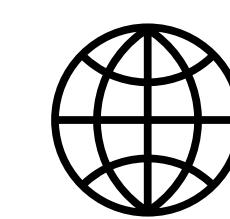




# Trimming MSAs

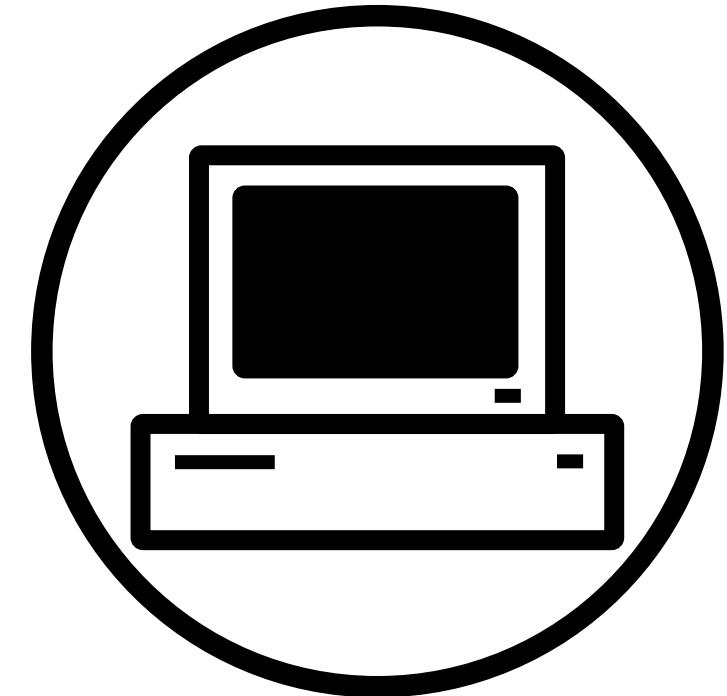
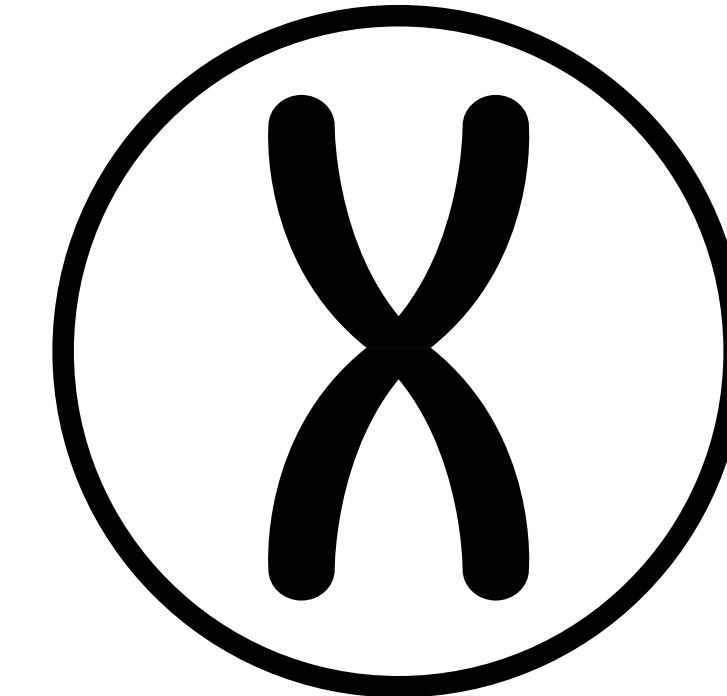
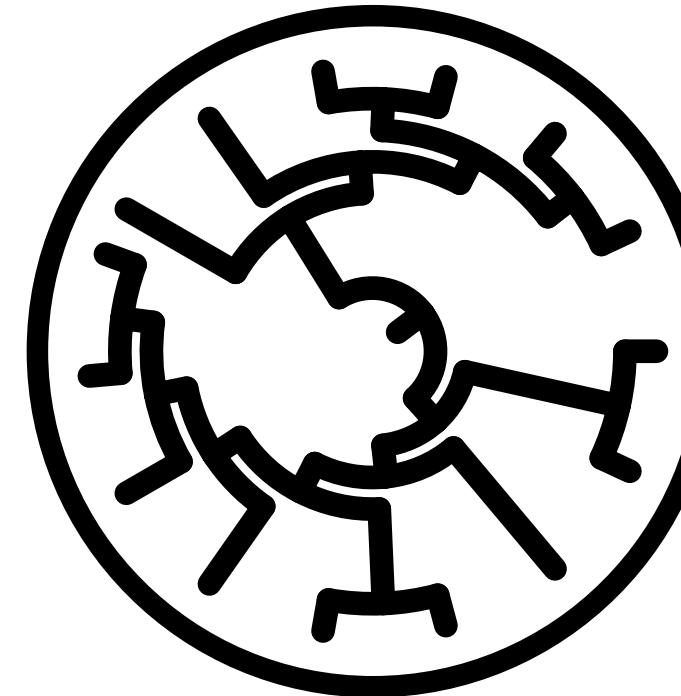


@JLSteenwyk



<https://jlsteenwyk.com/>

# Outline



- Housekeeping
- Why trim? A brief history
- Trimming becomes a contentious topic
- ClipKIT implements a novel approach



@JLSteenwyk

# Raw course materials are available via GitHub

The screenshot shows a GitHub profile for the user **JLSteenwyk**. The profile features a large circular portrait of a smiling young man, Jacob L. Steenwyk. Below the portrait, his name and GitHub handle are displayed, along with his professional background: Omics & Software Eng, HHMI Awardee, LSRF & Berkeley Science Fellow, UC-Berkeley, and HHMI Gilliam Fellow at Vanderbilt. A prominent "Edit profile" button is located just below his name. The GitHub interface includes a search bar, navigation tabs for Overview, Repositories (41), Projects, Packages, and Stars (16), and a pinned section containing five public repositories: ClipKIT, PhyKIT, BioKIT, orthosnap, and orthofisher. Each repository card includes a brief description, programming language (Python or R), star count, and fork count. At the bottom, a heatmap visualizes the user's contributions over the last year, showing a high density of activity in March, April, May, and October.

**Jacob L. Steenwyk**  
JLSteenwyk

Omics & Software Eng | HHMI  
Awardee of the LSRF & Berkeley Science  
Fellow at UC-Berkeley | Previous  
HHMI Gilliam Fellow at Vanderbilt

Edit profile

342 contributions in the last year

Contribution settings ▾ 2024

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

Mon Wed Fri

2023 2022

# Raw course materials are available via GitHub

The screenshot shows a GitHub repository page for '2024\_phylogenomics\_workshop'. The repository is public and owned by 'JLSteenwyk'. The 'Code' tab is selected. The repository has 1 branch ('main') and 0 tags. The commit history shows the following changes:

File / Commit Message	Author	Date
formatting changes to trimming rmd	JLSteenwyk	5ecd9ab · 2 days ago
gene_trees_challenge		implemented comments from Karin and Gemma · 2 days ago
partitioning_and_concatenation		implemented comments from Karin and Gemma · 2 days ago
trimming		formatting changes to trimming rmd · 2 days ago
.DS_Store		updated with new presentation · 3 days ago
20120115-IMG_0297.webp		updated README · 3 days ago
README.md		updated partitioning and concatenation using ... · 3 days ago
Trimming.key		implemented comments from Karin and Gemma · 2 days ago
alignment_information_content.txt		updated partitioning and concatenation using ... · 3 days ago
gene_trees_challenge.key		updated partitioning and concatenation using ... · 3 days ago
partitioning_and_concatenation.key		updated with new presentation · 3 days ago

The repository has 0 stars, 0 forks, and 1 watching. The 'About' section describes it as a 2024 workshop on phylogenomics organized by Evomics. It also links to [evomics.org/2024-workshop-on-phylogenomics/](https://evomics.org/2024-workshop-on-phylogenomics/). The 'Releases' and 'Packages' sections are currently empty.



@JLSteenwyk

# I recommend using links available on the website

## EVOLUTION AND GENOMICS

Intensive and immersive training opportunities

WORKSHOPS   LEARNING   PEOPLE   APPLY   INFORMATION

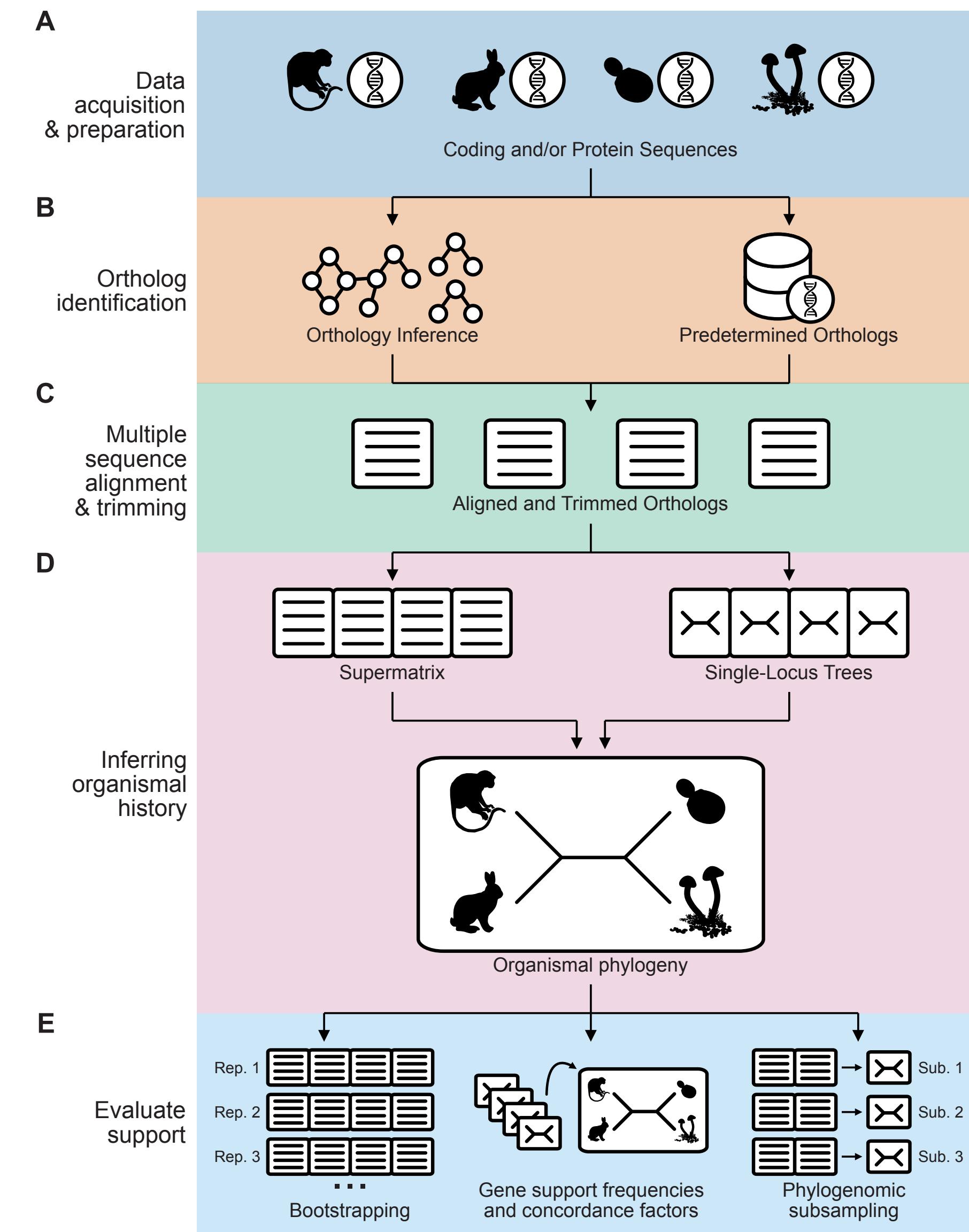
Week 1 : 21-27 January, 2024

DATE	DAY	TIME	PRESENTER	TOPIC	LOCATION
Jan 21	Sunday	18 – 22	Everyone	Reception	Hotel Zlaty Andel
Jan 22	Monday	09 – 12	Anna Karnkowska	Introduction & Orientation, City Information	Town Theatre
	Monday	14 – 17	Workshop Team: Gemma Martínez-Redondo & Karin Steffen	Lab introduction & Unix	House of Prelate
	Monday	19 – 22	Everyone	Scientific speed networking	Krumlov mill
Jan 23	Tuesday	09 – 12	Rosa Fernández	Introduction to Phylogenomics	Town Theatre
	Tuesday	14 – 17	Workshop Team: Marina Marcet-Houben & Jacob L Steenwyk	Alignment and Multiple Sequence Alignment Trimming	House of Prelate
	Tuesday	19 – 22	Workshop Team: Michał & Eduard	Tree Visualization & Tree Challenge	House of Prelate
Jan 24	Wednesday	09 – 12	Marina Marcet-Houben	Introduction to Phylogenetics, and Orthology and Paralogy	Town Theatre
	Wednesday	14 – 17	Workshop Team: Marina Marcet-Houben & Jacob L Steenwyk	Orthology and Paralogy Prediction lab & Gene trees challenge	House of Prelate
	Wednesday	19 – 22	Workshop Team: Jacob L Steenwyk & Karin Steffen	Partitioning and Concatenation Laboratory workshop & data workshop	House of Prelate
Jan 25	Thursday	09 – 12	Olivier Gascuel	State-of-the-art methods and software used in phylogenomic inference	Town Theatre

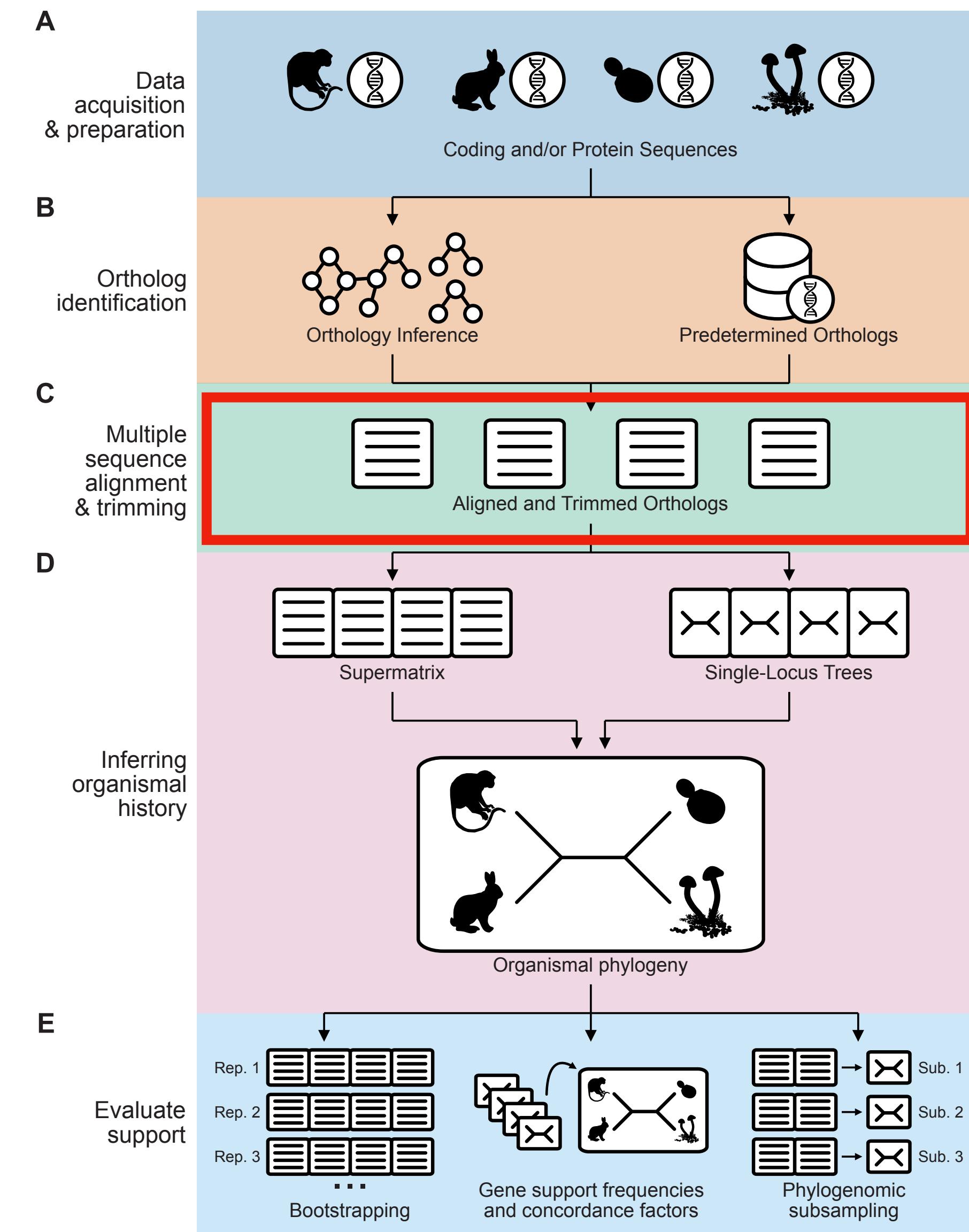


Karin Steffen

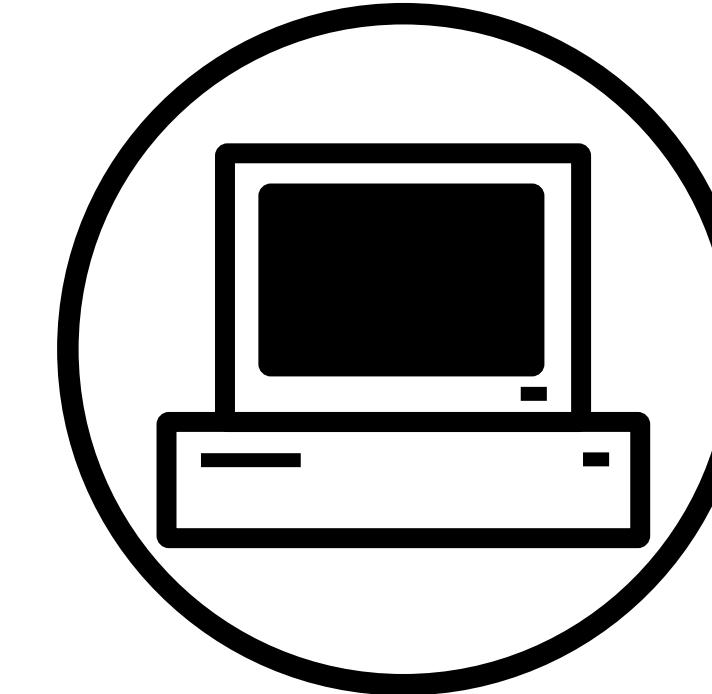
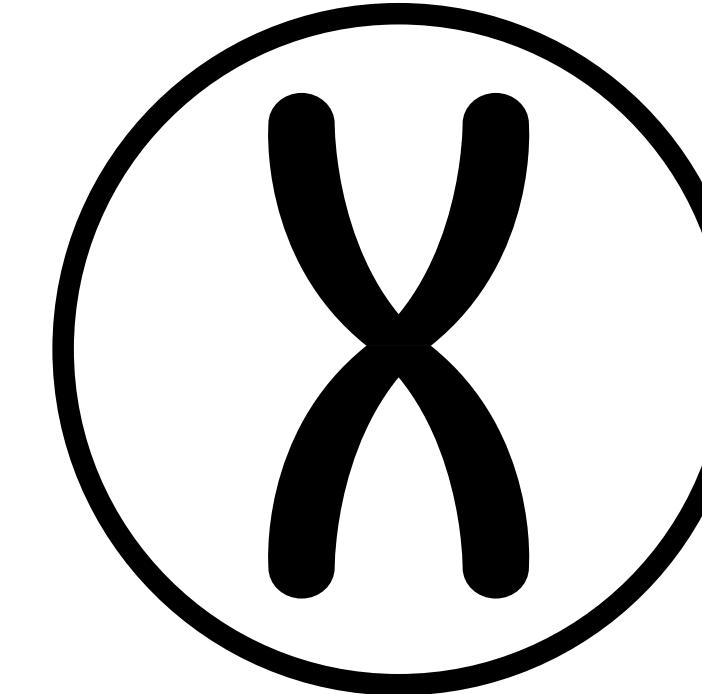
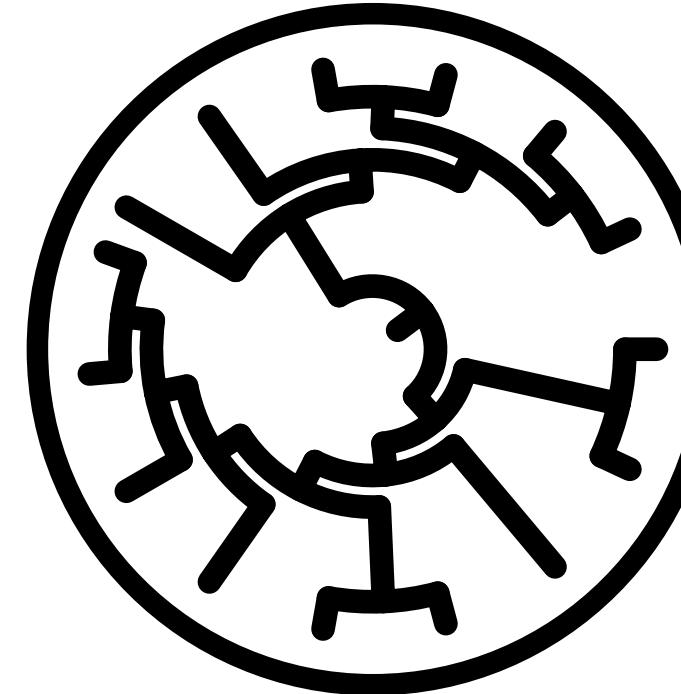
# Facilitating phylogenomic workflows and beyond



# Facilitating phylogenomic workflows and beyond



# Outline



- Housekeeping
- **Why trim? A brief history**
- Trimming becomes a contentious topic
- ClipKIT implements a novel approach



@JLSteenwyk

# Variation in conservation

Highly conserved

<i>Komagataella_pastori</i>	EDAKKEEAIVRHDMVMAHVHTFGKTCPAAGIIHLGATSCYVTDNADLIFRDAYDILIPKLVMVIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRSTLWLQELLWDLRMQRARNDIGLRGAKGTTGTOASFLS
<i>Komagataella_populi</i>	EDAKKEEAIVRHDMVMAHVHTFGKTCPAAGIIHLGATSCYVTDNADLIFRDAYDILIPKLVMVIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRSTLWLQELLWDLRMQRARNDIGLRGAKGTTGTOASFLS
<i>Ogataea_polymerpha</i>	EAAKKEEAIVRHDMVMAHVHVFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDVLIPKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRMQRARNDIGLRGVKGTTGTOASFLS
<i>Ogataea_henricii</i>	EAKAKKEEAIVRHDMVMAHVHVFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDILIPKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRMQRARNDIGLRGVKGTTGTOASFLS
<i>Ogataea_pini</i>	EAKAKKEEAIVRHDMVMAHVHVFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDILIPKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRMQRARNDIGLRGVKGTTGTOASFLS
<i>Ambrosiozyma_monospor</i>	EAKAKKEEAIVRHDMVMAHVHTFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDVLIPKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRMQRARDDIGLRGVKGTTGTOASFLS
<i>Ambrosiozyma_vanderk</i>	EAKAKKEEAIVRHDMVMAHVHVFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDVLIPKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARNDIGLRGVKGTTGTOASFLS
<i>Ambrosiozyma_ambrosi</i>	EAKAKKEEAIVRHDMVMAHVHTFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDILIPKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARNDIGLRGVKGTTGTOASFLS
<i>Ambrosiozyma_oregone</i>	EAKAKKEEAIVRHDMVMAHVHTFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDILIPKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARNDIGLRGVKGTTGTOASFLS
<i>Ambrosiozyma_philent</i>	EAKAKKEEAIVRHDMVMAHVHTFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDILIPKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARNDIGLRGVKGTTGTOASFLS
<i>Kregervanrija_delfte</i>	EIAKIEESKVVRHDVMVMAHVHTFGOTCPAAAGIIHLGATSCYVTDNADLIFRDAYDILISKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARDDIGLRGVKGTTGTOASFLS
<i>Kregervanrija_fluxuu</i>	ETAKIEESKVVRHDVMVMAHVHTFGOTCPAAAGIIHLGATSCYVTDNADLIFRDAYDILIGKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARDDIGLRGVKGTTGTOASFLS
<i>Pichia_membranifacie</i>	EAAKVEESKVVRHDVMVMAHVHVFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDILIAKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARDDIGLRGVKGTTGTOASFLS
<i>Pichia_terricola</i>	EDAKIEESKVVRHDVMVMAHVHVFGETCPAAAGIIHLGATSCYVTDNADLIFRDAYDILIGKLVNINIDRLSKFALEYKDLPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARDDIGLRGVKGTTGTOASFLS
<i>Brettanomyces_custer</i>	EAAKKEEARVRHDVMVMAHVHVFGDTCPAAAGIIHLGATSCYVTDNADLIFRDAYDILISYLLIEKIVNVIDRLSKFALKYKDMPVLGWTHFQPAQLTTVGKRACLWLQELLWDLRNFERARDIGLRGAKGTTGTOASFLS
<i>Brettanomyces_anomal</i>	EAAKKEEARVRHDVMVMAHVHVFGDTCPAAAGIIHLGATSCYVTDNADLIFRDAYDILISYLLIEKLVNVIDRLSKFALKYKDMPVLGWTHFQPAQLTTVGKRACLWLQELLWDLRNFERARDIGLRGTKGTTGTOASFLS
<i>Brettanomyces_bruxel</i>	EAAKKEEARVRHDVMVMAHVHVFGDTCPAAAGIIHLGATSCYVTDNADLIFRDAYDILISYLLIEKLVNVIDRLSKFALKYKDMPVLGWTHFQPAQLTTVGKRACLWLQELLWDLRNFERARDIGLRGTKGTTGTOASFLS
<i>Wickerhamiella-versa</i>	QAASKOEAI VRHDVMVMAHVHVFGEECPAAAGIIHLGATSCYVTDNADLIFLRRGLDLLLPKLASVIDRLSQFAYKYKDLPTLGWTHFQPAQLTTVGKRATLWLQELLWDLRNRRARDIGLRGVKGTTGTOASFLS
<i>Starmerella_apicola</i>	EGATKOEAI VRHDVMVMAHVHVFGEECPAAAGIIHLGATSCYVTDNADLIFLRRGLDLLLPKLASVIDRLSQFALAYKDMPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARNDIGLRGVKGTTGTOASFLS
<i>Starmerella_bombicola</i>	EGAKKOEAI VRHDVMVMAHVHVFGEECPAAAGIIHLGATSCYVTDNADLIFLRRGLDLLLPKLASVIDRLSQFALAYKDMPVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERARNDIGLRGVKGTTGTOASFLS



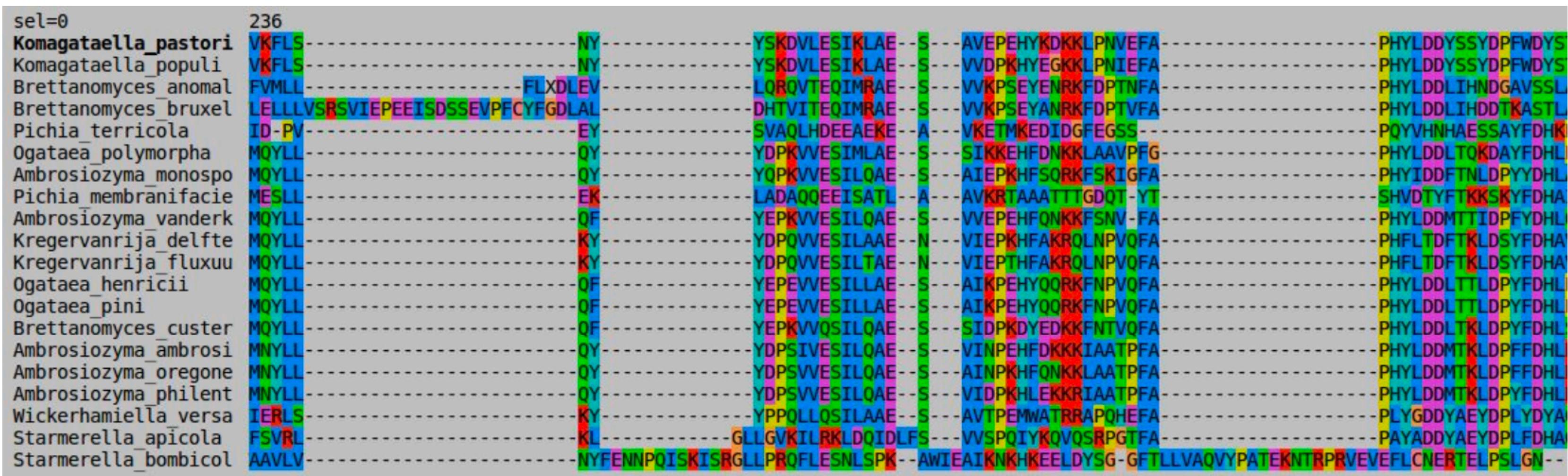
@JLSteenwyk

# Variation in conservation

Highly conserved



Highly variable



@JLSteenwyk

# Sites in alignments may be “unfit”



# Sites in alignments may be “unfit”

Highly divergent sites in alignments can be caused by:

- 1) Erroneously inferred site homology and

# Sites in alignments may be “unfit”

Highly divergent sites in alignments can be caused by:

- 1) Erroneously inferred site homology and
- 2) Saturation of multiple substitutions

# Sites in alignments may be “unfit”

Highly divergent sites in alignments can be caused by:

- 1) Erroneously inferred site homology and
- 2) Saturation of multiple substitutions

For over 30 years, it has been common practice in molecular phylogenetic to remove these sites because they are thought to lack phylogenetic signal

(Lake, 1991, Molecular Biology and Evolution)

# Other sources of error

## Biological

- Difficult to align regions may be structurally disordered and evolve under relaxed selection



@JLSteenwyk

# Other sources of error

## Biological

- Difficult to align regions may be structurally disordered and evolve under relaxed selection

## Analytical

- Errors in
  - Genome assembly,
  - Gene annotation, &
  - Alignment errors

# Trimming becomes controversial in 2015

## Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference

Ge Tan, Matthieu Muffato, Christian Ledergerber, Javier Herrero, Nick Goldman, Manuel Gil,  
Christophe Dessimoz  Author Notes

*Systematic Biology*, Volume 64, Issue 5, September 2015, Pages 778–791,  
<https://doi.org/10.1093/sysbio/syv033>

**Published:** 01 June 2015 **Article history ▾**



@JLSteenwyk

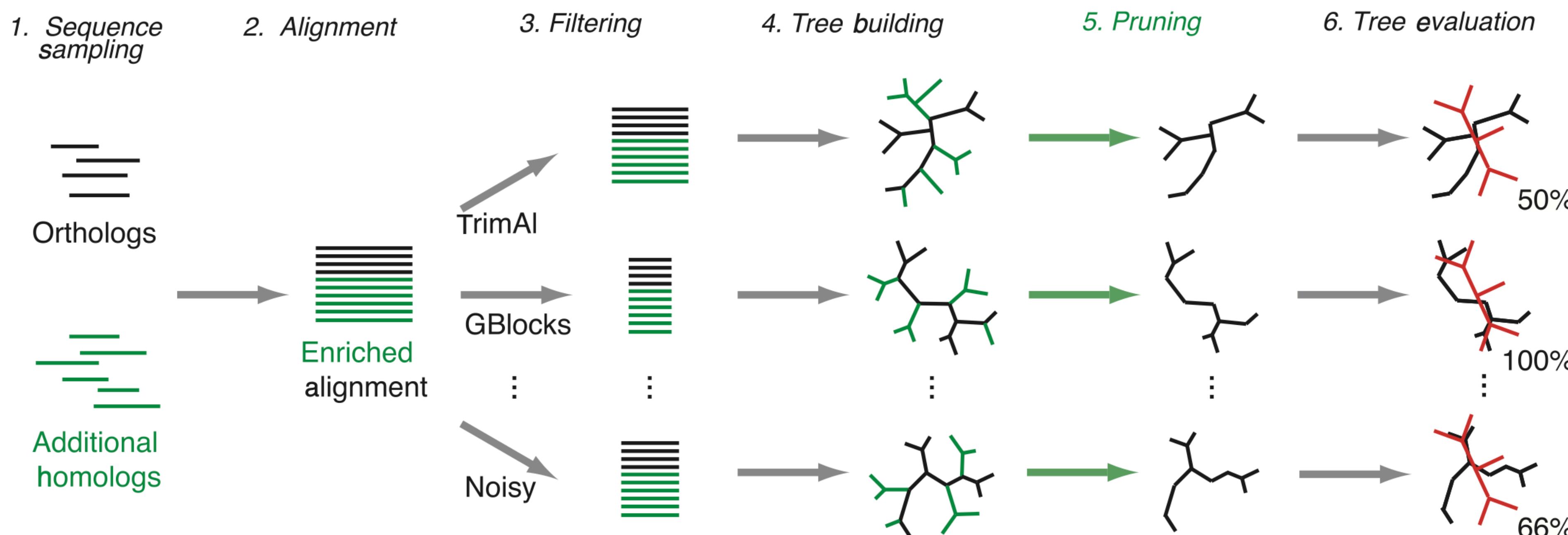
# Methods examined; most remove divergent sites

Method	Sites removed	Reference
<b>Gblocks</b>	<b>Gap-rich and variable</b>	Talavera and Castresana (2007)
<b>TrimAl</b>	<b>Gap-rich and variable</b>	Capella-Gutiérrez et al. (2009)
<b>Noisy</b>	<b>Homoplastic sites</b>	Dress et al. (2008)
<b>Aliscore</b>	<b>Random-like sites</b>	Kück et al. (2010)
<b>BMGE</b>	<b>High entropy sites</b>	Criscuolo and Ribaldo (2010)
<b>Zorro</b>	Sites with low posterior	Wu et al. (2012)
<b>Guidance</b>	Sensitive to guide tree	Penn et al. (2010)



@JLSteenwyk

# Testing the impact of trimming



@JLSteenwyk

# Sequence sampling using orthologs (& additional homologs)

## 1. Sequence sampling



Orthologs

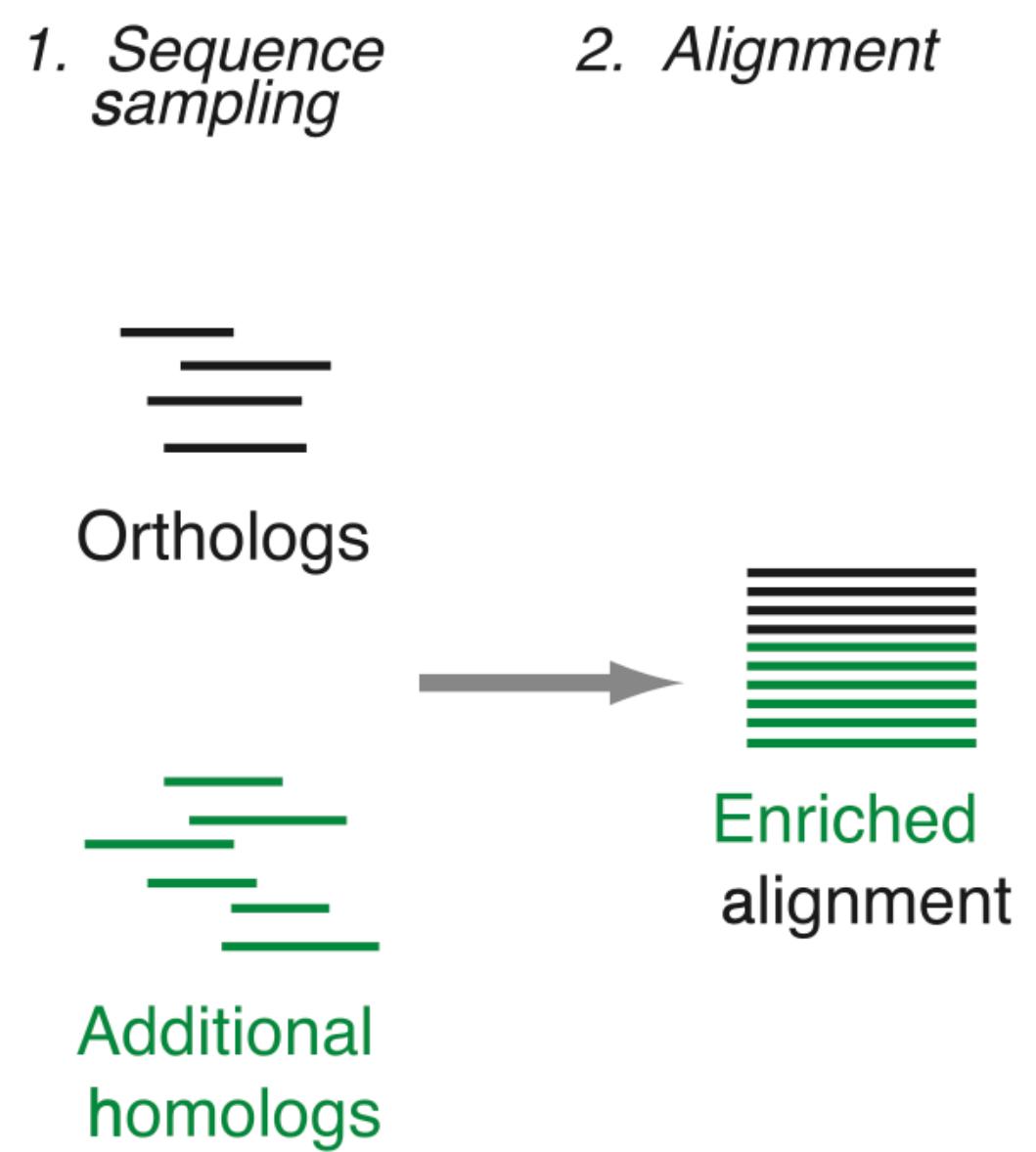


Additional homologs

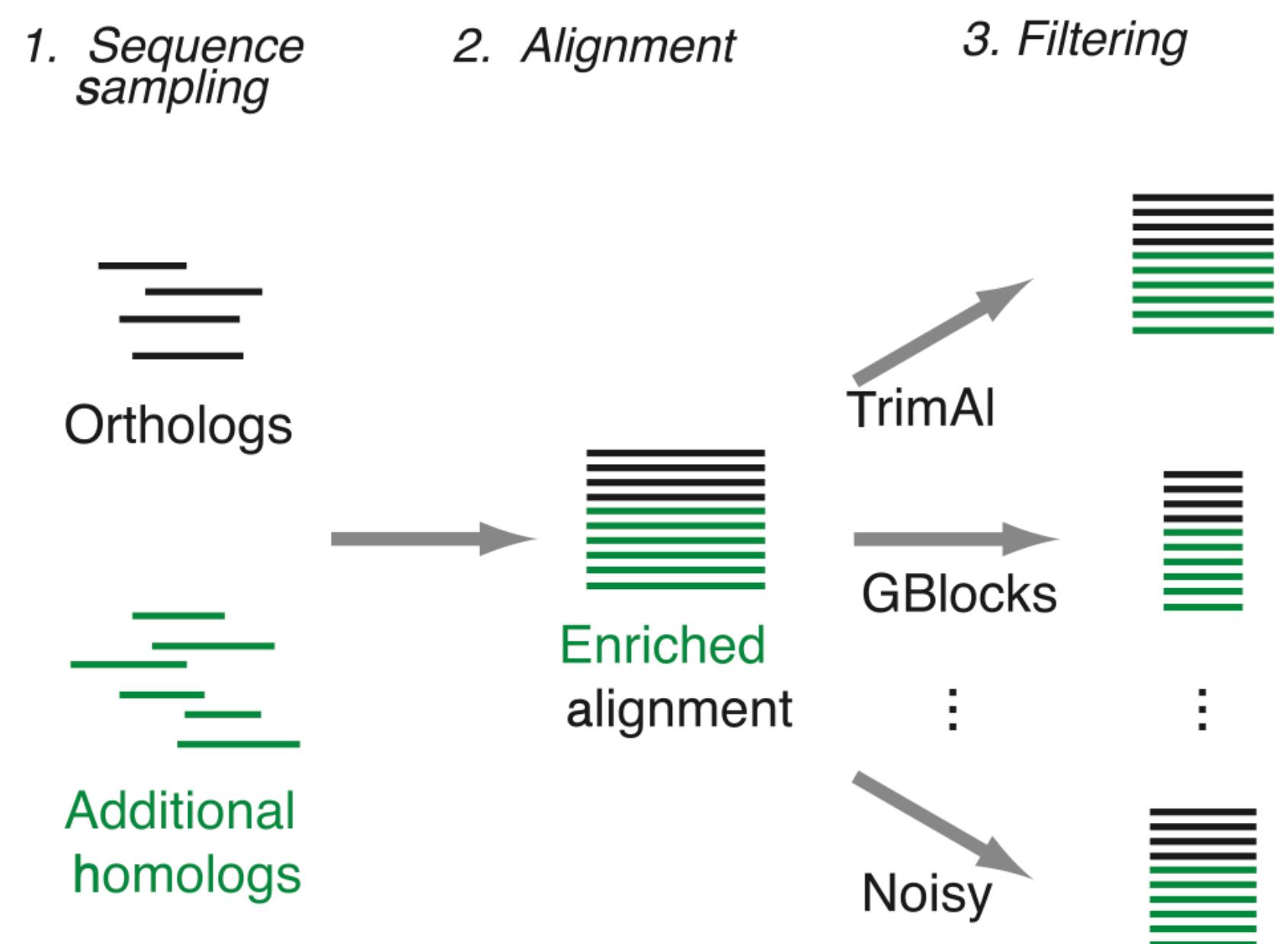


@JLSteenwyk

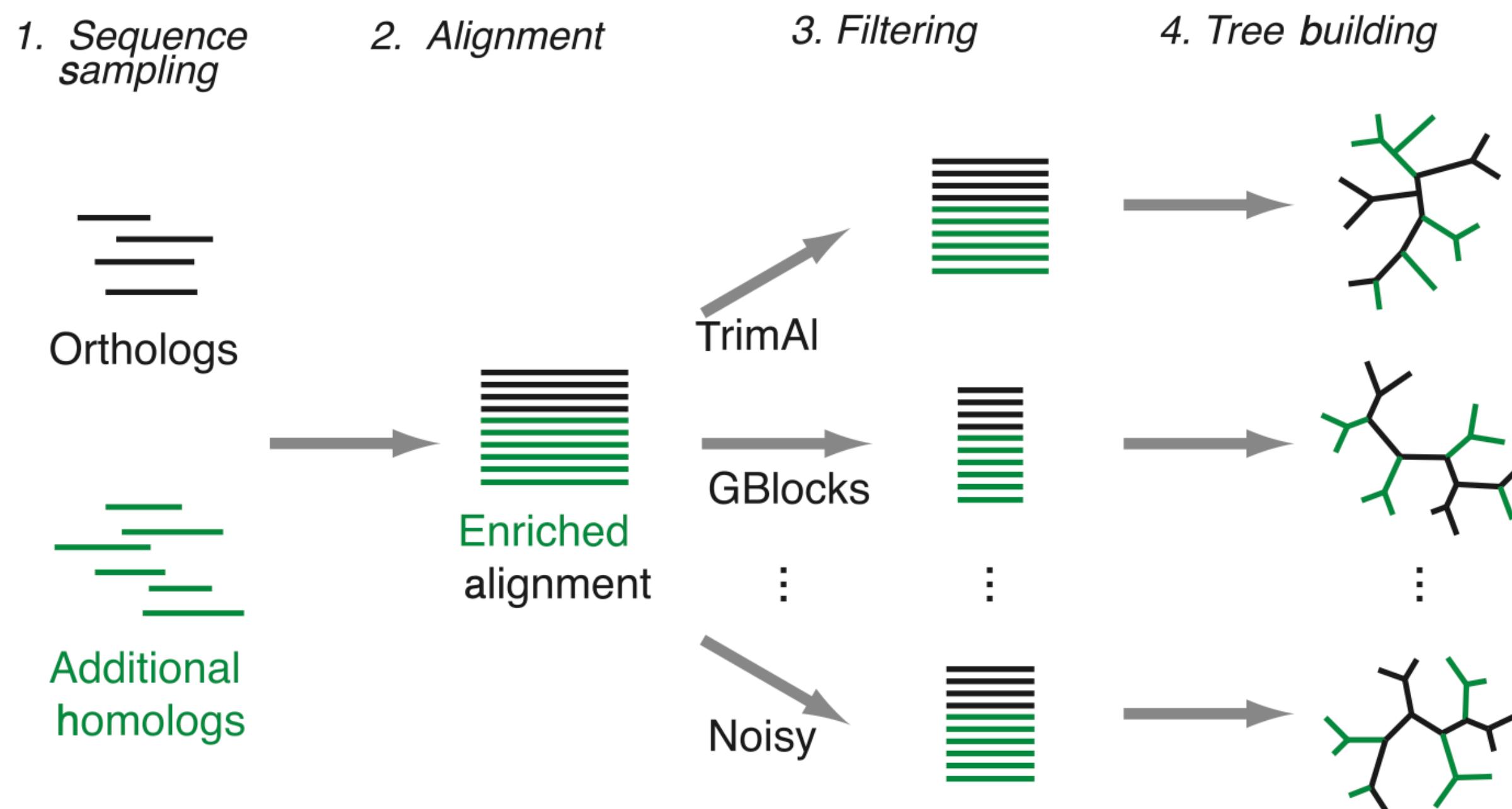
# Aligning sequences



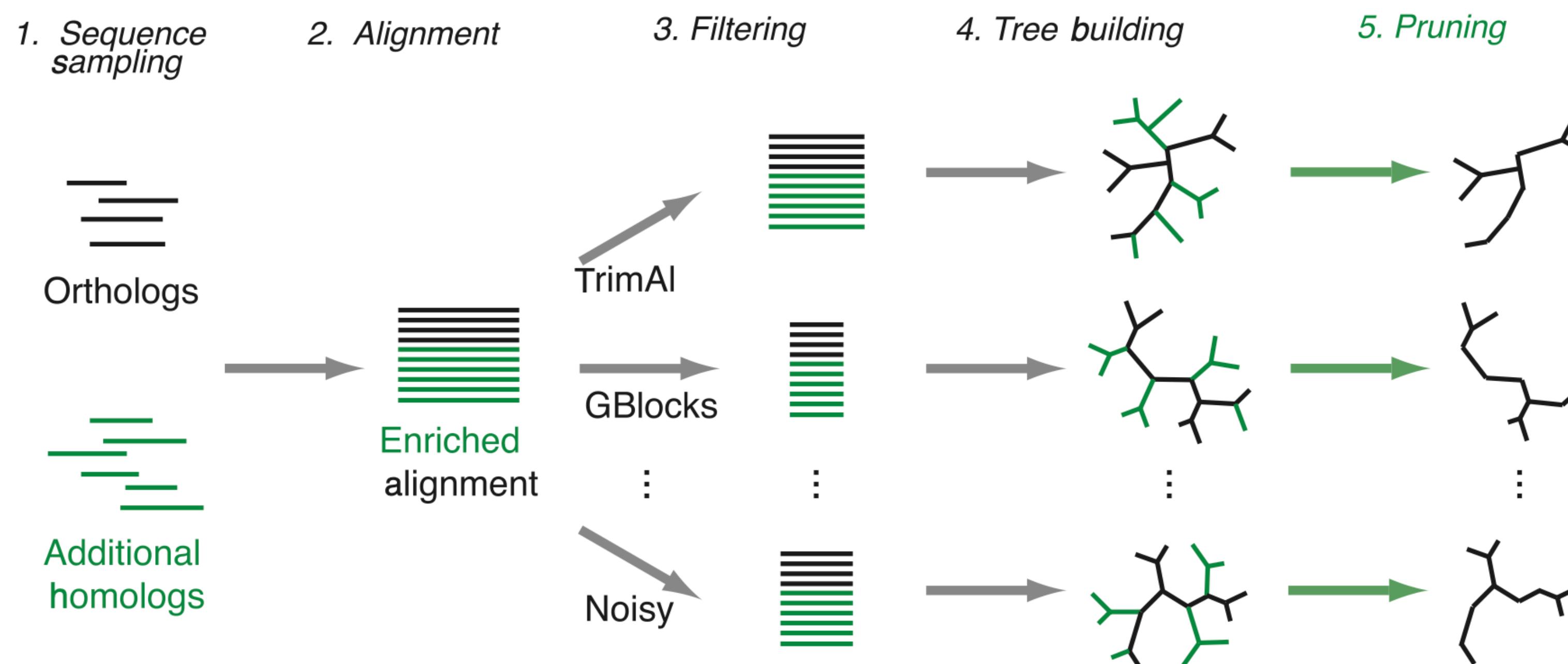
# Trimming using diverse software



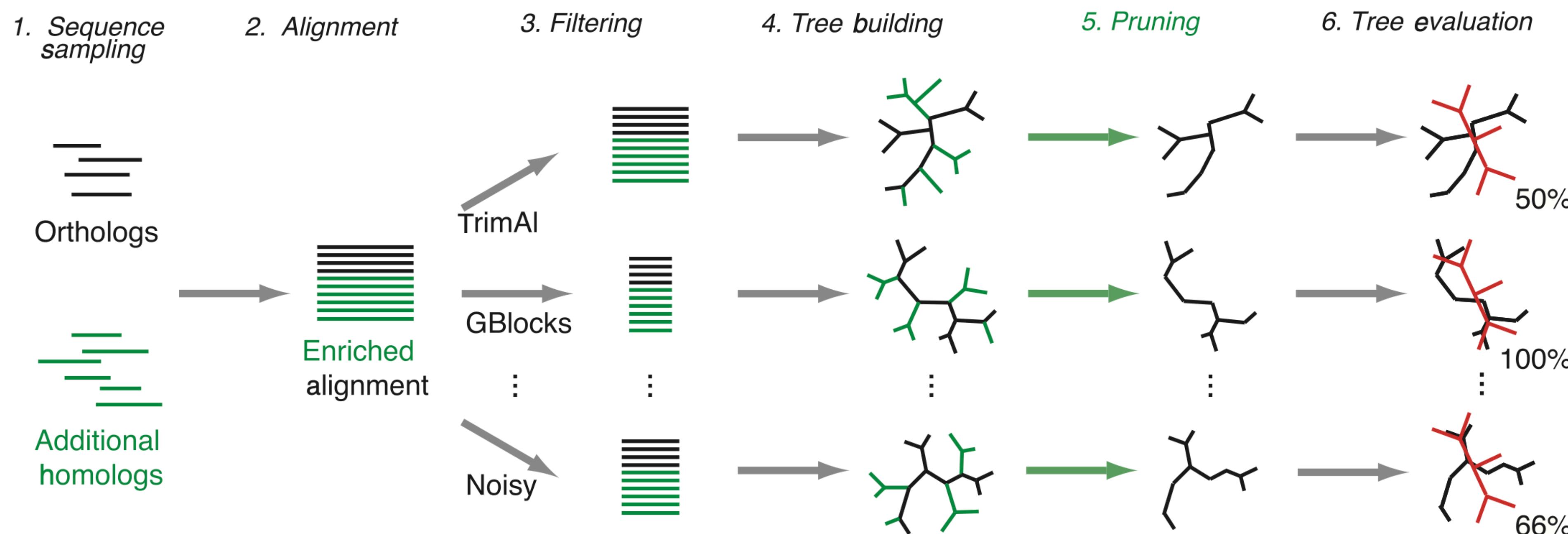
# Inferring trees from trimmed MSAs



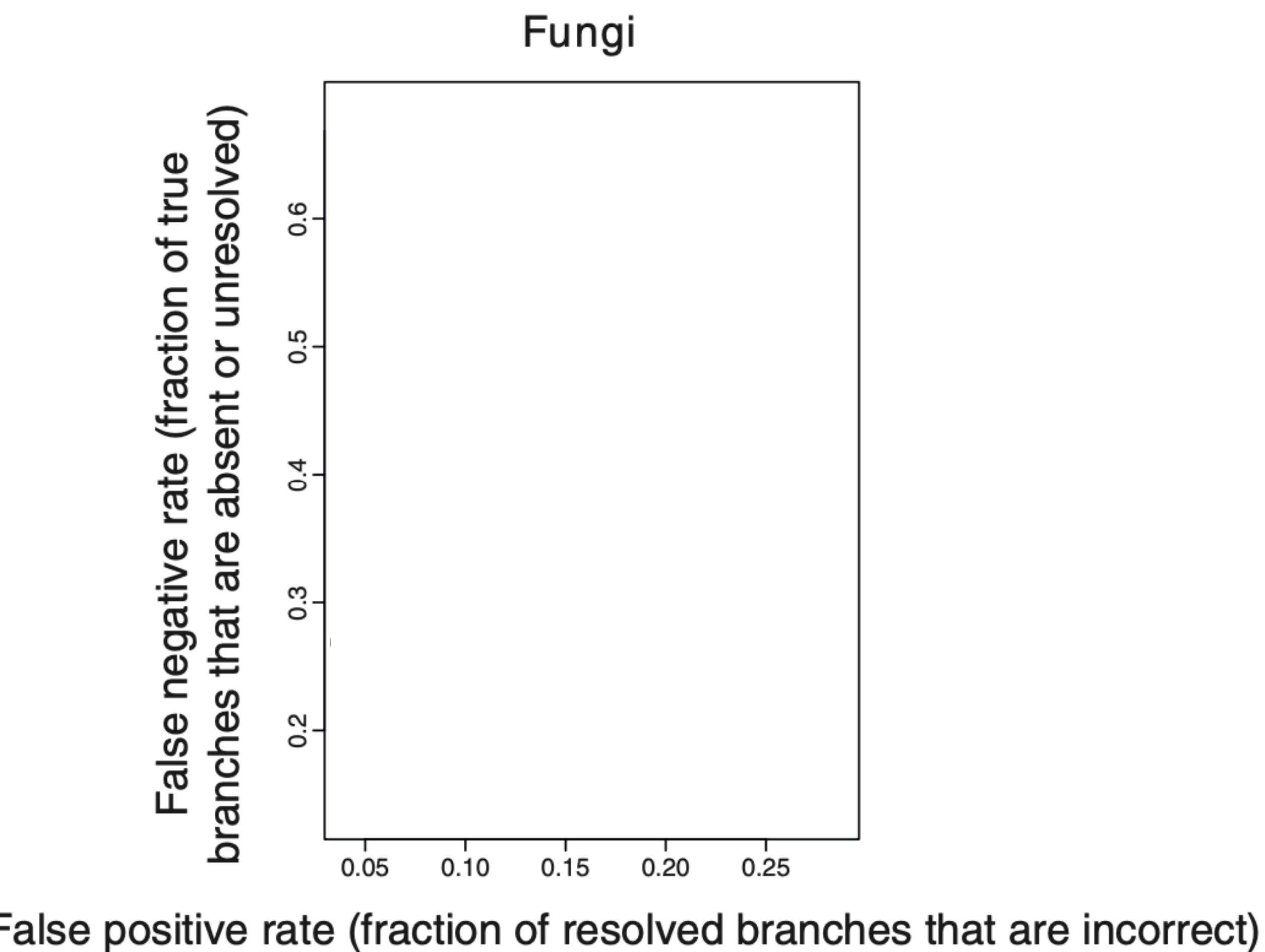
# Tree pruning to organisms with “incontestable” relationships



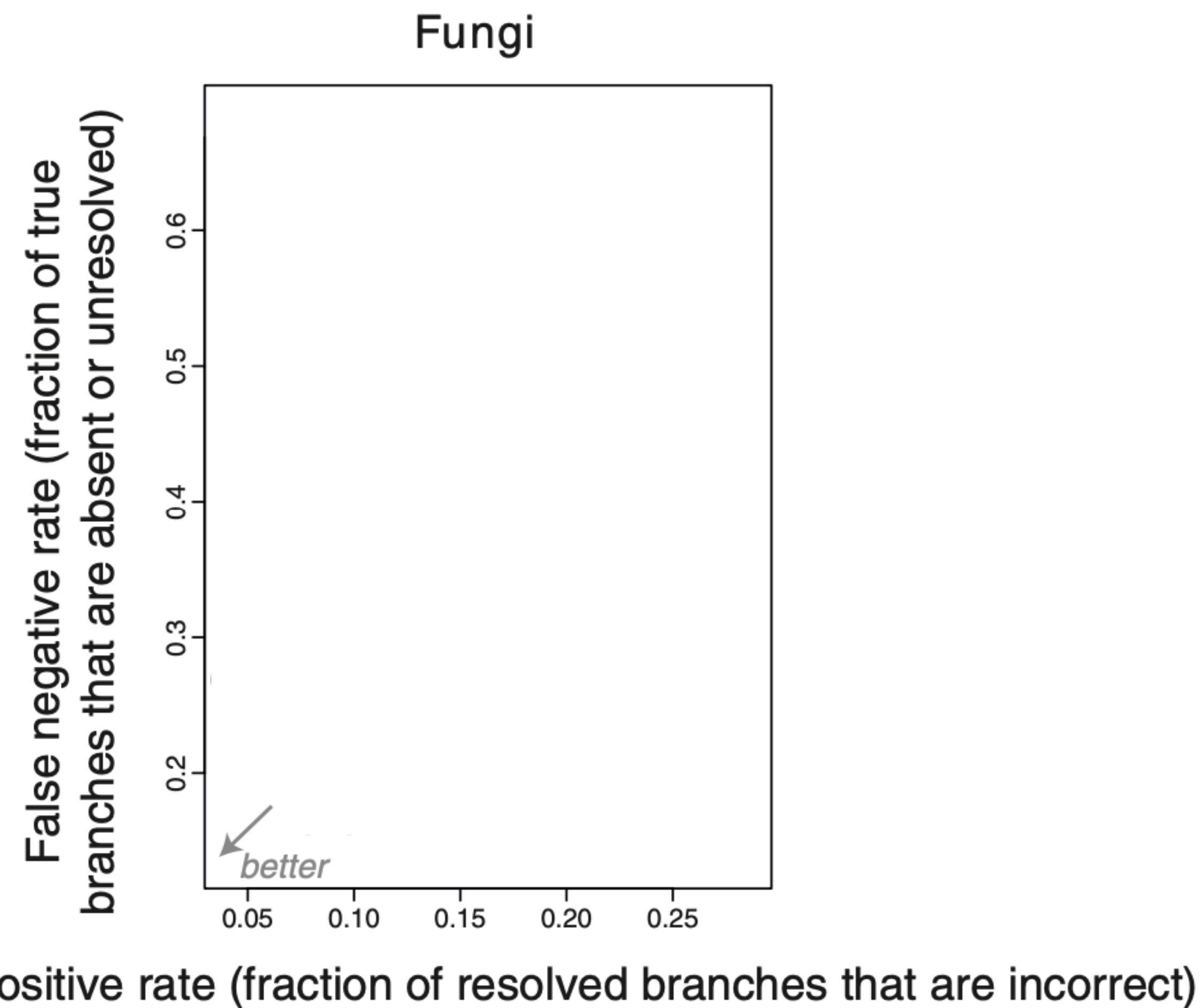
# Comparing inferred tree to “incontestable” tree



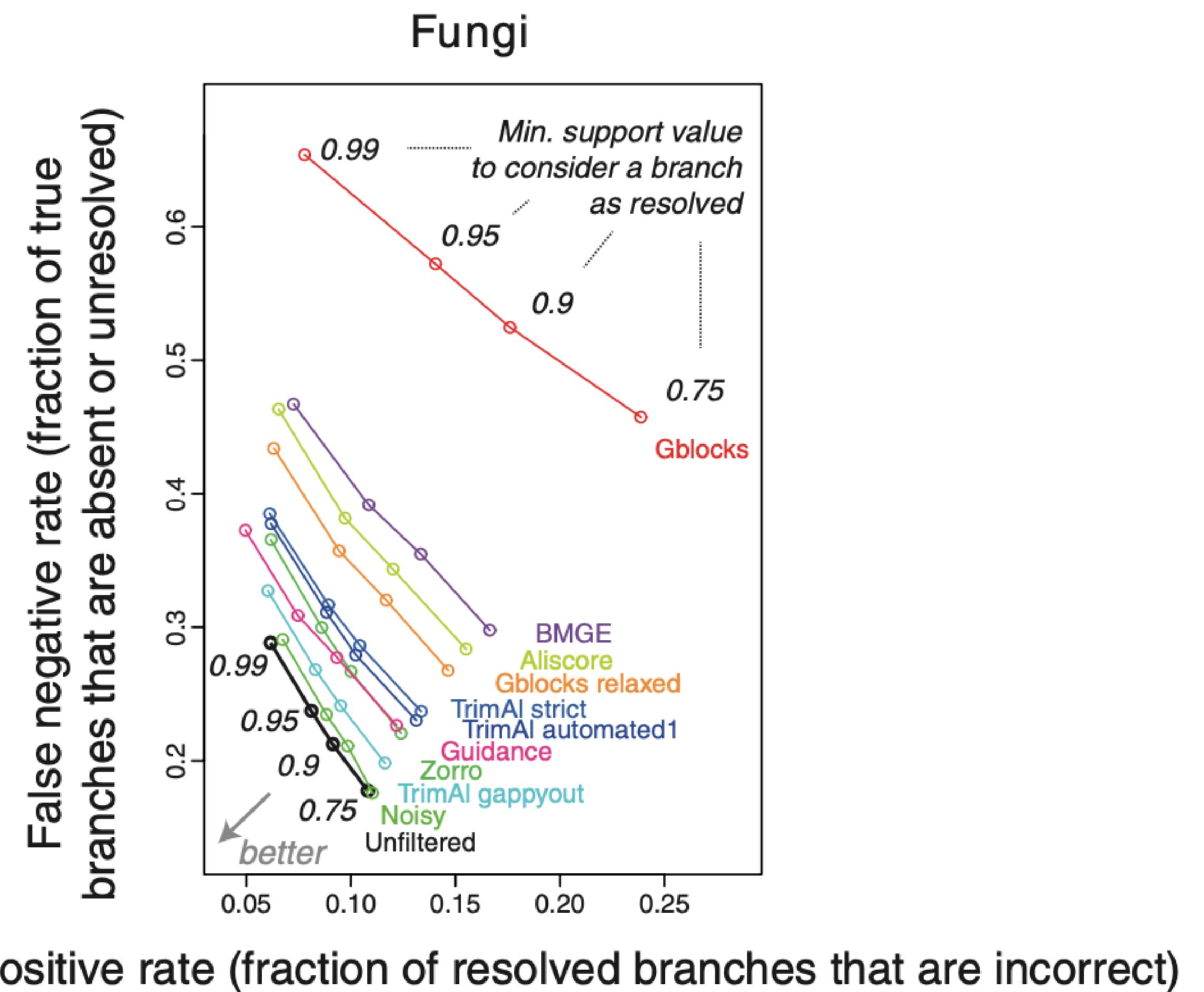
# Testing the impact of trimming



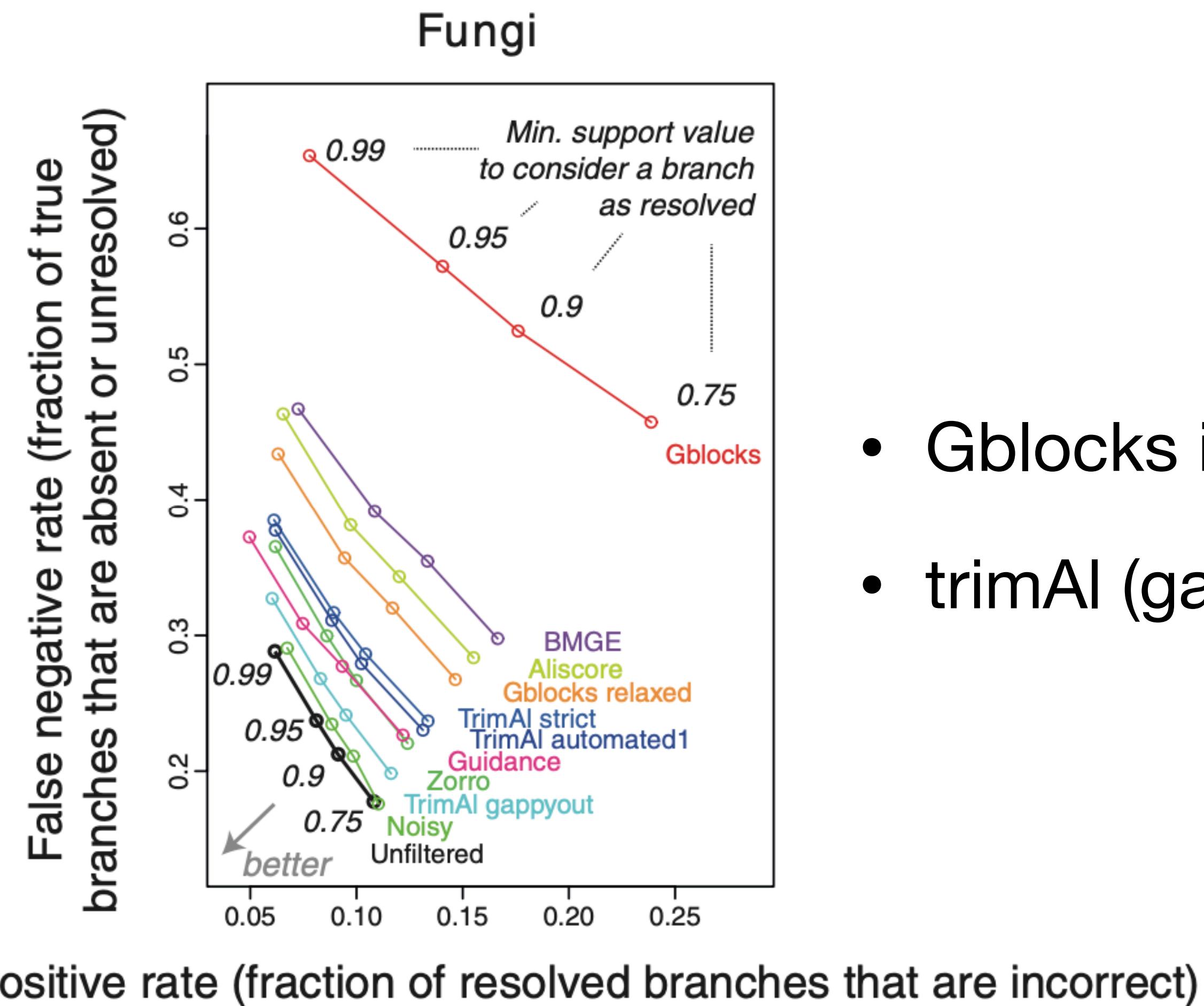
# Testing the impact of trimming



# Testing the impact of trimming

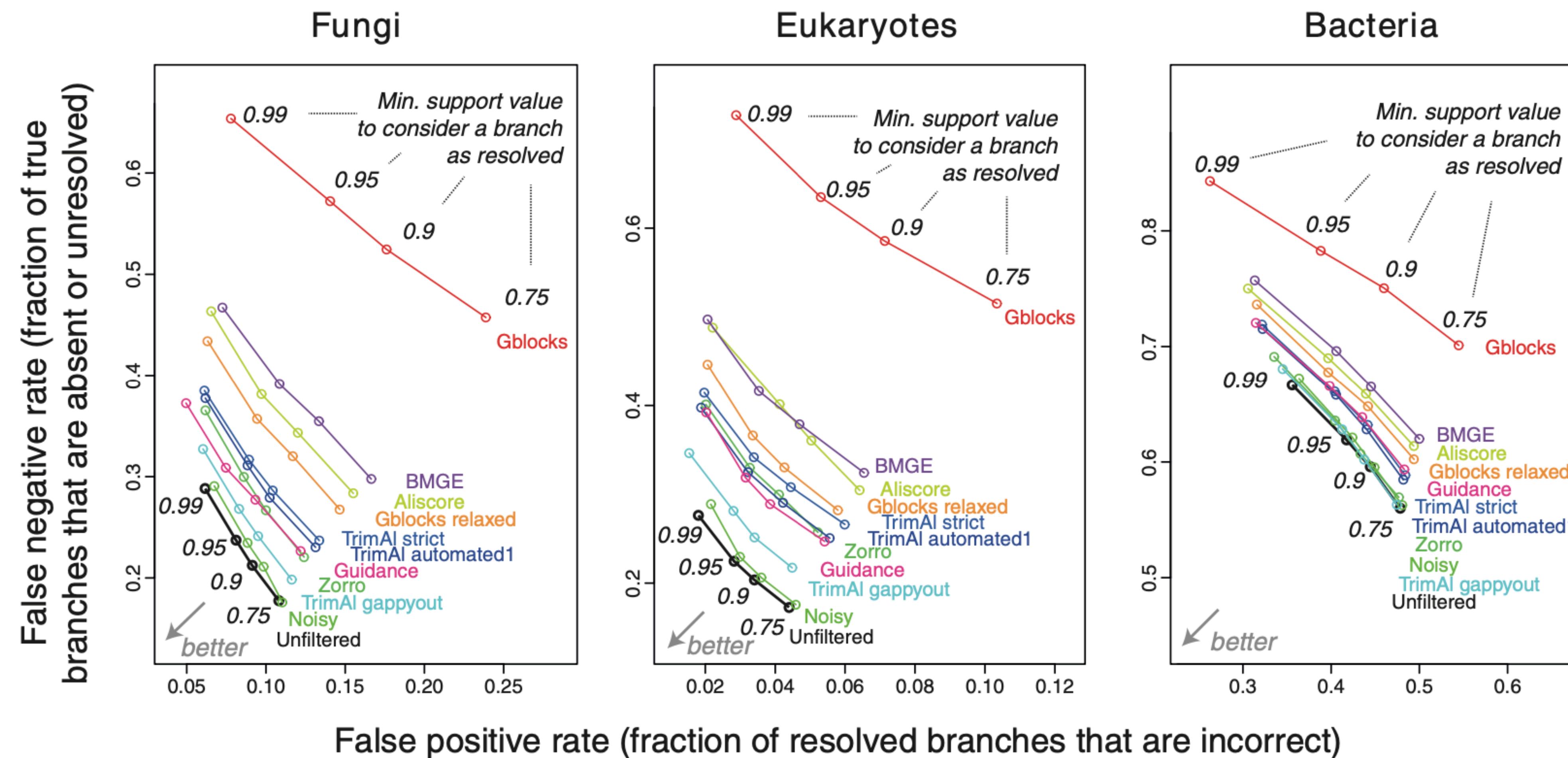


# Testing the impact of trimming



- Gblocks is an aggressive trimmer
- trimAI (gappyout) conducts “lighter” trimming

# Testing the impact of trimming



# Take home message



@JLSteenwyk

# Take home message

- Alignment trimming often resulted in lower phylogenetic signal in an alignment

# Take home message

- Alignment trimming often resulted in lower phylogenetic signal in an alignment
- The more aggressive the trimmer, the worse it performed

# Take home message

- Alignment trimming often resulted in lower phylogenetic signal in an alignment
- The more aggressive the trimmer, the worse it performed

“Although our results suggest that **light filtering** (up to 20% of alignment positions) has little impact on tree accuracy and may save some computation time, contrary to widespread practice, we do not generally recommend the use of current alignment filtering methods for phylogenetic inference”



# Take home message

- Alignment trimming often resulted in lower phylogenetic signal in an alignment
- The more aggressive the trimmer, the worse it performed

“Although our results suggest that **light filtering** (up to 20% of alignment positions) has little impact on tree accuracy and may save some computation time, contrary to widespread practice, we do not generally recommend the use of current alignment filtering methods for phylogenetic inference”

**This suggest current methods remove sites with phylogenetic signal**

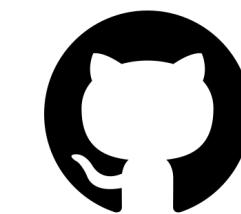
What if we kept sites with phylogenetic signal?

# ClipKit

---

the alignment trimming toolkit

Jacob L. Steenwyk, Thomas J. Buida III,  
Yuanning Li, Xing-Xing Shen, Antonis Rokas



JLSteenwyk



@JLSteenwyk

# ClipKIT has several modes

- based on keeping parsimony informative sites and sites that aren't gappy-rich

# ClipKIT has several modes

- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
  - kpi (keeps only parsimony informative sites)

# ClipKIT has several modes

- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
  - kpi (keeps only parsimony informative sites)
  - kpic (keeps parsimony informative sites and constant sites)



@JLSteenwyk

# ClipKIT removes sites that aren't parsimony informative

## Untrimmed

```
>1  
A-GTAT  
>2  
A-G-AT  
>3  
A-G-TA  
>4  
AGA-TA  
>5  
ACa-T-
```

# ClipKIT removes sites that aren't parsimony informative

Untrimmed	kpi
>1	>1
A-GTAT	A-GTAT
>2	>2
A-G-AT	A-G-AT
>3	>3
A-G-TA	A-G-TA
>4	>4
AGA-TA	AGA-TA
>5	>5
ACa-T-	ACa-T-

# ClipKIT removes sites that aren't parsimony informative

Untrimmed	kpi	kpic
>1	>1	>1
A-GTAT	A-GTAT	A-GTAT
>2	>2	>2
A-G-AT	A-G-AT	A-G-AT
>3	>3	>3
A-G-TA	A-G-TA	A-G-TA
>4	>4	>4
AGA-TA	AGA-TA	AGA-TA
>5	>5	>5
ACa-T-	ACa-T-	ACa-T-

# ClipKIT has several modes

- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
  - kpi (keeps only parsimony informative sites)
  - kpic (keeps parsimony informative sites and constant sites)
  - smart-gap (dynamic gappyness threshold determination)

# ClipKIT has several modes

- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
  - kpi (keeps only parsimony informative sites)
  - kpic (keeps parsimony informative sites and constant sites)
  - smart-gap (dynamic gappyness threshold determination)
  - gappy (removes sites with >90% gaps)



@JLSteenwyk

# ClipKIT has several modes

- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
  - kpi (keeps only parsimony informative sites)
  - kpic (keeps parsimony informative sites and constant sites)
  - smart-gap (dynamic gappyness threshold determination)
  - gappy (removes sites with >90% gaps)
  - combinations of kpi/kpic and gappy-based trimming can be used
    - e.g., kpic-smart-gap or kpi-gappy



@JLSteenwyk

# ClipKIT has several modes

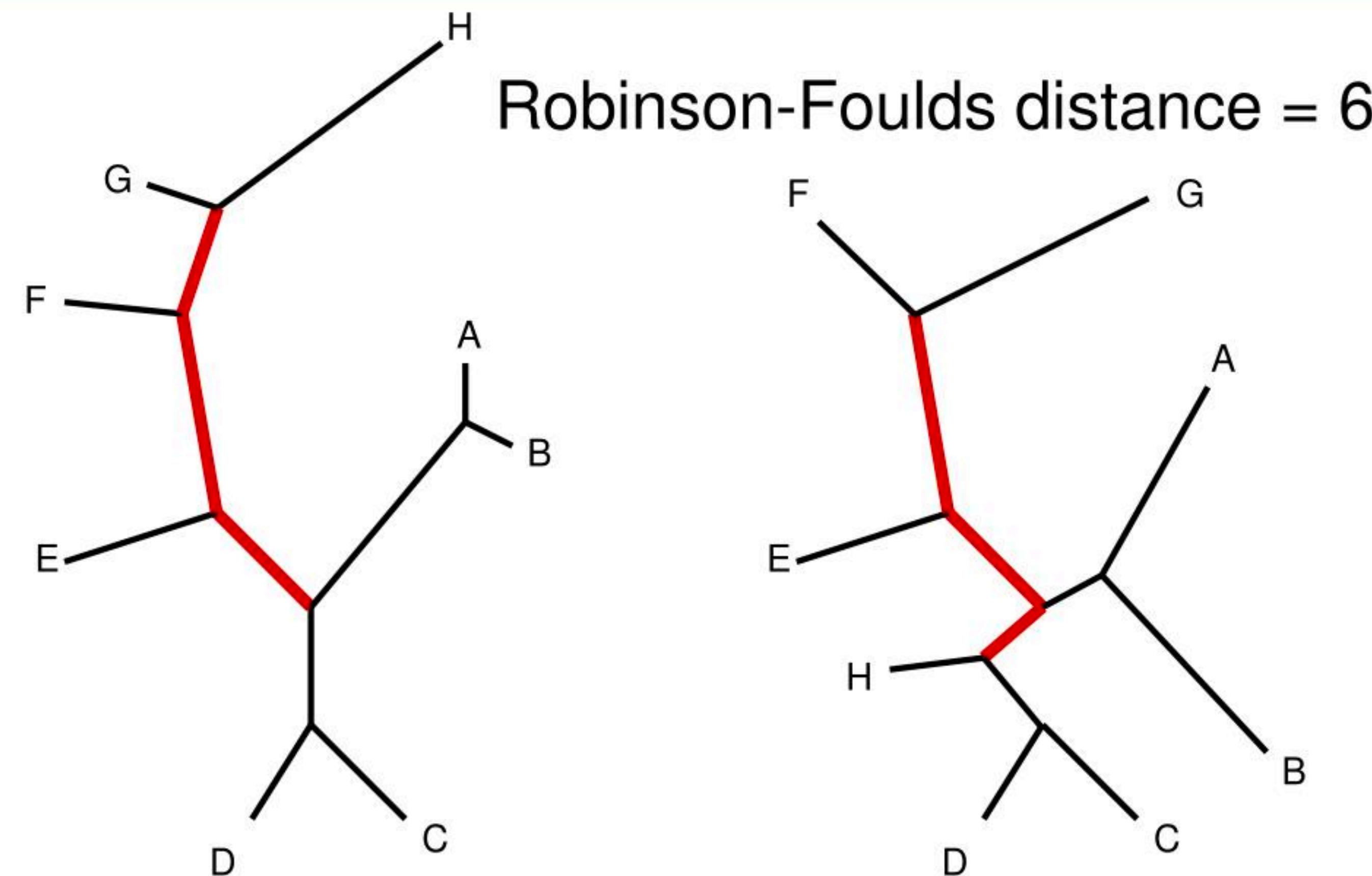
- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
  - kpi (keeps only parsimony informative sites)
  - kpic (keeps parsimony informative sites and constant sites)
  - smart-gap (dynamic gappyness threshold determination)
  - gappy (removes sites with >90% gaps)
  - combinations of kpi/kpic and gappy-based trimming can be used
    - e.g., kpic-smart-gap or kpi-gappy
- ClipKIT focuses on **keeping sites rich** in phylogenetic signal rather than identifying and removing those that lack signal



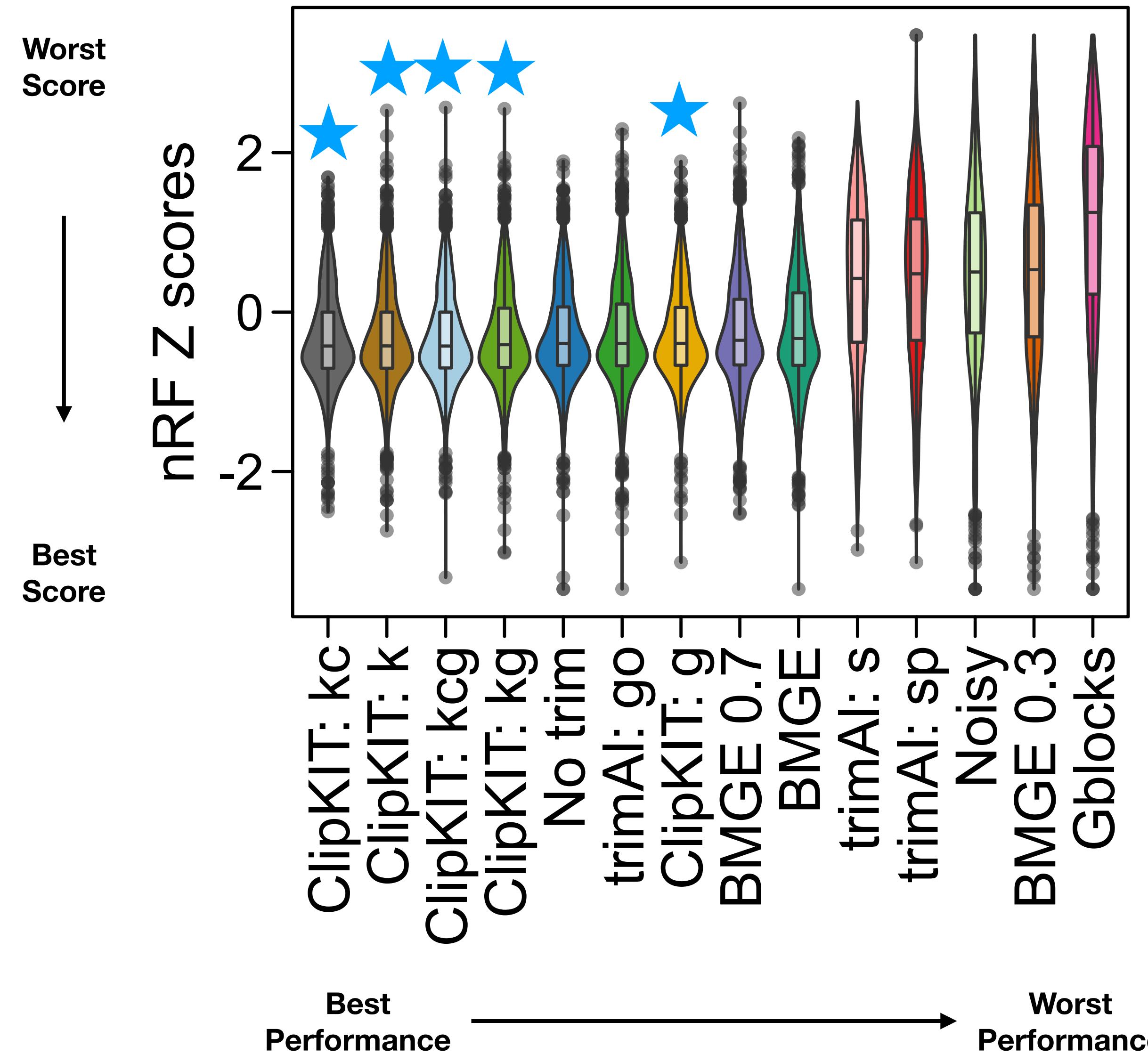
@JLSteenwyk

# Measuring accuracy between inferred & expected tree

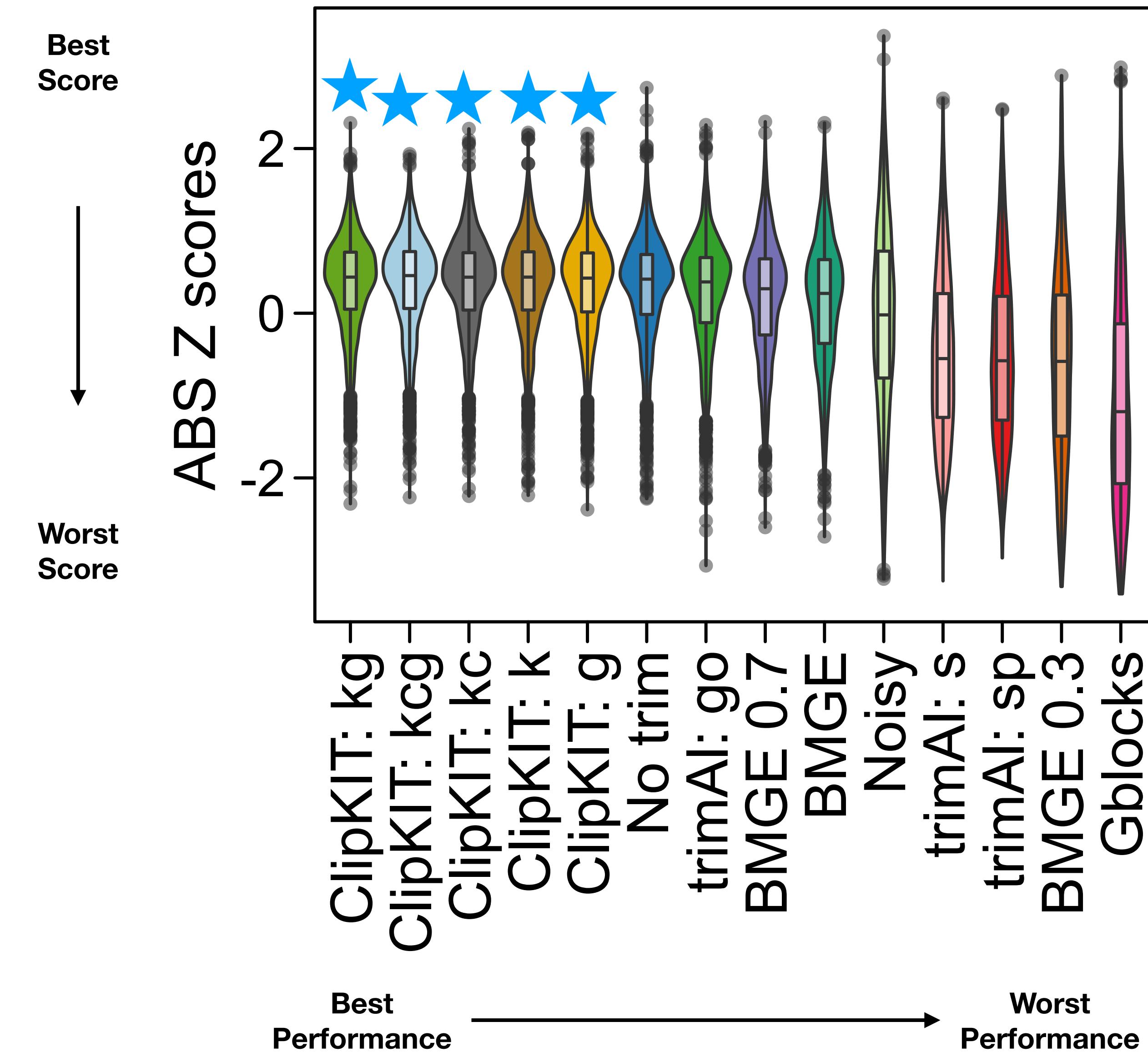
**Internal branches exist in one tree but not in the other**



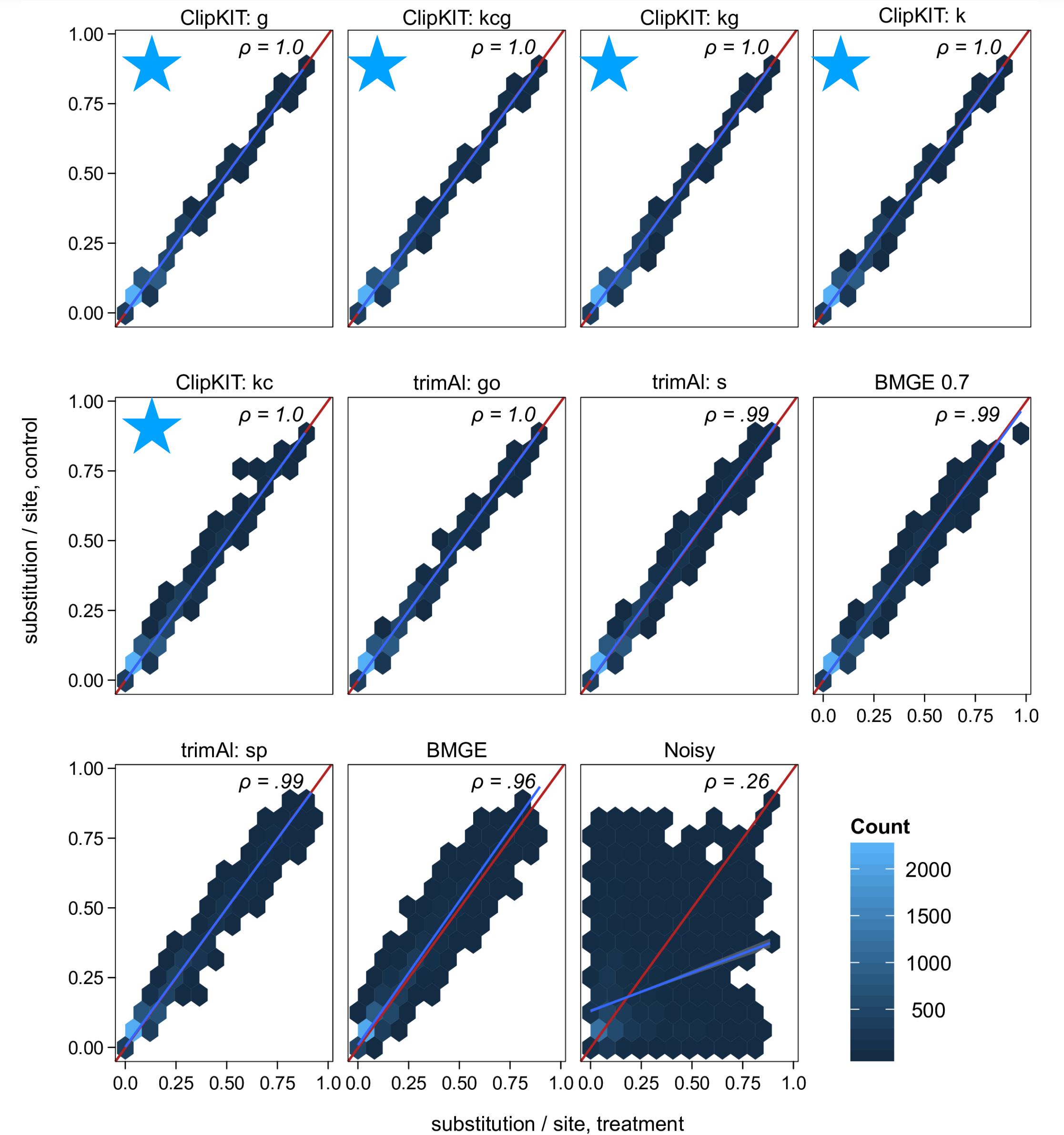
# Trees inferred using ClipKIT are accurate



# Trees inferred using ClipKIT are well supported

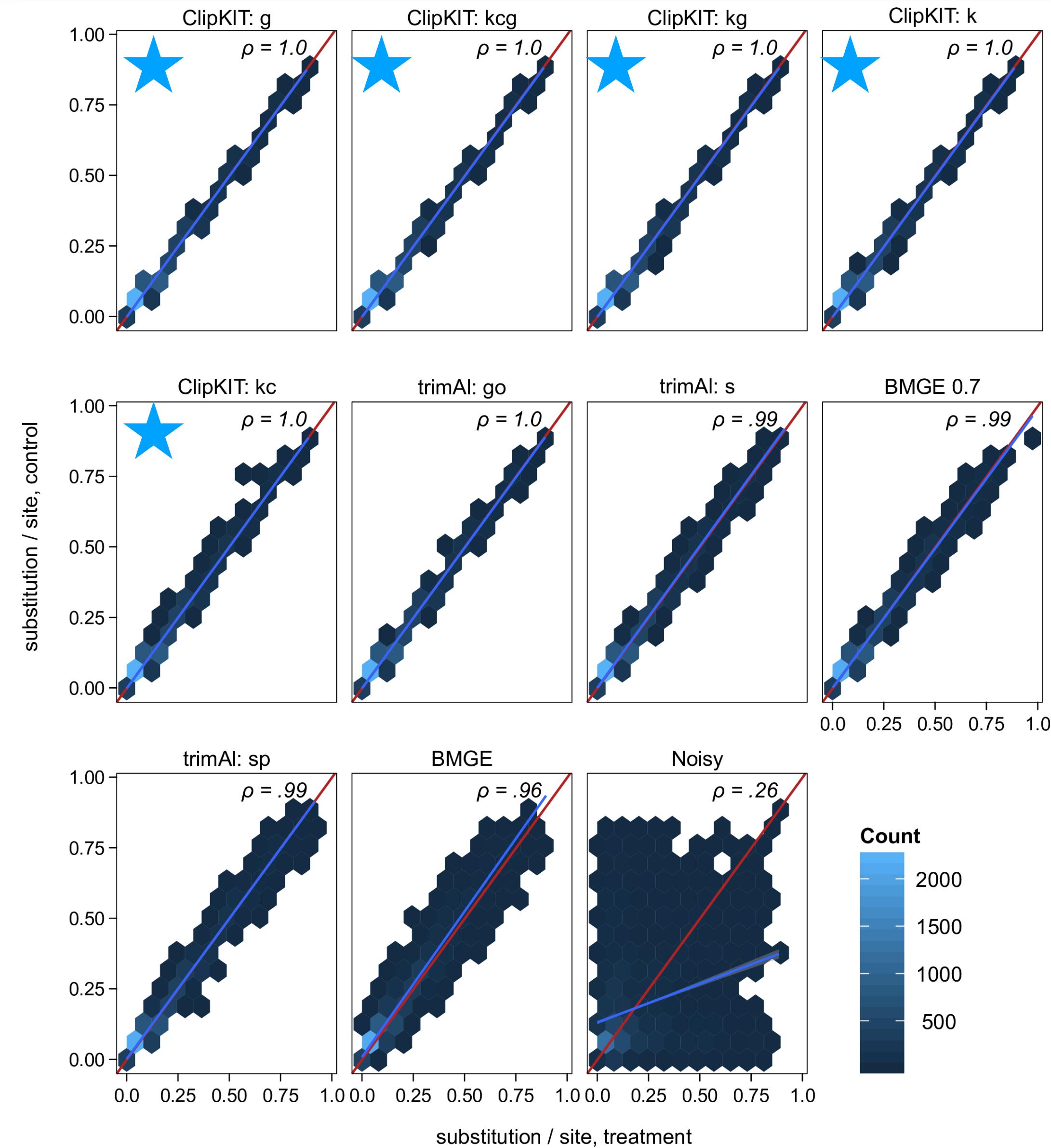


# Branch lengths estimates after trimming are typically accurate



**Best Performance** →  
→  
→ **Worst Performance**

# Branch lengths estimates after trimming are typically accurate



\* we also found ClipKIT trimmed alignments that were shorter than other methods still outperformed the other methods

# ClipKIT

<https://clipkit.genomelybio.com/>

Home

Help

CLI Docs

Other Software

Contact

Choose FASTA(s)

Mode

smart-gap (default) ▾

Trim FASTA(s)

Sequence Type

Auto detect ▾

ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference.

Steenwyk et al. PLoS Biology. doi: [10.1371/journal.pbio.3001007](https://doi.org/10.1371/journal.pbio.3001007)

# Thank you for your time and attention!

## King Lab

Becca Arruda  
Chrisa Staikou  
Alain Garcia De Las Bayonas  
Maxwell C. Coyle  
Josean Reyes-Rivera  
Michael Carver  
Stefany Gonzalez



**hhmi**

Howard Hughes  
Medical Institute



## Rokas Lab

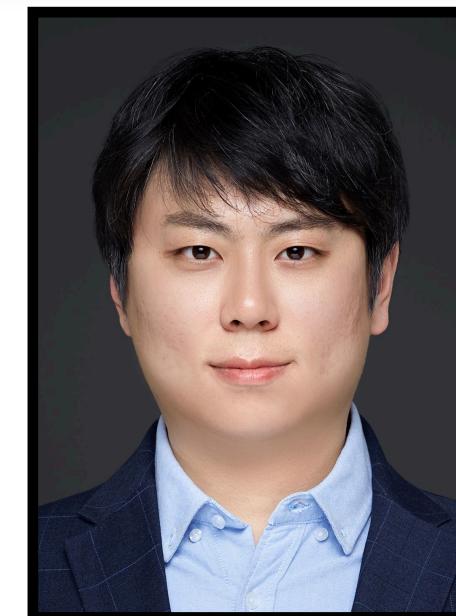
Megan Phillips  
Carla Gonçalves  
Matthew Mead  
Marie-Claire Harrison  
E. Anne Hatmaker  
Charu Balamurugan  
Thodoros Danis



Buida, J



O'Meara, T



Li, Y



Verbruggen, H



King, N



Coyle, M



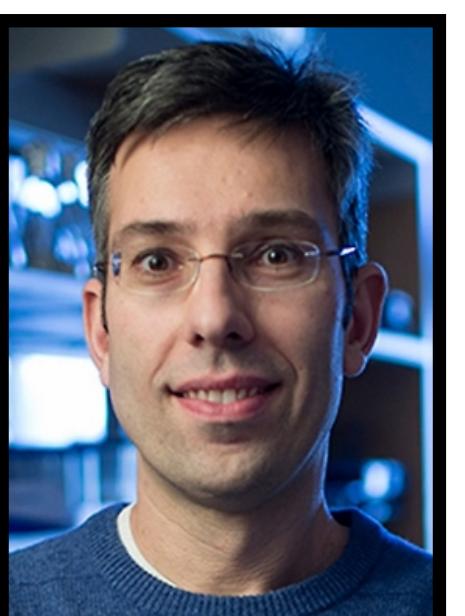
Geiser, D



Goldman, G

## Trainees

Saelin Bjornson  
Charu Balamurugan



Rokas, A



Hittinger, C



Berman, J



Shen, X

## Former Trainees

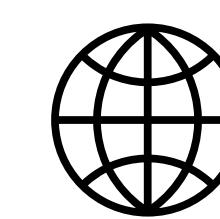
Olivia Zheng  
Megan Phillips

**Stay tuned for a silly quiz in the last 10 minutes!**

# Trimming MSAs



@JLSteenwyk

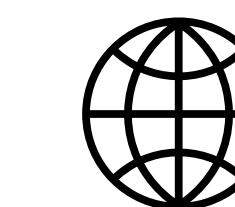


<https://jlsteenwyk.com/>

# Fun quiz, no winners...except each and every one of you!



@JLSteenwyk



<https://jlsteenwyk.com/>