

# Benchmarking Phylogenetic Trees in the Age of AI-Driven Protein Structures

---

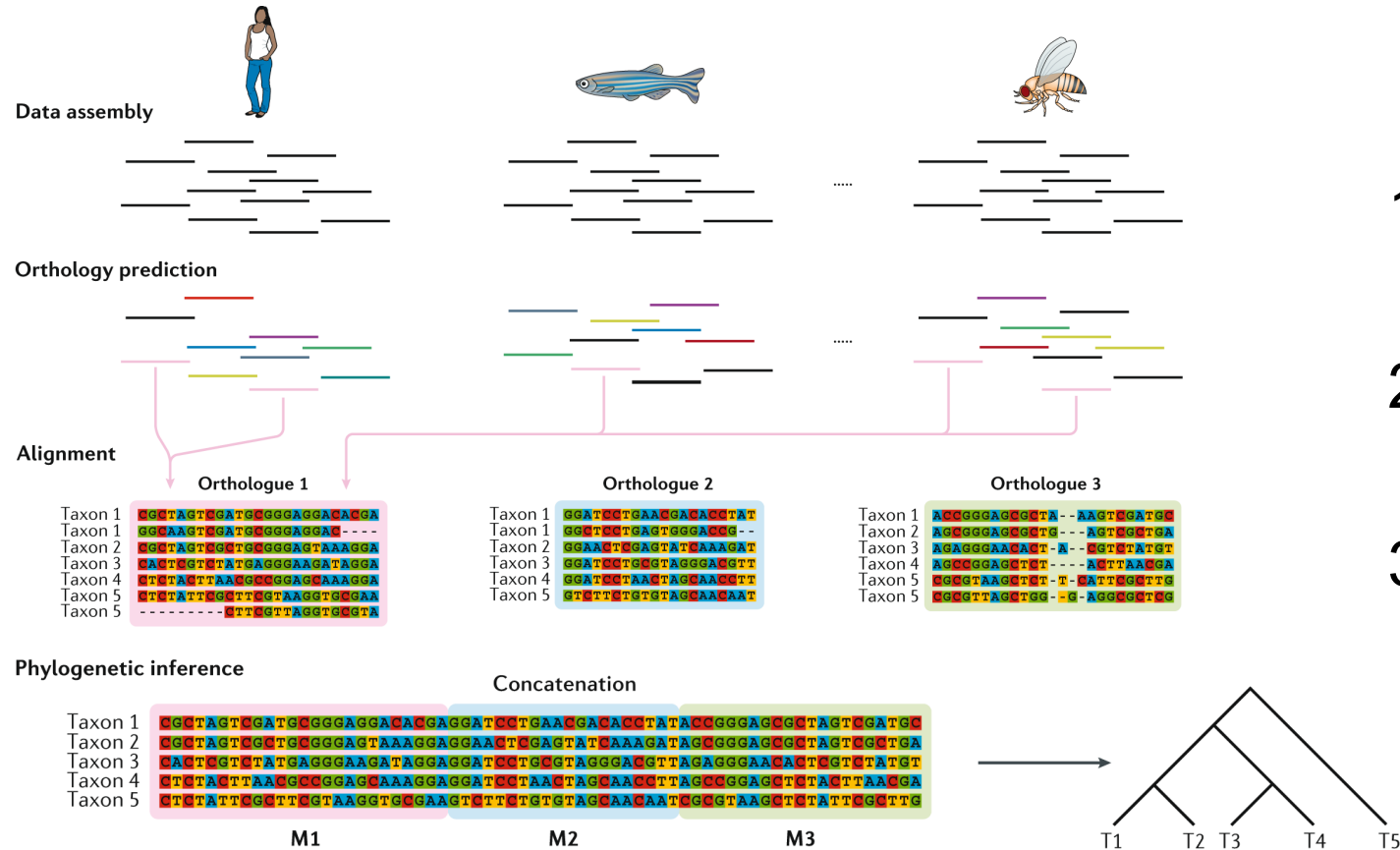
**Dongwook Kim**

Reviews in Quantitative Biology, UNIL

Nov 21, 2025



# What is phylogenomics? A standard workflow



1. Find the genes that tell evolutionary history
2. Align the genes to reveal their evolutionary context
3. Reconstruct their history with a statistical model

Kapli, P., Yang, Z., & Telford, M. J. (2020). *Nature Reviews. Genetics*, 21(7), 428–444.

# A story from 11 years ago: Who watches the watchmen?

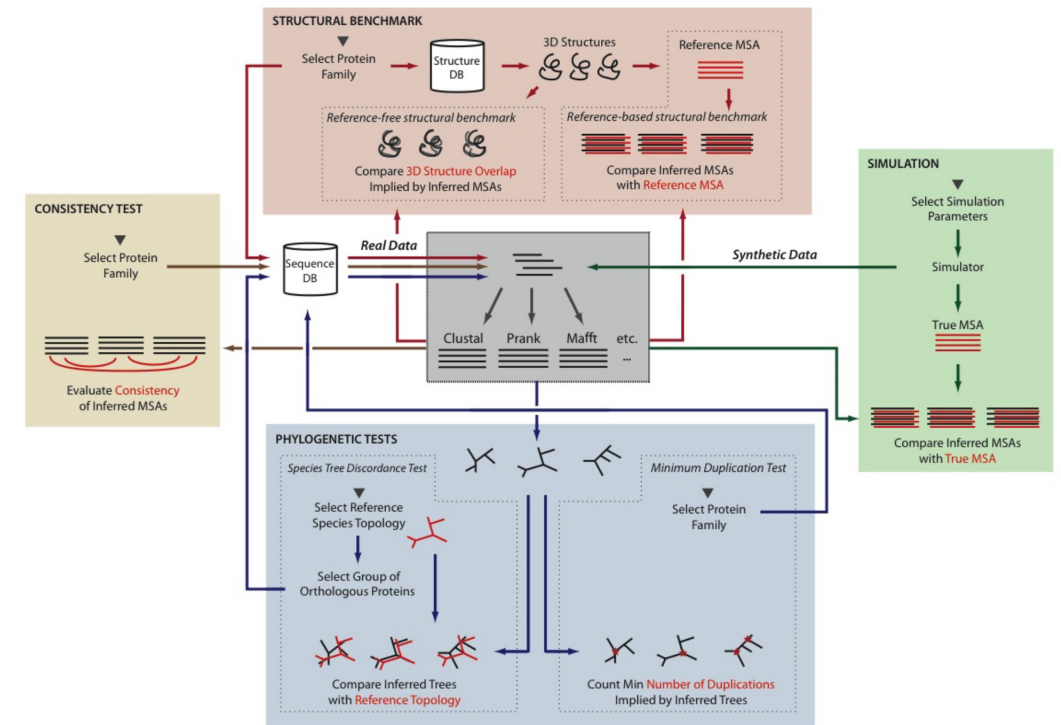
## Who Watches the Watchmen? An Appraisal of Benchmarks for Multiple Sequence Alignment

Stefano Iantorno<sup>1,2,\*</sup>, Kevin Gori<sup>3,\*</sup>, Nick Goldman<sup>3</sup>, Manuel Gil<sup>4</sup>, Christophe Dessimoz<sup>3,†</sup>

Your names might end up here!

### Acknowledgements

The authors thank Julie Thompson for helpful feedback on the manuscript. CD is supported by SNSF advanced researcher fellowship #136461. This article started as assignment for the graduate course “Reviews in Computational Biology” at the Cambridge Computational Biology Institute, University of Cambridge.



Iantorno, S., Gori, K., Goldman, N., Gil, M., & Dessimoz, C. (2014). *Methods in Molecular Biology* (Clifton, N.J.), 1079, 59–73.

# Phylogenetic inference has been a cornerstone of evolutionary biology

---

## MAFFT multiple sequence alignment software version 7: improvements in performance and usability

[K Katoh](#), [DM Standley](#) - *Molecular biology and evolution*, 2013 - [academic.oup.com](#)

We report a major update of the **MAFFT** multiple sequence alignment program. This version has several new features, including options for adding unaligned sequences into an existing ...

☆ Save [Cite](#) Cited by 45090 [Related articles](#) [All 17 versions](#) [Web of Science: 26712](#)

## [HTML] RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies

[A Stamatakis](#) - *Bioinformatics*, 2014 - [academic.oup.com](#)

... Since the last **RAxML** paper in 2006, it has been continuously ... new features and extensions of **RAxML**, such as a substantial ... user manual covering all new **RAxML** options is available. ...

☆ Save [Cite](#) Cited by 34850 [Related articles](#) [All 20 versions](#) [Web of Science: 26325](#)

## IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era

[BQ Minh](#), [HA Schmidt](#), [O Chernomor](#)... - *Molecular biology ...*, 2020 - [academic.oup.com](#)

... the release of **IQ-TREE** version 1.0 in 2014, we have continuously developed **IQ-TREE** to ... Here, we present **IQ-TREE** version 2 and highlight the key new features and improvements. ...

☆ Save [Cite](#) Cited by 13602 [Related articles](#) [All 22 versions](#) [Web of Science: 9574](#)

## Confidence limits on phylogenies: an approach using the bootstrap

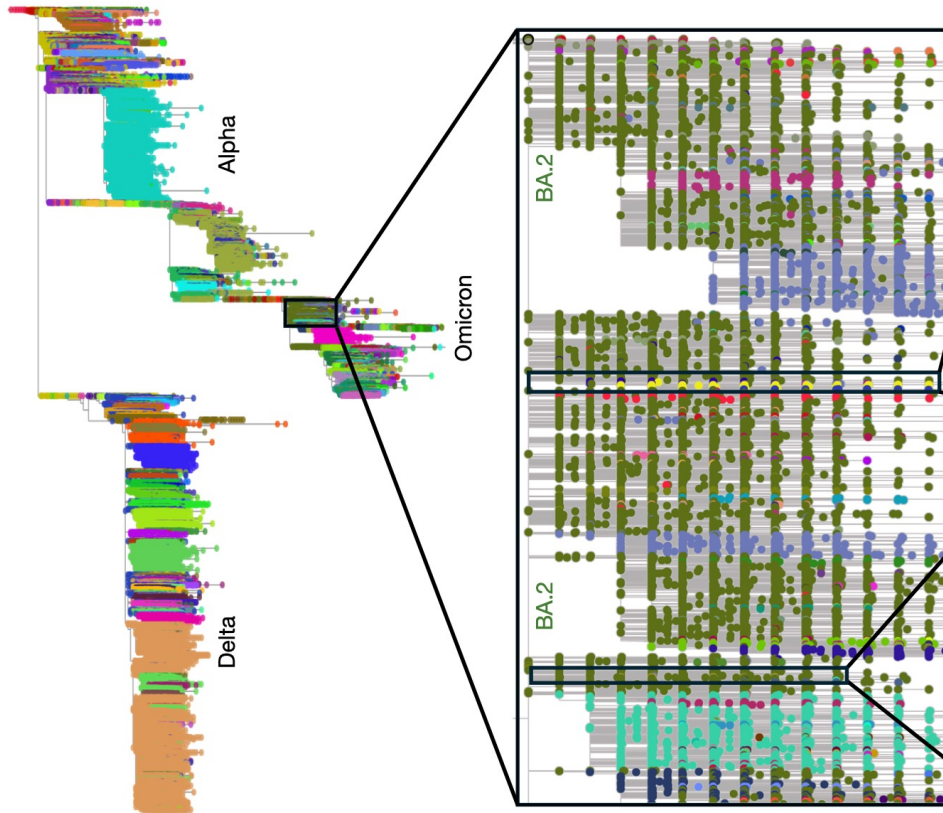
[J Felsenstein](#) - *evolution*, 1985 - [academic.oup.com](#)

... constitutes an adequate exploration of the size of the **confidence limits** on the estimate. ... can be inferred by computing the variance of this collection of  $t^*$  values, and the **confidence limits** ...

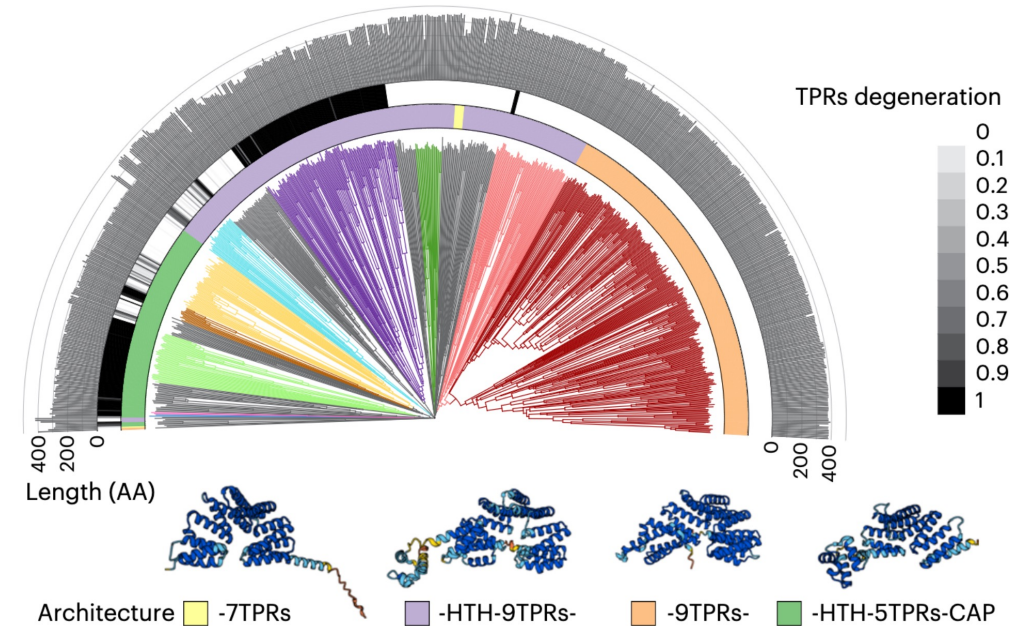
☆ Save [Cite](#) Cited by 53433 [Related articles](#) [All 18 versions](#) [Web of Science: 39148](#)

## But how can you be sure that your tree is correct?

# Current phylogenomics requires new assessment standards



Phylogeny of 2,072,111 SARS-CoV-2 genomes



Structural tree of bacterial RRNPPA receptors

De Maio, N., Ly-Trong, N., Martin, S., Minh, B. Q., & Goldman, N. (2025). *Nature*, 1–7.

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., & Dessimoz, C. (2025). *Nature Structural & Molecular Biology*, 1–11.

## **Key message to convey through this review paper would be...**

---

1. Revisit the conventional methods and introduce new methods to benchmark phylogenetic tree
2. Discuss their pros and cons from various perspectives (scalability, availability of references, etc.)
3. Suggest a scheme for tree benchmarking that reflects the current landscape



# Tree assessment methods to be reviewed: in a nutshell

---

|                                 | <i>Reference-free</i>                             | <i>Reference-based</i>                            |
|---------------------------------|---------------------------------------------------|---------------------------------------------------|
| <i>Statistical implications</i> | <div>Bootstrapping<br/>Bayesian inference</div>   | <div>Robinson-Foulds<br/>Quartet similarity</div> |
| <i>Biological implications</i>  | <div>Minimum duplication<br/>Ultrametricity</div> | <div>Taxonomic congruence</div>                   |

# Bootstrapping – How consistent are your branches?

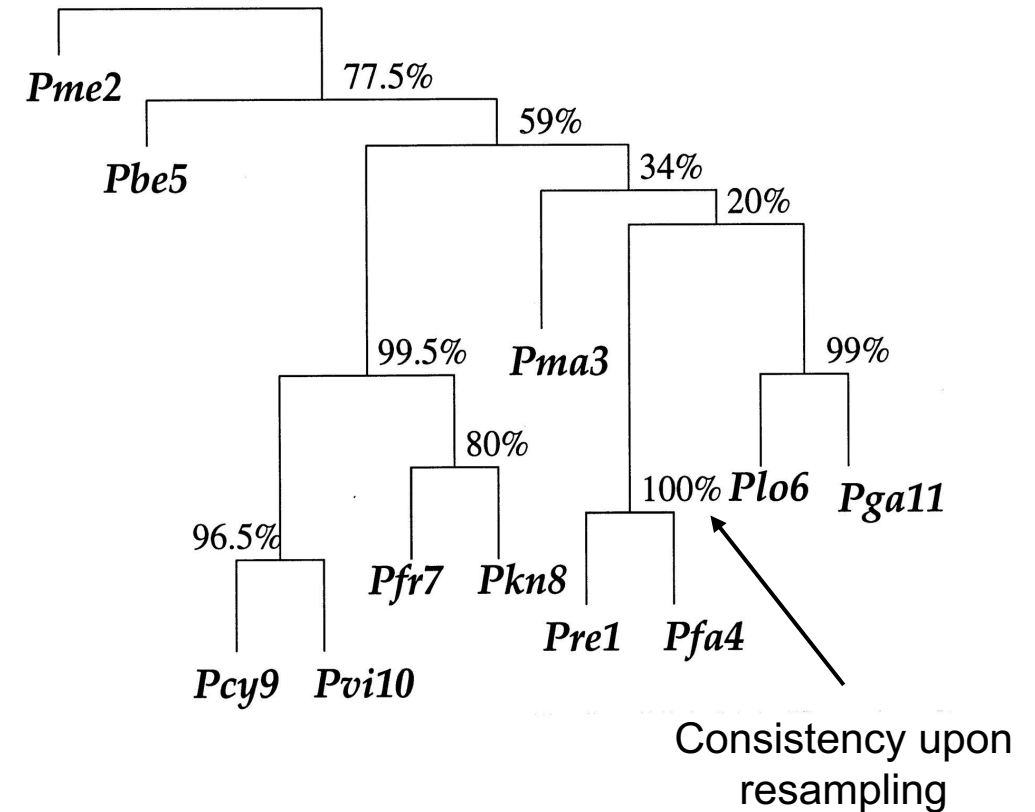
|                | Site: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|----------------|-------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|
| Species        |       |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |    |    |
| 1 Pre (Chimp)  |       | C | T | T | G | A | G | A | A | A | A  | T  | T  | C  | T  | T  | A  | G  | A  | T  | A  |
| 2 Pme (Lizard) |       | T | C | T | A | A | A | A | G | A | T  | T  | A  | T  | A  | T  | A  | G  | A  | T  | A  |
| 3 Pma (Human)  |       | T | T | T | A | A | G | G | A | A | A  | T  | T  | C  | T  | T  | A  | A  | A  | T  | T  |
| 4 Pfa (Human)  |       | T | T | T | G | A | G | A | A | A | A  | T  | T  | C  | T  | T  | A  | G  | A  | T  | A  |
| 5 Pbe (Rodent) |       | T | T | T | A | A | G | A | A | A | A  | T  | T  | T  | A  | T  | A  | A  | A  | T  | A  |
| 6 Plo (Bird)   |       | T | T | T | A | A | G | A | A | A | A  | C  | T  | C  | A  | C  | A  | A  | A  | T  | C  |
| 7 Pfr (Monkey) |       | C | T | T | A | A | G | A | A | G | A  | T  | T  | C  | T  | T  | A  | G  | G  | A  | A  |
| 8 Pkn (Monkey) |       | C | T | T | A | A | G | A | A | A | G  | T  | T  | C  | T  | T  | A  | G  | A  | T  | A  |
| 9 Pcy (Monkey) |       | C | T | C | A | T | G | A | A | A | A  | T  | T  | C  | T  | T  | A  | G  | A  | T  | A  |
| 10 Pv (Human)  |       | C | T | T | A | T | G | A | A | A | A  | T  | T  | C  | T  | C  | G  | G  | A  | T  | A  |
| 11 Pga (Bird)  |       | T | T | T | A | A | G | A | A | A | A  | T  | T  | T  | T  | C  | A  | A  | A  | T  | C  |

Test 1: 5 14 18 6 19 1 20 9 3 8 15 13 13 19 19 11 18 19 14 7

Test 2: 14 16 19 5 1 19 18 5 5 19 15 9 1 3 18 4 20 15 18 3

...

Test N: 11 4 9 17 10 3 16 20 6 18 19 7 19 16 14 8 8 12 14 14



Efron, B., Halloran, E., & Holmes, S. (1996). *Proceedings of the National Academy of Sciences of the United States of America*, 93(23), 13429–13434.



# Bayesian inference: How probable are your branches?

Likelihood of data  
(alignment) given tree

Prior probability  
of trees

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem

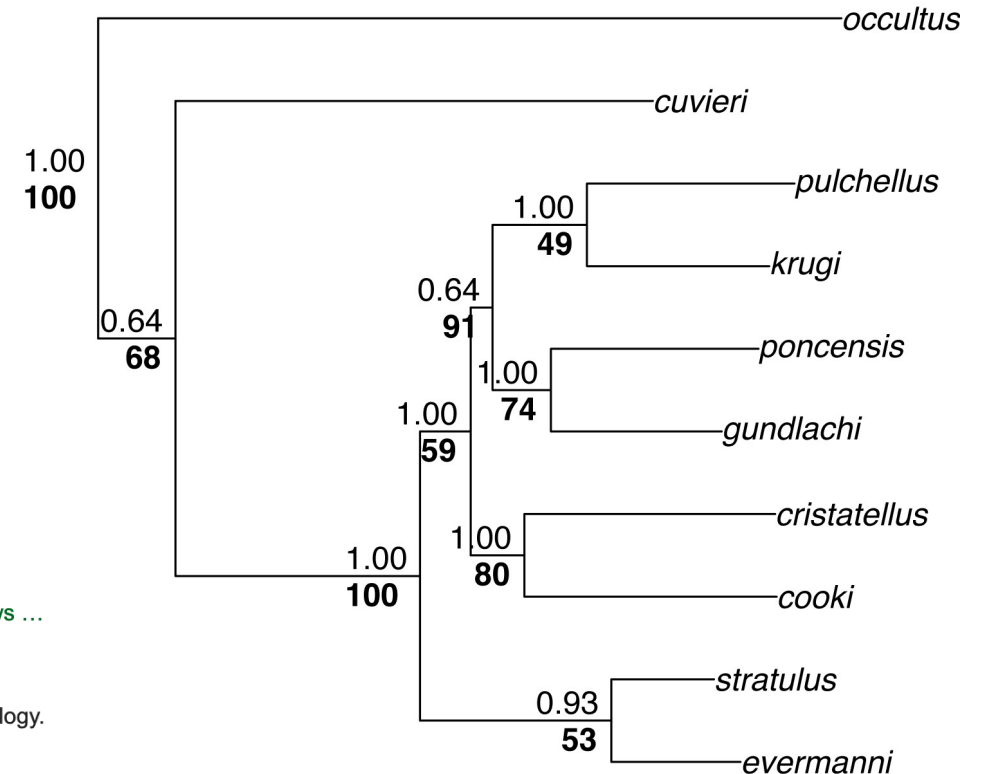
Posterior probability of  
tree/branches

[\[PDF\] MRBAYES: Bayesian inference of phylogenetic trees](#)

[JP Huelsenbeck, F Ronquist - Bioinformatics, 2001 - complexityexplorer.s3.amazonaws.com](#)

The program MRBAYES performs Bayesian inference of phylogeny using a variant of Markov chain Monte Carlo. Availability: MRBAYES, including the source code, documentation, sample data files, and an executable, is available at <http://brahms.biology.rochester.edu/software.html>. Contact: [johnh@brahms.biology.rochester.edu](mailto:johnh@brahms.biology.rochester.edu)

☆ Save 📄 Cite Cited by 27214 Related articles All 14 versions Web of Science: 25110



<https://treethinkers.blogspot.com/2008/10/labeling-trees-posterior-probability.html>  
Huelsenbeck, J., & Ronquist, F. (2001). *Bioinformatics*, 17(8), 754–755.

# Bayesian inference: How probable are your branches?

Article | [Open access](#) | Published: 05 November 2025

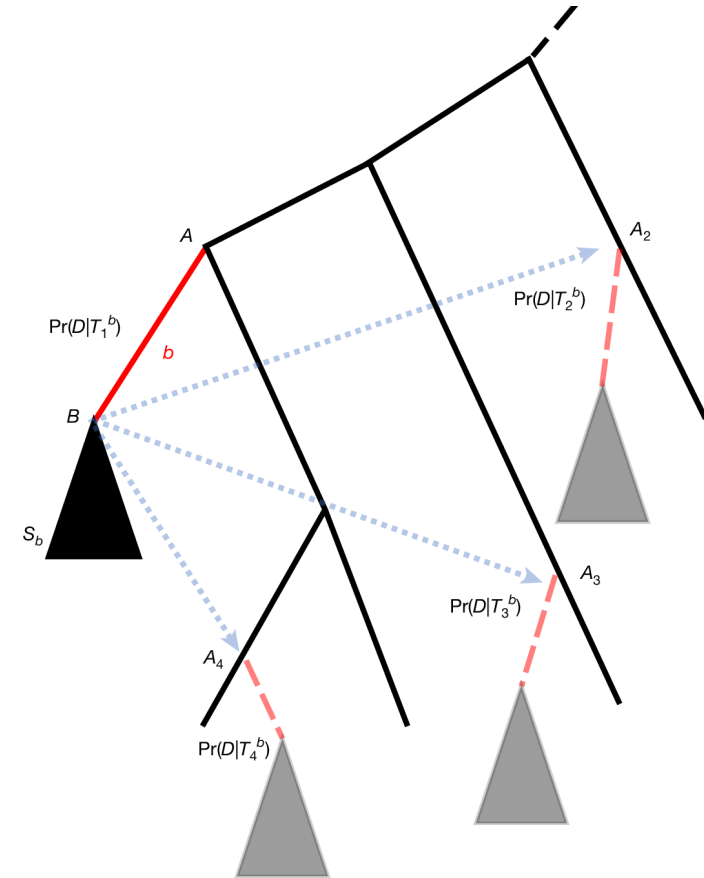
## Assessing phylogenetic confidence at pandemic scales

[Nicola De Maio](#) , [Nhan Ly-Trong](#), [Samuel Martin](#), [Bui Quang Minh](#) & [Nick Goldman](#)

[Nature](#) **647**, 472–478 (2025) | [Cite this article](#)

$$\text{SPRTA}(b) = \Pr(b|D, T \setminus b) = \frac{\Pr(D|T)}{\sum_{1 \leq i \leq I_b} \Pr(D|T_i^b)}$$

How probable is the branch, regarding its other possible positions?



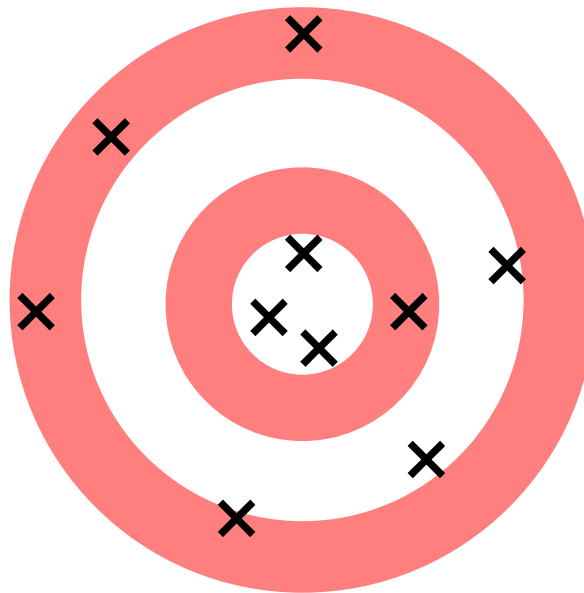
De Maio, N., Ly-Trong, N., Martin, S., Minh, B. Q., & Goldman, N. (2025). *Nature*, 1–7.

# Statistical consistency ~~implies accuracy~~ is blind to systematic bias

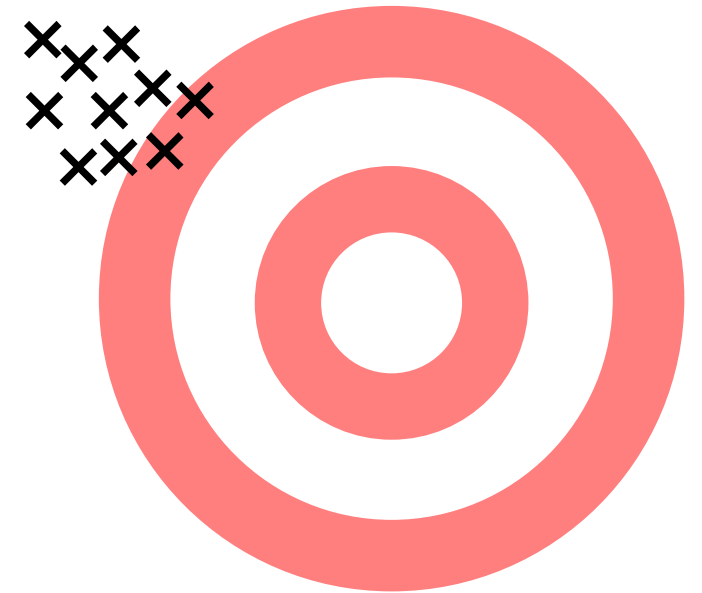
---

**Statistics**

**Tree A**



**Tree B**

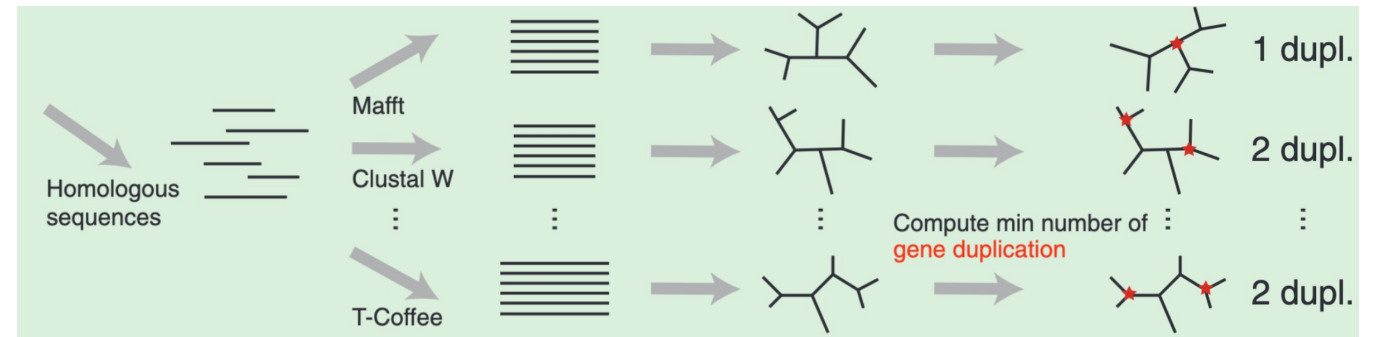
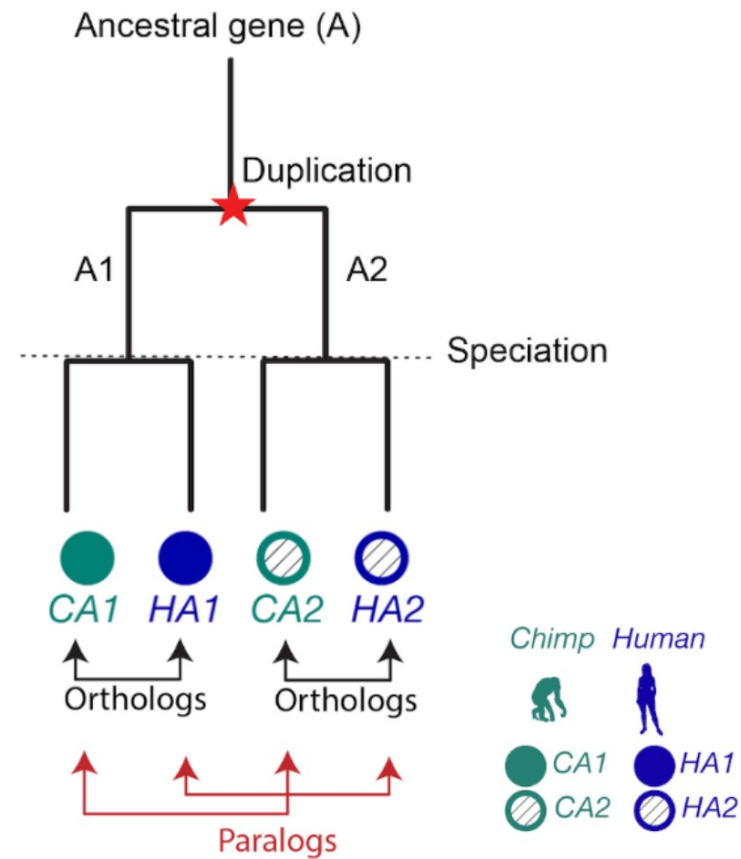


# This is when you need some biological insights

---

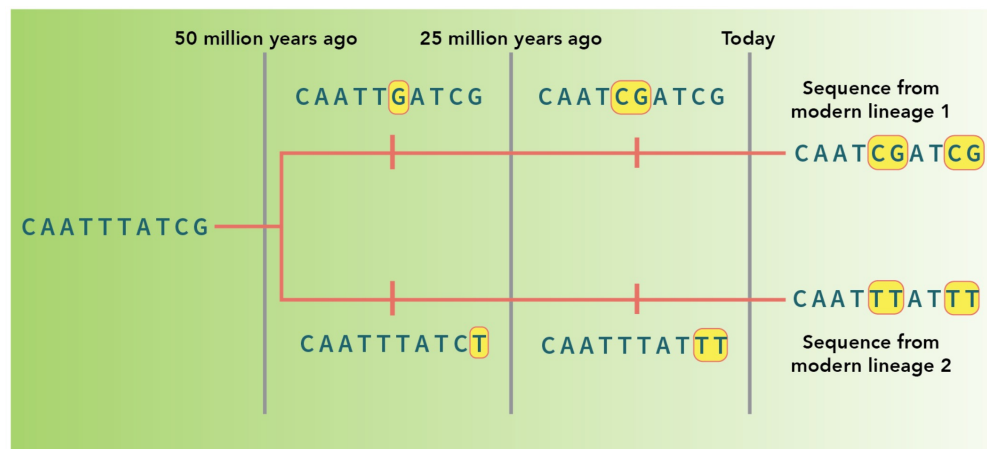
|                                 | <i>Reference-free</i>                                    | <i>Reference-based</i>                                   |
|---------------------------------|----------------------------------------------------------|----------------------------------------------------------|
| <i>Statistical implications</i> | <div>Bootstrapping</div> <div>Bayesian inference</div>   | <div>Robinson-Foulds</div> <div>Quartet similarity</div> |
| <i>Biological implications</i>  | <div>Minimum duplication</div> <div>Ultrametricity</div> | <div>Taxonomic congruence</div>                          |

# Minimum duplication test of gene family evolution

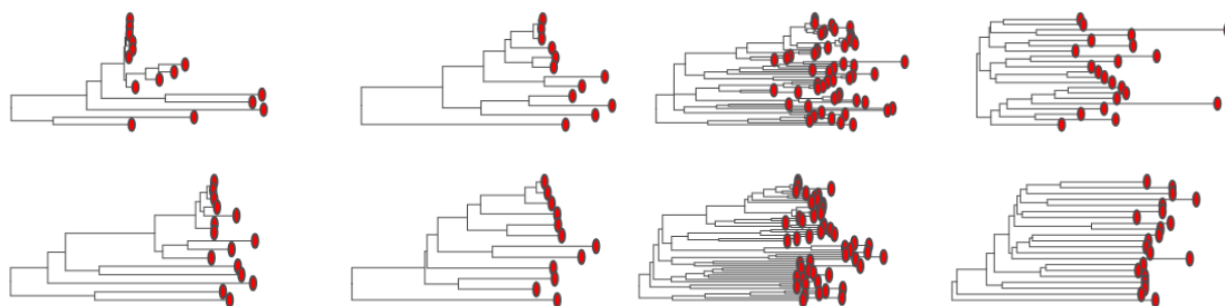


“...holding all else constant, the gene tree with the least number of duplication nodes is the most likely.”

# Molecular clock: we all live in the present



<https://evolution.berkeley.edu/molecular-clocks/>



LUCA  
(Last Universal  
Common Ancestor)

~4.2B years



Your neighbor's  
dog



You



A germ named  
"Jimmy" who lives  
in your gut

Root-to-tip variance, or **ultrametricity**

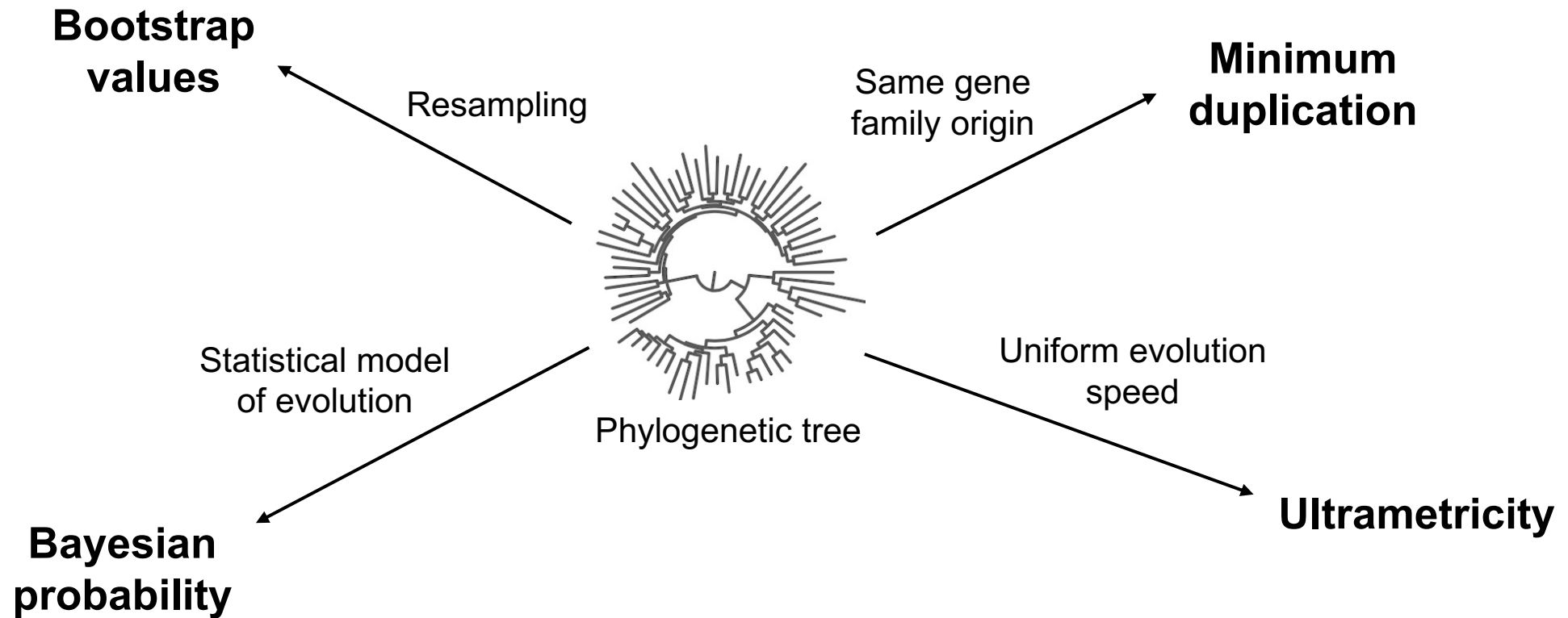
$$E(\text{rootdist}) = \sum_{i=1}^{n\text{leaves}} \text{dist}(l_i, \text{root}) / n\text{leaves}$$

$$S_{\text{norm}}(\text{rootdist}) = \sum_{i=1}^{n\text{leaves}} (\text{dist}(l_i, \text{root}) / E(\text{rootdist}) - 1)^2 / (n\text{leaves} - 1)$$

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., & Dessimoz, C. (2025). *Nature Structural & Molecular Biology*, 1–11.

# Summary on the reference-free methods

---





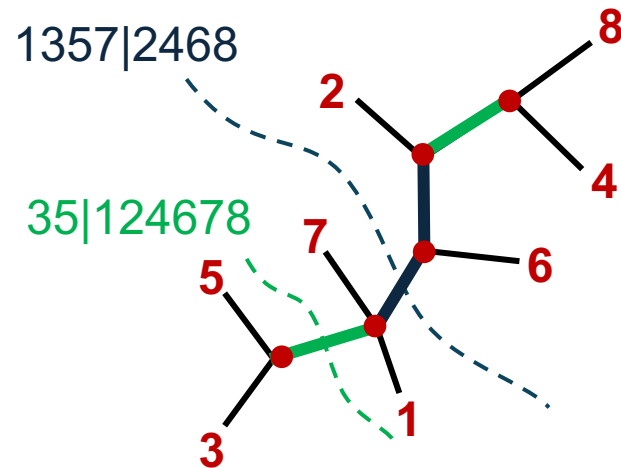
# In most cases, people want to stay on the solid ground

---

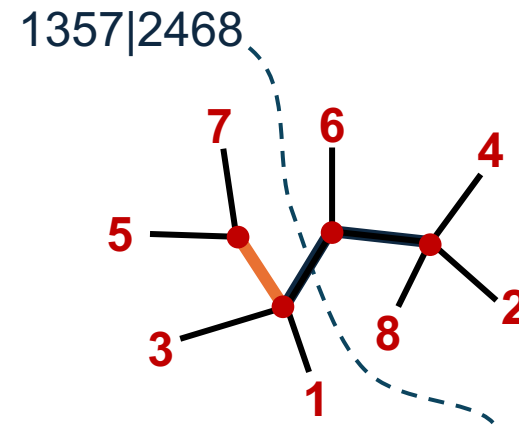
|                                 | <i>Reference-free</i>                                    | <i>Reference-based</i>                                   |
|---------------------------------|----------------------------------------------------------|----------------------------------------------------------|
| <i>Statistical implications</i> | <div>Bootstrapping</div> <div>Bayesian inference</div>   | <div>Robinson-Foulds</div> <div>Quartet similarity</div> |
| <i>Biological implications</i>  | <div>Minimum duplication</div> <div>Ultrametricity</div> | <div>Taxonomic congruence</div>                          |

# How to assess the trees if the ground-truth is given?

- Robinson-Foulds distance



$T_1$  = Your tree



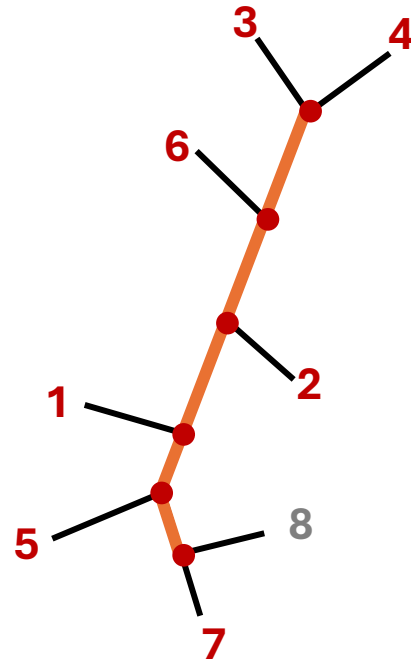
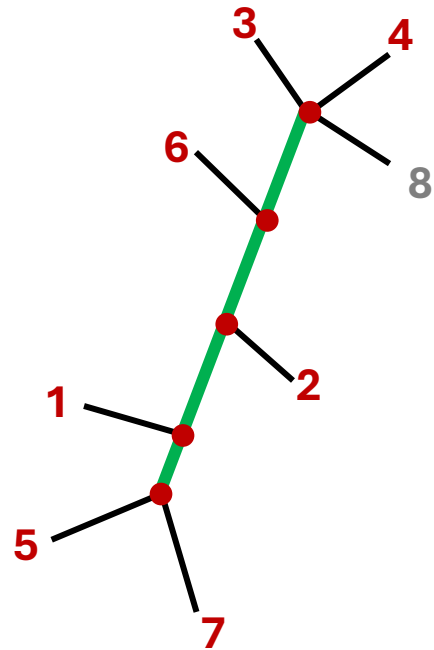
$T_2$  = Ground truth

$$\text{RF}(T_1, T_2) = \# \text{ of non-common splits} = 3$$

Slides partially imported from <https://www.slideserve.com/neviah/triplet-and-quartet-distances-between-trees-of-arbitrary-degree>  
Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. 53(1-2), 131-147.

# RF distance is straightforward, but accompanies internal flaws

---



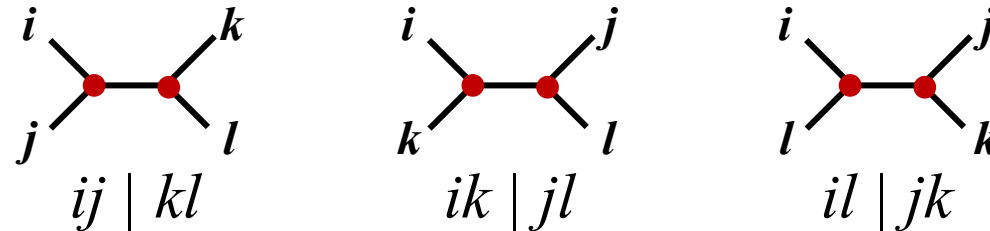
RF = 9  
New RF = 0

- Very sensitive to outliers
- Low resolution
- Topology-centric, lacks biological rationale

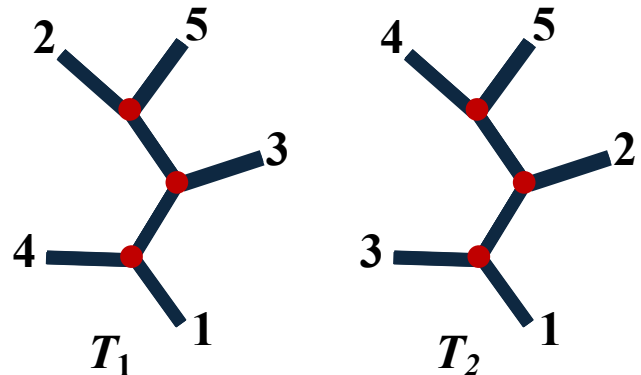
Slides partially imported from <https://www.slideserve.com/neviah/triplet-and-quartet-distances-between-trees-of-arbitrary-degree>  
Briand, S., Dessimoz, C., El-Mabrouk, N., Lafond, M., & Lobinska, G. (2020). *BMC Genomics*, 21(Suppl 10), 779.

# One (of many) other approaches: Quartet distance

- Consider all  $\binom{n}{4}$  quartets, i.e. topologies of subsets of 4 leaves  $\{i, j, k, l\}$



- Quartet distance = # of non-identical quartets



| Quartet          | $T_1$   | $T_2$   |
|------------------|---------|---------|
| $\{1, 2, 3, 4\}$ | 14   23 | 13   24 |
| $\{1, 2, 3, 5\}$ | 13   25 | 13   25 |
| $\{1, 2, 4, 5\}$ | 14   25 | 12   45 |
| $\{1, 3, 4, 5\}$ | 14   35 | 13   45 |
| $\{2, 3, 4, 5\}$ | 25   34 | 23   45 |

$$\text{QD}(T_1, T_2) = 4$$

Read more in...

**Robust Analysis of Phylogenetic Tree Space**

 MARTIN R. SMITH\*

Slides partially imported from <https://www.slideserve.com/neviah/triplet-and-quartet-distances-between-trees-of-arbitrary-degree>  
 Estabrook, G. F., McMorris, F. R., & Meacham, C. A. (1985). *Systematic Biology*, 34(2), 193–200.

# But we won't know the truth without a time machine...

---

- Most of the reference-based metrics require another phylogenetic tree to compare against.
- This can be only achieved in limited situations...
  - Entire evolutionary history is known (e.g. SARS-CoV-2)
  - Very robustly simulated evolution
- *“Assessing the accuracy of trees reconstructed from empirical data is notoriously difficult.”*

---

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., & Dessimoz, C. (2025). *Nature Structural & Molecular Biology*, 1–11.

# But we won't know the truth without a time machine...

---

|                                 | <i>Reference-free</i>                                    | <i>Reference-based</i>                                   |
|---------------------------------|----------------------------------------------------------|----------------------------------------------------------|
| <i>Statistical implications</i> | <div>Bootstrapping</div> <div>Bayesian inference</div>   | <div>Robinson-Foulds</div> <div>Quartet similarity</div> |
| <i>Biological implications</i>  | <div>Minimum duplication</div> <div>Ultrametricity</div> | <div>Taxonomic congruence</div>                          |

# Taxonomic congruence with the established taxonomy

- We can quantify the correctness of the topology, according to the known taxonomy (e.g. NCBI taxonomy)



**Homo sapiens**

Taxonomy ID: 9606 (for references in articles please use NCBI:txid9606)

current name

***Homo sapiens*** Linnaeus, 1758

Genbank common name: **human**

NCBI BLAST name: **primates**

Rank: **species**

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)

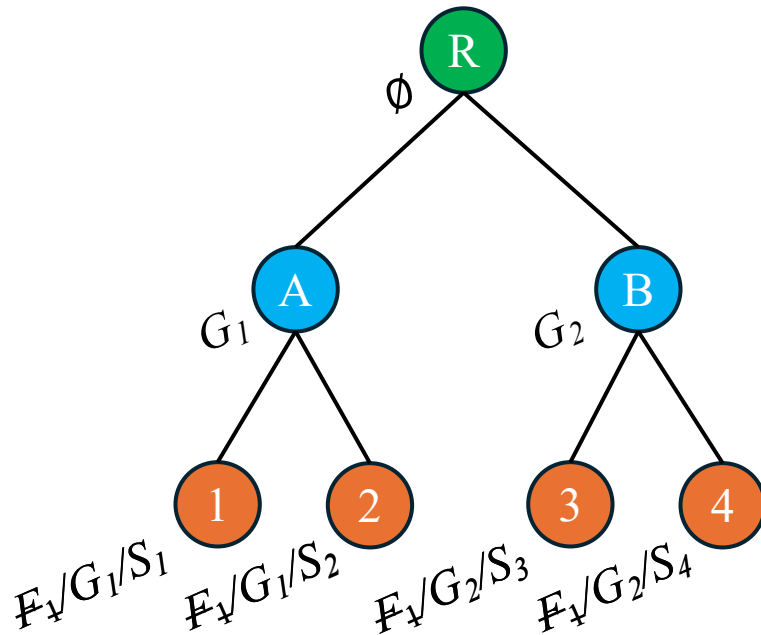
Lineage(*full*)

[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Deuterostomia](#); [Chordata](#); [Cranialata](#); [Vertebrata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Dipnotetrapodomorpha](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Boreoeutheria](#); [Euarchontoglires](#); [Primates](#); [Haplorrhini](#); [Simiiformes](#); [Catarrhini](#); [Hominoidea](#); [Hominidae](#); [Homininae](#); [Homo](#)

<https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>

# Taxonomic congruence with the established taxonomy

- TCS (Taxonomic Congruence Score)



1. Mask globally common entries
2. Calculate the intersections
3. Score the leaves
4. Average out the score

$$TCS_1 = TCS_2 = |A| + |R| = 1$$

$$TCS_3 = TCS_4 = |B| + |R| = 1$$

$$TCS = (TCS_1 + TCS_2 + TCS_3 + TCS_4) / 4 = 1$$

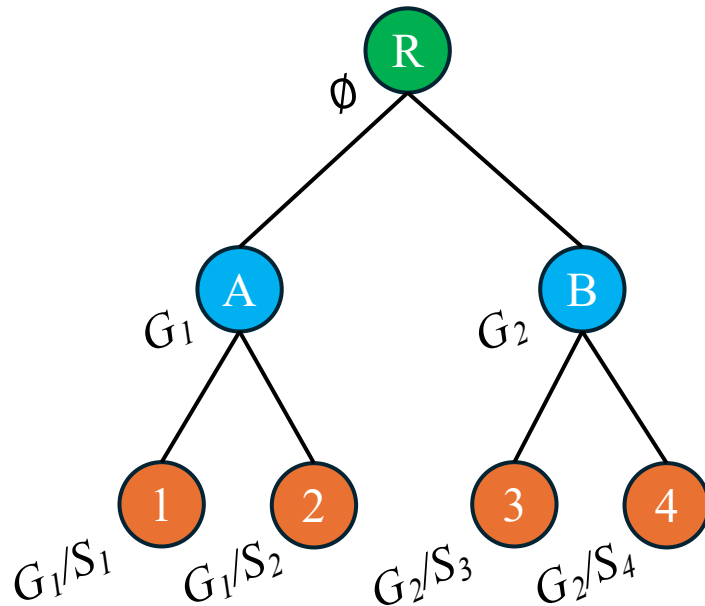
Tan, G., Gil, M., Löytynoja, A. P., Goldman, N., & Dessimoz, C. (2015). *Proceedings of the National Academy of Sciences of the United States of America*, 112(2), E99-100.

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., & Dessimoz, C. (2025). *Nature Structural & Molecular Biology*, 1-11.

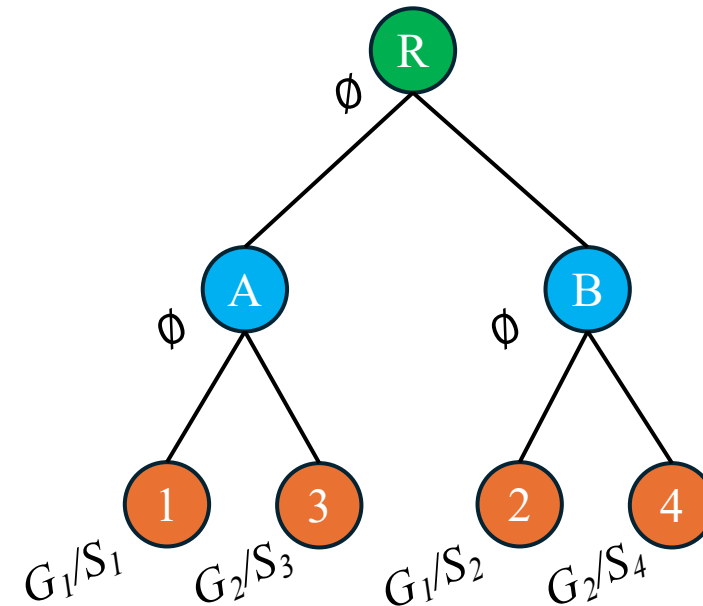


# Taxonomic congruence with the established taxonomy

- Congruent tree vs. Incongruent tree



$$TCS = 1$$

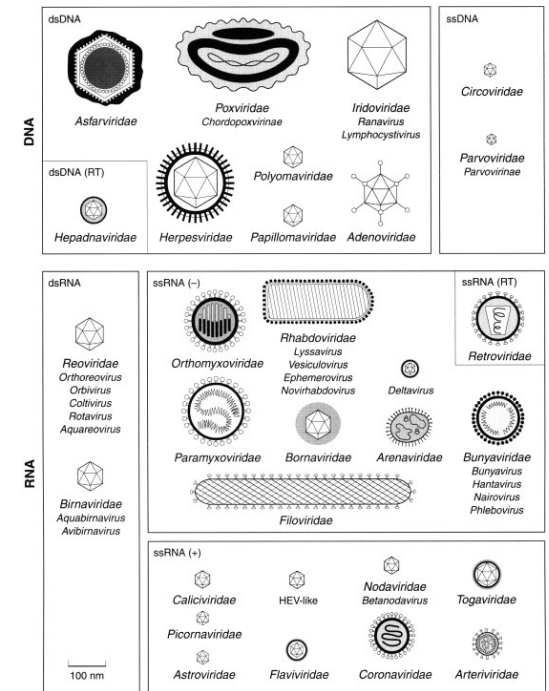


$$TCS = 0$$

Tan, G., Gil, M., Löytynoja, A. P., Goldman, N., & Dessimoz, C. (2015). *Proceedings of the National Academy of Sciences of the United States of America*, 112(2), E99-100.  
Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., & Dessimoz, C. (2025). *Nature Structural & Molecular Biology*, 1–11.

# Is this our golden ticket?

- No.
- Known taxonomy is not always robust (e.g. Viruses)
- Events happening across the species tree are not addressed, such as horizontal gene transfers



Fauquet, C. M. (1999). In *Encyclopedia of Virology* (pp. 1730–1756). Elsevier.

# Overview of the methods we discussed today

---

|                                 | <i>Reference-free</i>                             | <i>Reference-based</i>                            |
|---------------------------------|---------------------------------------------------|---------------------------------------------------|
| <i>Statistical implications</i> | <div>Bootstrapping<br/>Bayesian inference</div>   | <div>Robinson-Foulds<br/>Quartet similarity</div> |
| <i>Biological implications</i>  | <div>Minimum duplication<br/>Ultrametricity</div> | <div>Taxonomic<br/>congruence</div>               |

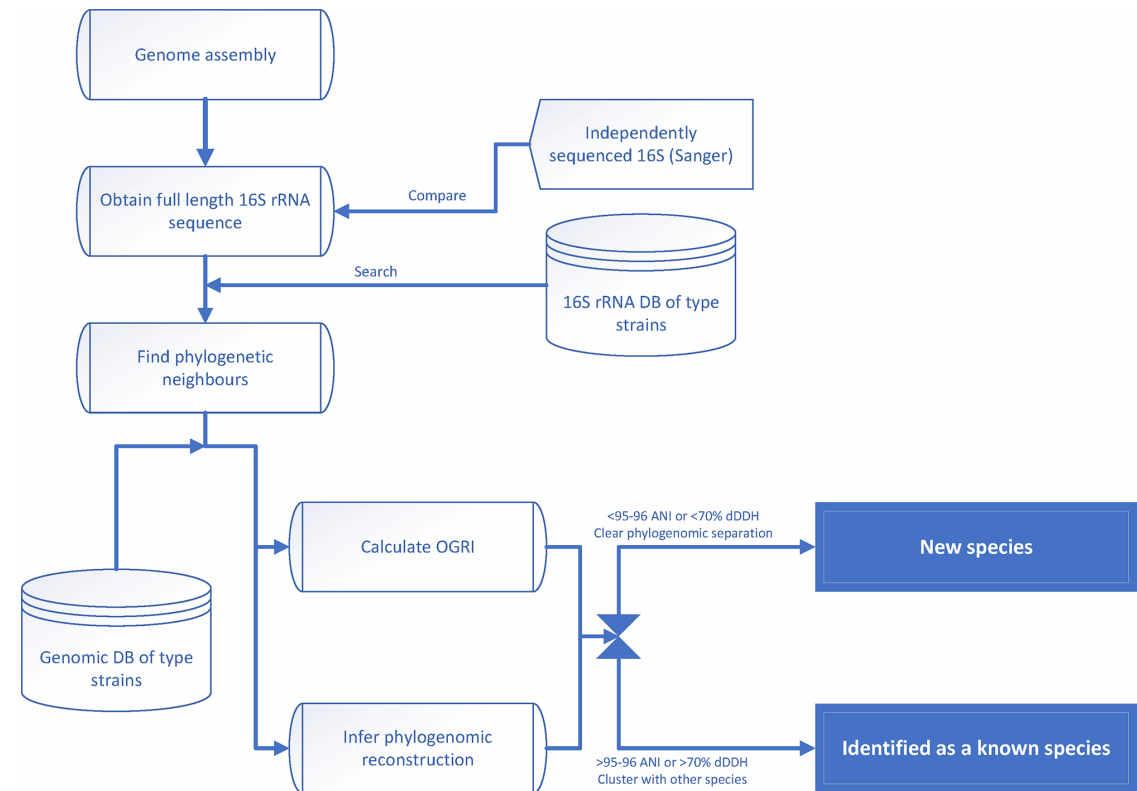
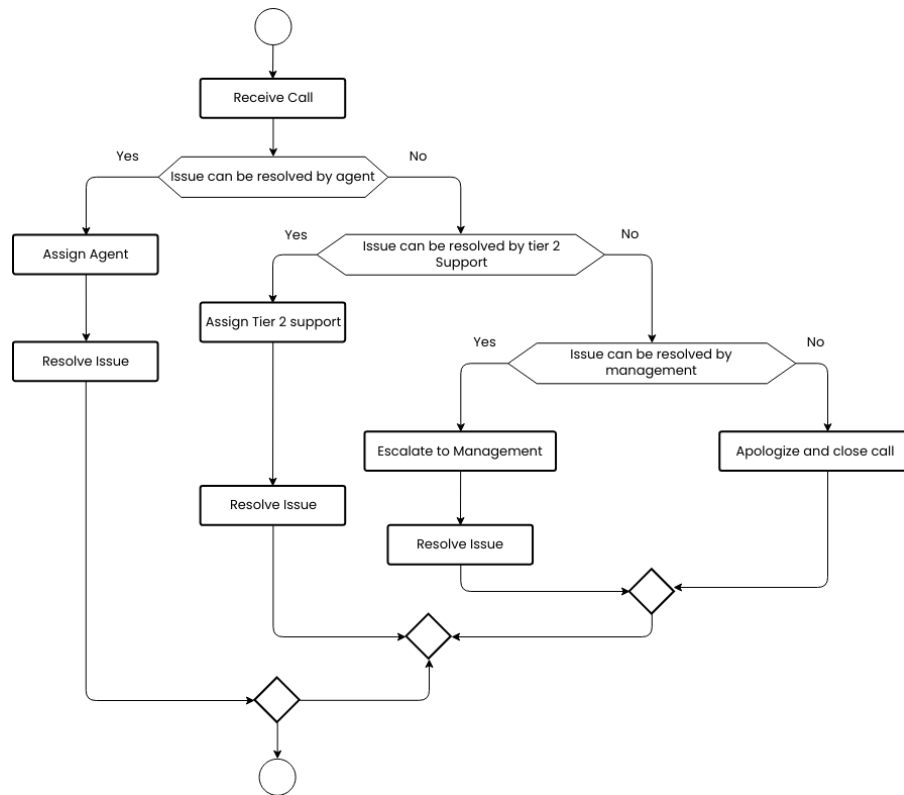
# Recap: Key message to convey through this review paper

---

1. Revisit the conventional methods and introduce new methods to benchmark phylogenetic tree
2. Discuss their pros and cons from various perspectives (scalability, availability of references, etc.)
3. Suggest a scheme for tree benchmarking that reflects the current landscape

# Proposed standard to benchmark trees in the new age?

- Contact center workflow for tree benchmarking?



**THANK YOU FOR LISTENING!**

---