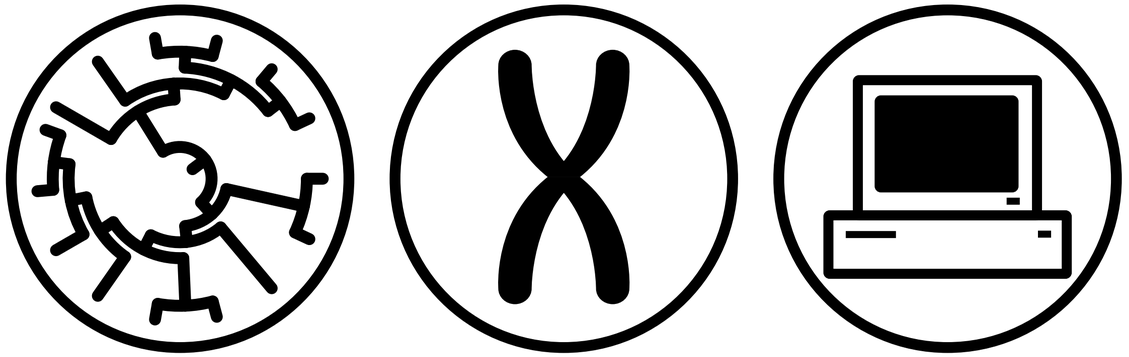
# Trimming MSAs

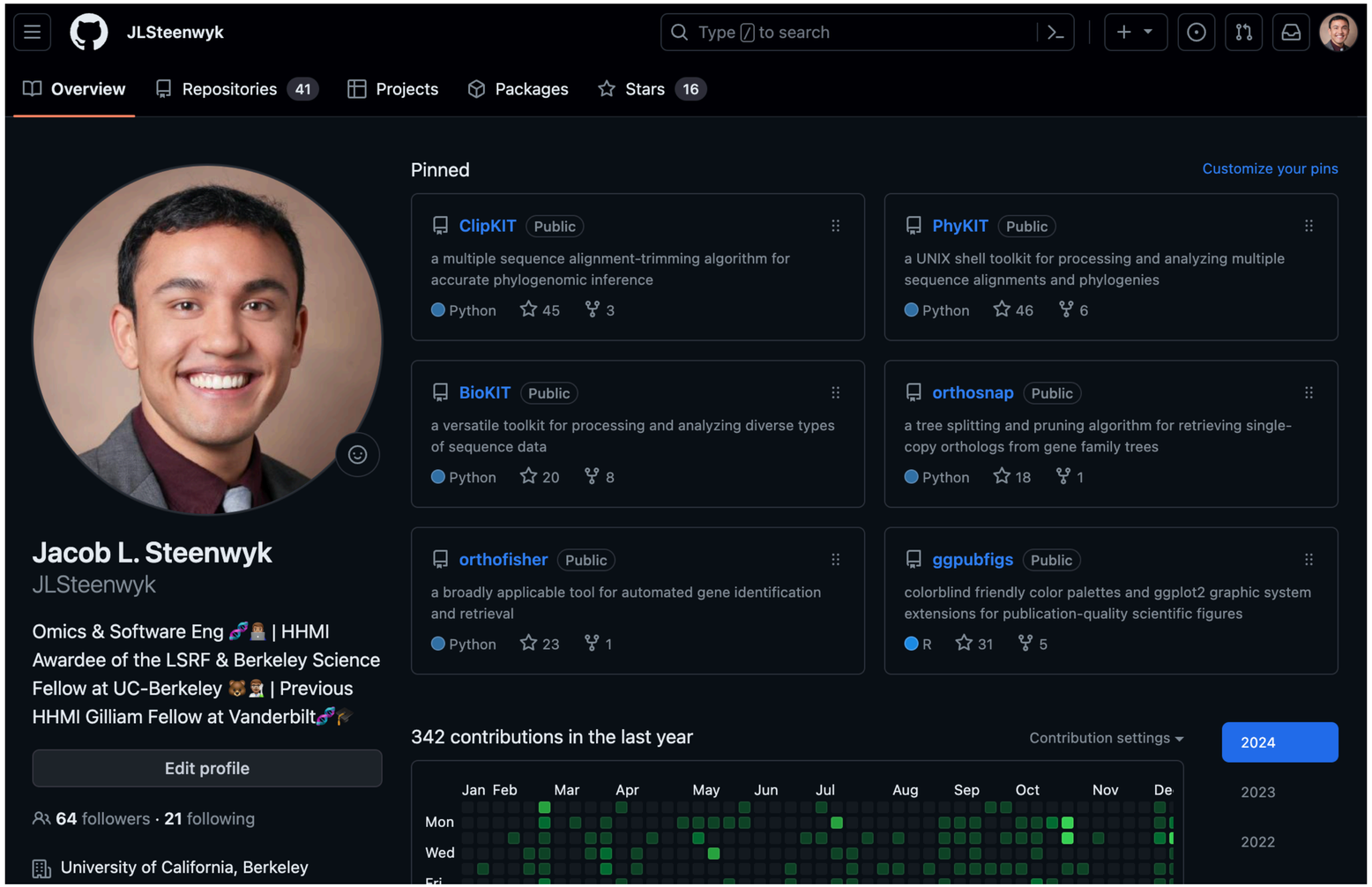
@JLSteenwyk https://jlsteenwyk.com/

# Outline



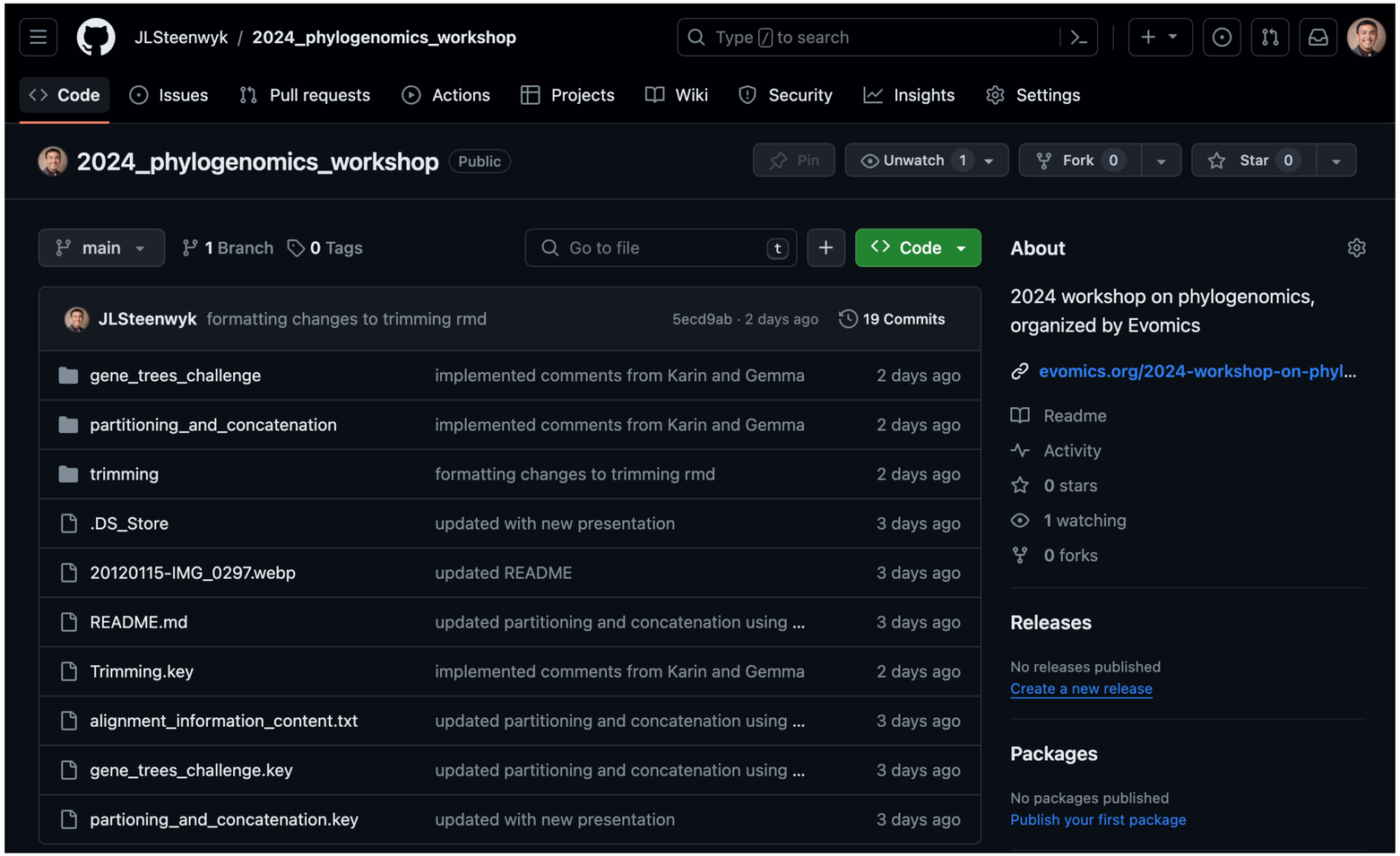
- Housekeeping
- Why trim? A brief history
- Trimming becomes a contentious topic
- ClipKIT implements a novel approach

# Raw course materials are available via GitHub




@JLSteenwyk

# Raw course materials are available via GitHub




@JLSteenwyk

# I recommend using links available on the website

## EVOLUTION AND GENOMICS

Intensive and immersive training opportunities

WORKSHOPS LEARNING PEOPLE APPLY INFORMATION

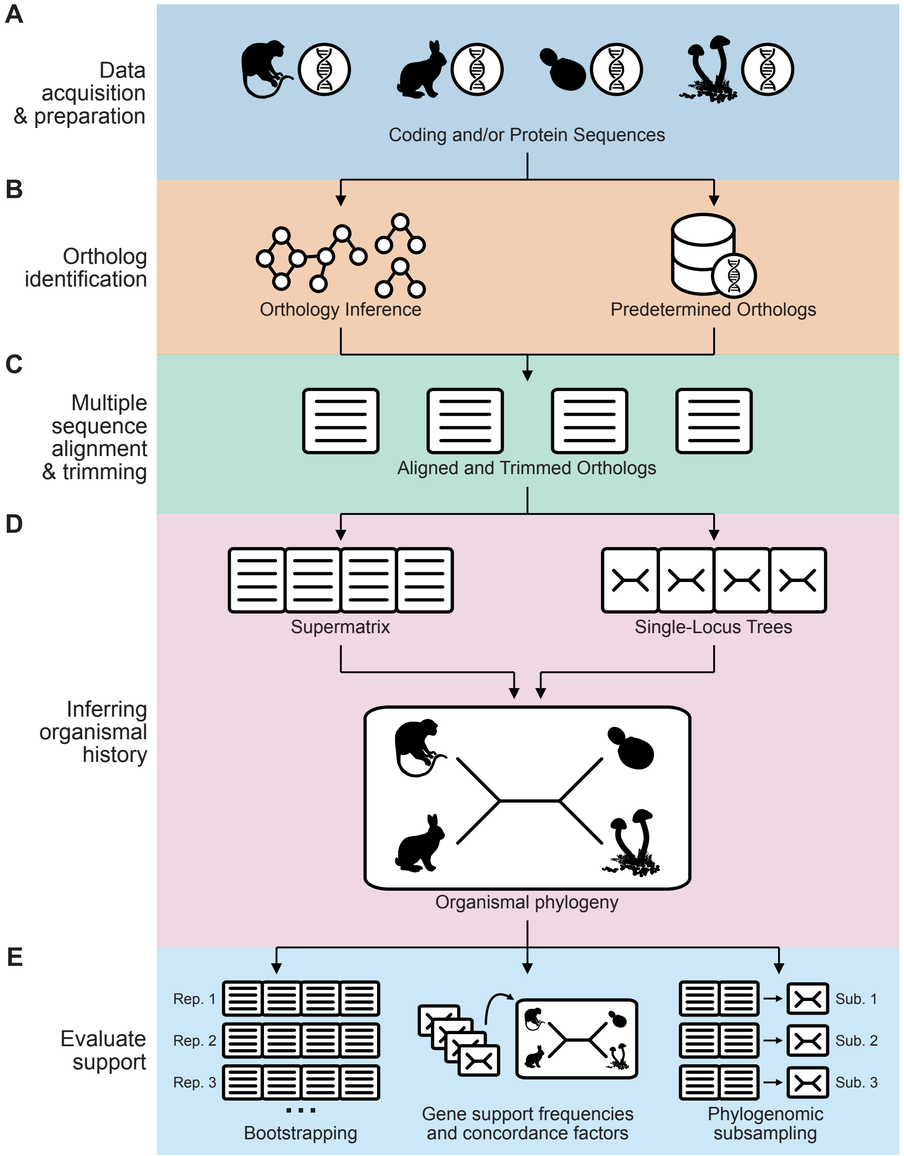
**Week 1 : 21-27 January, 2024**

| 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

@JLSteenwyk

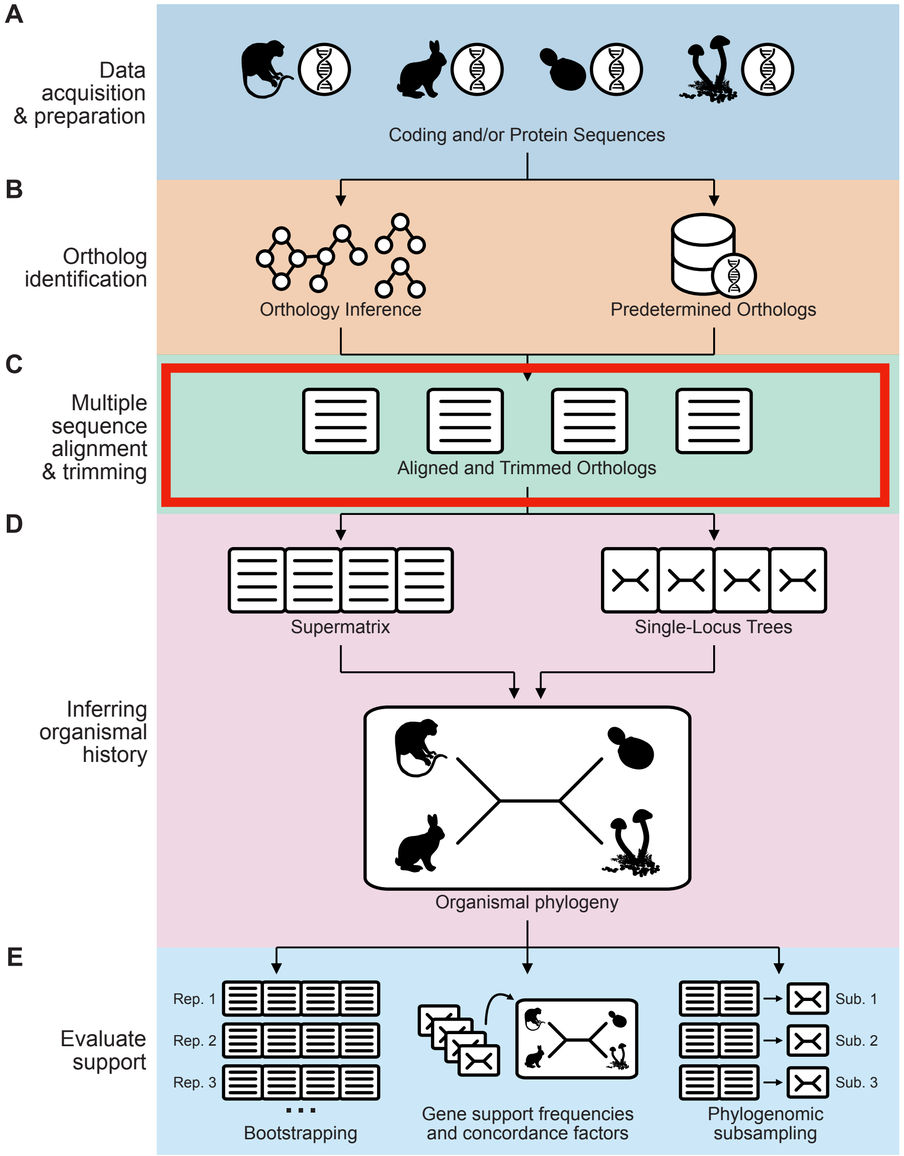
# Facilitating phylogenomic workflows and beyond




@JLSteenwyk

Bjornson et al. (2023), Preprints

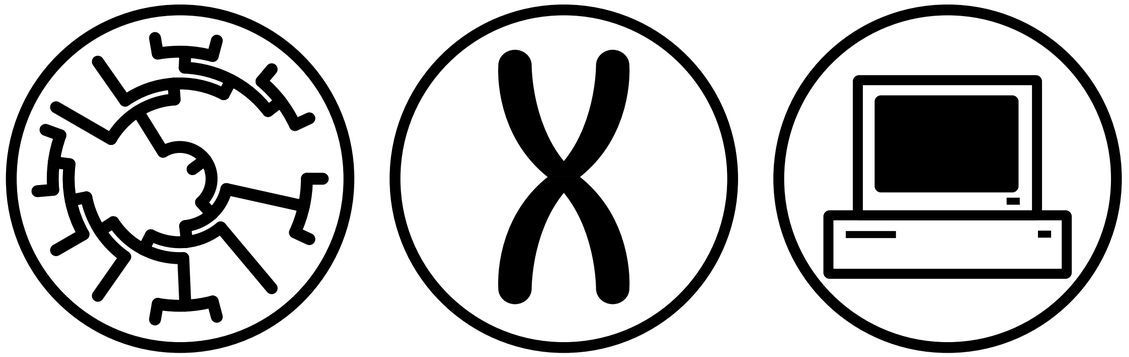
# Facilitating phylogenomic workflows and beyond




@JLSteenwyk

Bjornson et al. (2023), Preprints

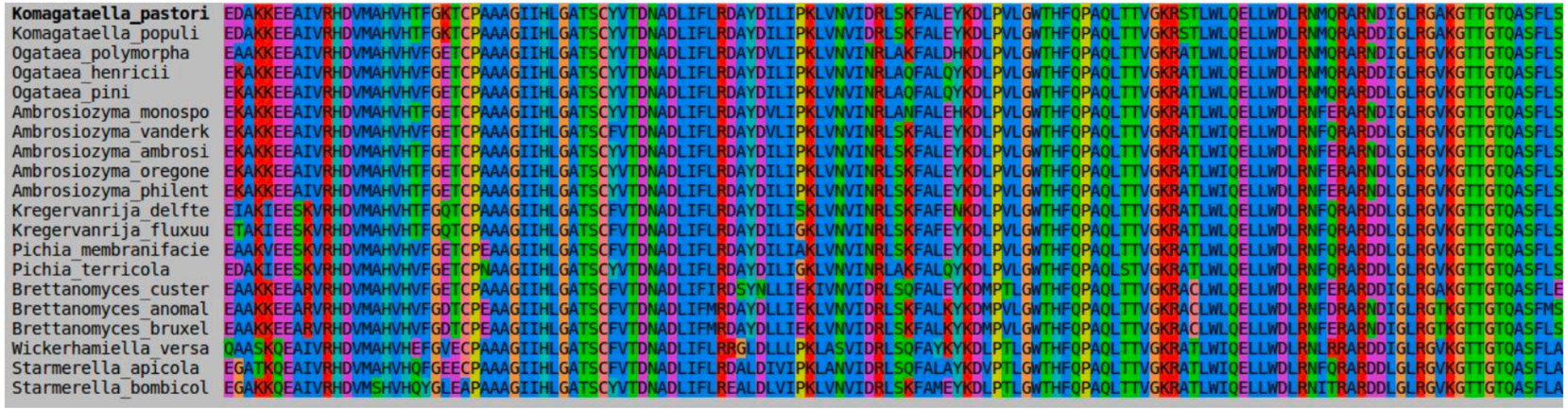
# Outline



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

# Variation in conservation

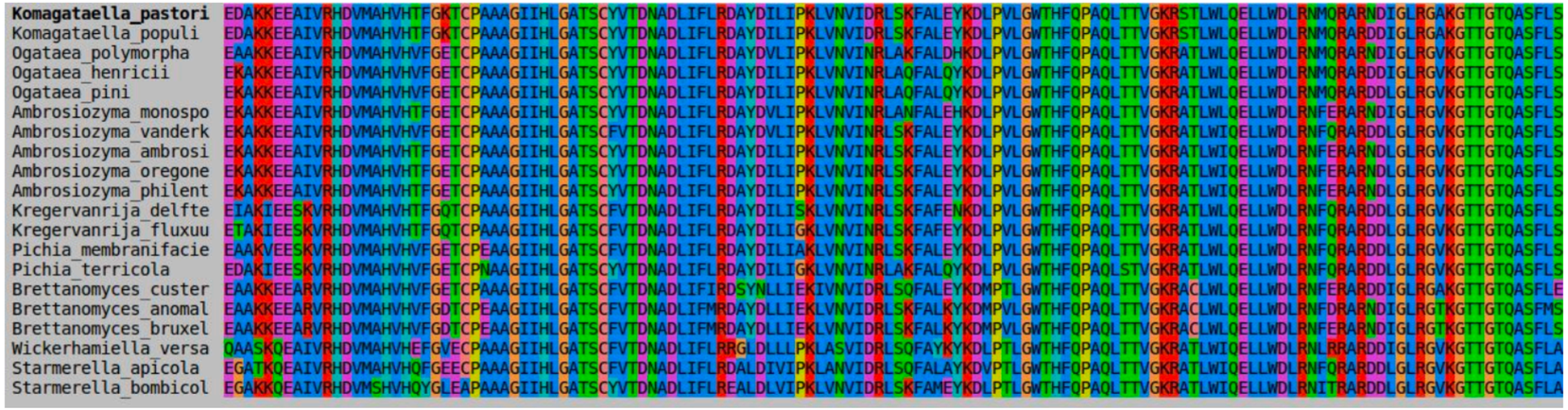
Highly conserved



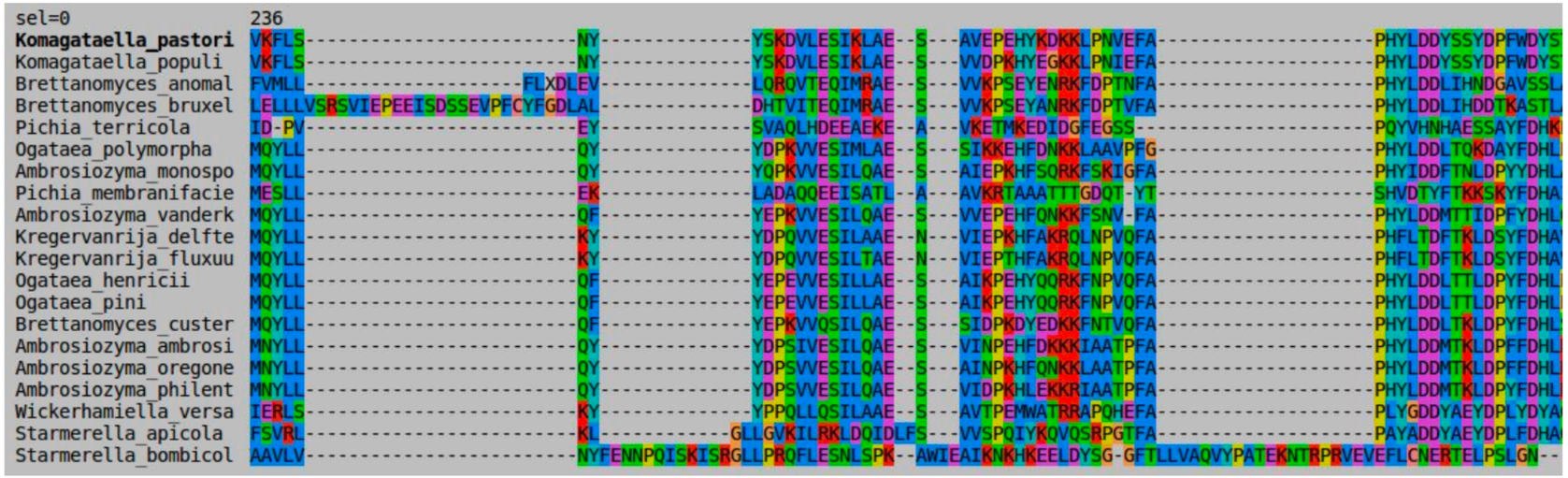

@JLSteenwyk

# Variation in conservation

Highly conserved




Highly variable




@JLSteenwyk

# Sites in alignments may be “unfit”

@JLSteenwyk

Talavera *et al.* (2007), Syst. Biol.; Capella-Gutierrez *et al.* (2009), Bioinformatics; Others

# Sites in alignments may be “unfit”

Highly divergent sites in alignments can be caused by:

1) Erroneously inferred site homology and

@JLSteenwyk

Talavera *et al.* (2007), Syst. Biol.; Capella-Gutierrez *et al.* (2009), Bioinformatics; Others

# 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

@JLSteenwyk

Talavera *et al.* (2007), Syst. Biol.; Capella-Gutierrez *et al.* (2009), Bioinformatics; Others

# 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)

@JLSteenwyk

Talavera *et al.* (2007), Syst. Biol.; Capella-Gutierrez *et al.* (2009), Bioinformatics; Others

# Other sources of error

Biological

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

# 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** ▾

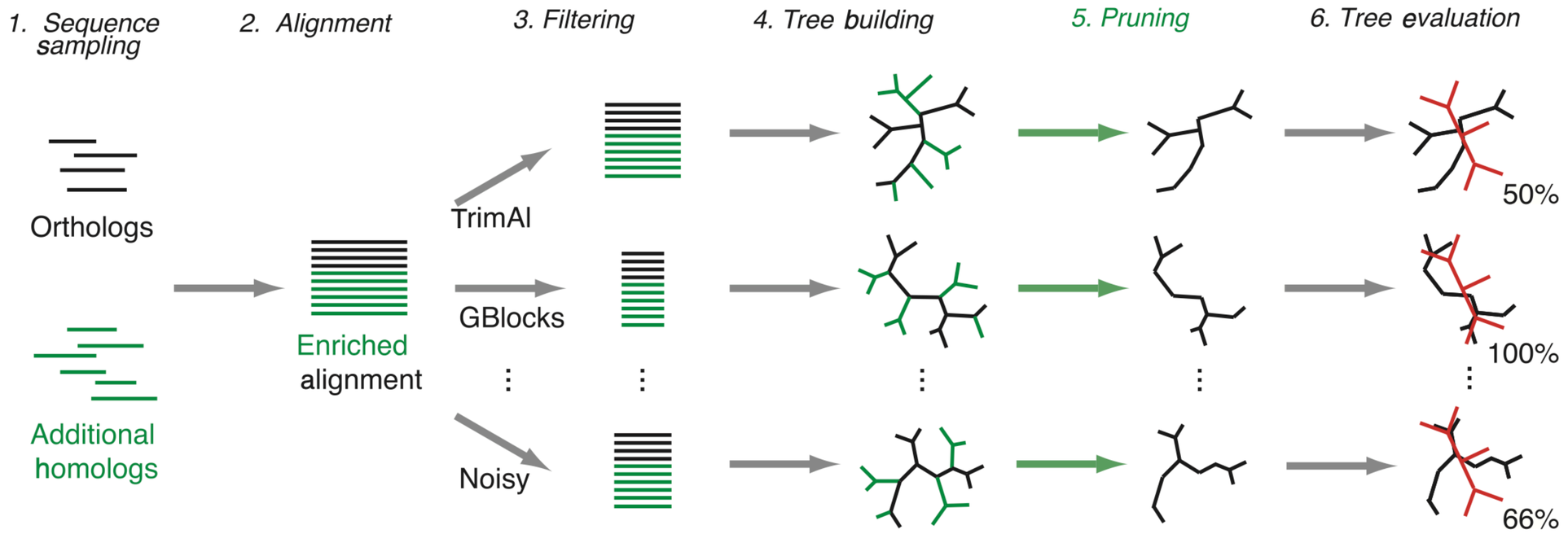
@JLSteenwyk

# Methods examined; most remove divergent sites

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

@JLSteenwyk

# Testing the impact of trimming




@JLSteenwyk

# Sequence sampling using orthologs (& additional homologs)

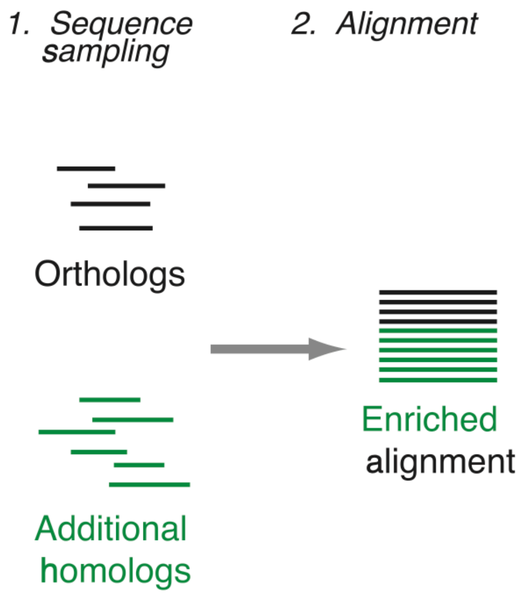
*1. Sequence sampling*

Orthologs

Additional homologs

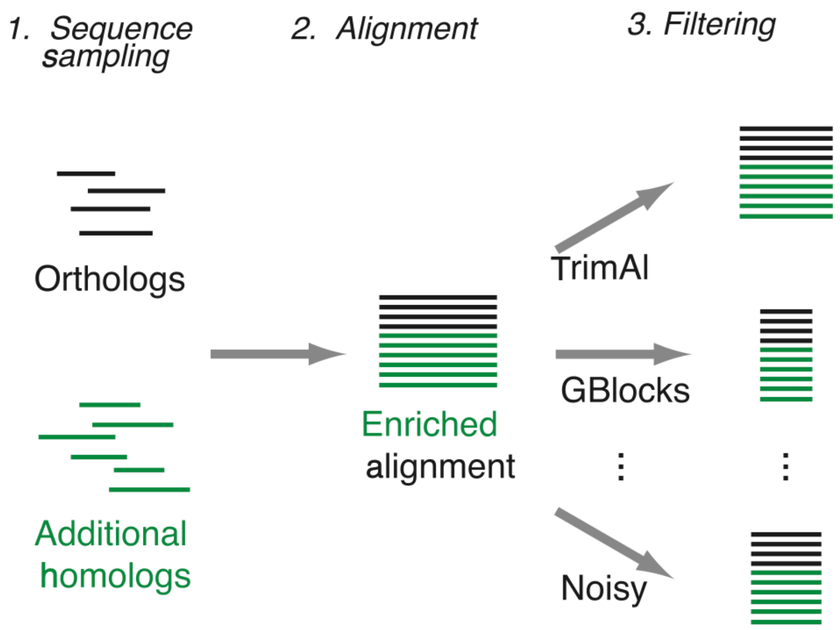
@JLSteenwyk

# Aligning sequences



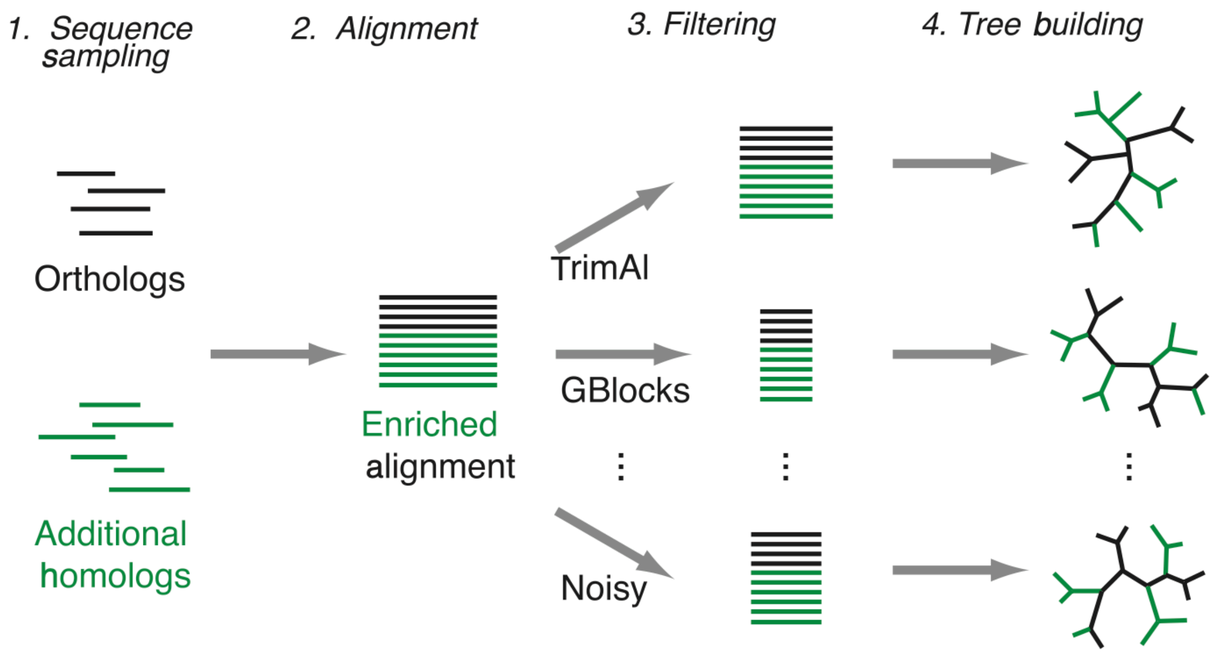

@JLSteenwyk

# Trimming using diverse software



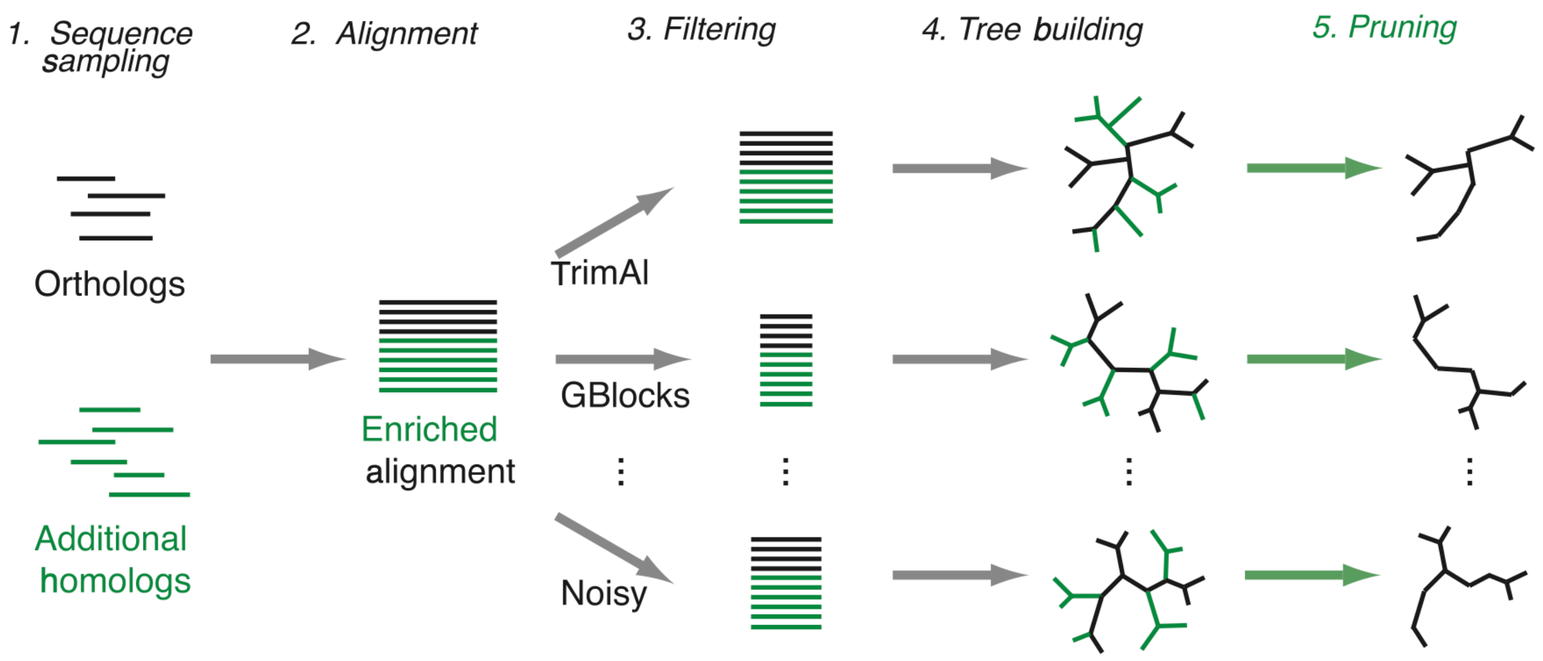

@JLSteenwyk

# Inferring trees from trimmed MSAs



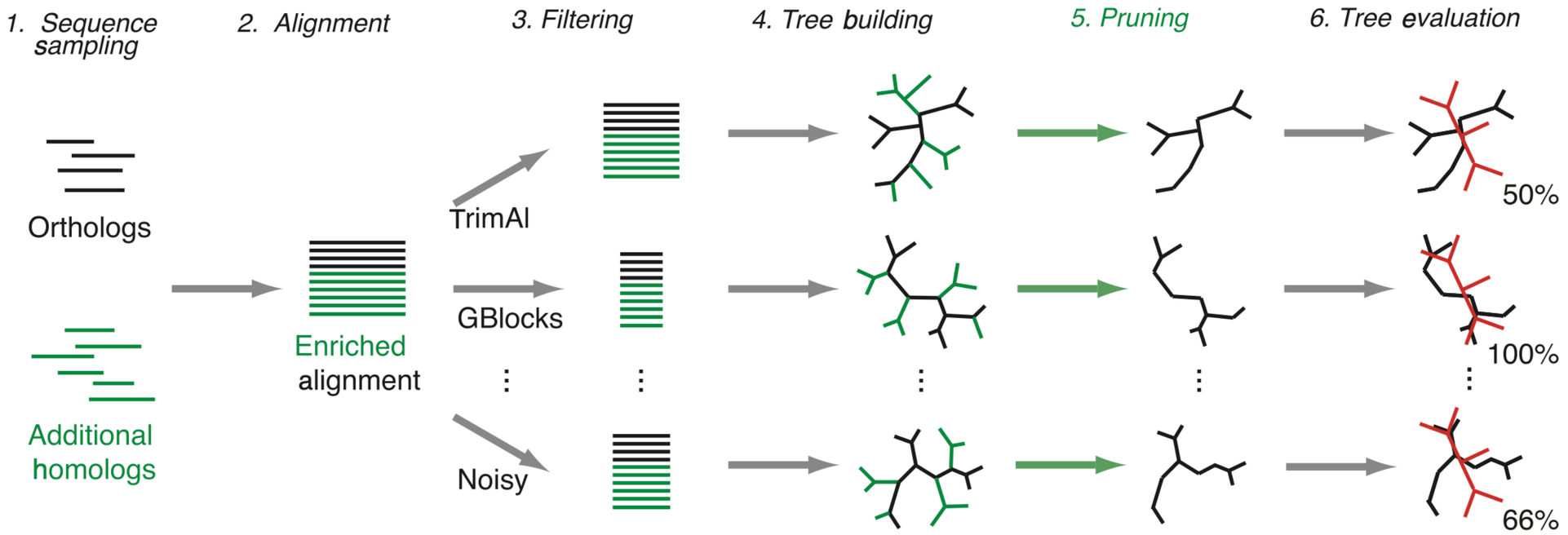

@JLSteenwyk

# Tree pruning to organisms with “incontestable” relationships



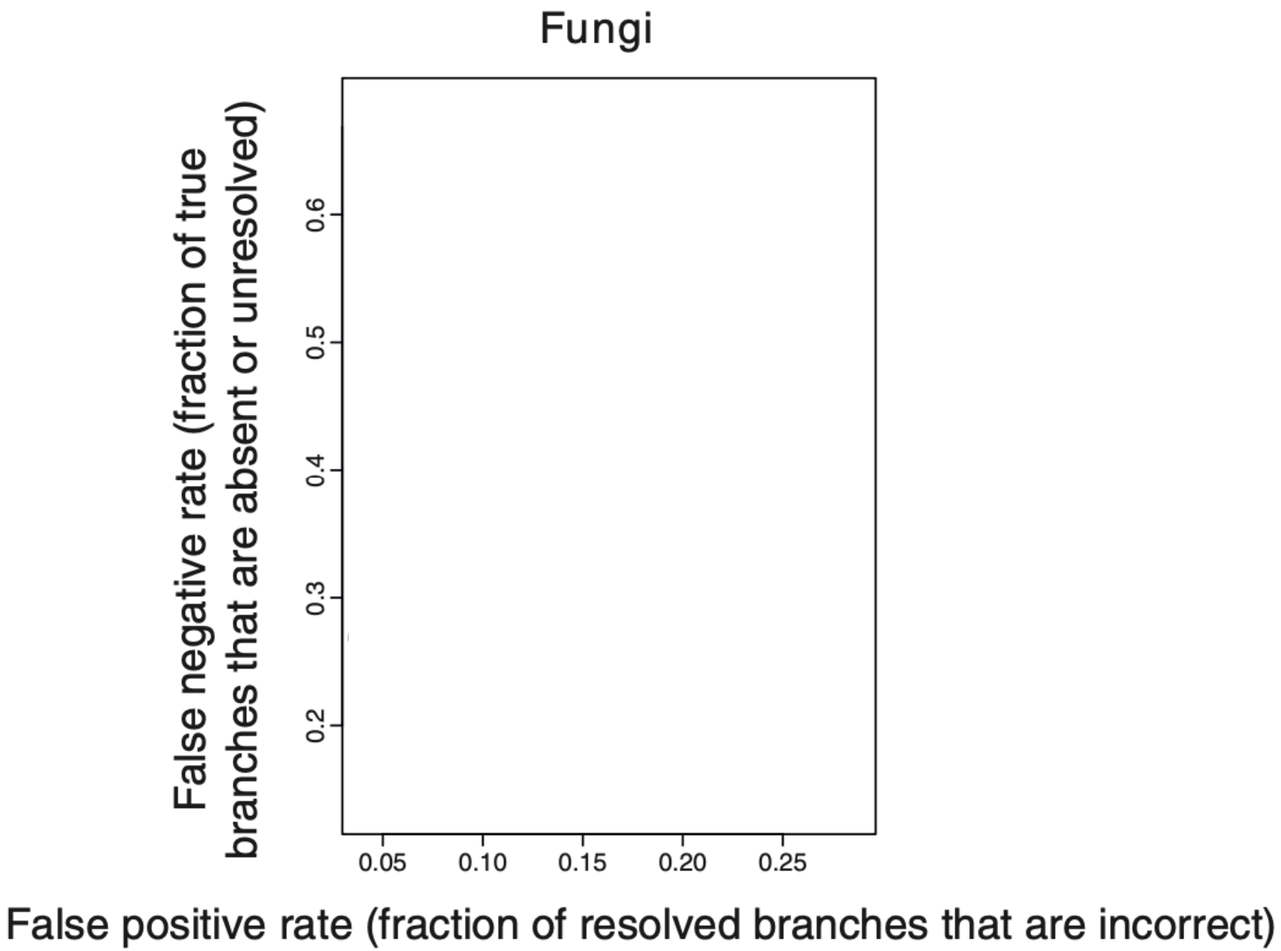

@JLSteenwyk

# Comparing inferred tree to “incontestable” tree



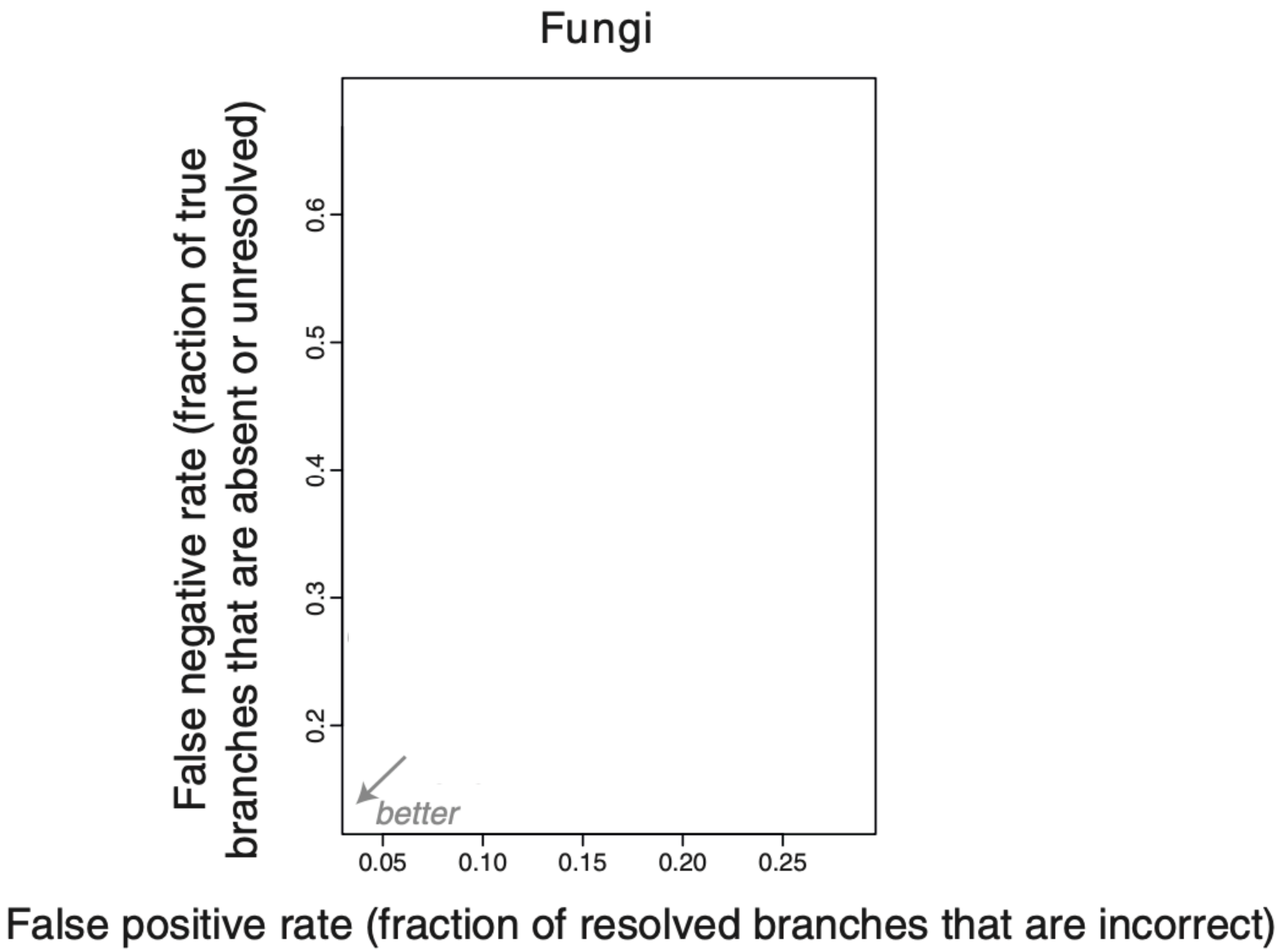

@JLSteenwyk

# Testing the impact of trimming



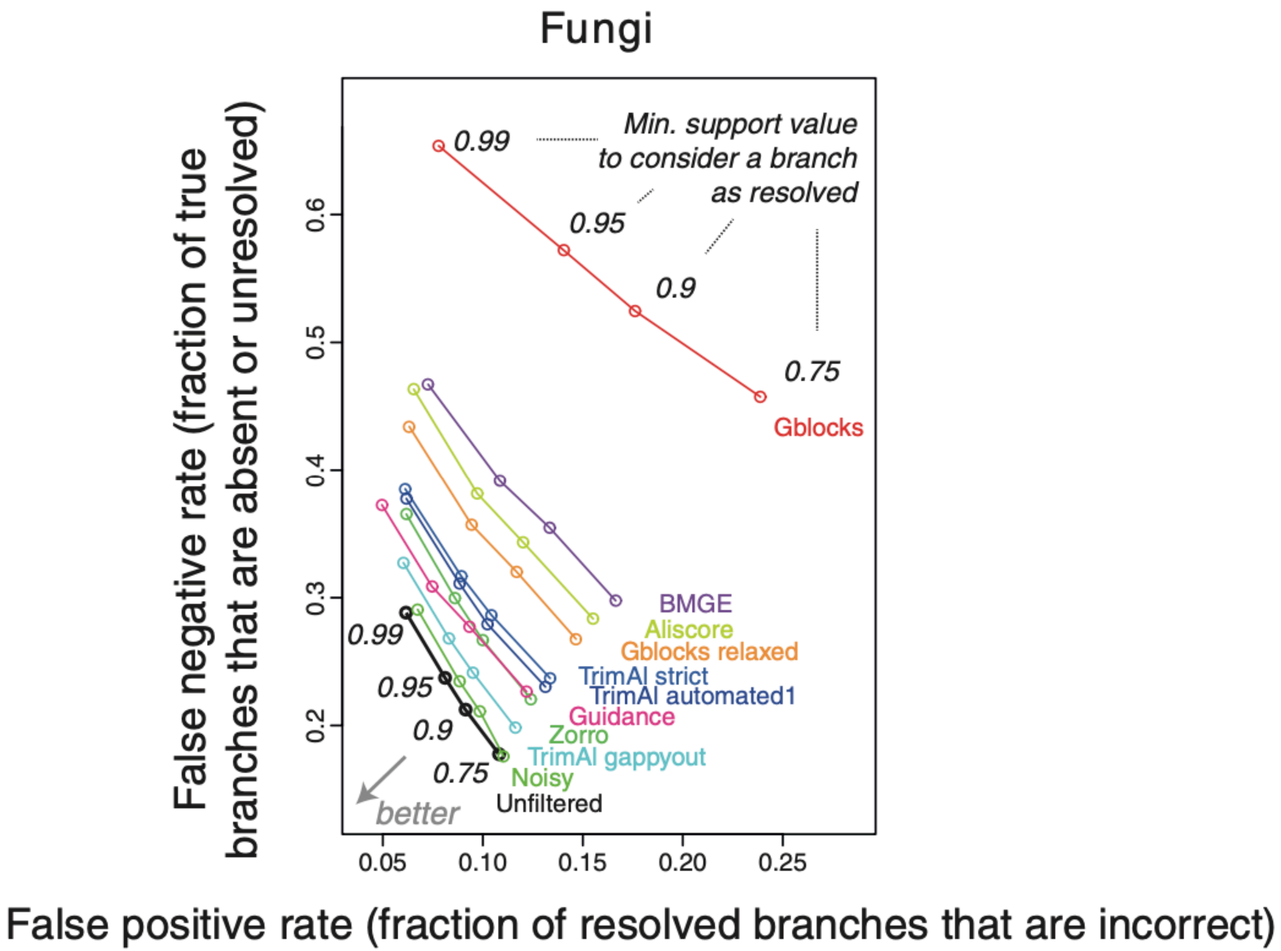

@JLSteenwyk

# Testing the impact of trimming



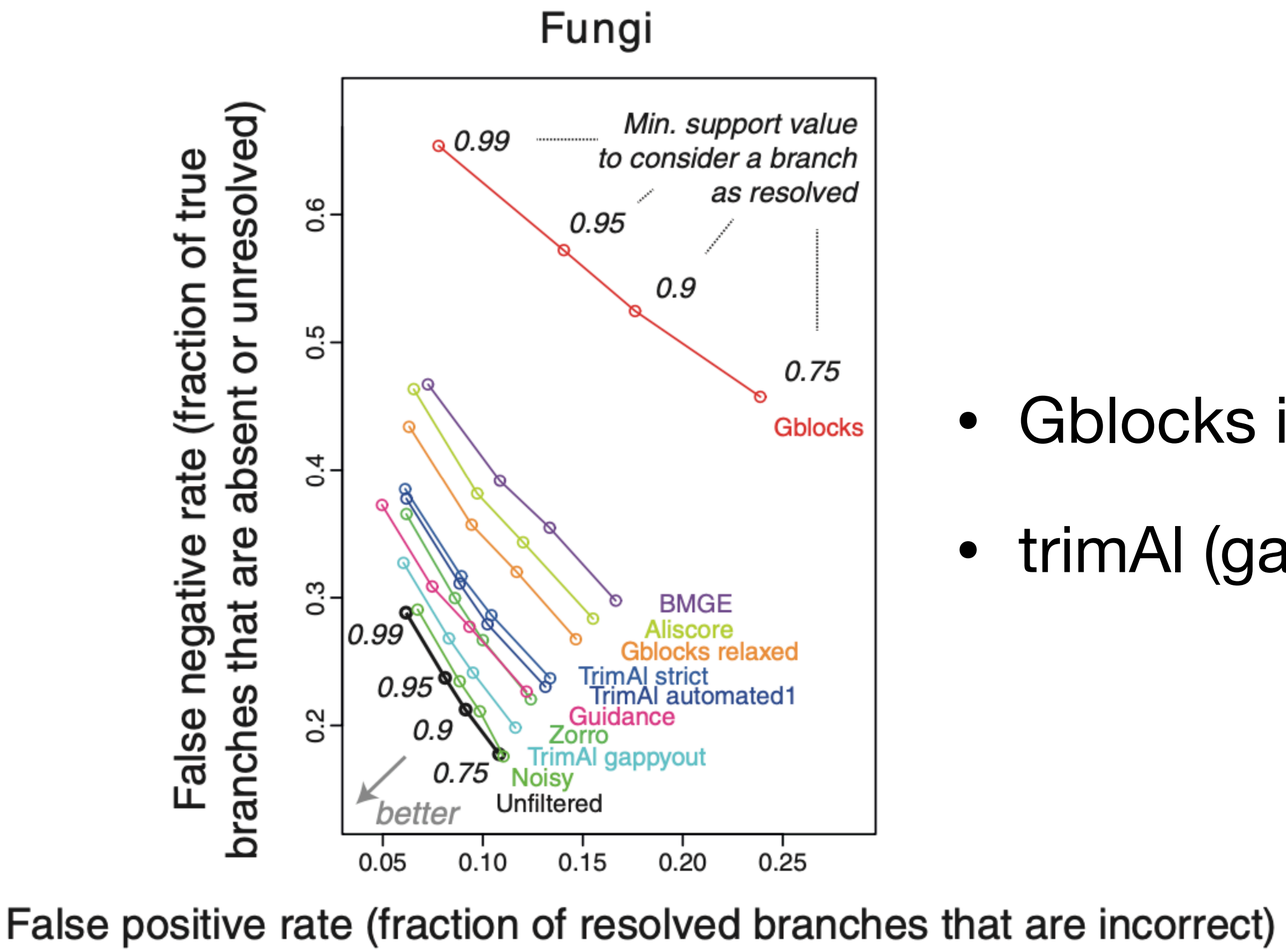

@JLSteenwyk

# Testing the impact of trimming




@JLSteenwyk

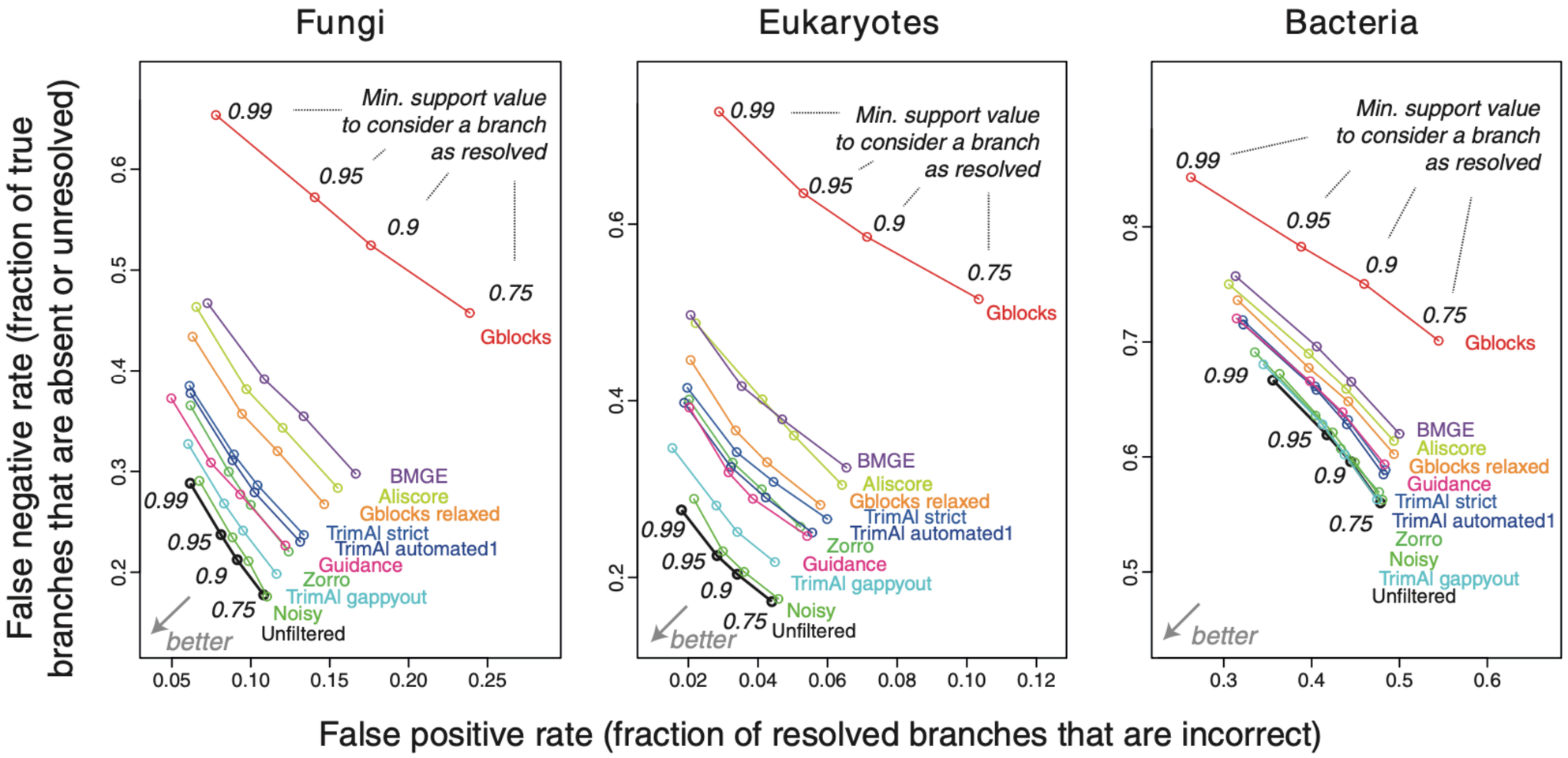
# Testing the impact of trimming




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

@JLSteenwyk

# Testing the impact of trimming




@JLSteenwyk

# Take home message

# 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)

# 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– |

@JLSteenwyk

# 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– |

@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)

# 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)

# ClipKIT has several modes

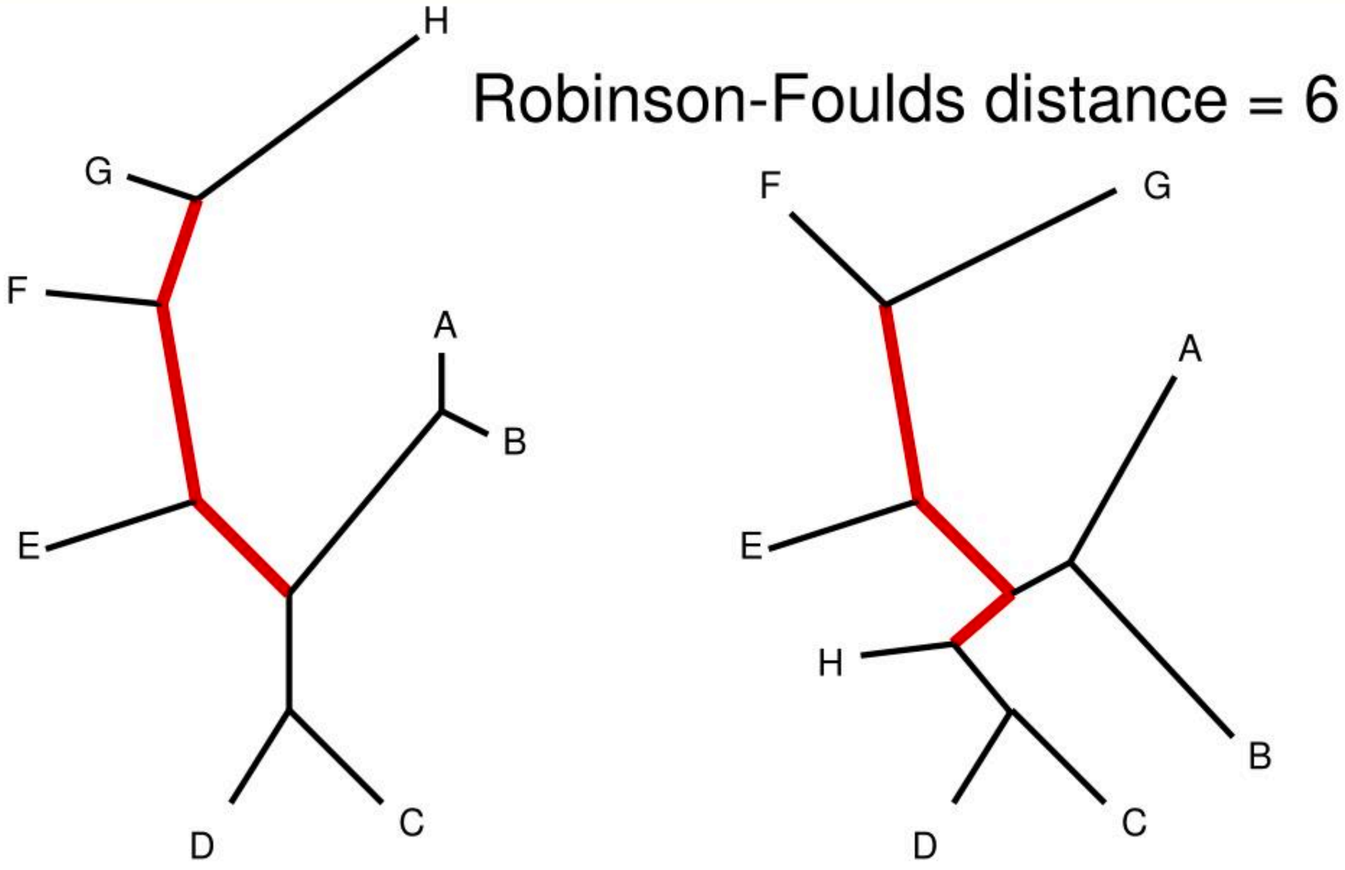
- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
  - kpi (keeps only parsimony informative sites)
  - kpic (keeps parsimony informative sites and constant sites)
  - smart-gap (dynamic gappyness threshold determination)
  - gappy (removes sites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

# ClipKIT 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

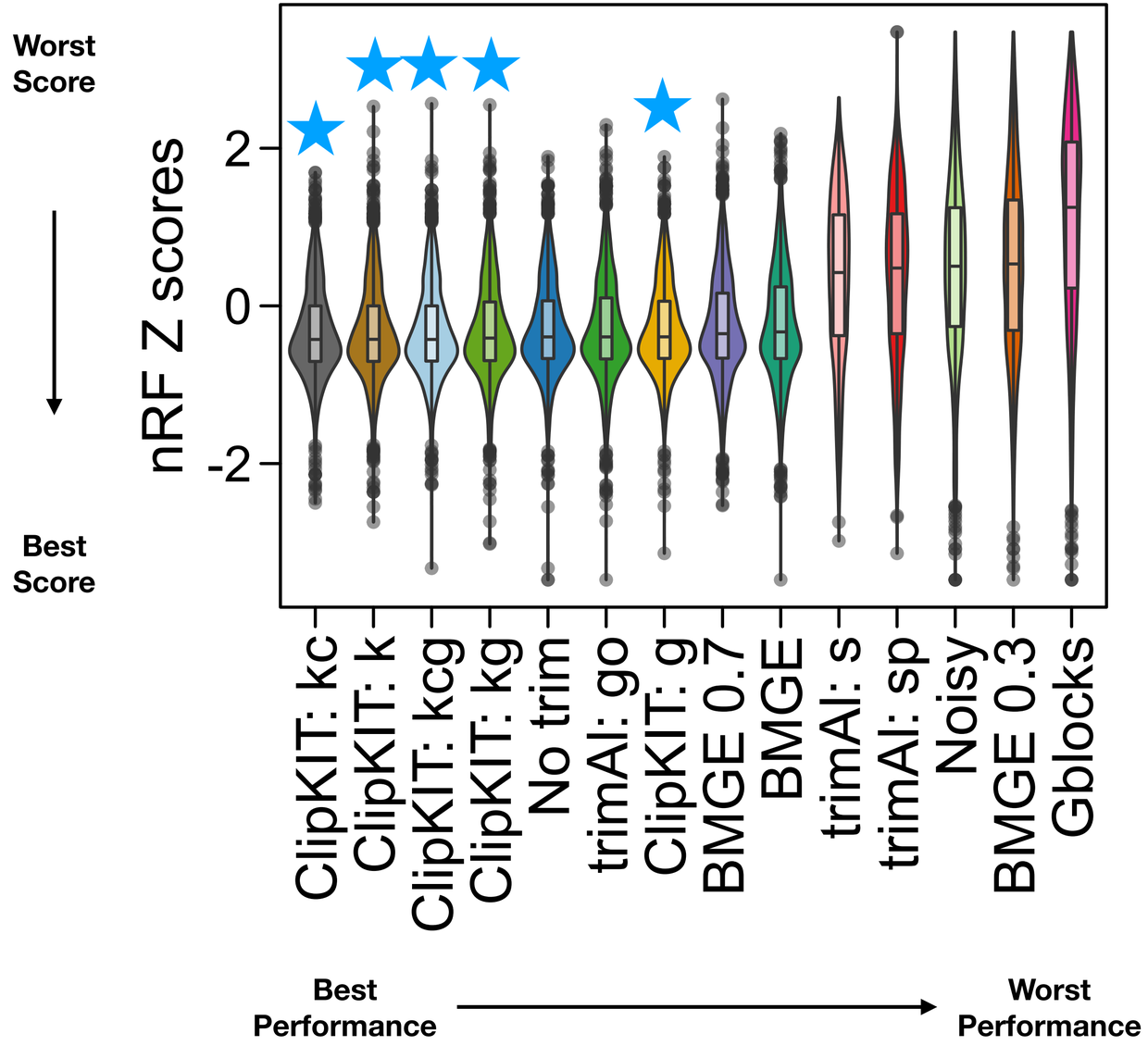
# Measuring accuracy between inferred & expected tree

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



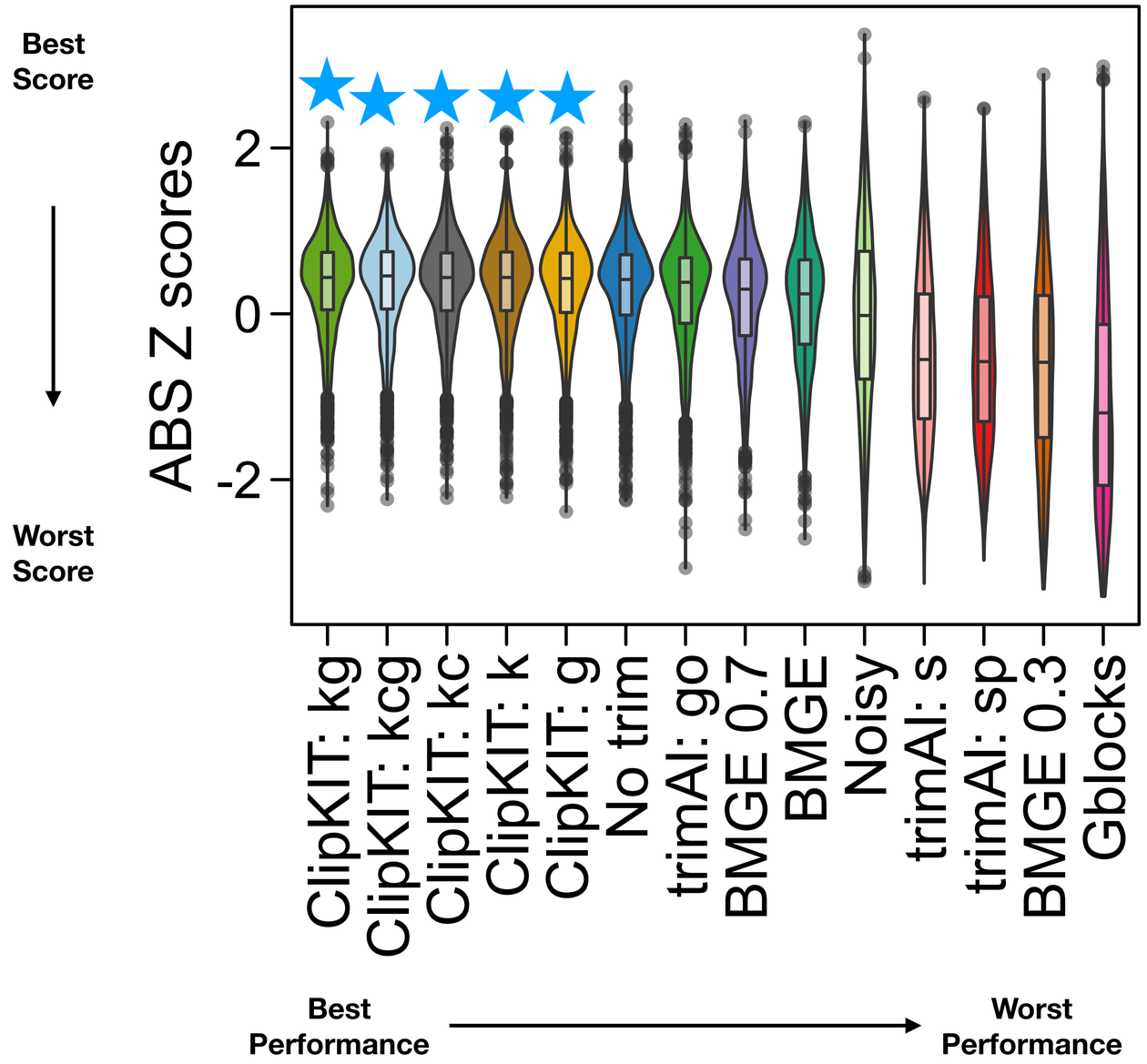

@JLSteenwyk

# Trees inferred using ClipKIT are accurate



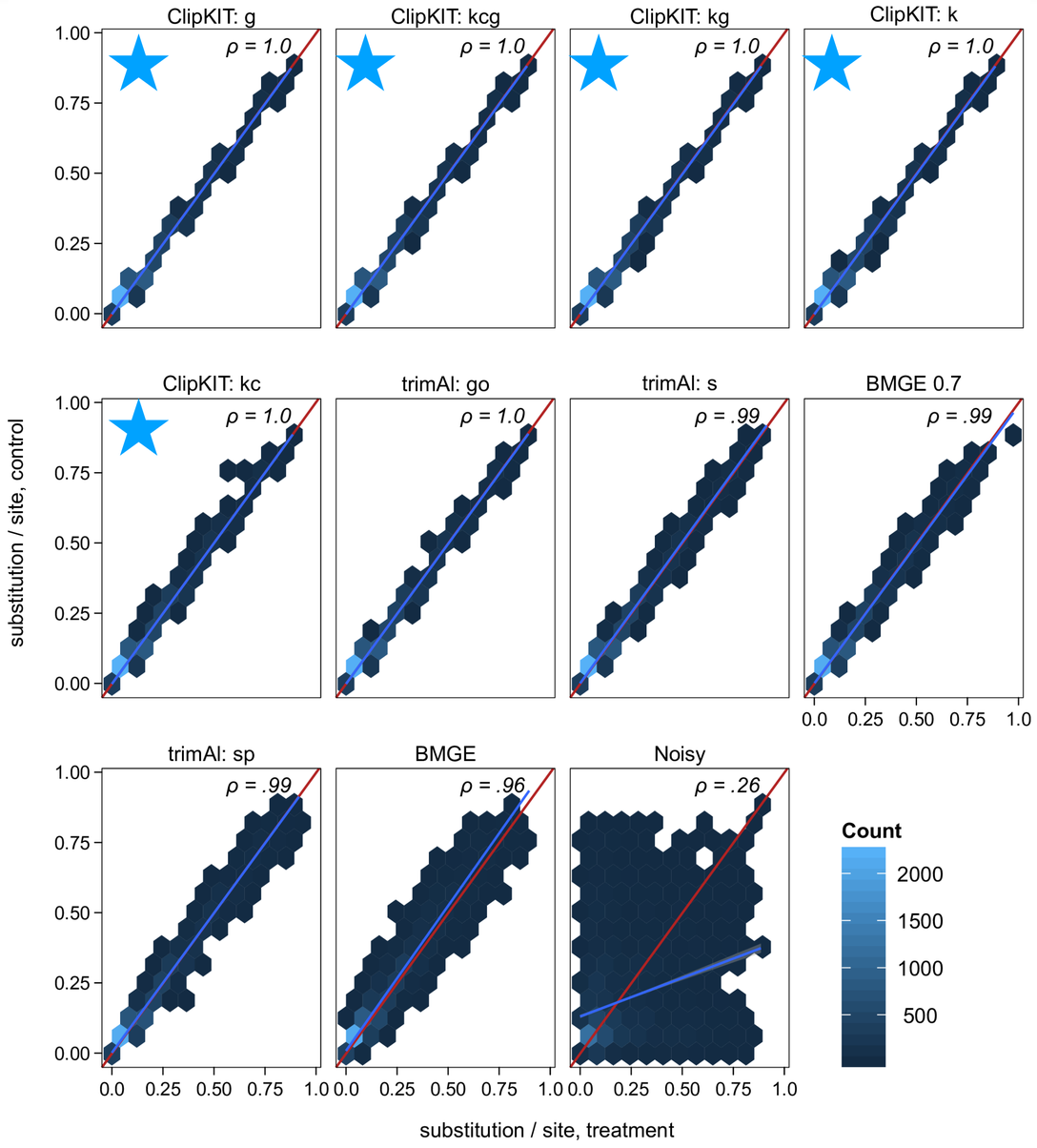

@JLSteenwyk

# Trees inferred using ClipKIT are well supported




@JLSteenwyk

# Branch lengths estimates after trimming are typically accurate

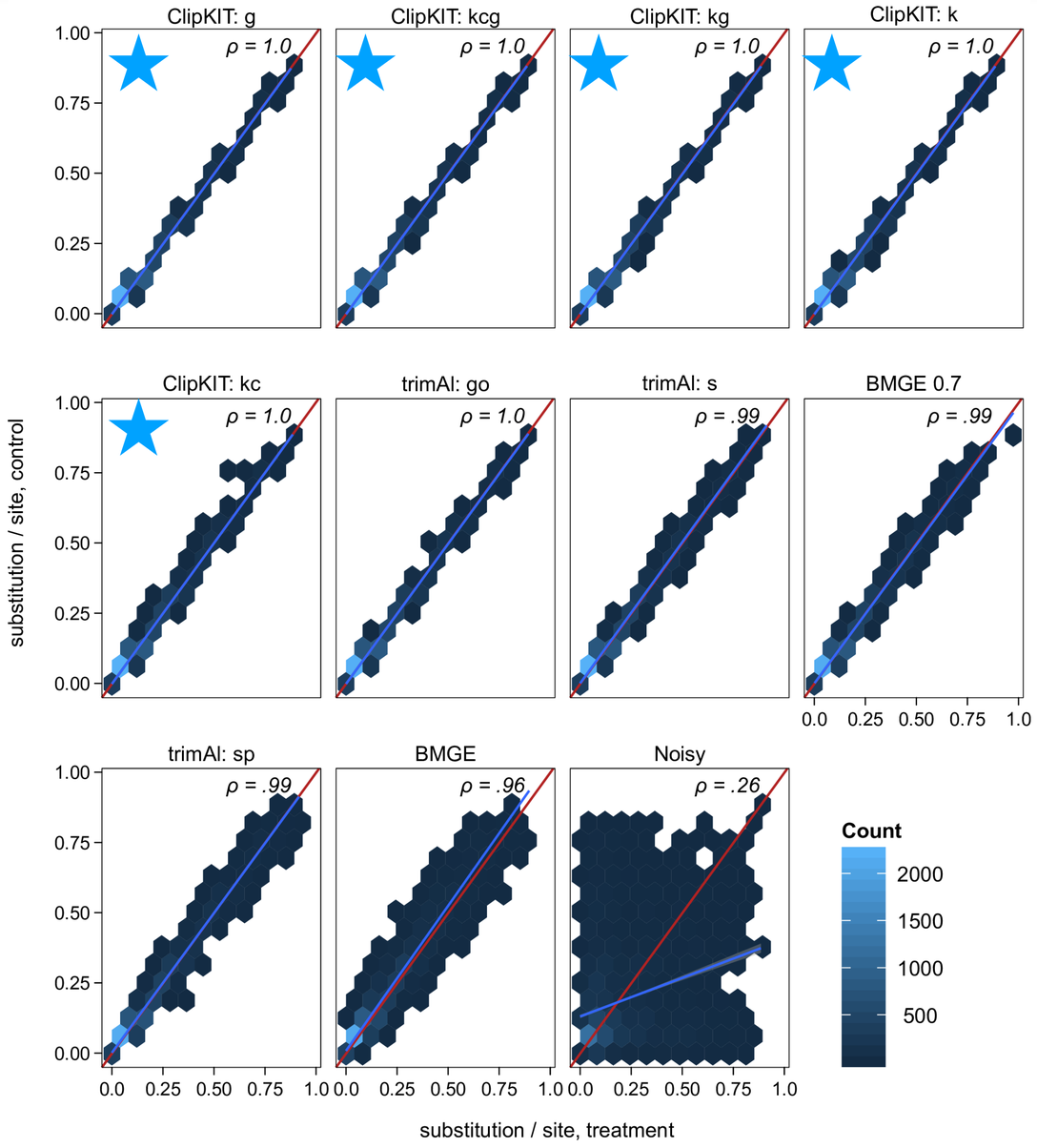



**Best Performance**

**Worst Performance**

@JLSteenwyk

# 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

@JLSteenwyk

https://clipkit.genomelybio.com/

# ClipKIT

Home | Help | CLI Docs | Other Software | Contact

Choose FASTA(s)

Mode

smart-gap (default)

Sequence Type

Auto detect

Trim FASTA(s)

ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference.
Steenwyk et al. PLoS Biology. doi: 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

**Rokas Lab**

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

**Trainees**

Saelin Bjornson
Charu Balamurugan

**Former Trainees**

Olivia Zheng
Megan Phillips

hhmi
Howard Hughes Medical Institute

Life Sciences RESEARCH FOUNDATION


Buida, J


O'Meara, T


Li, Y

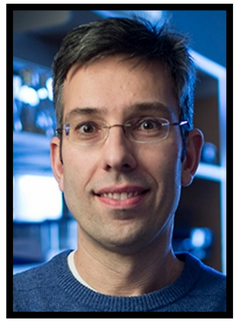

Verbruggen, H


King, N


Coyle, M


Geiser, D


Goldman, G


Rokas, A


Hittinger, C


Berman, J


Shen, X

**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/