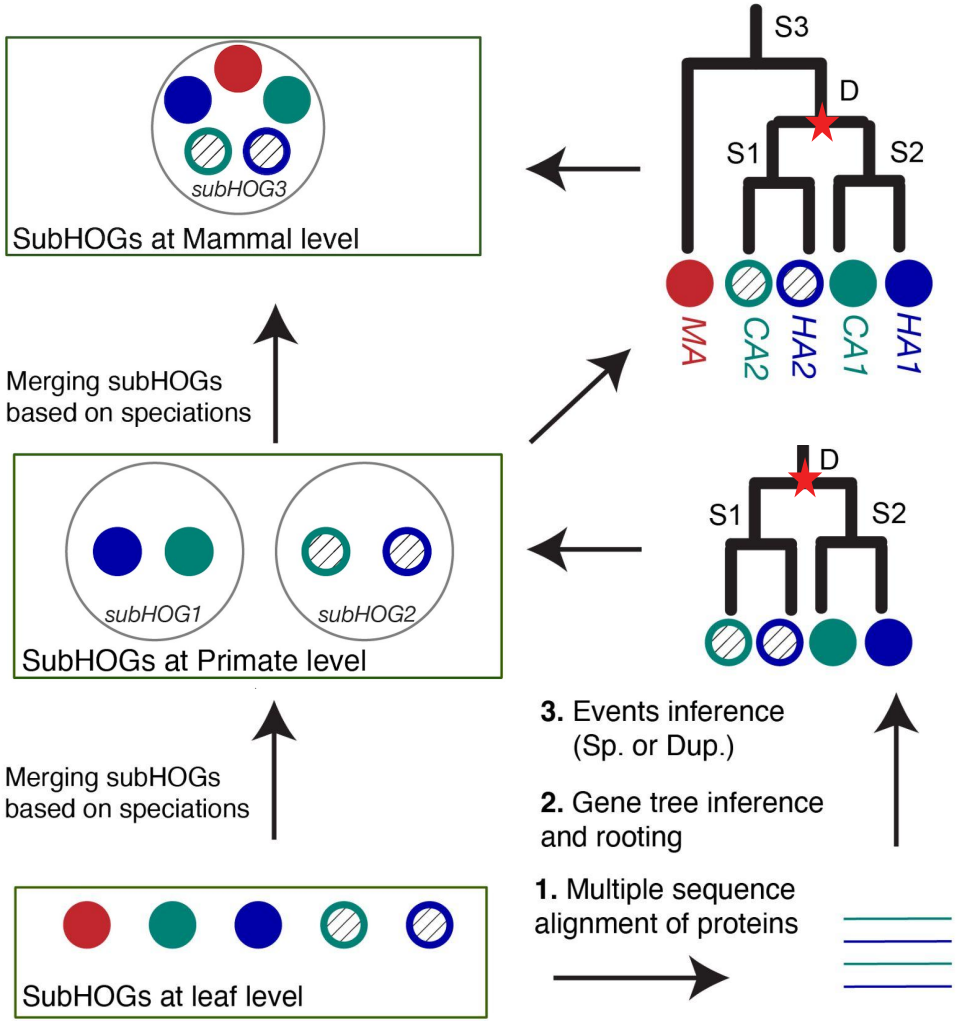
- Introduction
- Orthology inference with FastOMA
- Phylogeny inference from raw sequencing reads with read2tree

FastOMA, our new tool



HOG inference

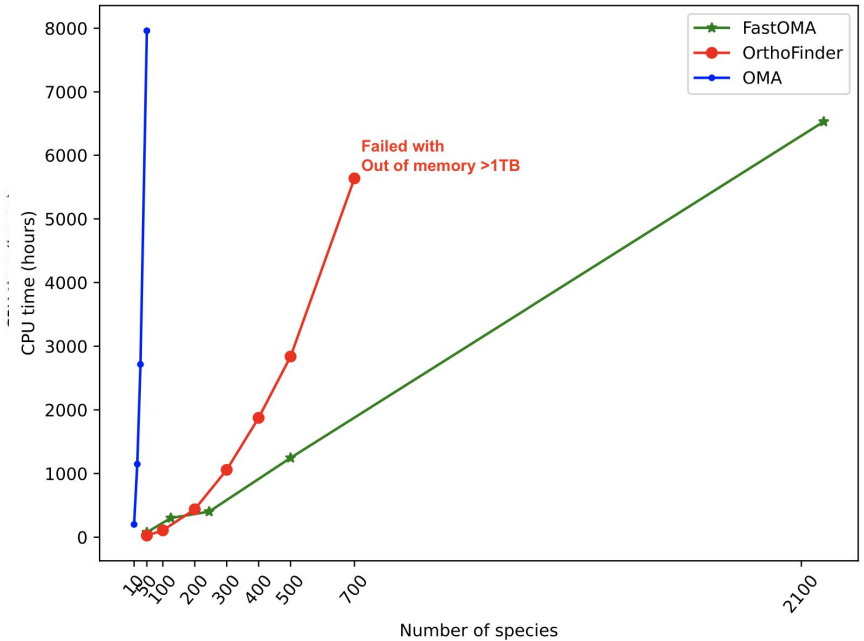




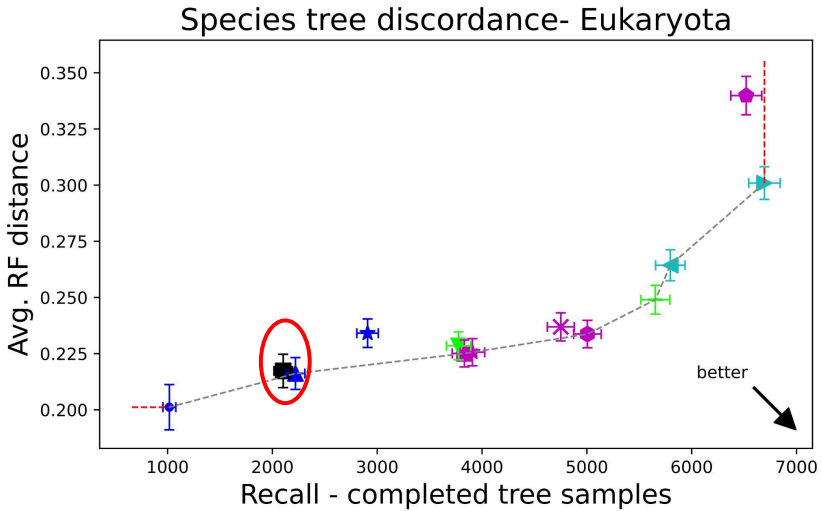
Orthology inference with FastOMA

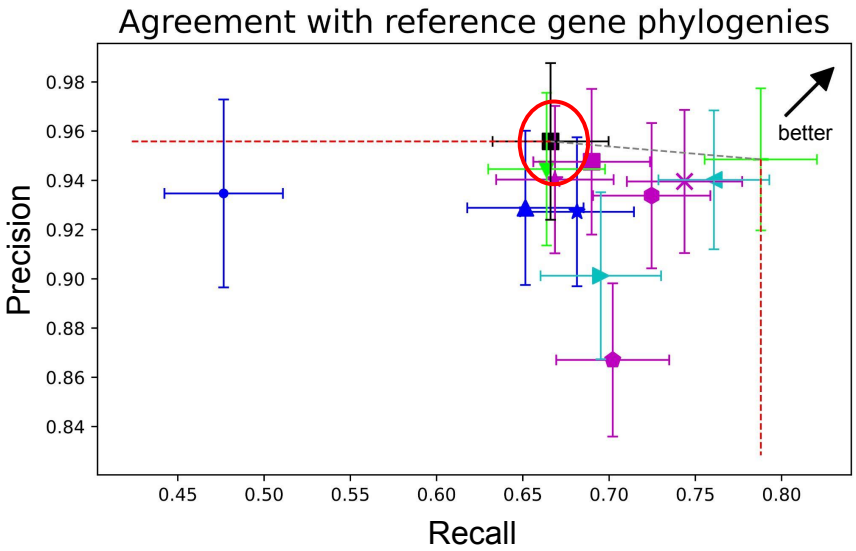
- UniProt reference proteomes
- 2180 eukaryotic species
- in a single day using 300 CPUs

github.com/DessimozLab/FastOMA



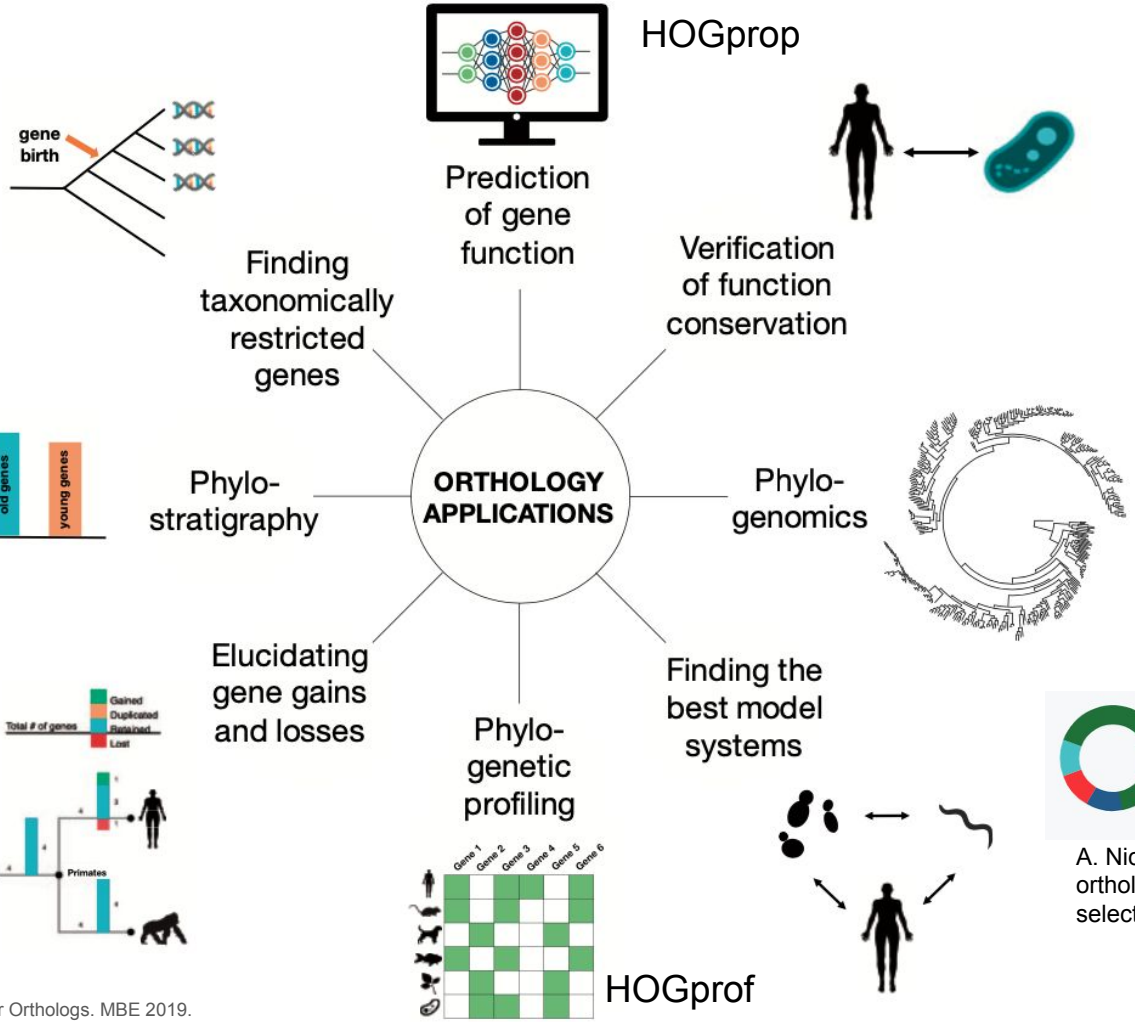
Quest for orthologs benchmarking





OMA_Pairs
OMA_Groups
OMA_GETHOGs
Domainoid+
InParanoid_Xenfix
OrthoMCL
OrthoInspector 3
sonicparanoid
PANTHER
Ensembl_Compara
Hieranoid_2
Orthofinder
FastOMA

State-of-the-art methods



OMAMO browser

A. Nicheperovich, … S. Majidian. OMAMO: orthology-based alternative model organism selection. **Bioinformatics**, 2022.

Advances and Applications in the Quest for Orthologs. MBE 2019.

Take-home messages (1/2)

- New genome assemblies are becoming available.
- Orthology is at the heart of studying evolution.
- Scalable methods are needed to handle huge amount of data.
- FastOMA is a scalable and reliable method for orthology inference.

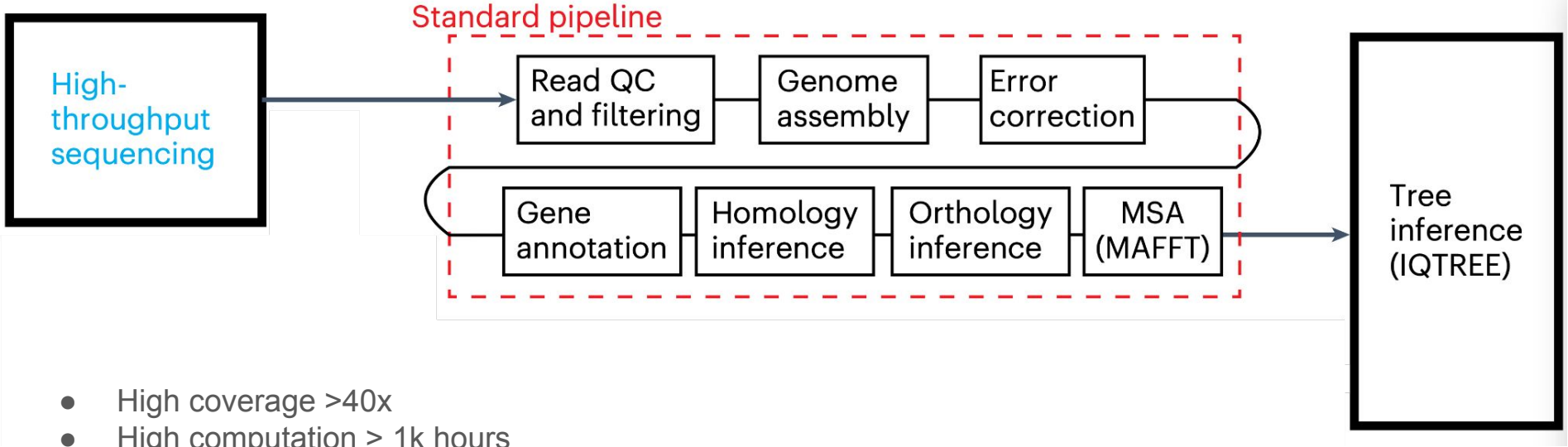
The basis of this approach is proteomes.

What if there is no high quality assembly?

Outline

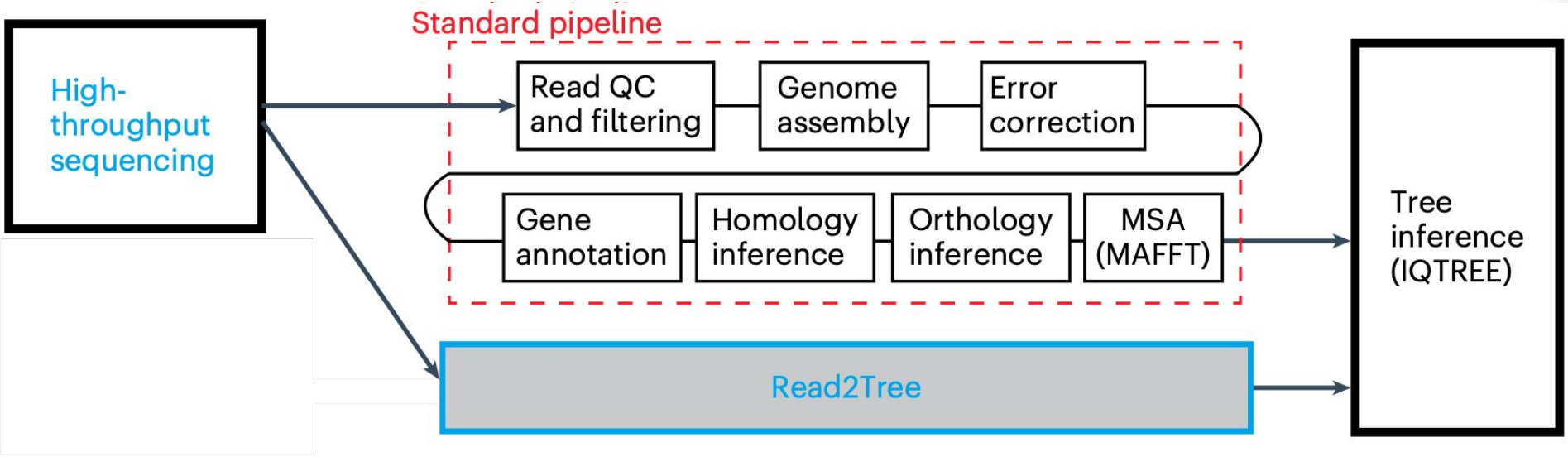
- Introduction
- Orthology inference with FastOMA
- Phylogeny inference from raw sequencing reads with read2tree

Phylogenetic tree: current status



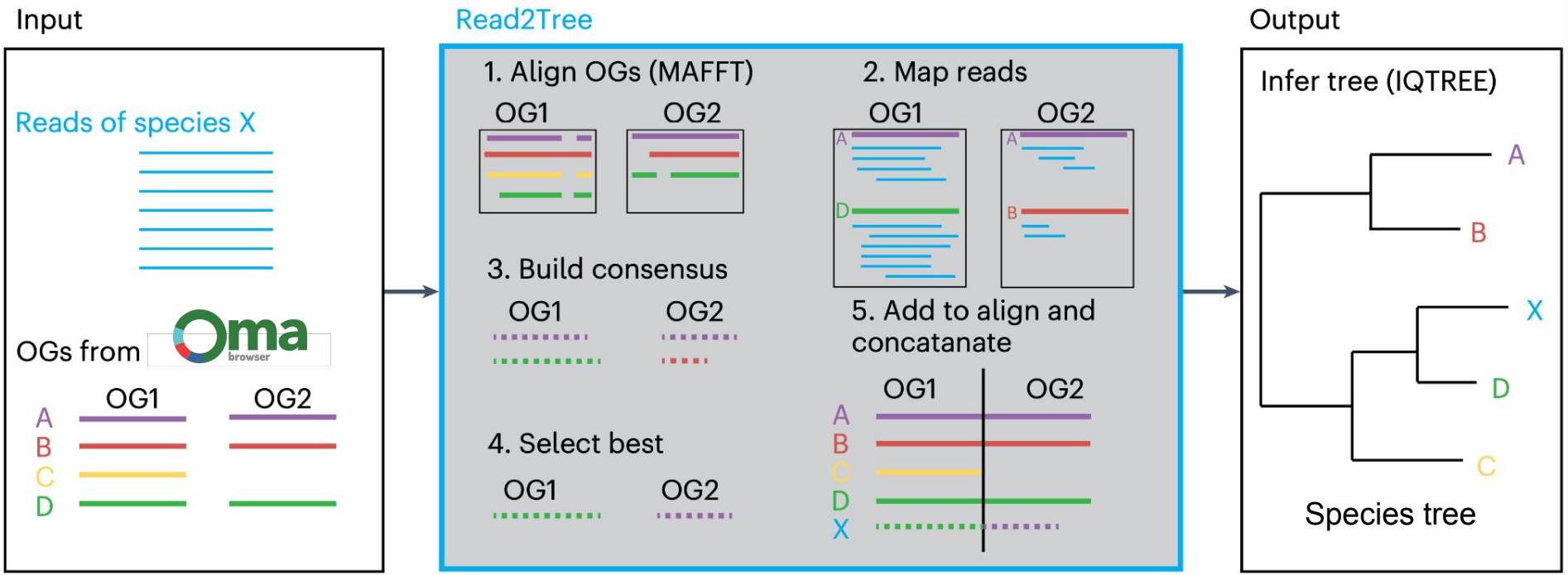
- High coverage >40x
- High computation > 1k hours
- High cost
- Orthogonal techs
- Expertise to close gaps, spots errors/contamination, annotation

Read2Tree: A faster approach



D. Dylus, A. Altenhoff, S. Majidian, FJ Sedlazeck, C. Dessimoz. Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. Nature Biotechnology, 2023.

Read2Tree: How it works?



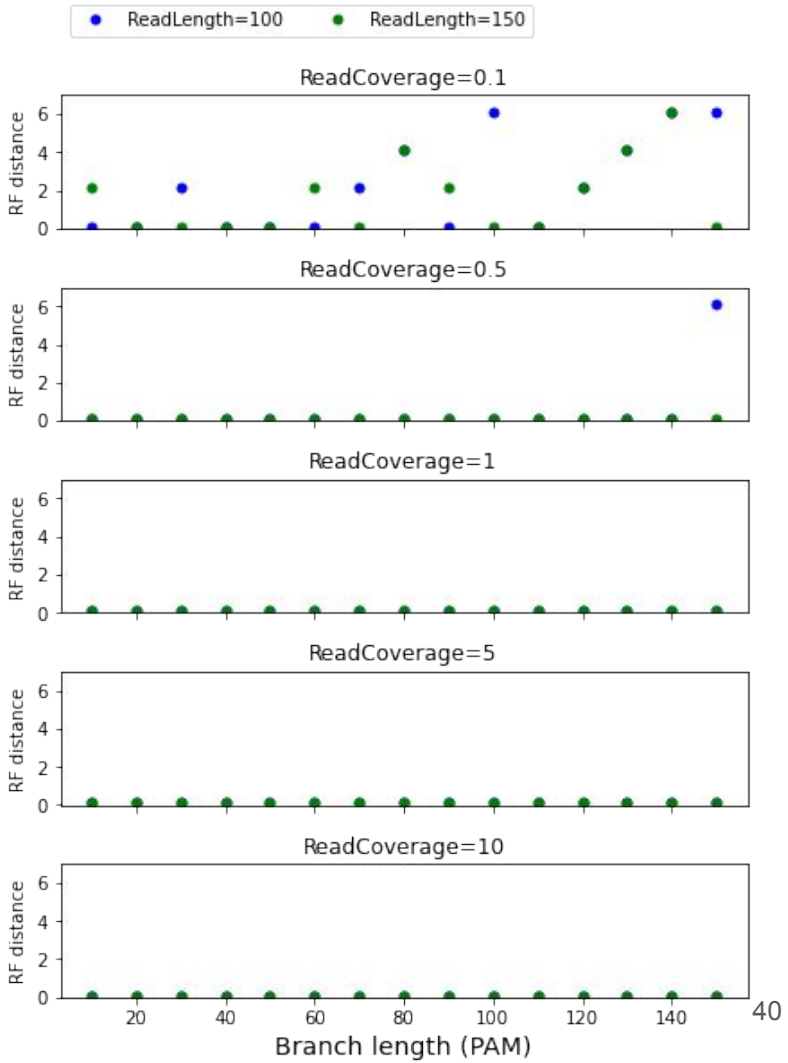
OG: **O**rthologous **G**roups (marker genes)

Read2Tree: inferring simulated species tree

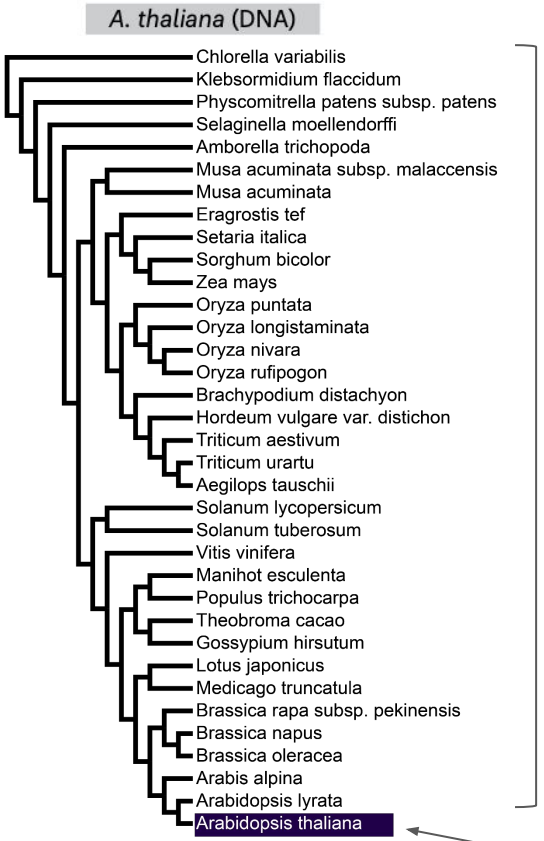
- 15 genomes containing 100 genes
- Simulated illumina DNA reads
  - Coverage values 0.1-10

- OG set based on 14 species
- Inferring the tree for one species

- Comparing inferred tree with true one
  - in terms of Robinson–Foulds (RF) distance
- 15 different branch length values



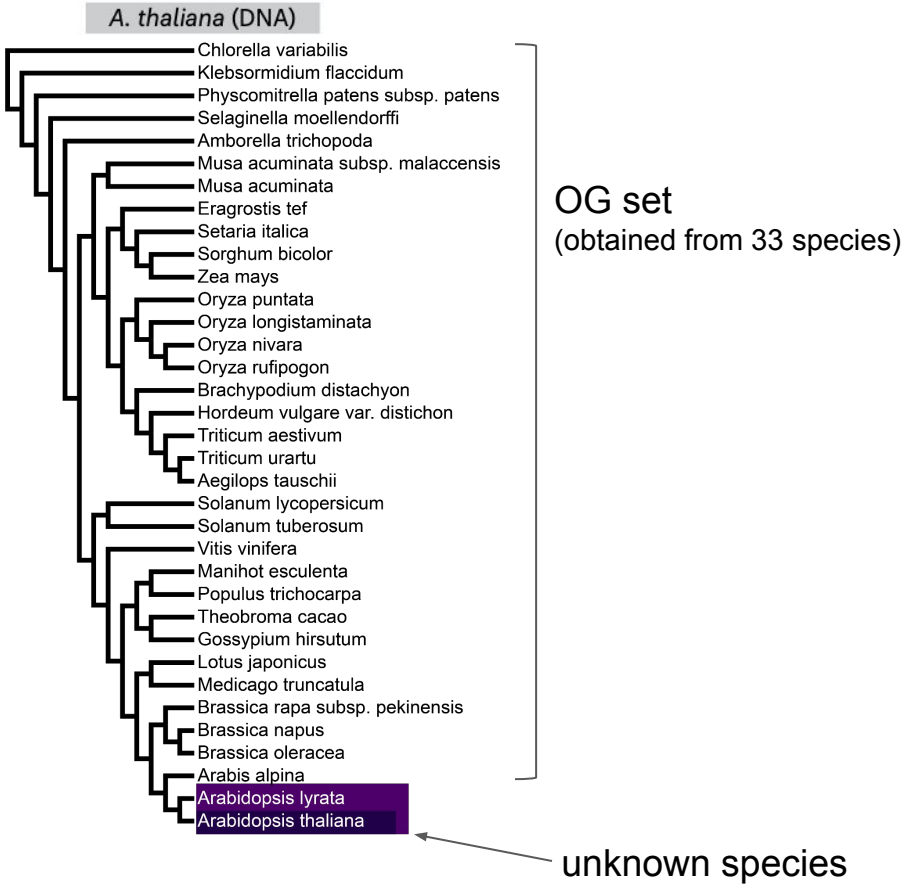
Read2Tree: Benchmarking



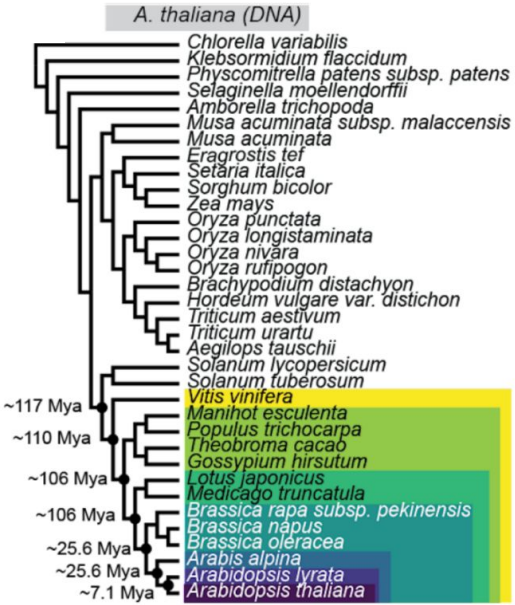
OG set
(obtained from 34 species)

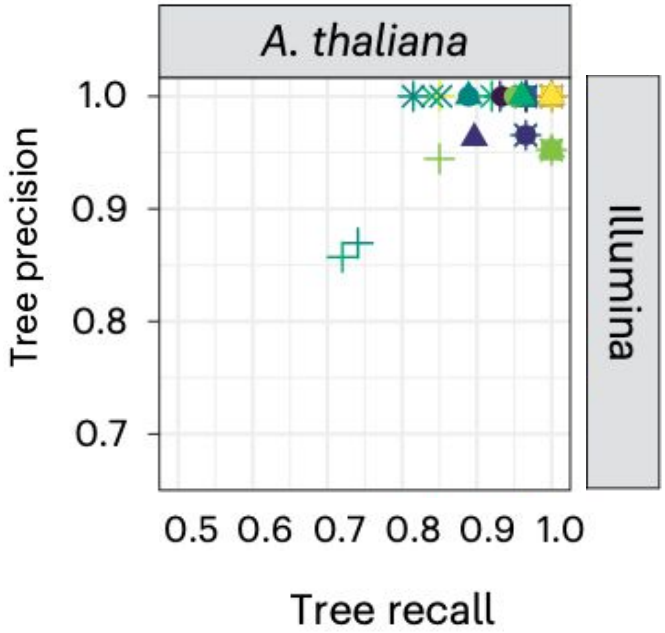
unknown species (sequencing read dataset)

Read2Tree: Benchmarking



Reconstructed tree accuracy





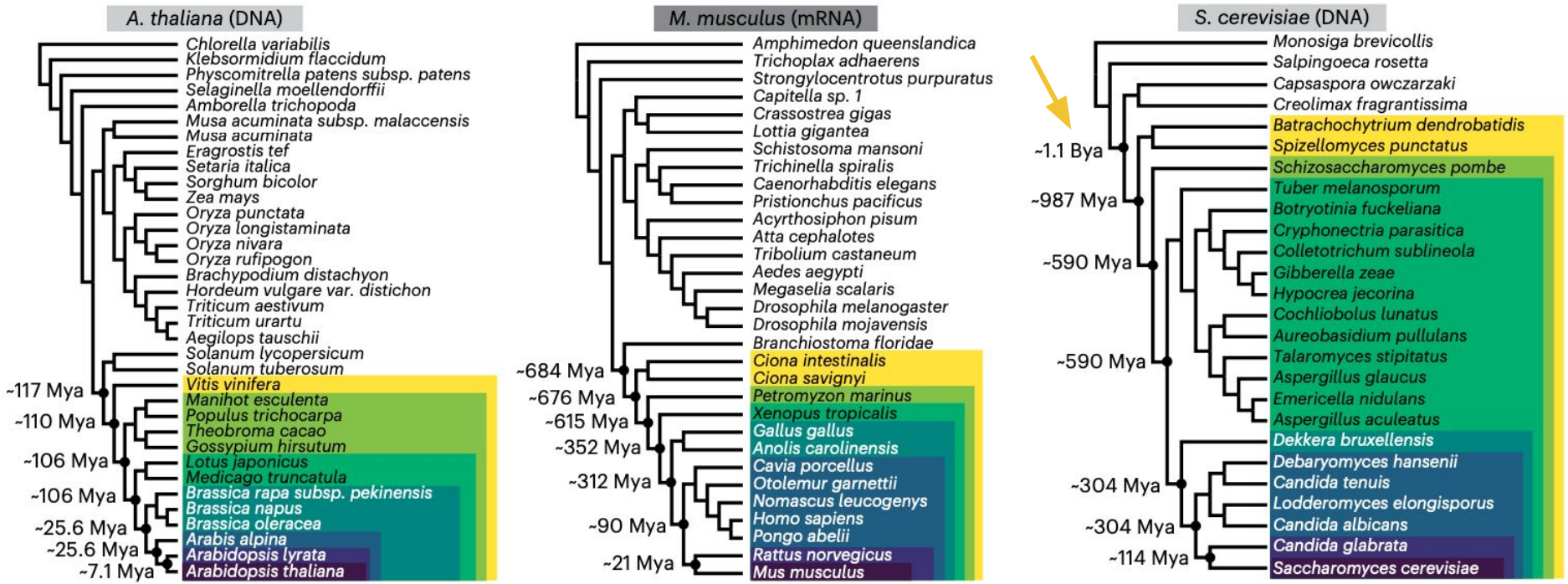
Sequencing coverage

- 0.2
- 0.5
- 1
- 5
- 10
- 20

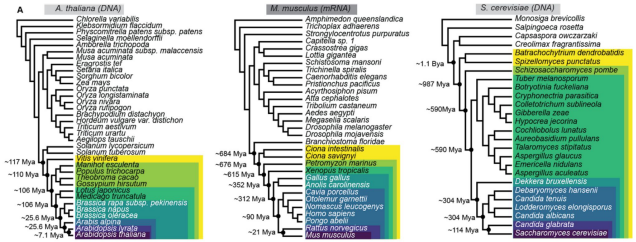
Species removal level

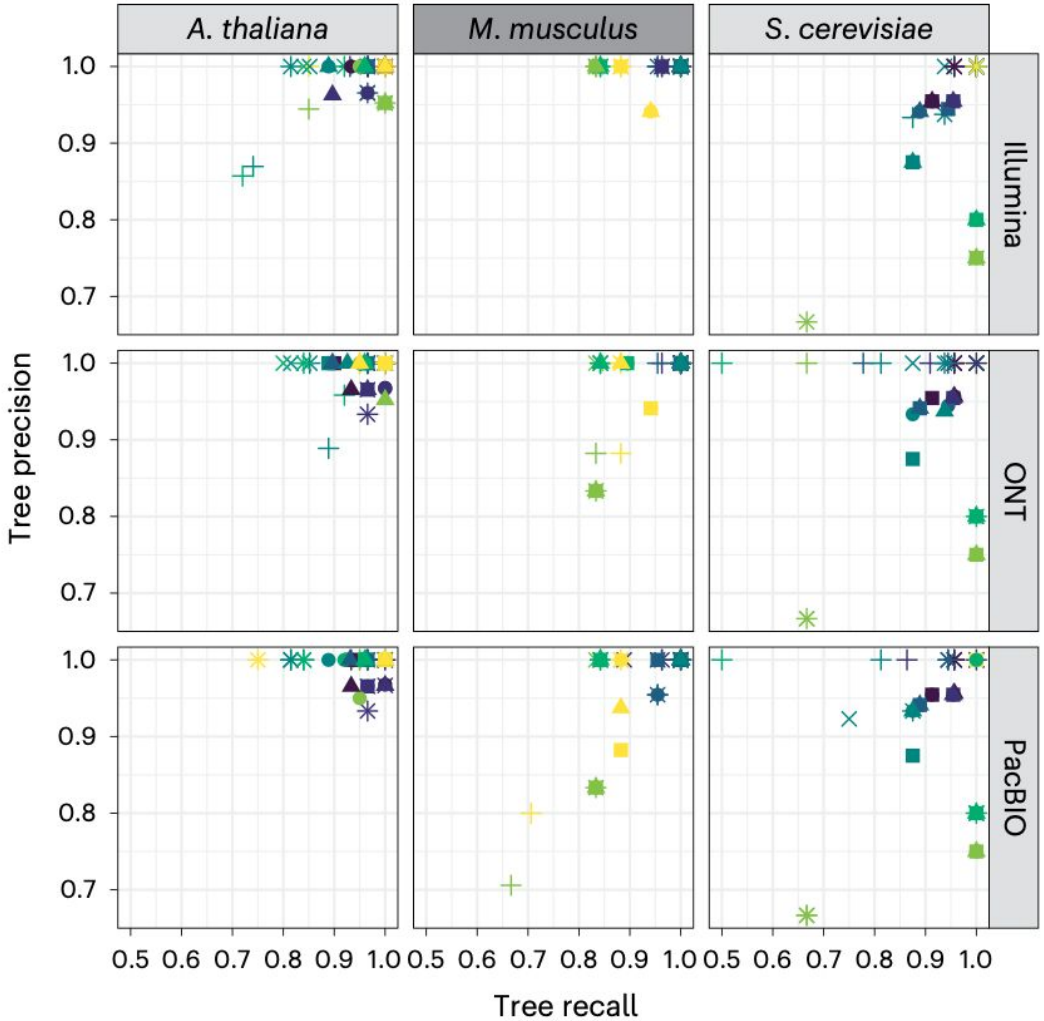
- 0
- 1
- 2
- 3
- 4
- 5
- 6

Read2Tree: Benchmarking

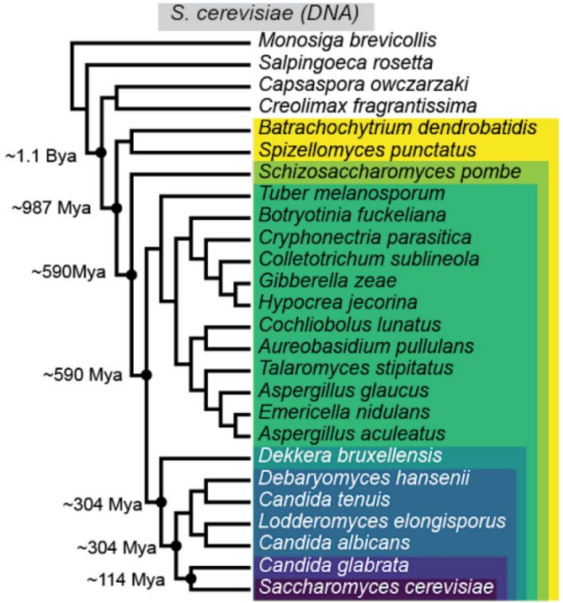


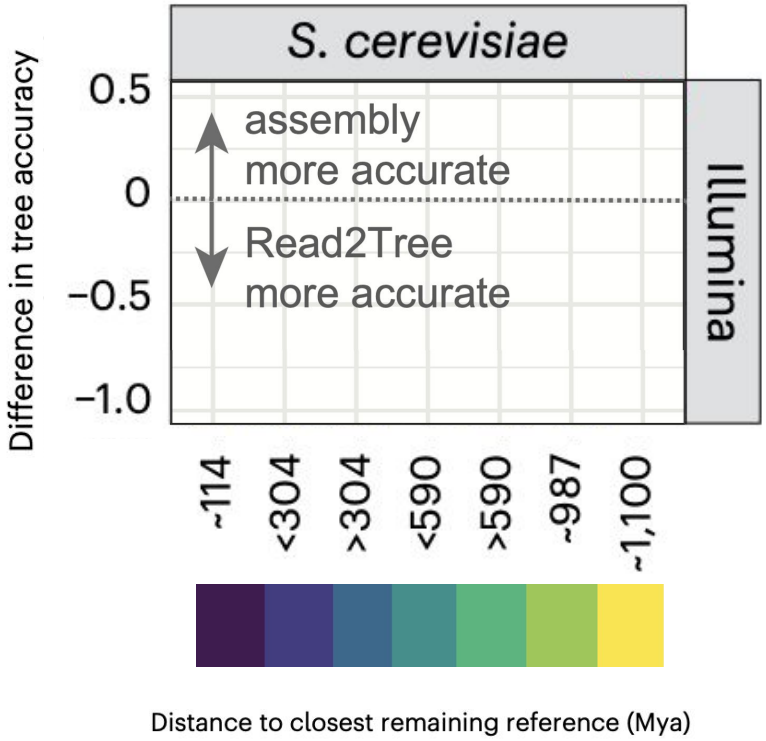
Reconstructed tree accuracy



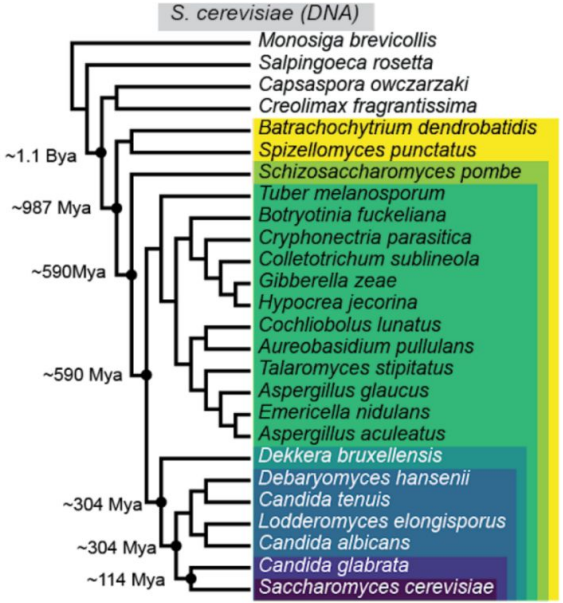


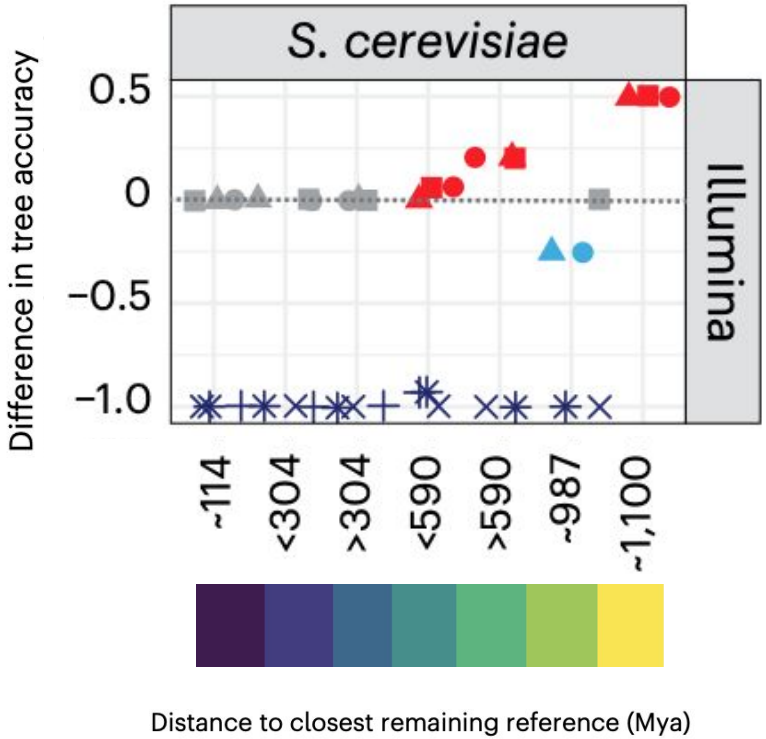
Read2Tree more accurate than assembly for tree inference





Read2Tree more accurate than assembly for tree inference

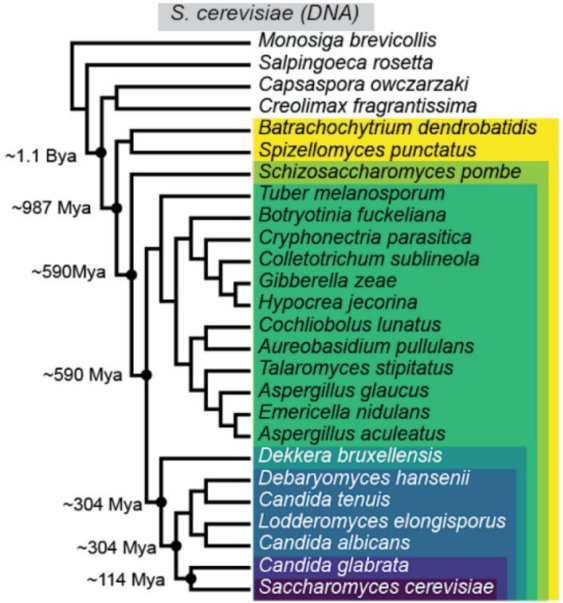


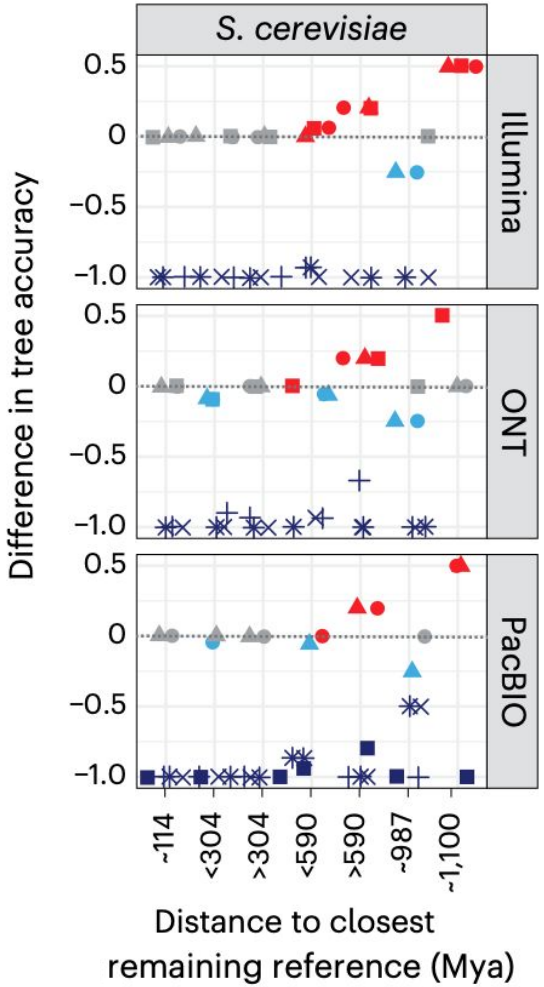


- Assembly better
- Equal
- Read2Tree better
- Only Read2Tree applicable (due to low coverage)

Sequencing coverage: +0.2 ✳1 ●10 ▽60 ×0.5 ■5 ▲20

Read2Tree more accurate than assembly for tree inference

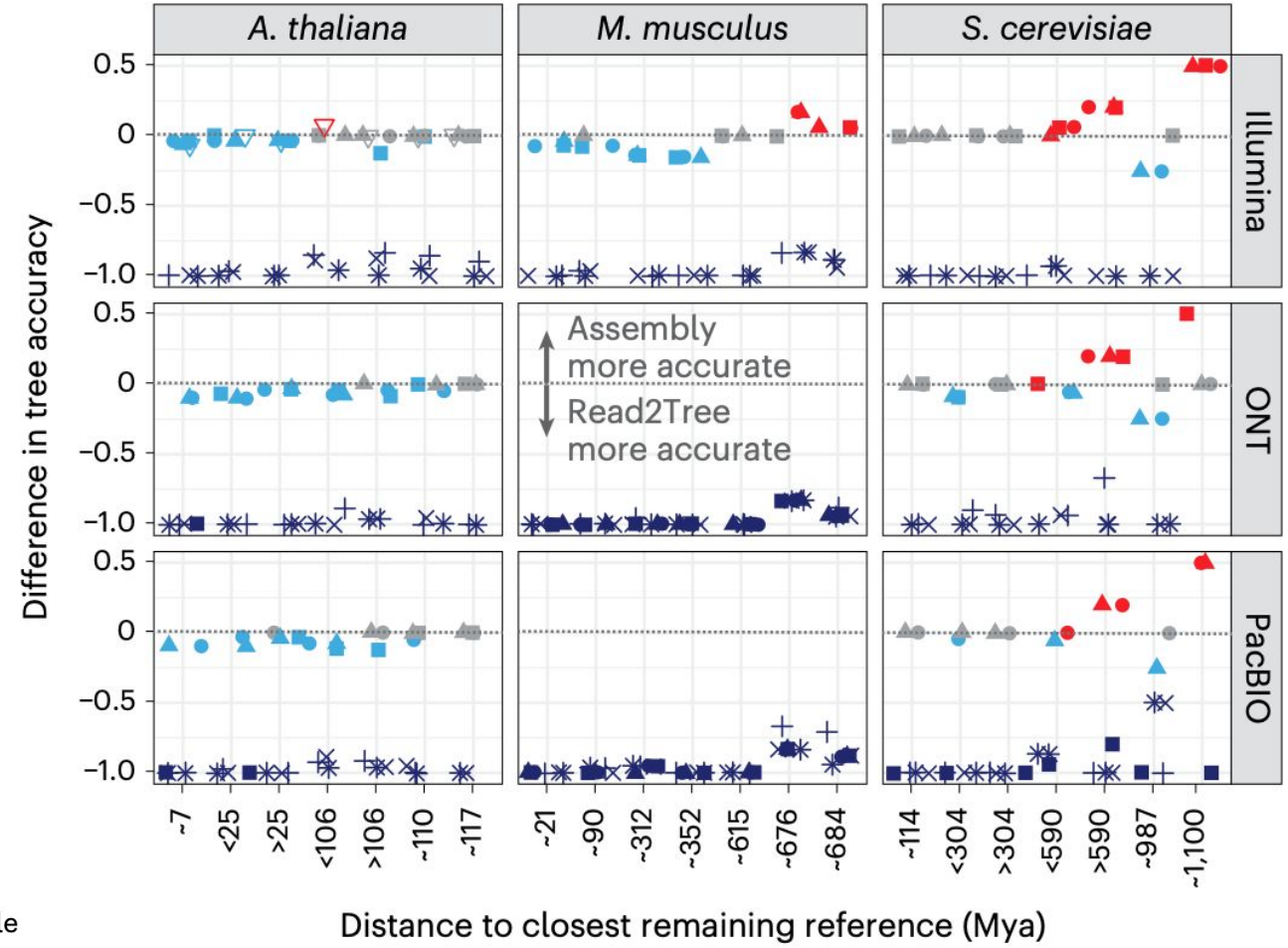




- Assembly better
- Equal
- Read2Tree better
- Only Read2Tree applicable (due to low coverage)

Sequencing coverage +0.2 ✳1 ●10 ▽60 ×0.5 ■5 ▲20

Read2Tree more accurate than assembly for tree inference

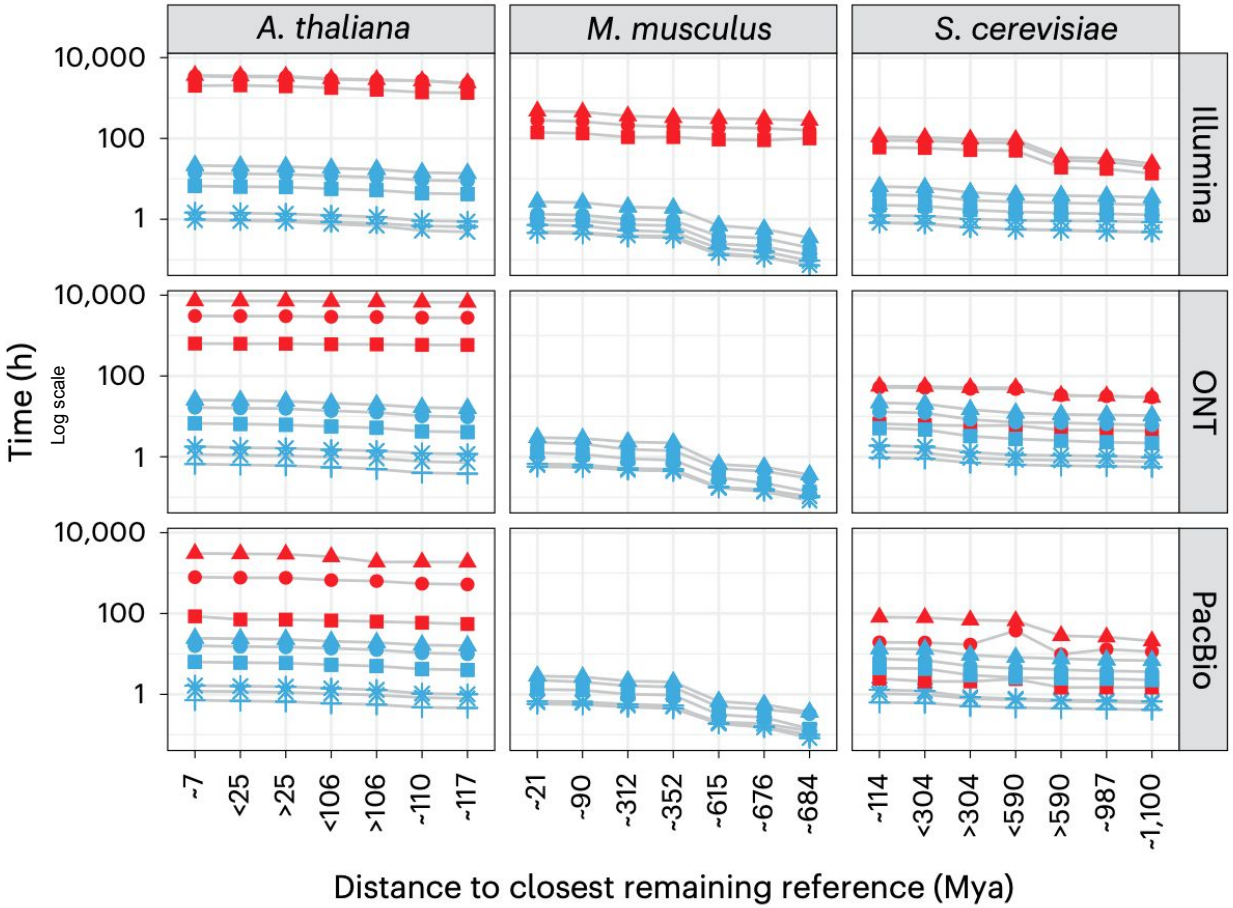


Sequencing coverage: +0.2 ✳1 ●10 ▽60 ×0.5 ■5 ▲20

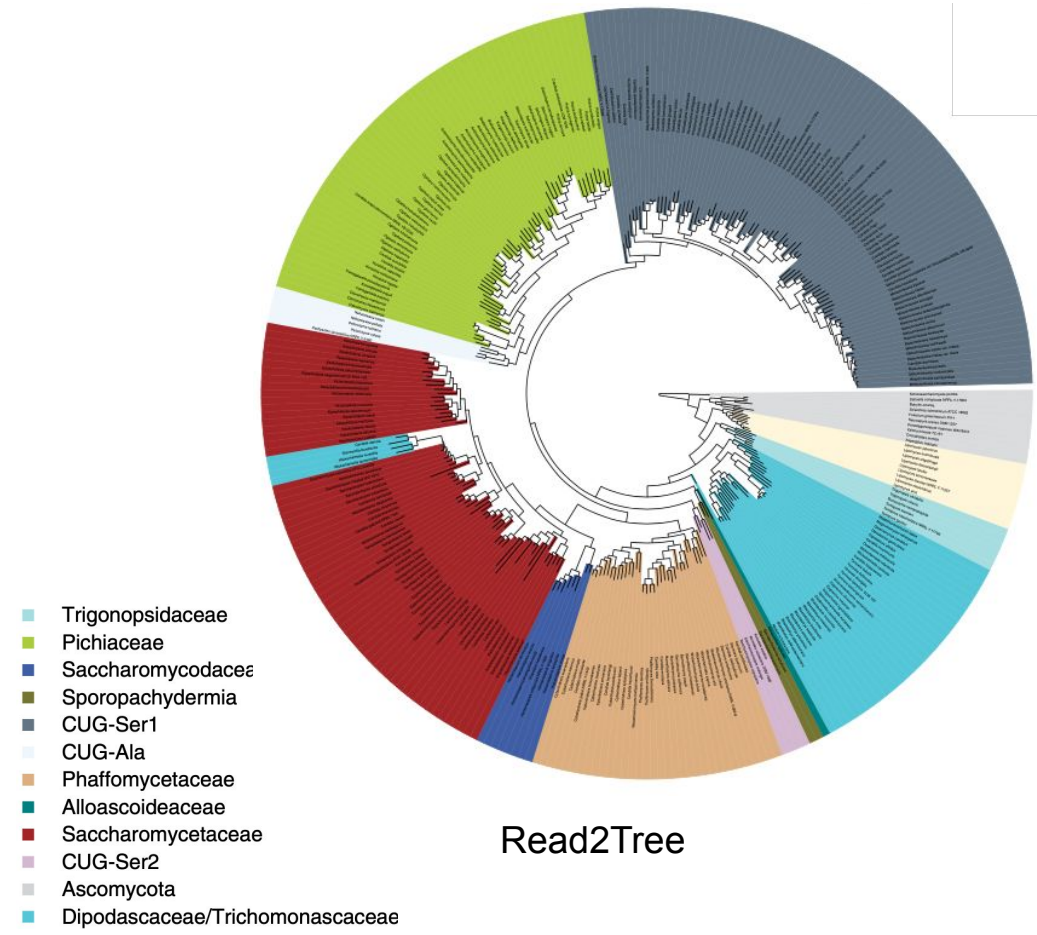
- Assembly better
- Read2Tree better
- Equal
- Only Read2Tree applicable

MASH results are not shown.

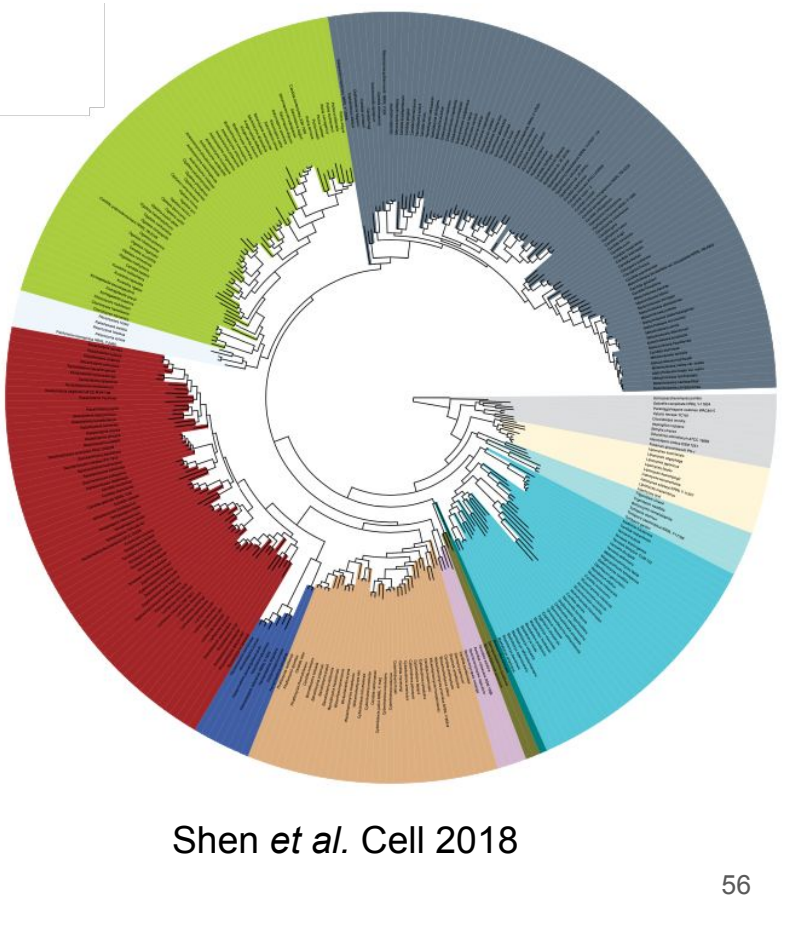
Read2Tree is orders of magnitude faster



Read2Tree reproduces state-of-the-art yeast phylogeny



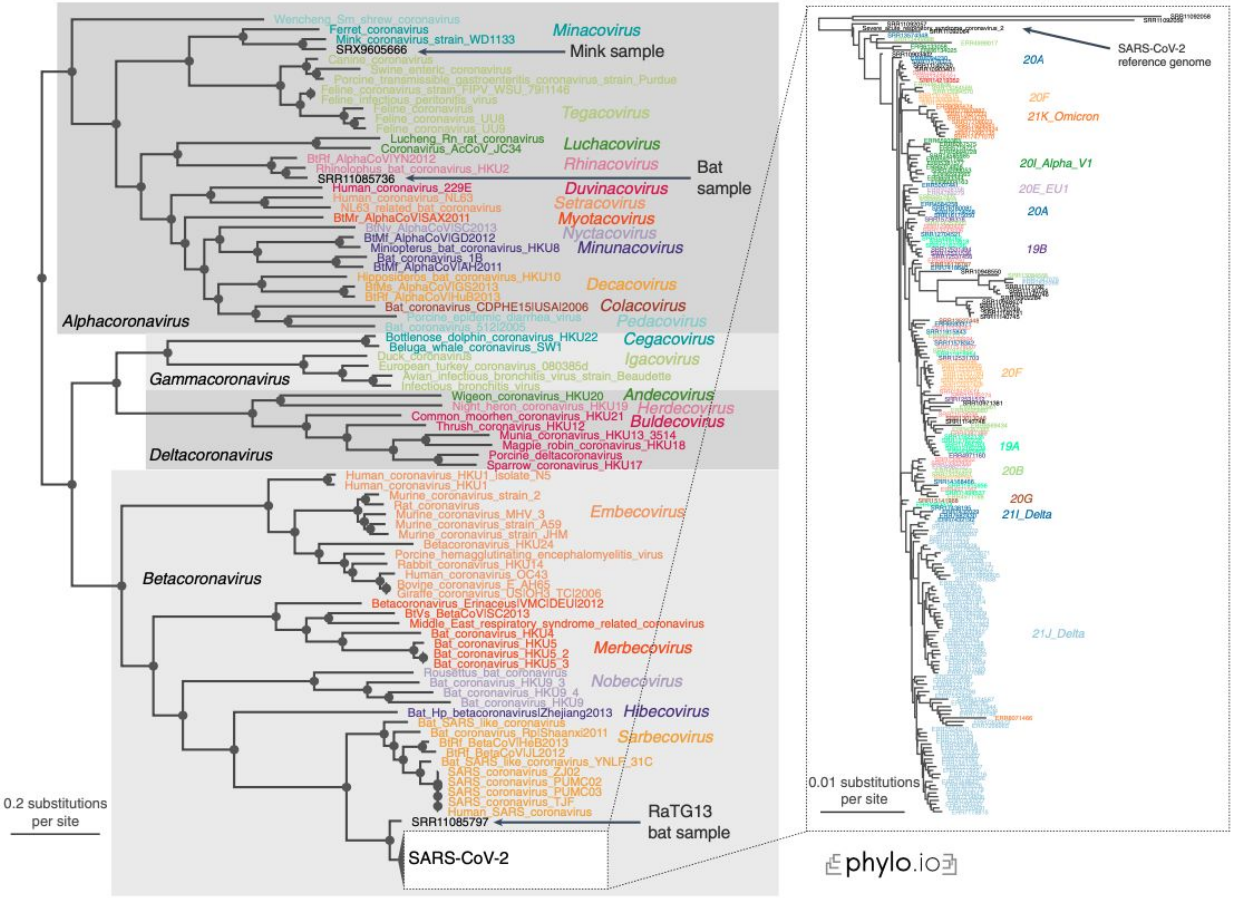
Read2Tree



Shen *et al.* Cell 2018

Read2Tree: Coronavirus

- Rapid identification and placement of viruses
- Inferring tree using raw sequencing data
  - Data download takes most time!
- Main genera
  - Alpha-, Beta-, Gamma- and Delta-
  - subgenera



Read2Tree: Summary

nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01753-4

Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree

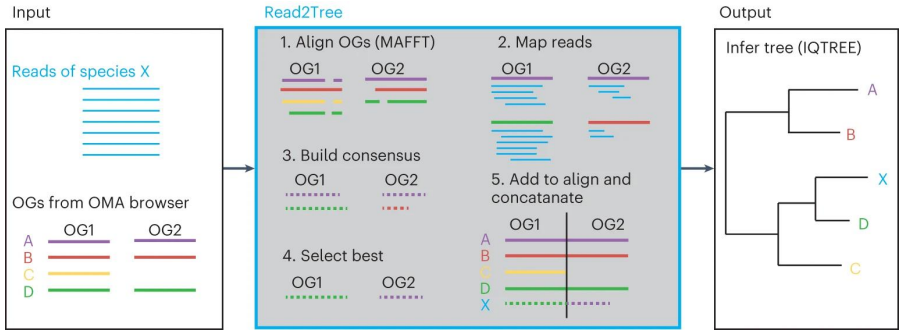
Received: 18 April 2022

Accepted: 16 March 2023

David Dylus[1,2,8], Adrian Altenhoff[2,3], Sina Majidian[1,2], Fritz J. Sedlazeck[4,5] & Christophe Dessimoz[1,2,6,7]

- Rapid phylogenetic tree reconstruction
- Scales from low coverage to large numbers of samples
- Eases comparative genomics from small to large labs
  - Potentially removes biases along the way
  - Low coverage

github.com/DessimozLab/read2tree



- **Future directions:**
  - Metagenomics samples (multiple samples)
  - Single cell genomics applications

Thank you !

comparative genomics LAB

UNIL | Université de Lausanne

SIB Swiss Institute of Bioinformatics

Zürcher Hochschule für Angewandte Wissenschaften

FNSNF Swiss National Science Foundation

zh aw

Baylor College of Medicine

HGSC HUMAN GENOME SEQUENCING CENTER