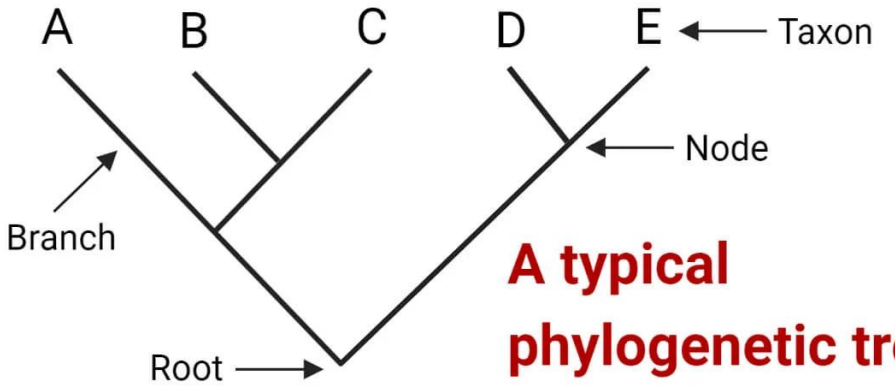
Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree

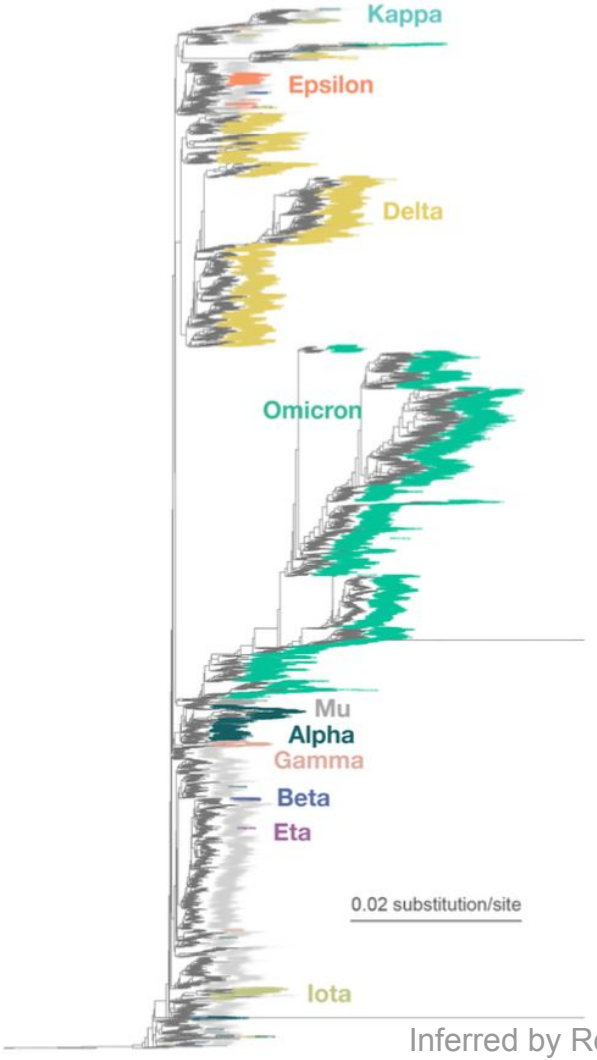
Sina Majidian

June 23, 2023

Phylogenetic tree analysis

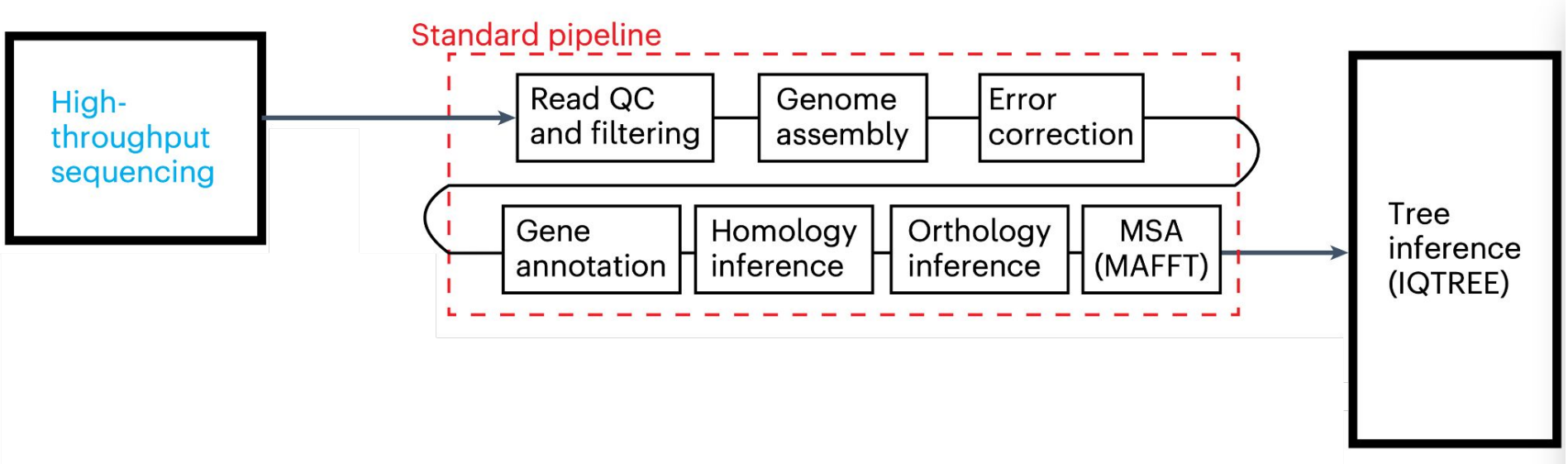


A typical phylogenetic tree



Inferred by Read2Tree

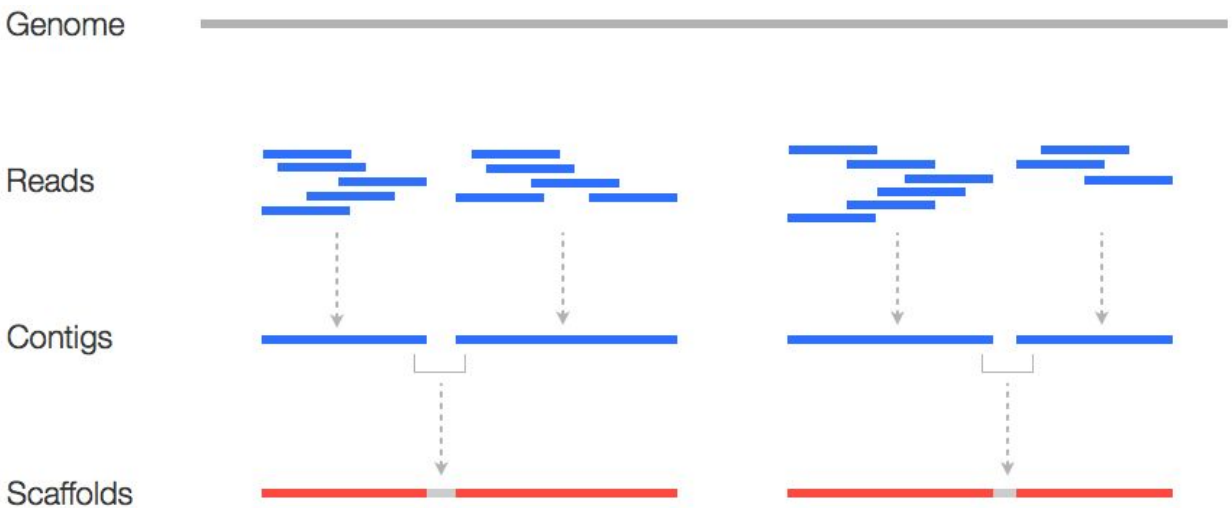
Phylogenetic tree: current status



Phylogenetic tree: current status

1. Assembly

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



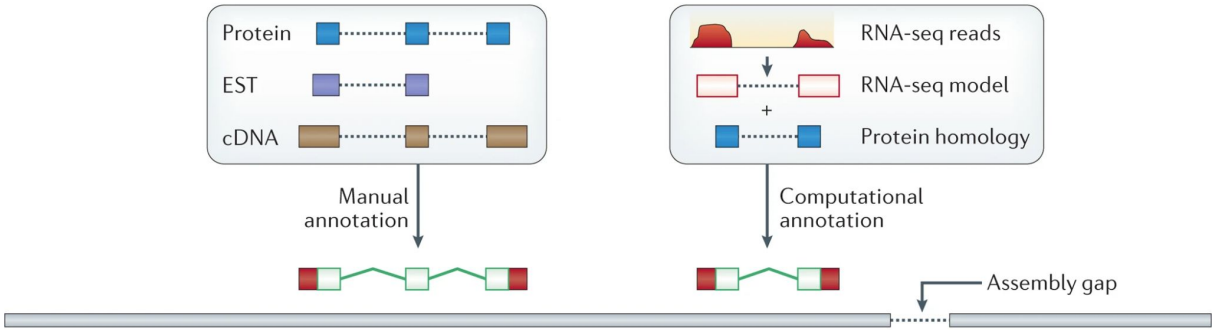
Phylogenetic tree: current status

1. Assembly
   seq., tech, coverage
2. Error correction
   often orthogonal sequencing
   (eg. Illumina)

3. Annotation

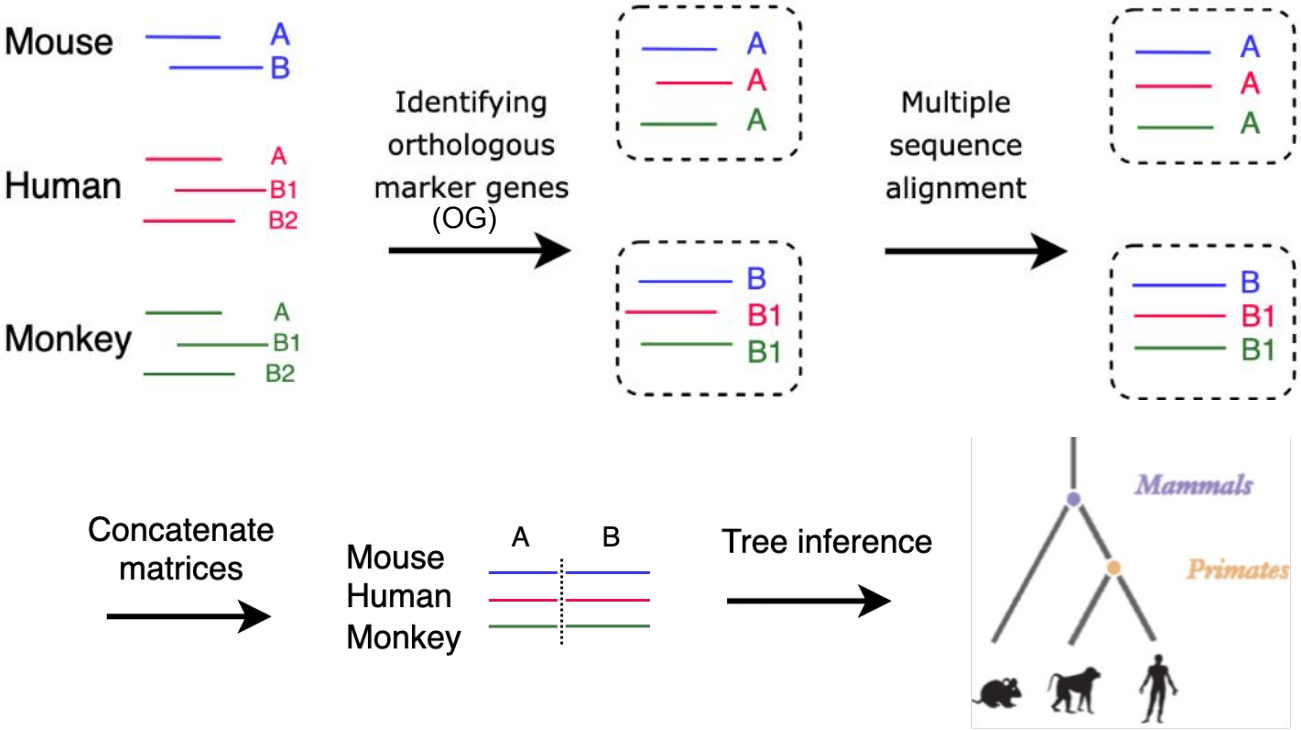
→ **proteome**

MAKER
Annotate this!

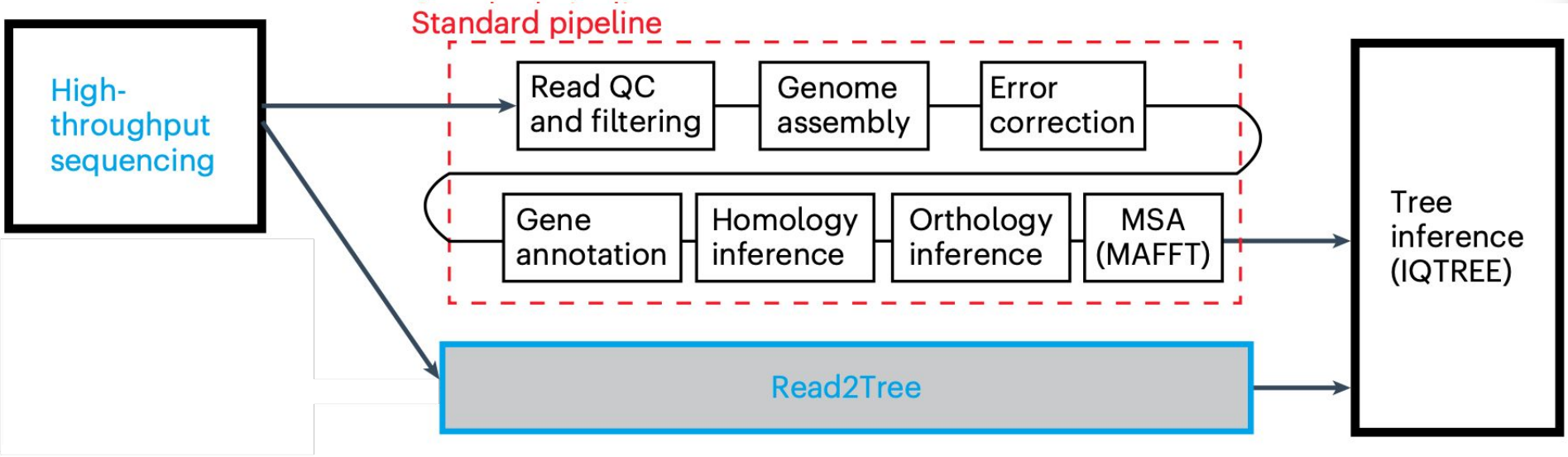


The state of play in higher eukaryote gene annotation, Nature review genetics .

Inferring species trees: the standard pipeline



Read2Tree: A faster approach

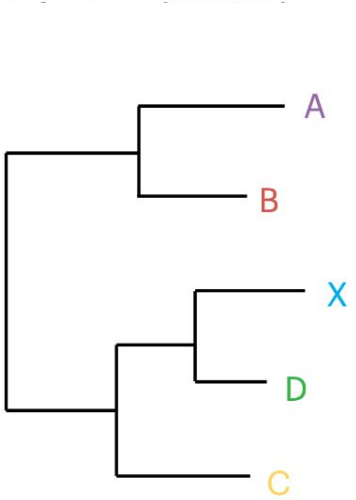


Read2Tree: How it works?

Input

Reads of species X

Output

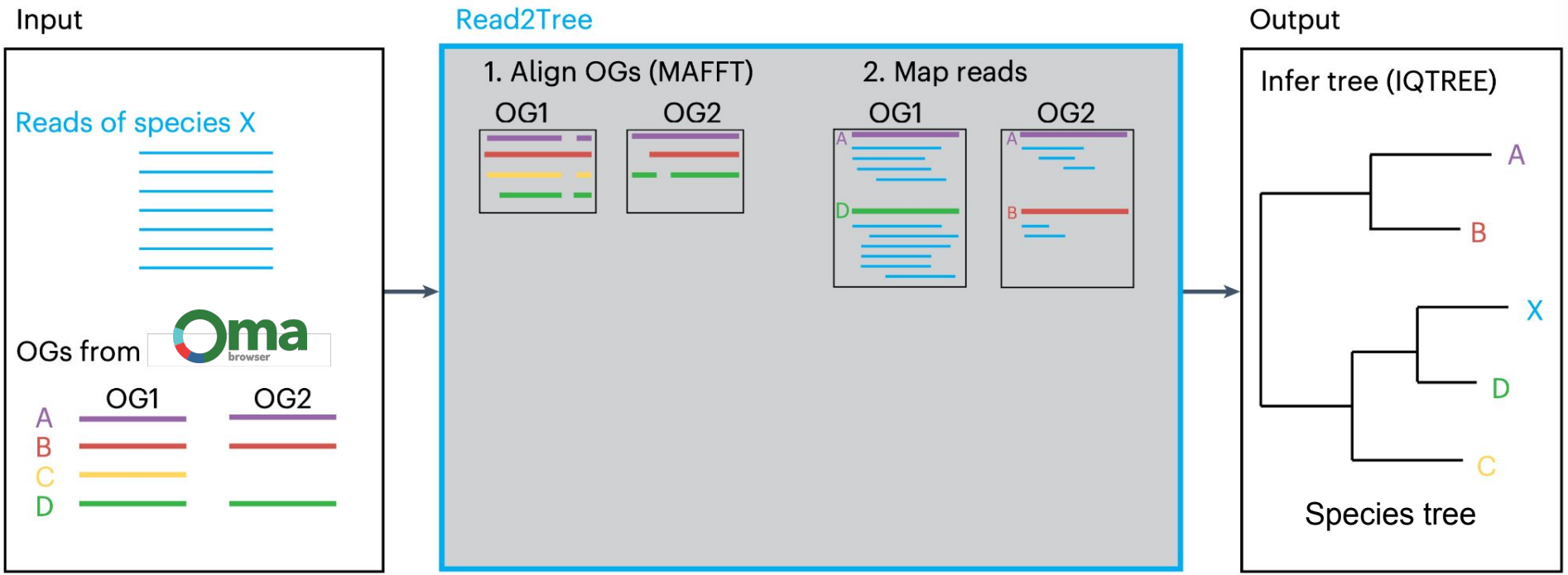


Species tree

Read2Tree: How it works?

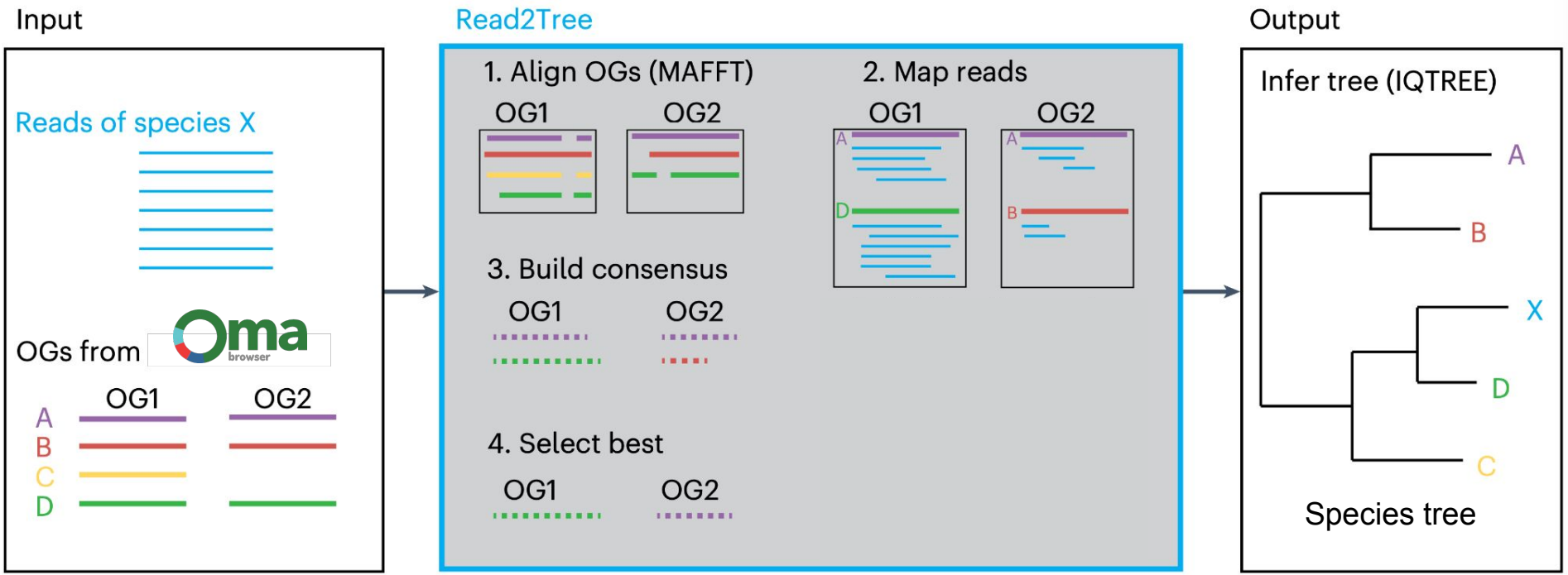
OG: **O**rthologous **G**roups (marker genes)

Read2Tree: How it works?



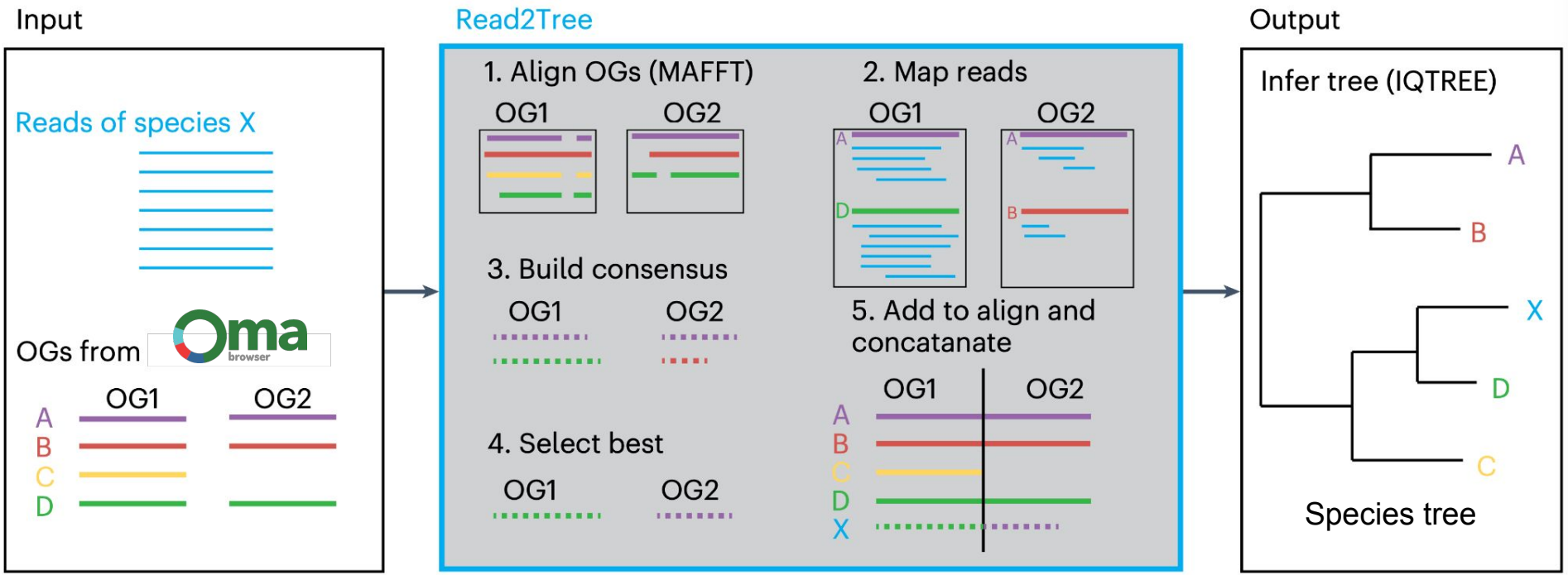
OG: **O**rthologous **G**roups (marker genes)

Read2Tree: How it works?



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

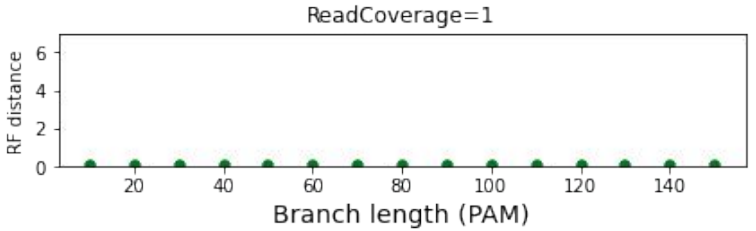
Read2Tree: How it works?



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

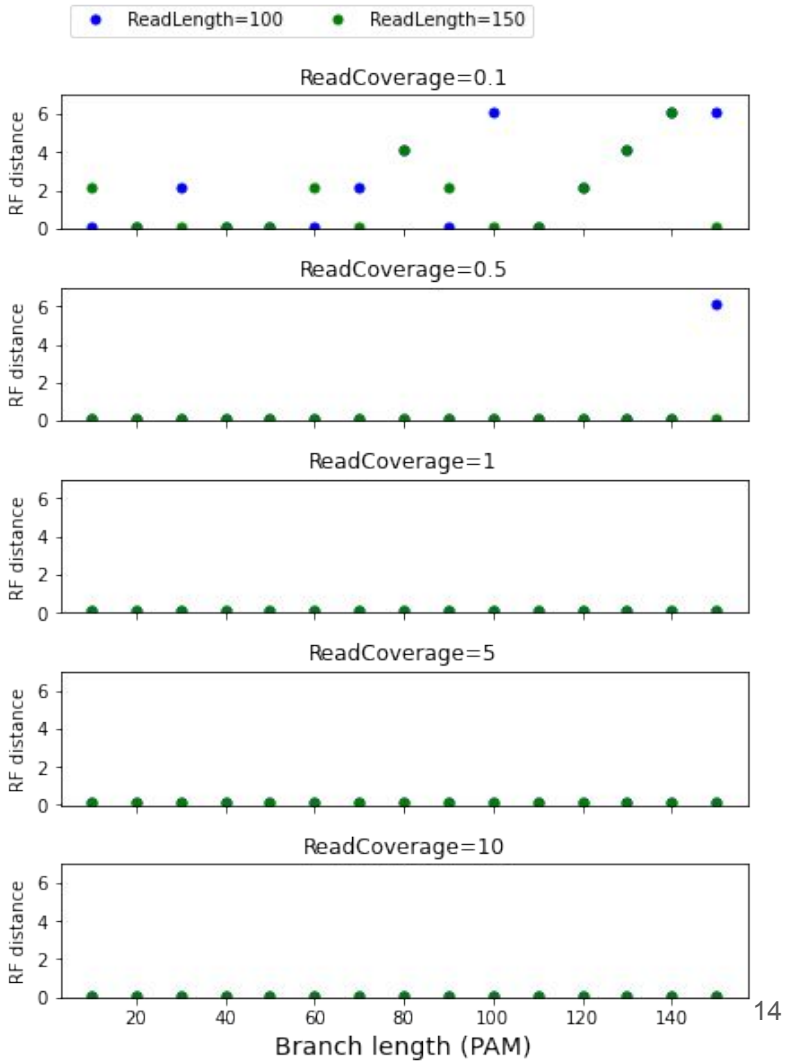
- Comparing inferred tree with true one
  - in terms of Robinson–Foulds (RF) distance
- 15 different branch length values

Read2Tree: inferring simulated species tree

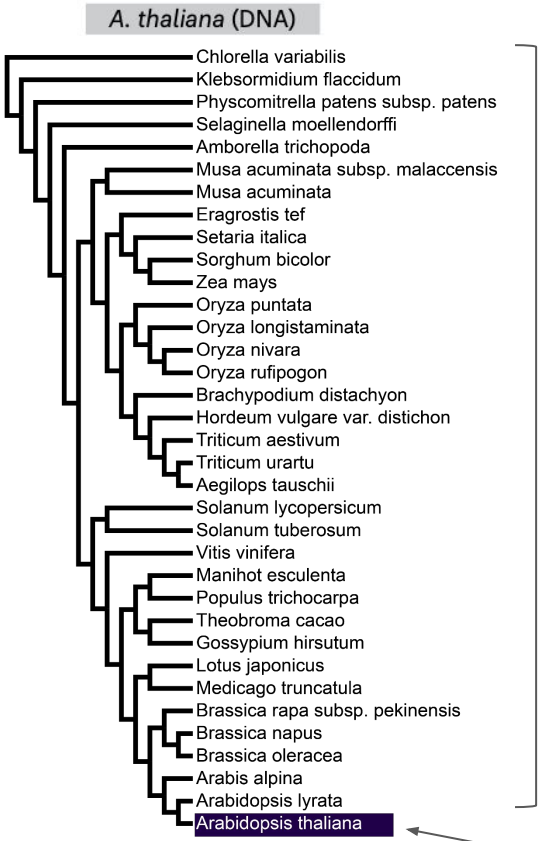
- 15 genomes containing 100 genes
- Simulated illumina DNA reads
  - Coverage values 0.1-10

- 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



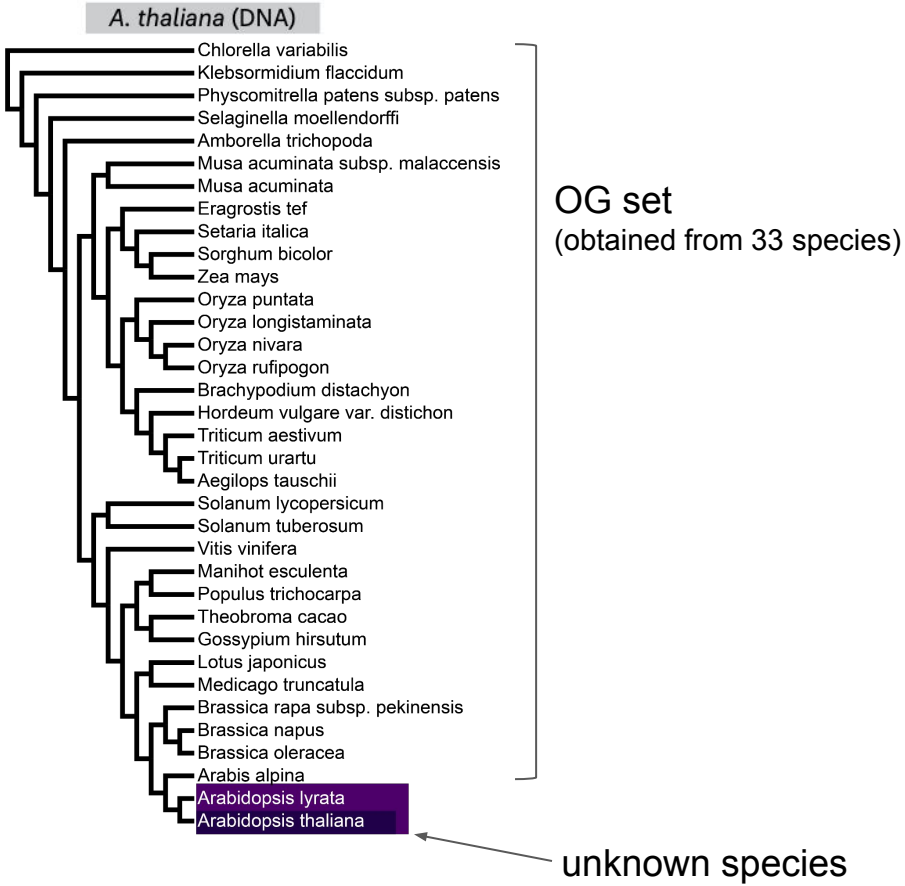
Read2Tree: Benchmarking



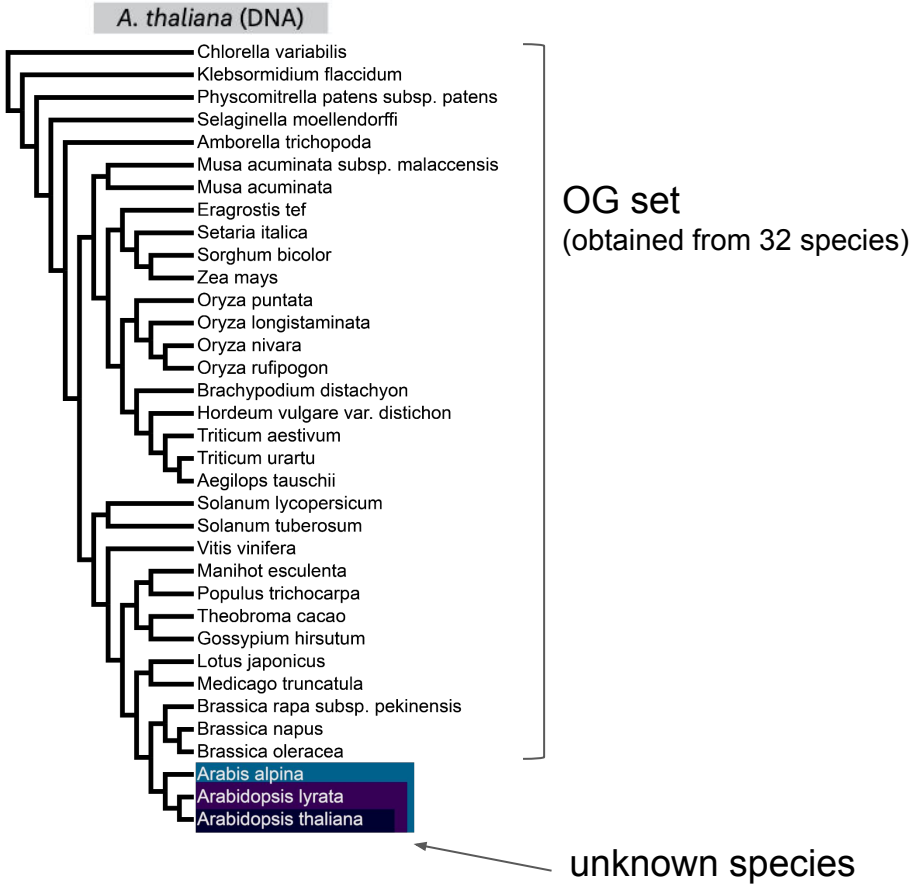
OG set
(obtained from 34 species)

unknown species (sequencing read dataset)

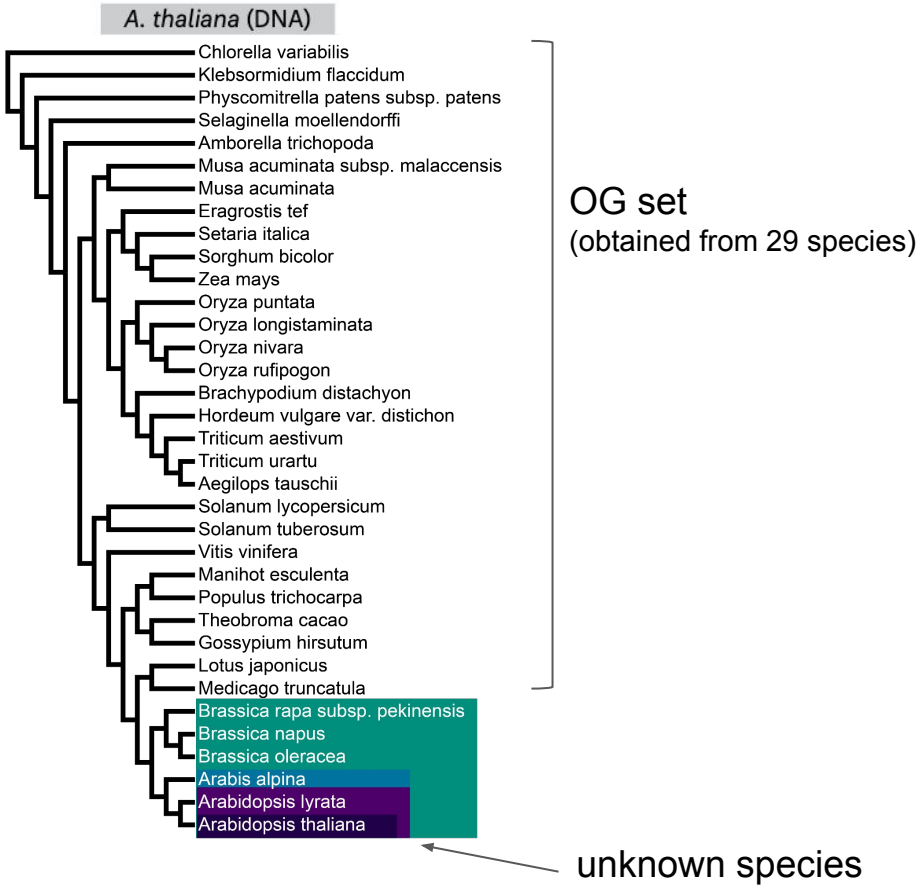
Read2Tree: Benchmarking



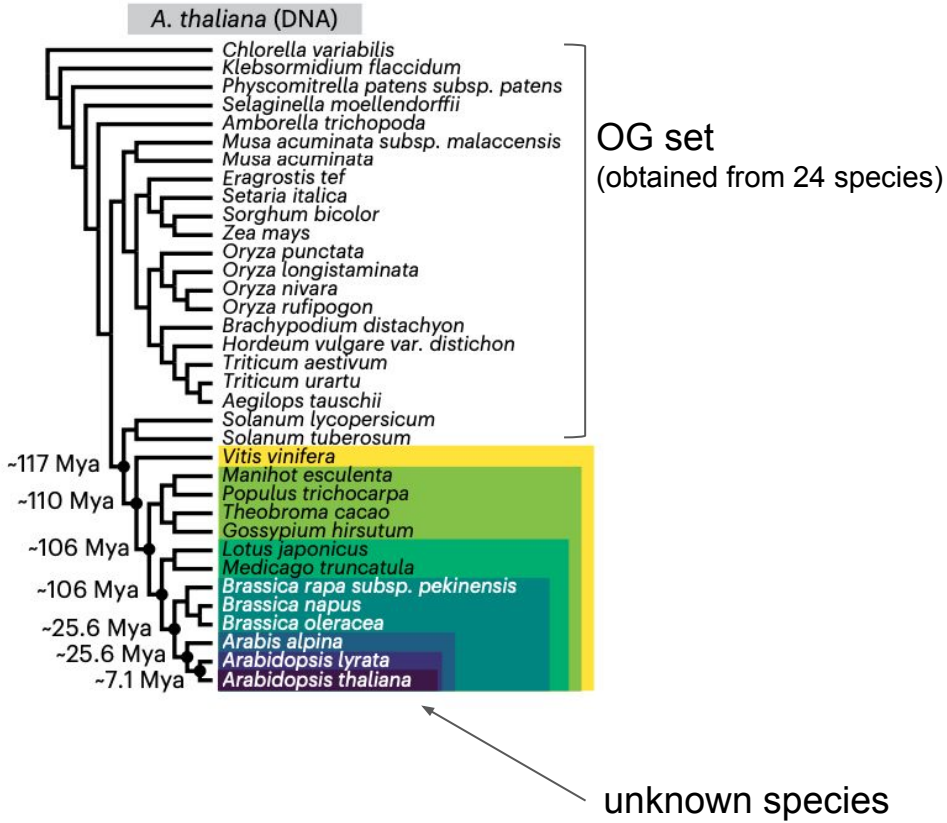
Read2Tree: Benchmarking



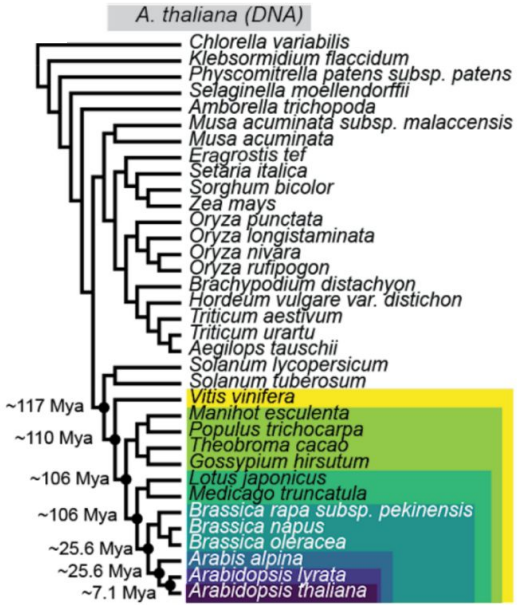
Read2Tree: Benchmarking

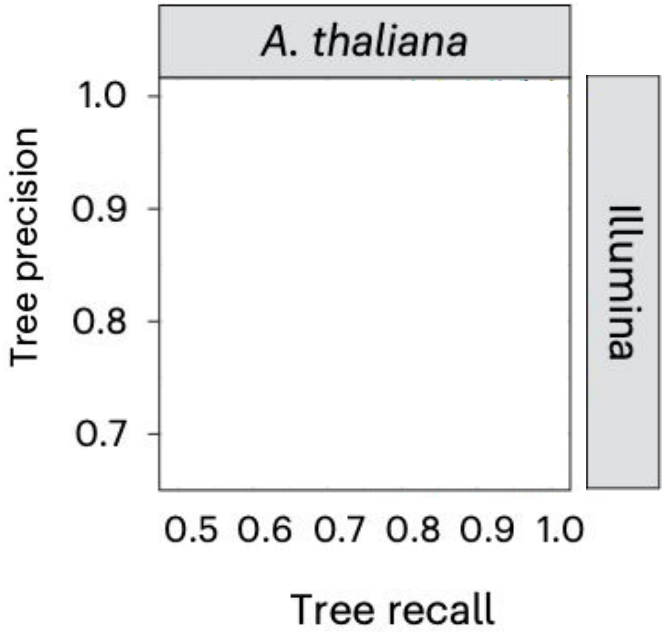


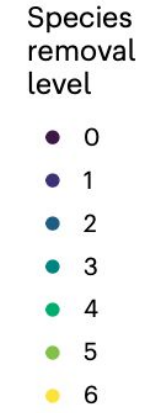
Read2Tree: Benchmarking impact of reference OGs



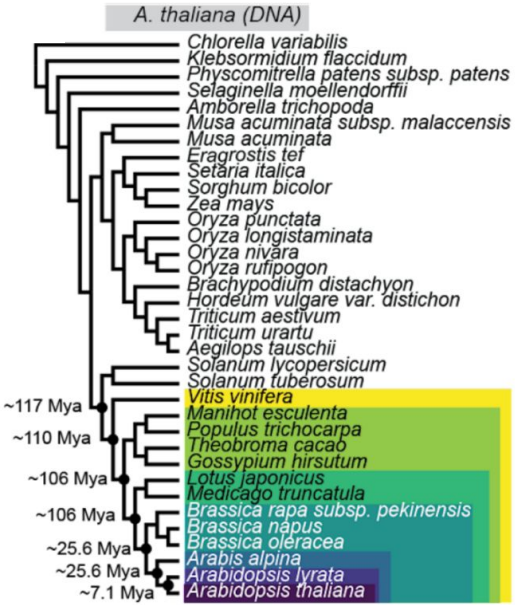
Reconstructed tree accuracy

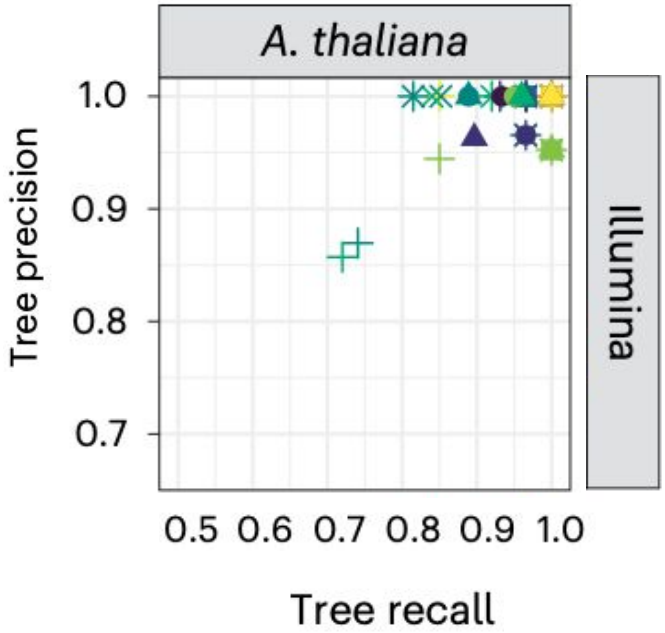






Reconstructed tree accuracy





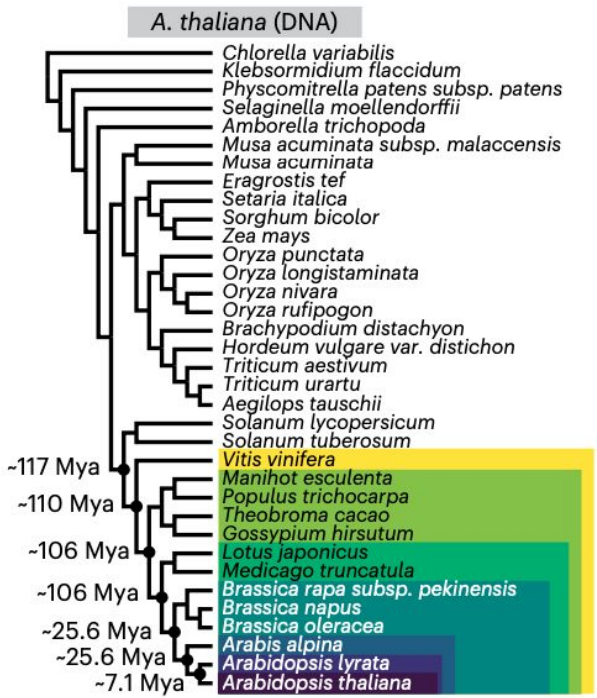
Sequencing coverage

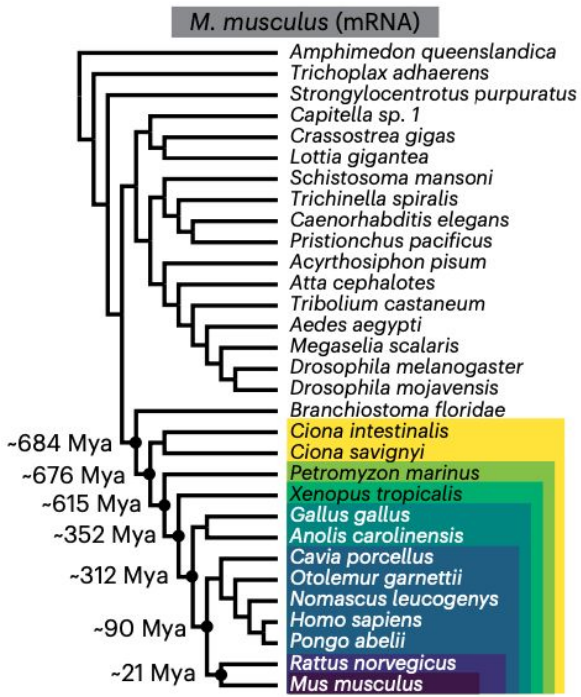
- 0.2
- 0.5
- 1
- 5
- 10
- 20

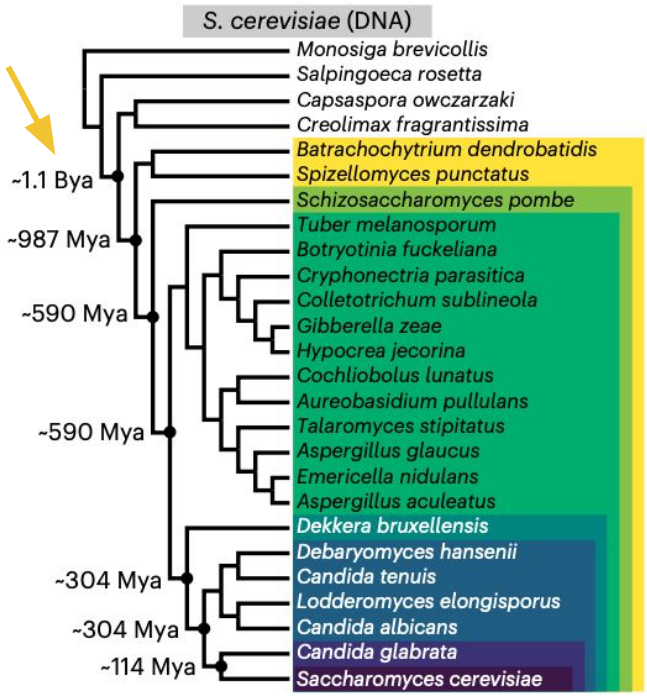
Species removal level

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

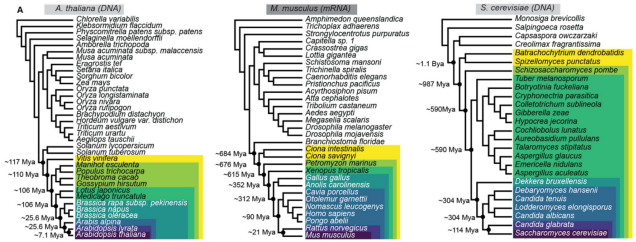
Read2Tree: Benchmarking

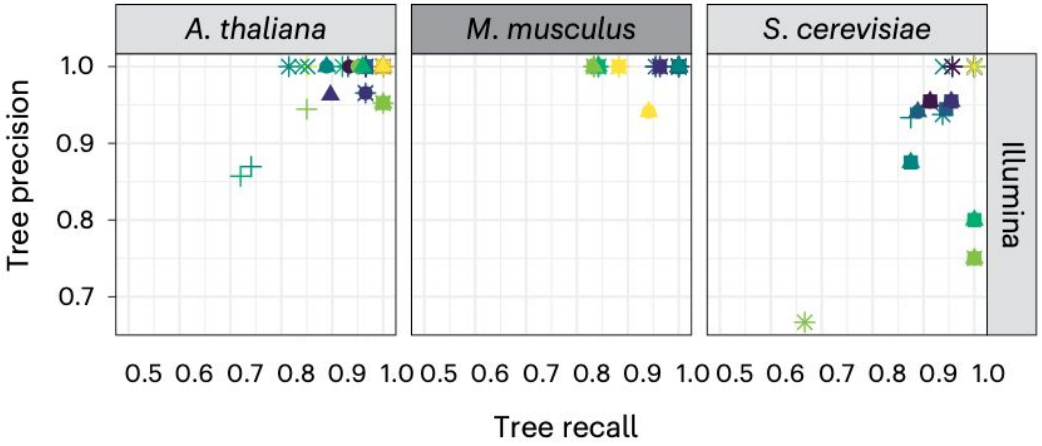




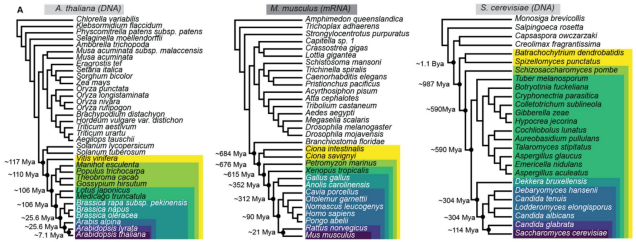


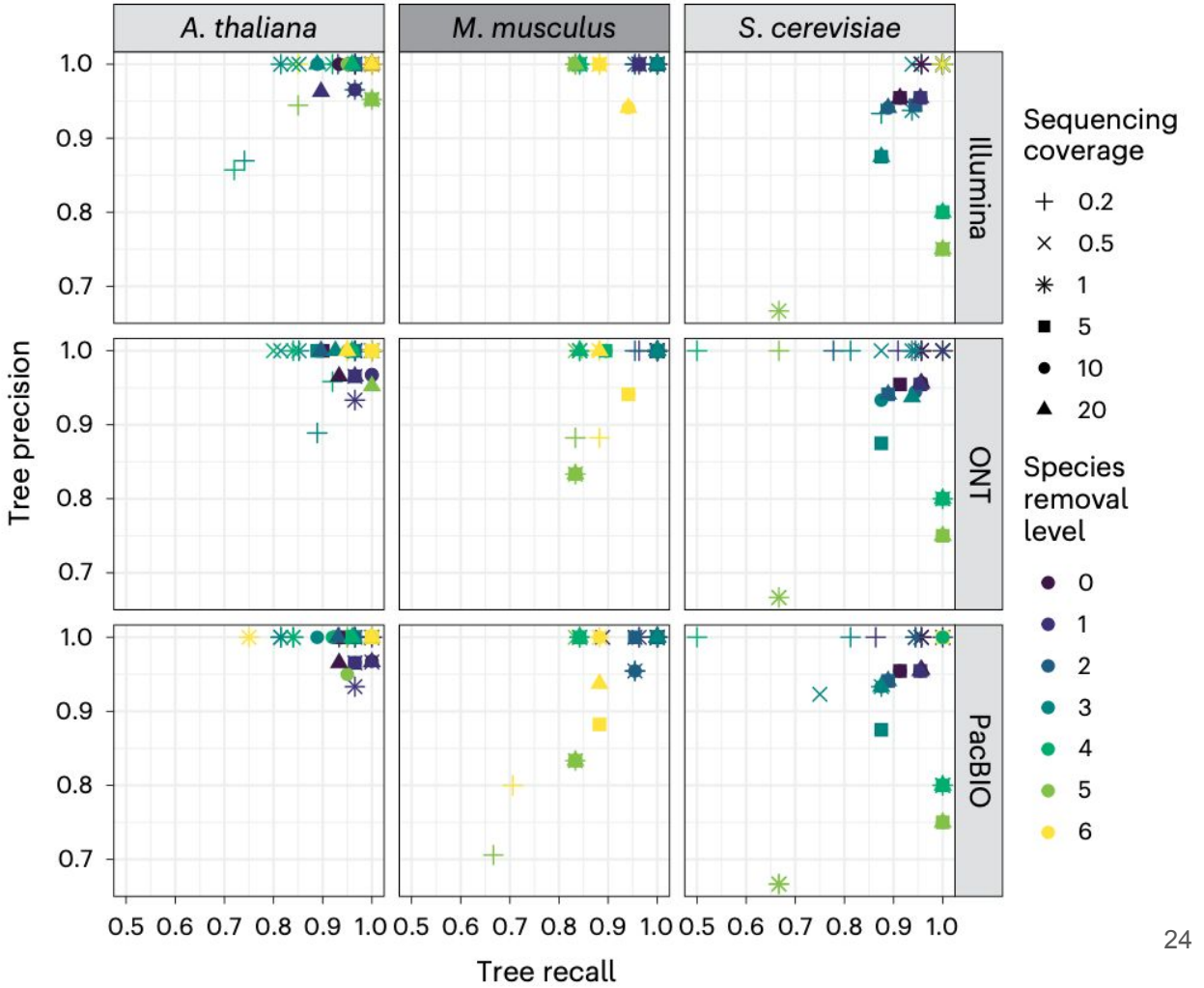
Reconstructed tree accuracy



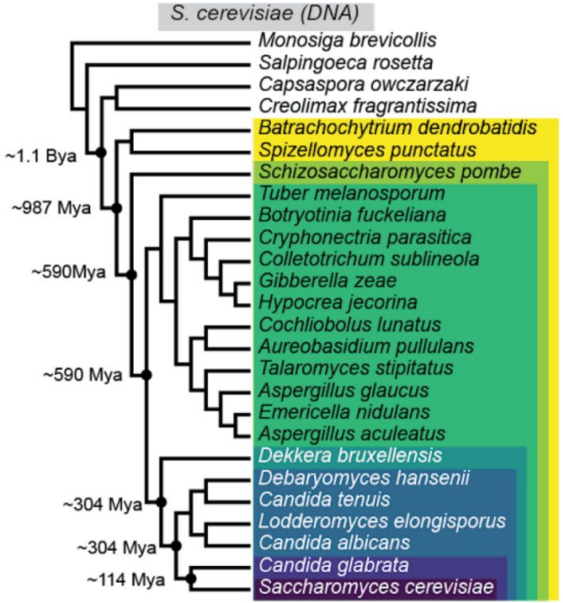


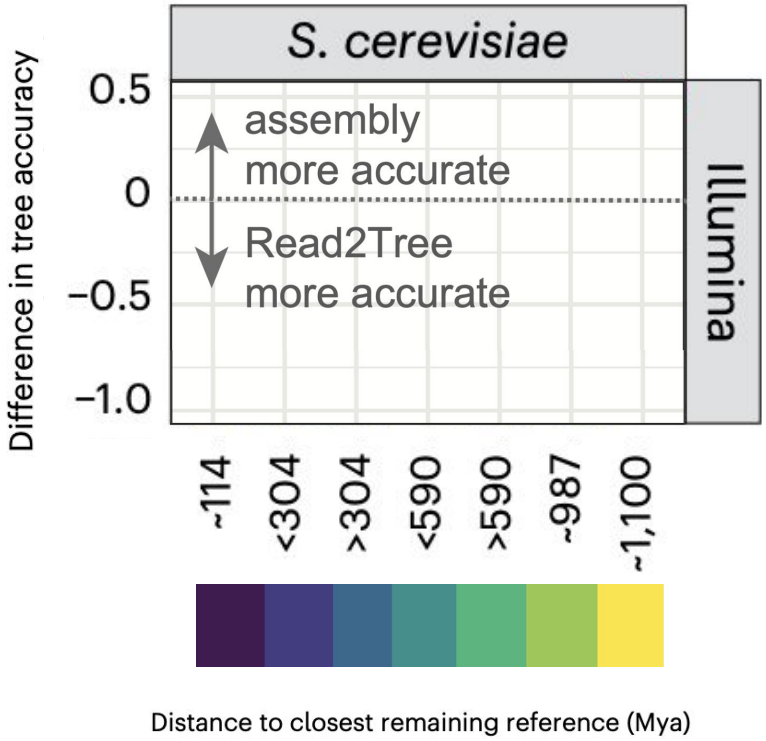
Reconstructed tree accuracy



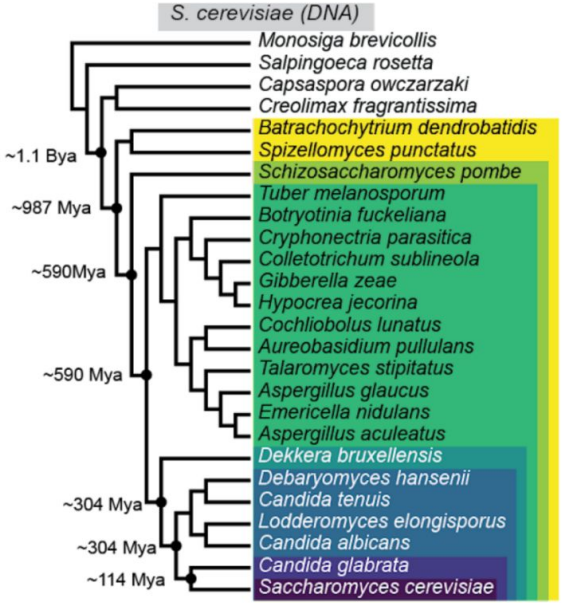


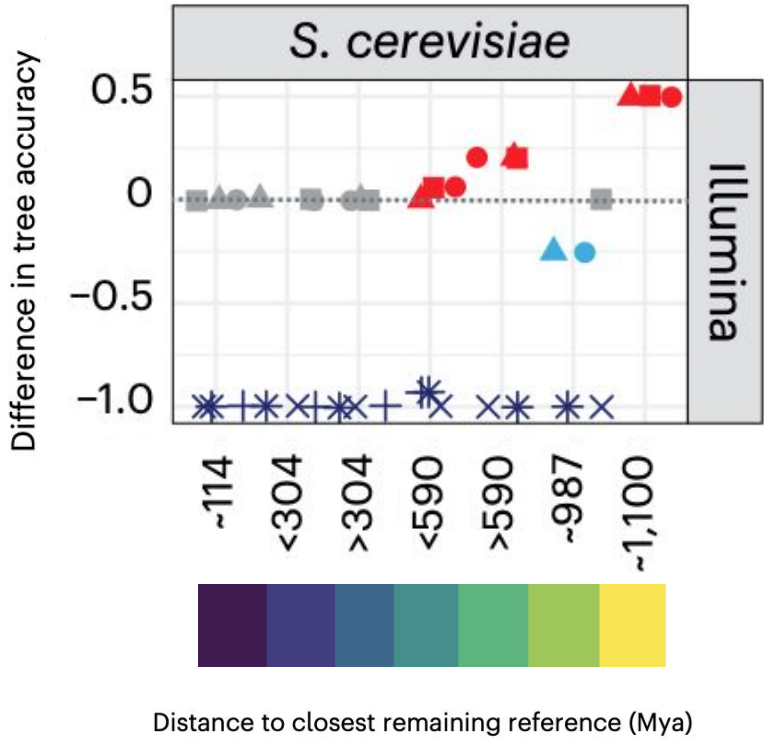
Read2Tree more accurate than assembly for tree inference





Read2Tree more accurate than assembly for tree inference

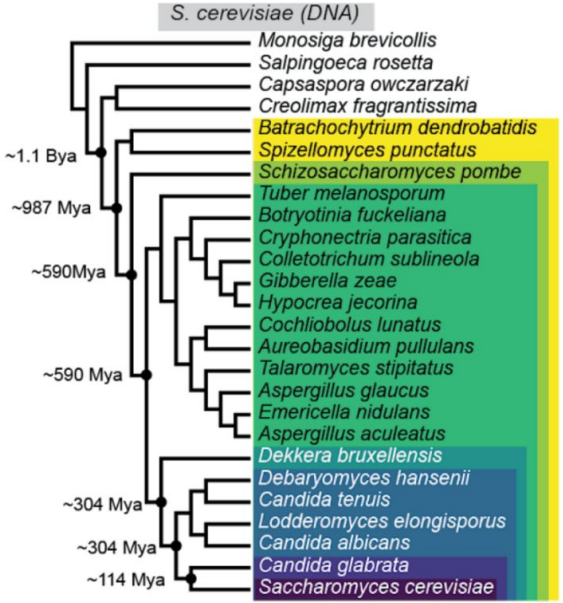


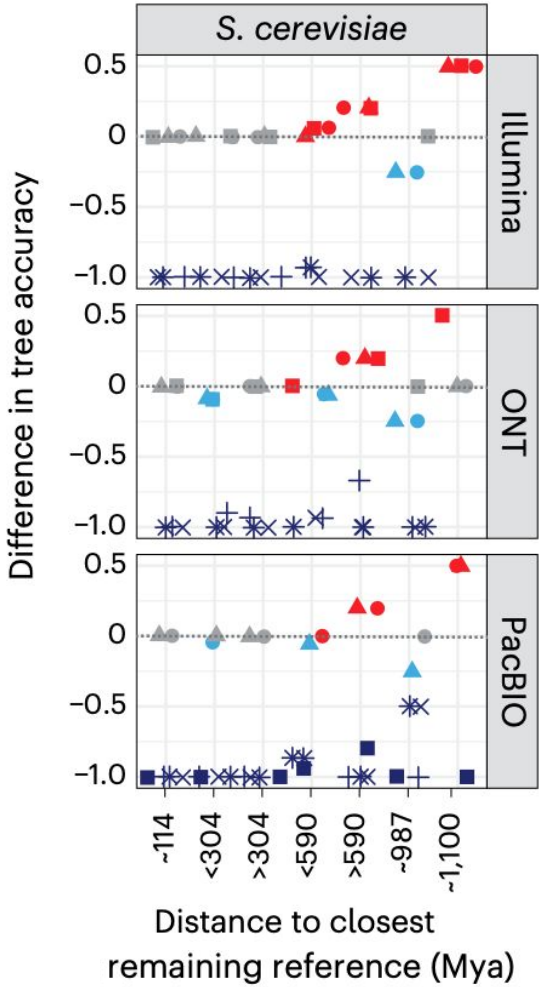


- Assembly better
- Equal
- Read2Tree better
- Only Read2Tree applicable (due to low coverage)

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

Read2Tree more accurate than assembly for tree inference

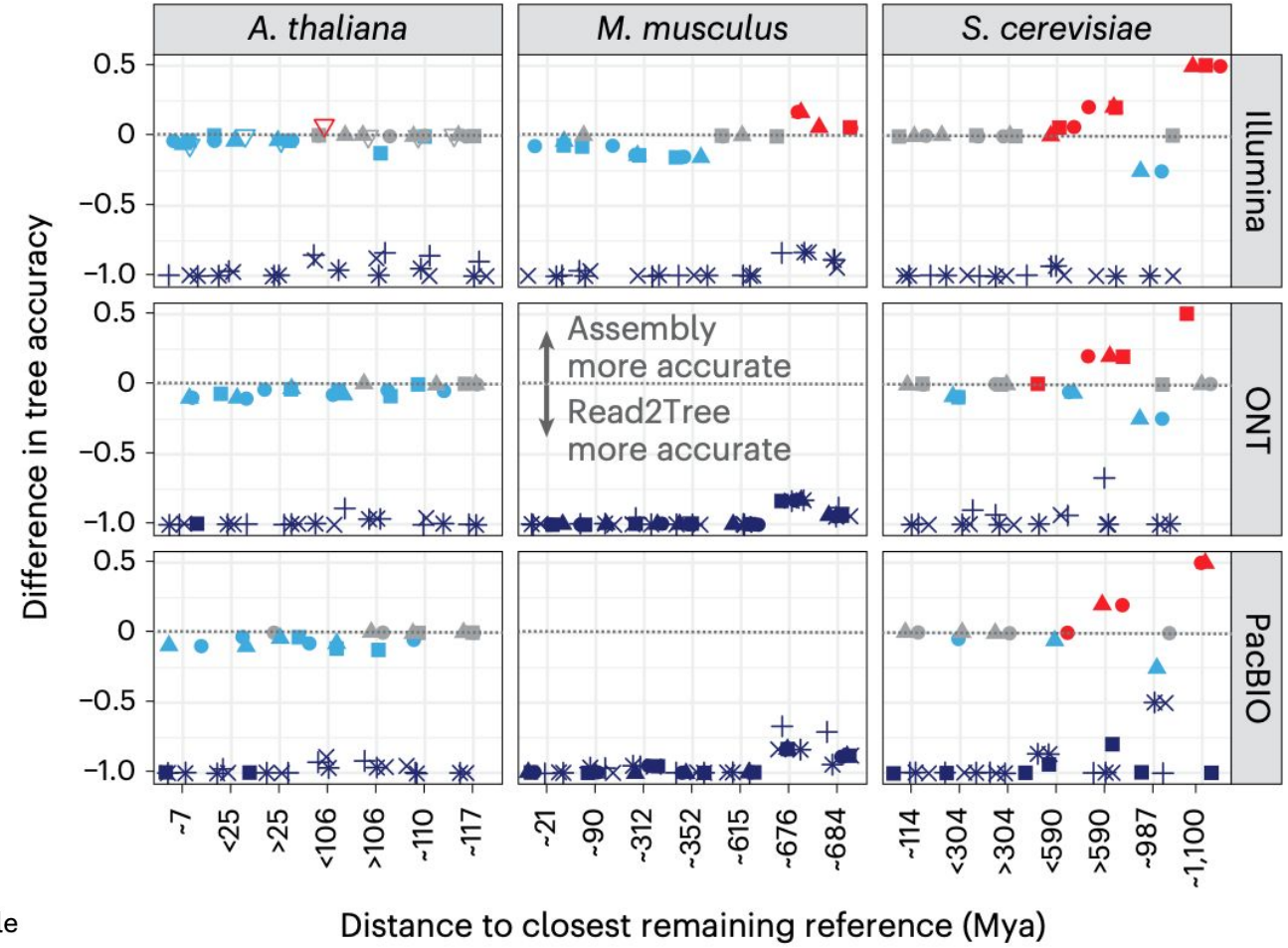




- Assembly better
- Equal
- Read2Tree better
- Only Read2Tree applicable (due to low coverage)

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

Read2Tree more accurate than assembly for tree inference

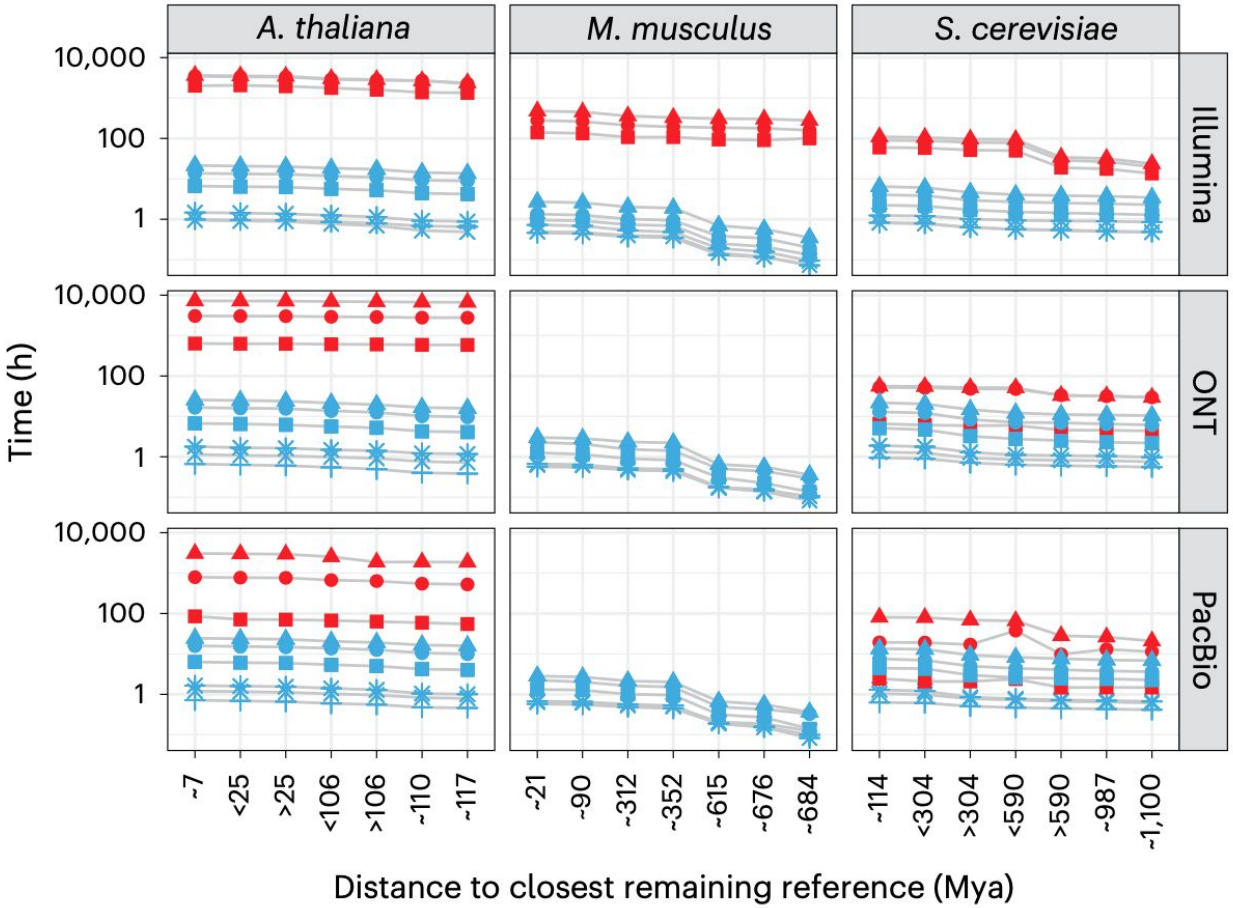


Sequencing coverage: +0.2 ✳1 ●10 ▽60 ×0.5 ■5 ▲20

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

MASH results are not shown.

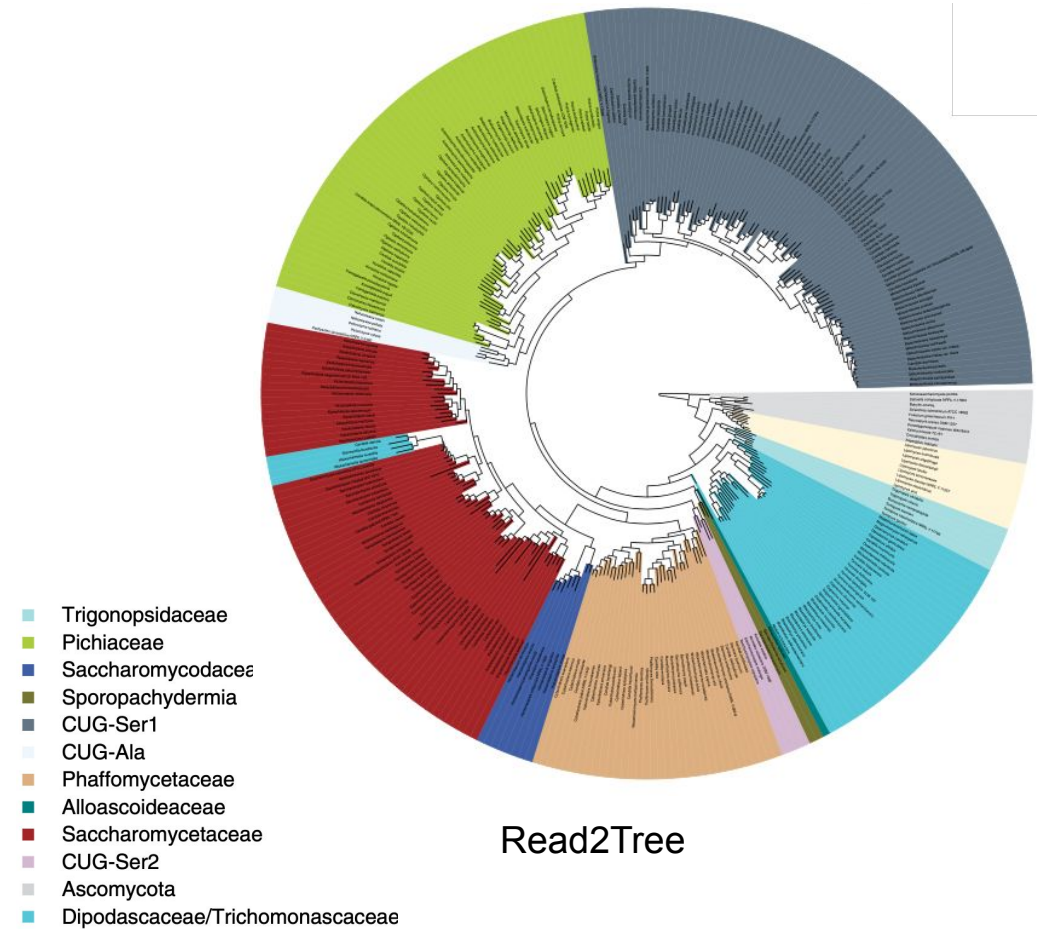
Read2Tree is orders of magnitude faster



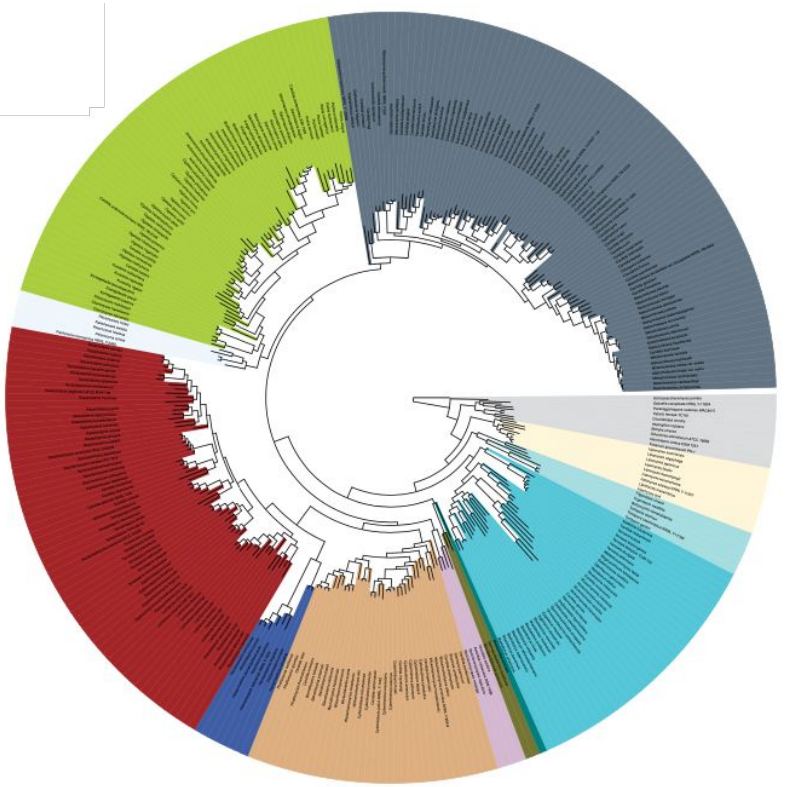
assembly

read2tree

Read2Tree reproduces state-of-the-art yeast phylogeny



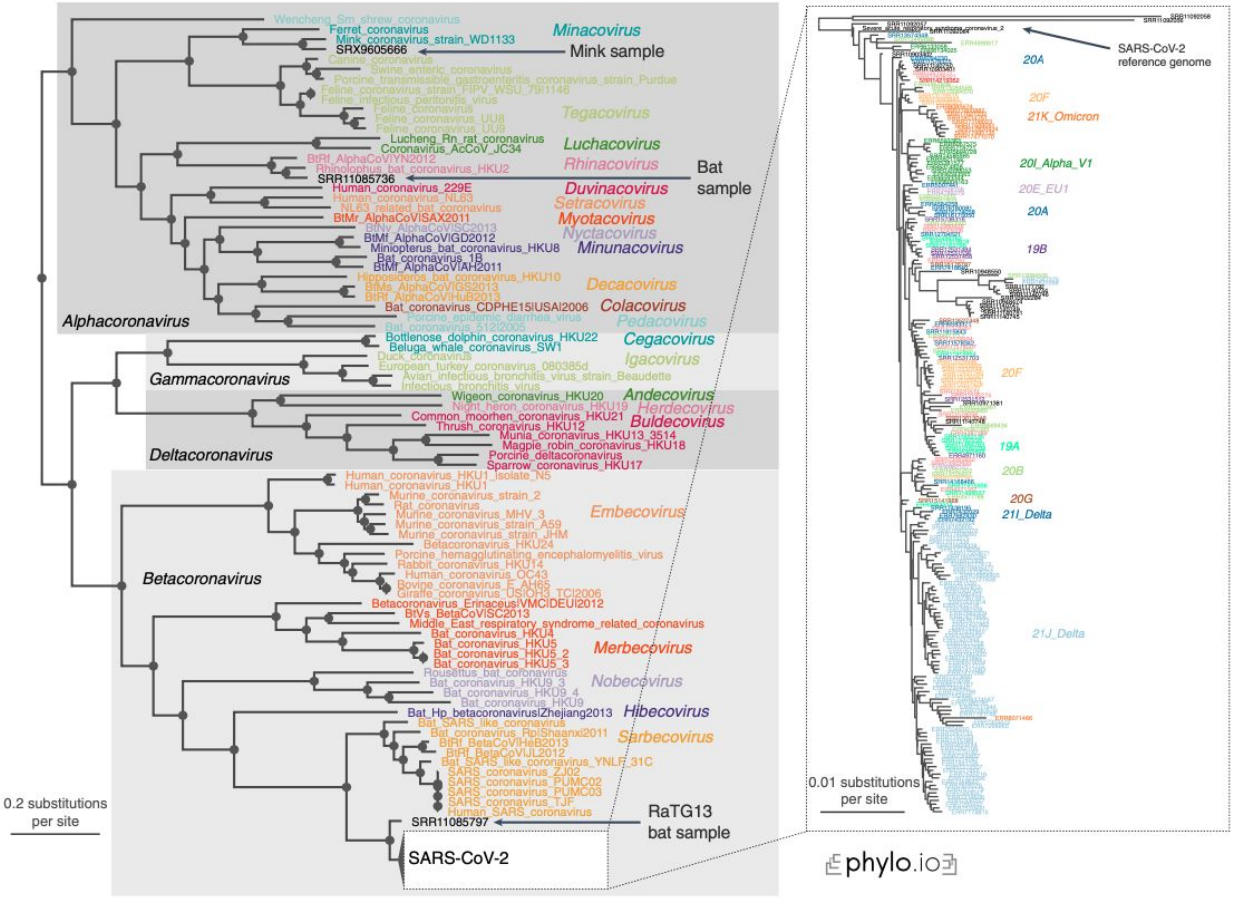
Read2Tree



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

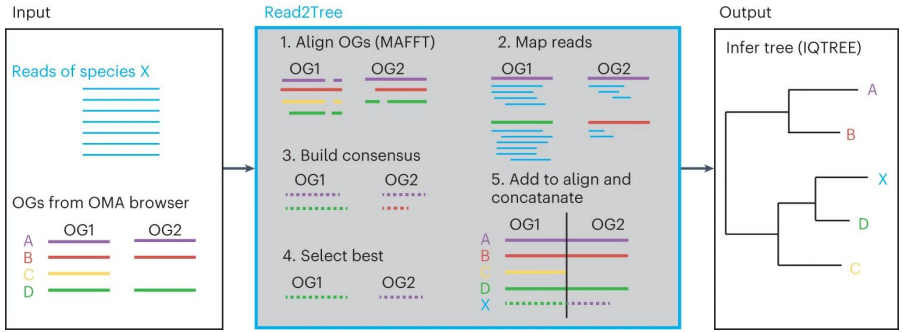
Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree

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



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

Thank you!

Acknowledgments

UNIL | Université de Lausanne

Baylor College of Medicine

HGSC HUMAN GENOME SEQUENCING CENTER

Swiss National Science Foundation

Adrian Altenhoff

David Dylus

Fritz Sedlazeck

Christophe Dessimoz