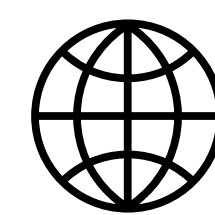


Trimming MSAs

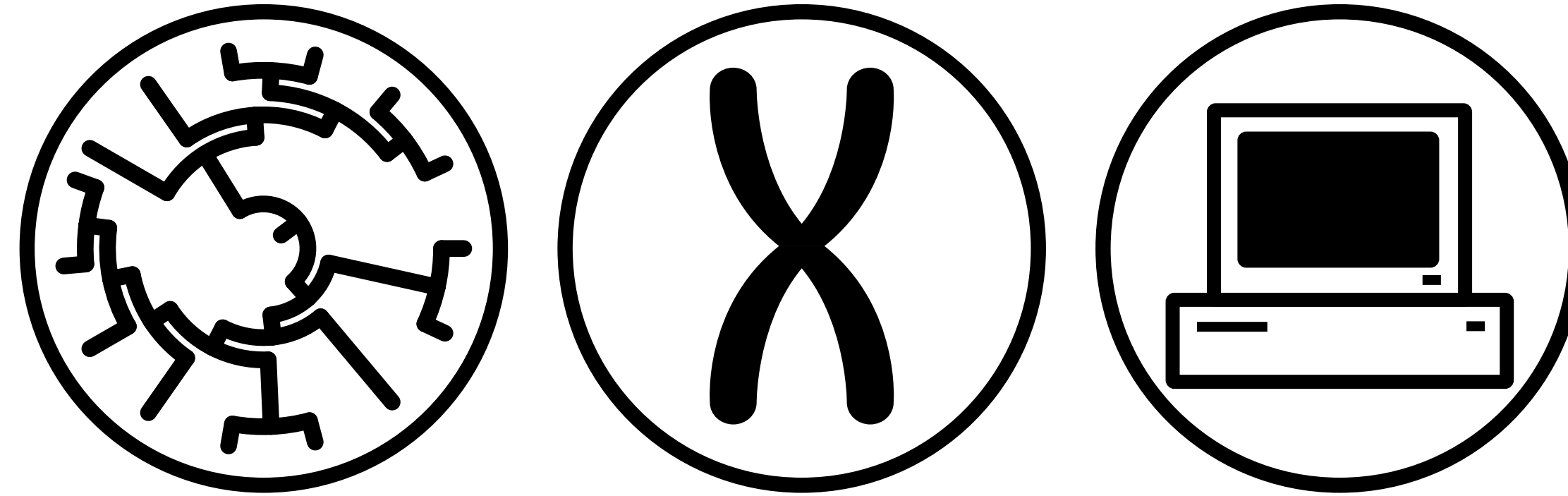


@JLSteenwyk



<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

Type to search

>

+


Overview

Repositories41

Projects

Packages

Stars16



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

64 followers · 21 following

University of California, Berkeley

Pinned

Customize your pins

ClipKITPublic

a multiple sequence alignment-trimming algorithm for accurate phylogenomic inference

Python453

BioKITPublic

a versatile toolkit for processing and analyzing diverse types of sequence data

Python208

orthofisherPublic

a broadly applicable tool for automated gene identification and retrieval

Python231

PhyKITPublic

a UNIX shell toolkit for processing and analyzing multiple sequence alignments and phylogenies

Python466

orthosnapPublic

a tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees

Python181

ggpubfigsPublic

colorblind friendly color palettes and ggplot2 graphic system extensions for publication-quality scientific figures

R315

342 contributions in the last year

Contribution settings

2024

2023

2022

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	De
Mon												
Wed												
Fri												

Raw course materials are available via GitHub

JLSteenwyk / 2024_phylogenomics_workshop

Type to search

<> Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Settings

2024_phylogenomics_workshop

Public

Pin

Unwatch 1

Fork 0

Star 0

main 1 Branch 0 Tags

Go to file

+<> Code

About

JLSteenwyk

formatting changes to trimming rmd

5ecd9ab · 2 days ago

19 Commits

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
partioning_and_concatenation.key	updated with new presentation	3 days ago

About

2024 workshop on phylogenomics, organized by Evomics

[evomics.org/2024-workshop-on-phyl...](#)

Readme

Activity

0 stars

1 watching

0 forks

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

I recommend using links available on the website

EVOLUTION AND GENOMICS

Intensive and immersive training opportunities

WORKSHOPS

LEARNING

PEOPLE

APPLY

INFORMATION

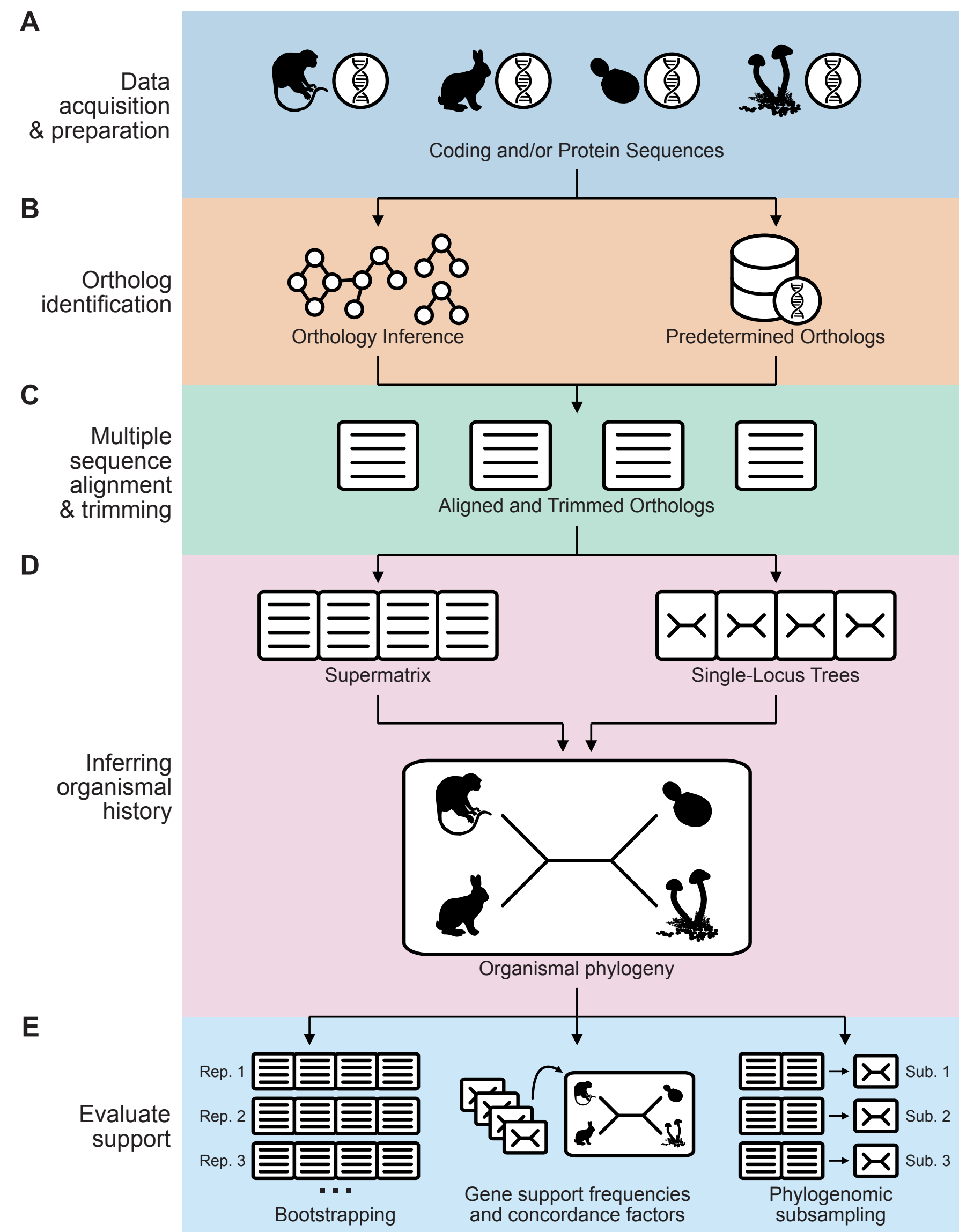
Week 1 : 21-27 January, 2024

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

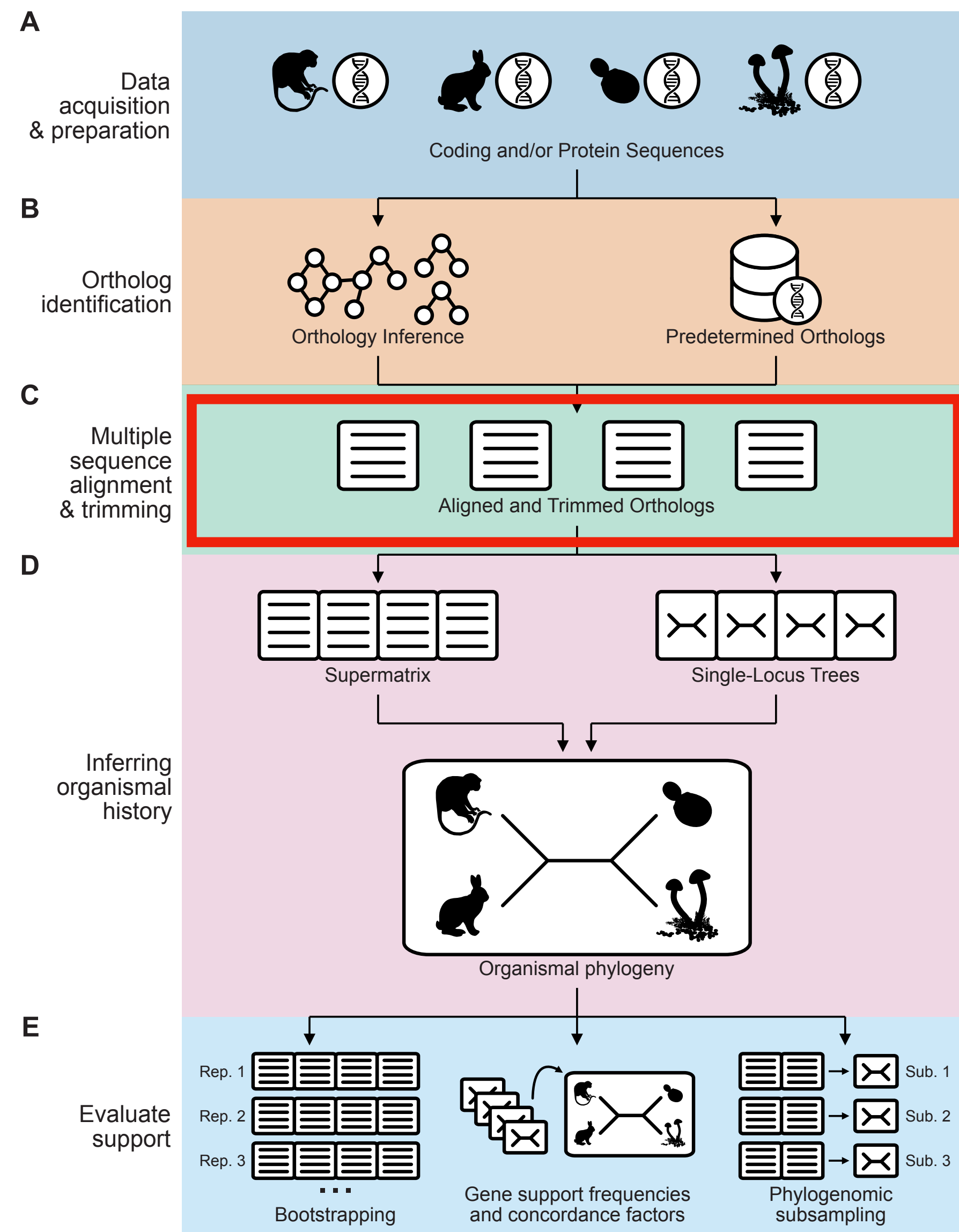


Karin Steffen

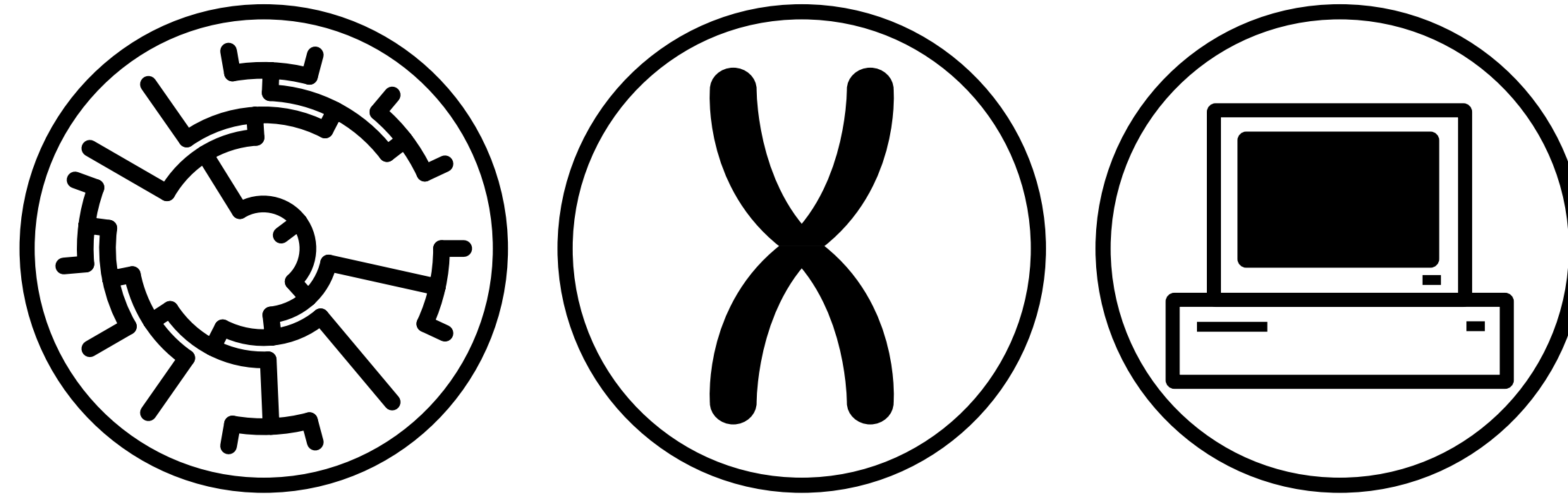
Facilitating phylogenomic workflows and beyond



Facilitating phylogenomic workflows and beyond



Outline



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

Variation in conservation

Highly conserved

Komagataella_pastori	EDAKKEEAIVRHVMAHVHTFGKTC	PAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNVIDRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRSTLWLQELLWDLRNMORARNDIGLRGAKGTTGTQASFLS
Komagataella_populi	EDAKKEEAIVRHVMAHVHTFGKTC	PAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNVIDRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRSTLWLQELLWDLRNMORARDDIGLRGAKGTTGTQASFLS
Ogataea_polymorpha	EAAKKEEAIVRHVMAHVHVFGET	CPAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNINRLAKFALDHKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNMORARNDIGLRGVKGTGTGTQASFLS
Ogataea_henricii	EKAKKEEAIVRHVMAHVHVFGET	CPAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNINRLAQFALQYKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNMORARDDIGLRGVKGTGTGTQASFLS
Ogataea_pini	EKAKKEEAIVRHVMAHVHVFGET	CPAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNINRLAQFALQYKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNMORARDDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_monosporo	EKAKKEEAIVRHVMAHVHTFGET	CPAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNINRLANFALEHKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARNDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_vanderk	EKAKKEEAIVRHVMAHVHVFGET	CPAAAGIIHLGATS	CFVTDNADLIFLRDAYDILI	PKLVNINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_ambrosi	EKAKKEEAIVRHVMAHVHTFGET	CPAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARNDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_oregone	EKAKKEEAIVRHVMAHVHTFGET	CPAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARNDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_philent	EKAKKEEAIVRHVMAHVHTFGET	CPAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	PKLVNINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARNDIGLRGVKGTGTGTQASFLS
Kregervanrija_delfte	EIAKIEESKVRHDMAHVHTFGOT	CPAAAGIIHLGATS	CFVTDNADLIFLRDAYDILI	SKLVNINRLSKFAFENKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Kregervanrija_fluxuu	ETAKIEESKVRHDMAHVHTFGOT	CPAAAGIIHLGATS	CFVTDNADLIFLRDAYDILI	GKLVNINRLSKFAFENKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Pichia_membranifacie	EAAKVEESKVRHDMAHVHVFGET	CPEAAGIIHLGATS	CFVTDNADLIFLRDAYDILI	AKLVNINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Pichia_terricola	EDAKIEESKVRHDMAHVHVFGET	CPAAAGIIHLGATS	CYVTDNADLIFLRDAYDILI	GKLVNINRLAKFALQYKDL	PVLGWTHFQPAQLSTVGKRATLWLQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Brettanomyces_custer	EAAKKEEARVRHDMAHVHVFGET	CPAAAGIIHLGATS	CFVTDNADLIFIRDSYNLLIEKIVNVIDRLS	QFALEYKDMP	TLGWTHFQPAQLTTVGKRACLWLQELLWDLRNFERRARDDIGLRGAKGTTGTQASFLE
Brettanomyces_anomal	EAAKKEEARVRHDMAHVHVFGET	CPEAAGIIHLGATS	CFVTDNADLIFMRDAYDILLIEKLVNVIDRLS	QFALKYKDMP	PVLGWTHFQPAQLTTVGKRACLWLQELLWDLRNFERRARNDIGLRGTGTTGTQASFMS
Brettanomyces_bruzel	EAAKKEEARVRHDMAHVHVFGET	CPEAAGIIHLGATS	CFVTDNADLIFMRDAYDILLIEKLVNVIDRLS	QFALKYKDMP	PVLGWTHFQPAQLTTVGKRACLWLQELLWDLRNFERRARNDIGLRGTGTTGTQASFLS
Wickerhamiella_versa	QAASKQEAIVRHDMAHVHEFGVECP	PAAAGIIHLGATS	CFVTDNADLIFLRGLDLLPKLASVIDRLS	QFAYKYKDLPTL	GWTHFQPAQLTTVGKRATLWIQELLWDLRNLRRARDDIGLRGVKGTGTGTQASFLA
Starmerella_apicola	EGATKQEAIVRHDMAHVHVFGECP	PAAAGIIHLGATS	CFVTDNADLIFLRDALDIVIPKLANVIDRLS	QFALAYKDVPTL	GWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLA
Starmerella_bombicol	EGAKKQEAIVRHDMASHVHVOYGLE	PAAAGIIHLGATS	CYVTDNADLIFLRDALDIVIPKLVNVIDRLS	QFAMEYKDLPTL	GWTHFQPAQLTTVGKRATLWIQELLWDLRNLRRARDDIGLRGVKGTGTGTQASFLA

Variation in conservation

Highly conserved

Komagataella_pastori	EDAKKEEAIVRHDMAHVHTFGKTC	PAAAGIIHLGATSCYVTDNADLIFLRDAYDILIPKLVNVIDRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRSTLWLQELLWDLRNMORARNDIGLRGAKGTTGTQASFLS	
Komagataella_populi	EDAKKEEAIVRHDMAHVHTFGKTC	PAAAGIIHLGATSCYVTDNADLIFLRDAYDILIPKLVNVIDRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRSTLWLQELLWDLRNMORARDDIGLRGAKGTTGTQASFLS	
Ogataea_polymorpha	EAAKKEEAIVRHDMAHVHVFGGET	CPAAAGIIHLGATSCYVTDNADLIFLRDAYDVLI	PKLVNVINRLAKFALDHKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNMORARNDIGLRGVKGTGTGTQASFLS
Ogataea_henricii	EKAKKEEAIVRHDMAHVHVFGGET	CPAAAGIIHLGATSCYVTDNADLIFLRDAYDILI	PKLVNVINRLAQFALQYKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNMORARDDIGLRGVKGTGTGTQASFLS
Ogataea_pini	EKAKKEEAIVRHDMAHVHVFGGET	CPAAAGIIHLGATSCYVTDNADLIFLRDAYDILI	PKLVNVINRLAQFALQYKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNMORARDDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_monosporo	EKAKKEEAIVRHDMAHVHTFGGET	CPAAAGIIHLGATSCYVTDNADLIFLRDAYDVLI	PKLVNVINRLANFALEHKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARNDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_vanderk	EKAKKEEAIVRHDMAHVHVFGGET	CPAAAGIIHLGATSCFVTDNADLIFLRDAYDVLI	PKLVNVINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_ambrosii	EKAKKEEAIVRHDMAHVHTFGGET	CPAAAGIIHLGATSCYVTDNADLIFLRDAYDILI	PKLVNVINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARNDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_oregonense	EKAKKEEAIVRHDMAHVHTFGGET	CPAAAGIIHLGATSCYVTDNADLIFLRDAYDILI	PKLVNVINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARNDIGLRGVKGTGTGTQASFLS
Ambrosiozyma_philrentii	EKAKKEEAIVRHDMAHVHTFGGET	CPAAAGIIHLGATSCYVTDNADLIFLRDAYDILI	PKLVNVINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARNDIGLRGVKGTGTGTQASFLS
Kregervanrija_delftensis	EIAKIEESKVRHDMAHVHTFGGTC	PAAAGIIHLGATSCFVTDNADLIFLRDAYDILI	SKLVNVINRLSKFAFENKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Kregervanrija_fluxuosa	ETAKIEESKVRHDMAHVHTFGGTC	PAAAGIIHLGATSCFVTDNADLIFLRDAYDILI	GKLVNVINRLSKFAFENKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Pichia_membranifaciens	EAAKVEESKVRHDMAHVHVFGGET	CPEAAGIIHLGATSCFVTDNADLIFLRDAYDILI	IAKLVNVINRLSKFALEYKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Pichia_terricola	EDAKIEESKVRHDMAHVHVFGGET	CPAAAGIIHLGATSCYVTDNADLIFLRDAYDILI	GKLVNVINRLAKFALQYKDL	PVLGWTHFQPAQLTTVGKRATLWLQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLS
Brettanomyces_custersii	EAAKKEEAIVRHDMAHVHVFGGET	CPAAAGIIHLGATSCFVTDNADLIFIRDSYNLLIEKIVNVIDRLS	QFALEYKDMPTL	GWTHFQPAQLTTVGKRACLWLQELLWDLRNFERRARDDIGLRGAKGTTGTQASFLS
Brettanomyces_anomalus	EAAKKEEAIVRHDMAHVHVFGGDT	CPEAAGIIHLGATSCFVTDNADLIFMRDAYDILLIEKLVNVIDRLS	SKFALKYKDMPTL	GWTHFQPAQLTTVGKRACLWLQELLWDLRNFERRARNDIGLRGTGKTGTGTQASFLS
Brettanomyces_bruxellensis	EAAKKEEAIVRHDMAHVHVFGGDT	CPEAAGIIHLGATSCFVTDNADLIFMRDAYDILLIEKLVNVIDRLS	SKFALKYKDMPTL	GWTHFQPAQLTTVGKRACLWLQELLWDLRNFERRARNDIGLRGTGKTGTGTQASFLS
Wickerhamiella_versatilis	QAASKQEAIVRHDMAHVHVFGGEEC	PAAAGIIHLGATSCFVTDNADLIFLRDALDIVIPKLANVIDRLS	QFALAYKDVPTL	GWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLA
Starmerella_apicola	EGATKQEAIVRHDMAHVHVFGGEEC	PAAAGIIHLGATSCFVTDNADLIFLRDALDIVIPKLANVIDRLS	QFALAYKDVPTL	GWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLA
Starmerella_bombicicola	EGAKKQEAIVRHDMASHVHQYGLEA	PAAAGIIHLGATSCYVTDNADLIFLRDALDIVIPKLVNVIDRLS	SKFAMEYKDLPTL	GWTHFQPAQLTTVGKRATLWIQELLWDLRNFERRARDDIGLRGVKGTGTGTQASFLA

Highly variable

sel=0	236																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																														
-------	-----	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

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

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](#)

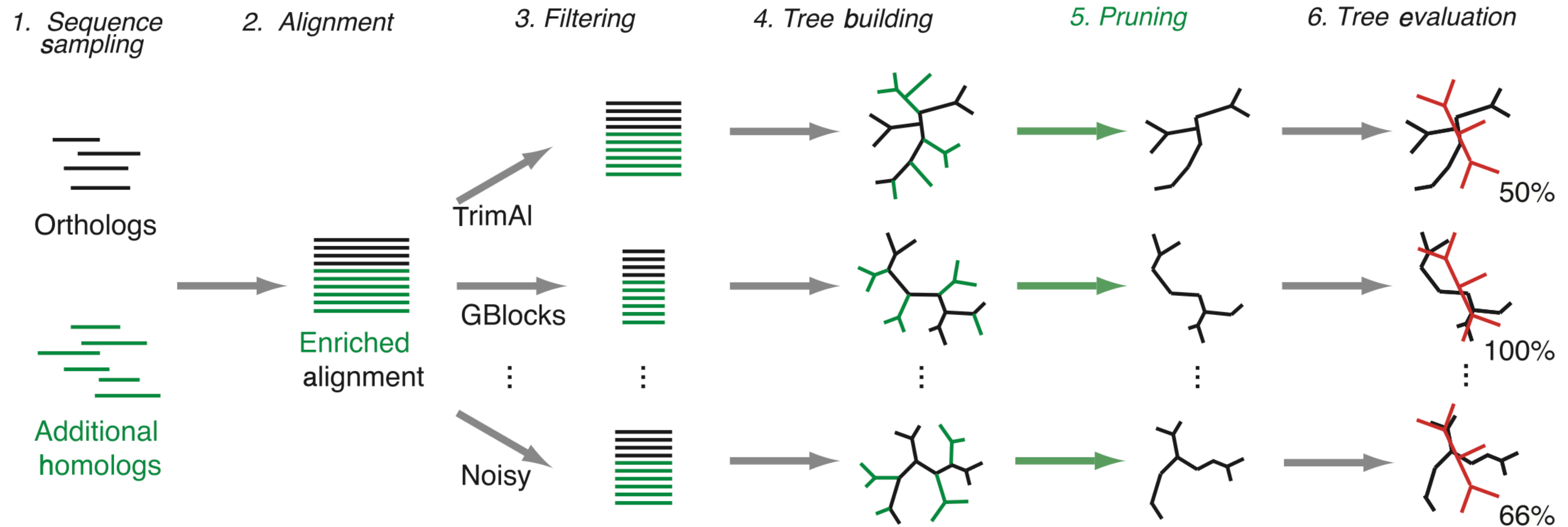
Systematic Biology, Volume 64, Issue 5, September 2015, Pages 778–791,
<https://doi.org/10.1093/sysbio/syv033>

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

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)

Testing the impact of trimming



Sequence sampling using orthologs (& additional homologs)

1. *Sequence sampling*



Orthologs



Additional
homologs

Aligning sequences

1. *Sequence sampling*

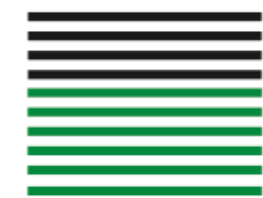
2. *Alignment*



Orthologs

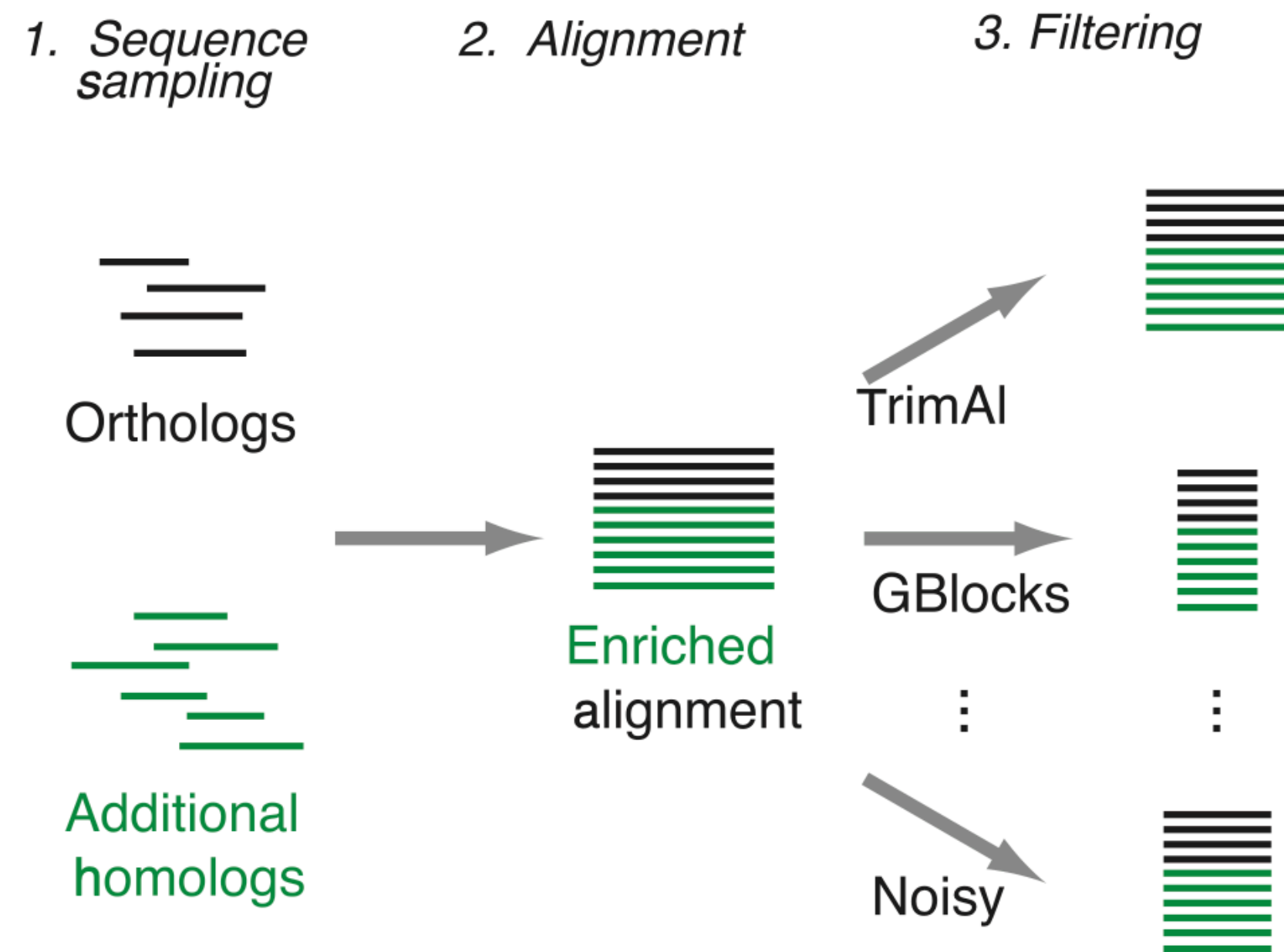


Additional
homologs

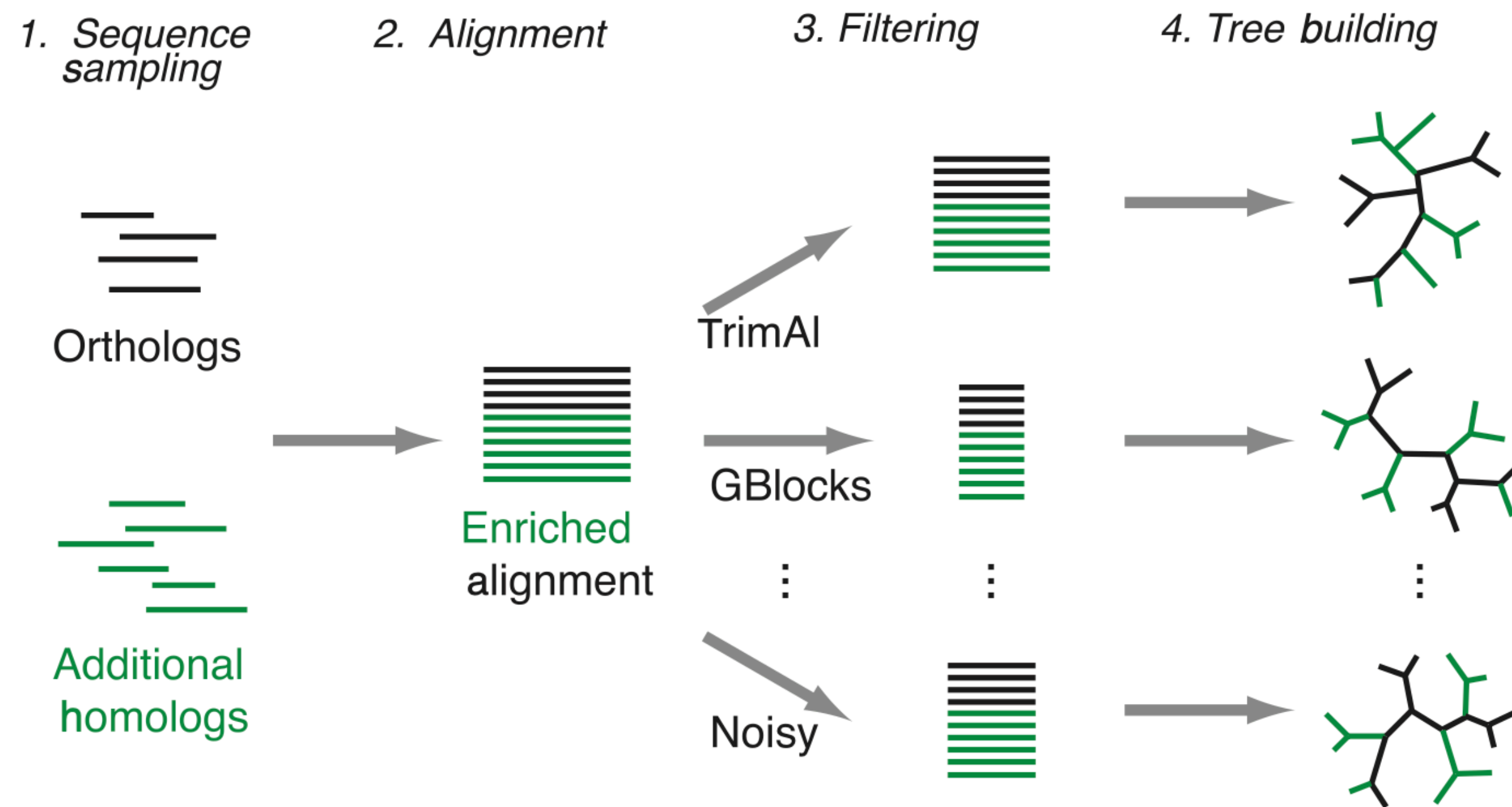


Enriched
alignment

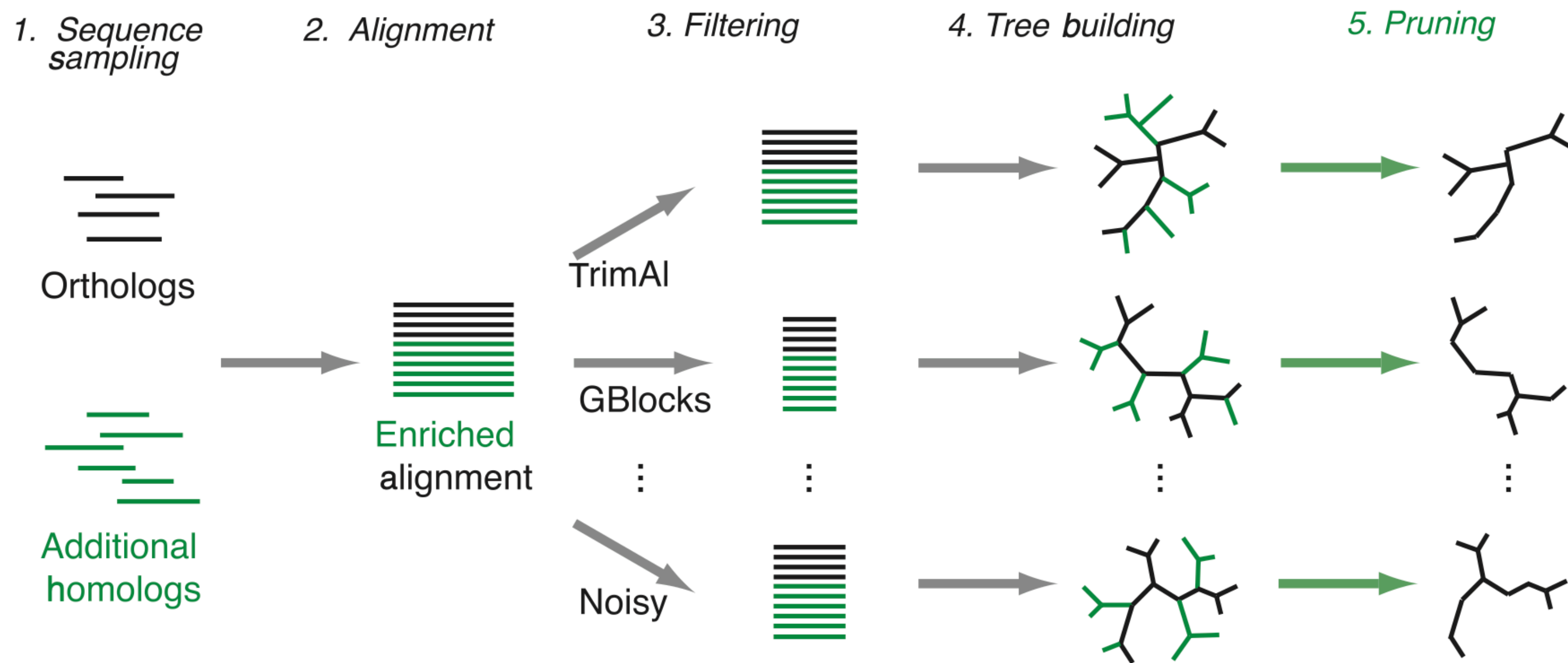
Trimming using diverse software



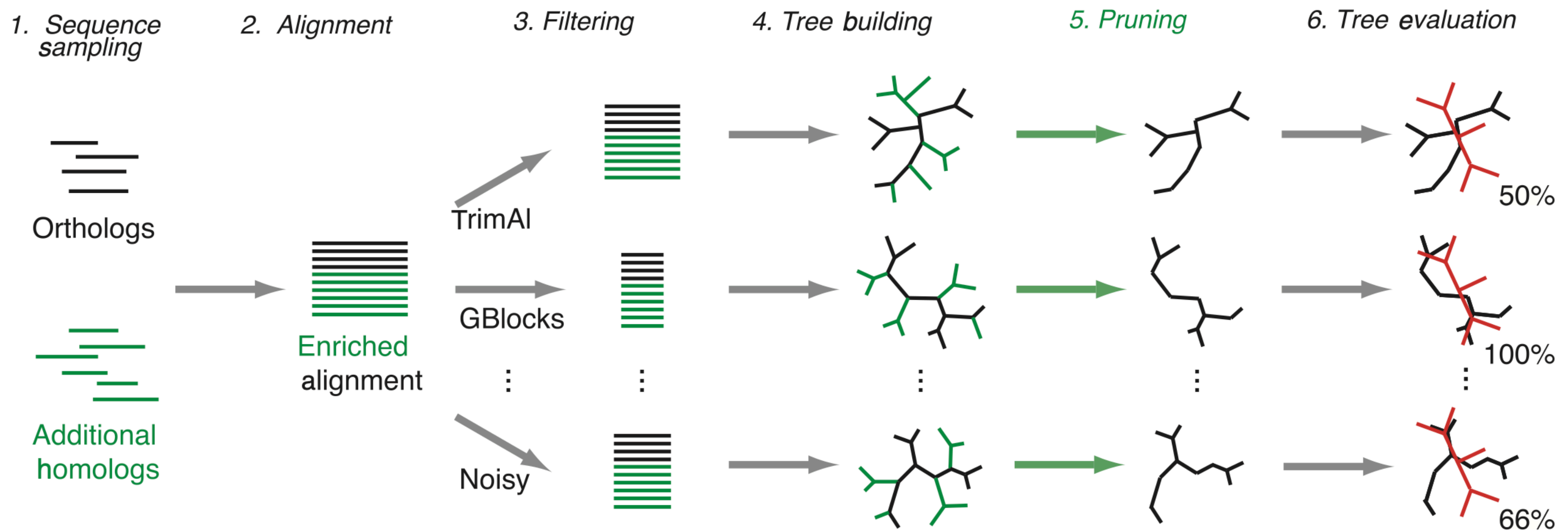
Inferring trees from trimmed MSAs



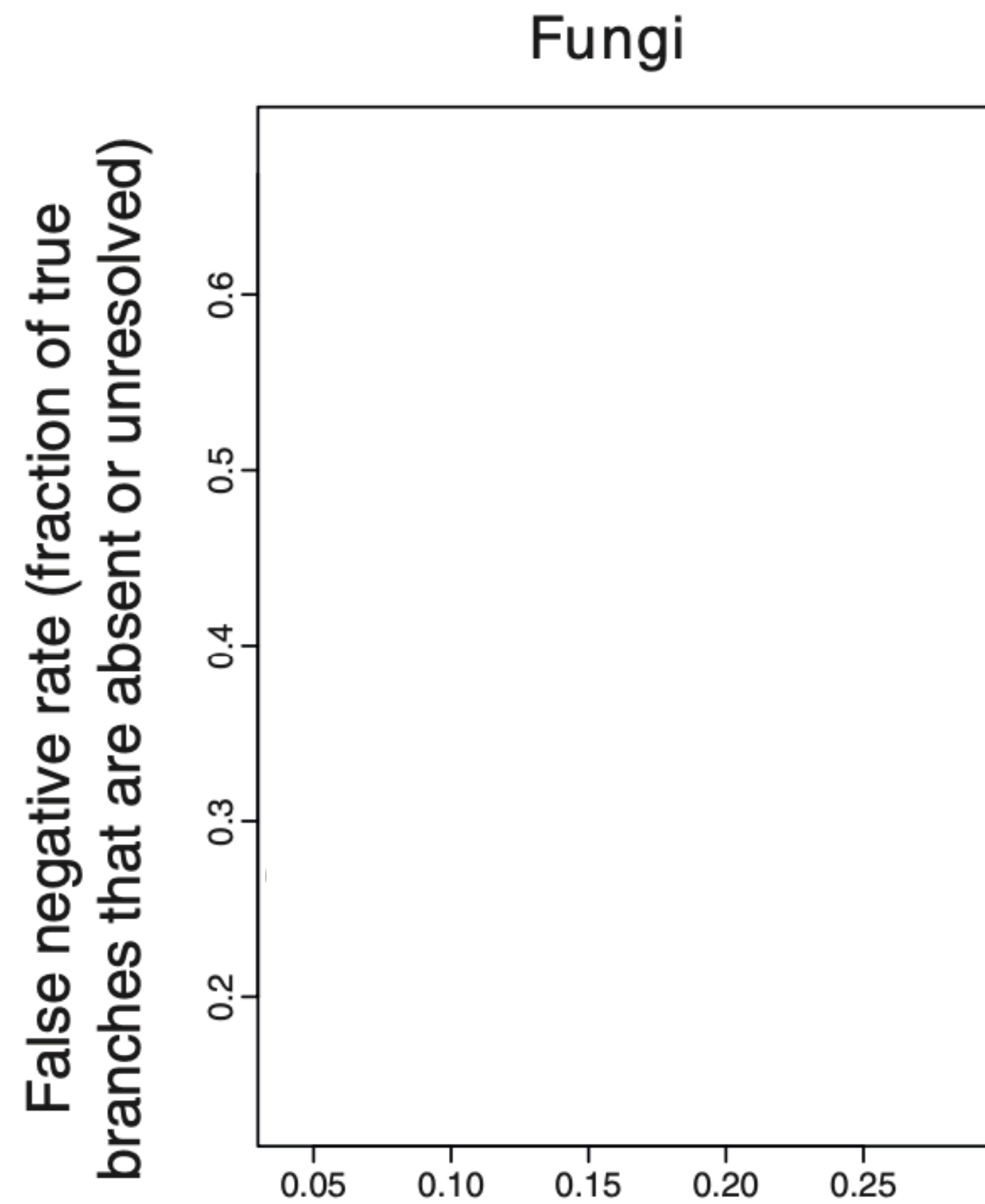
Tree pruning to organisms with “incontestable” relationships



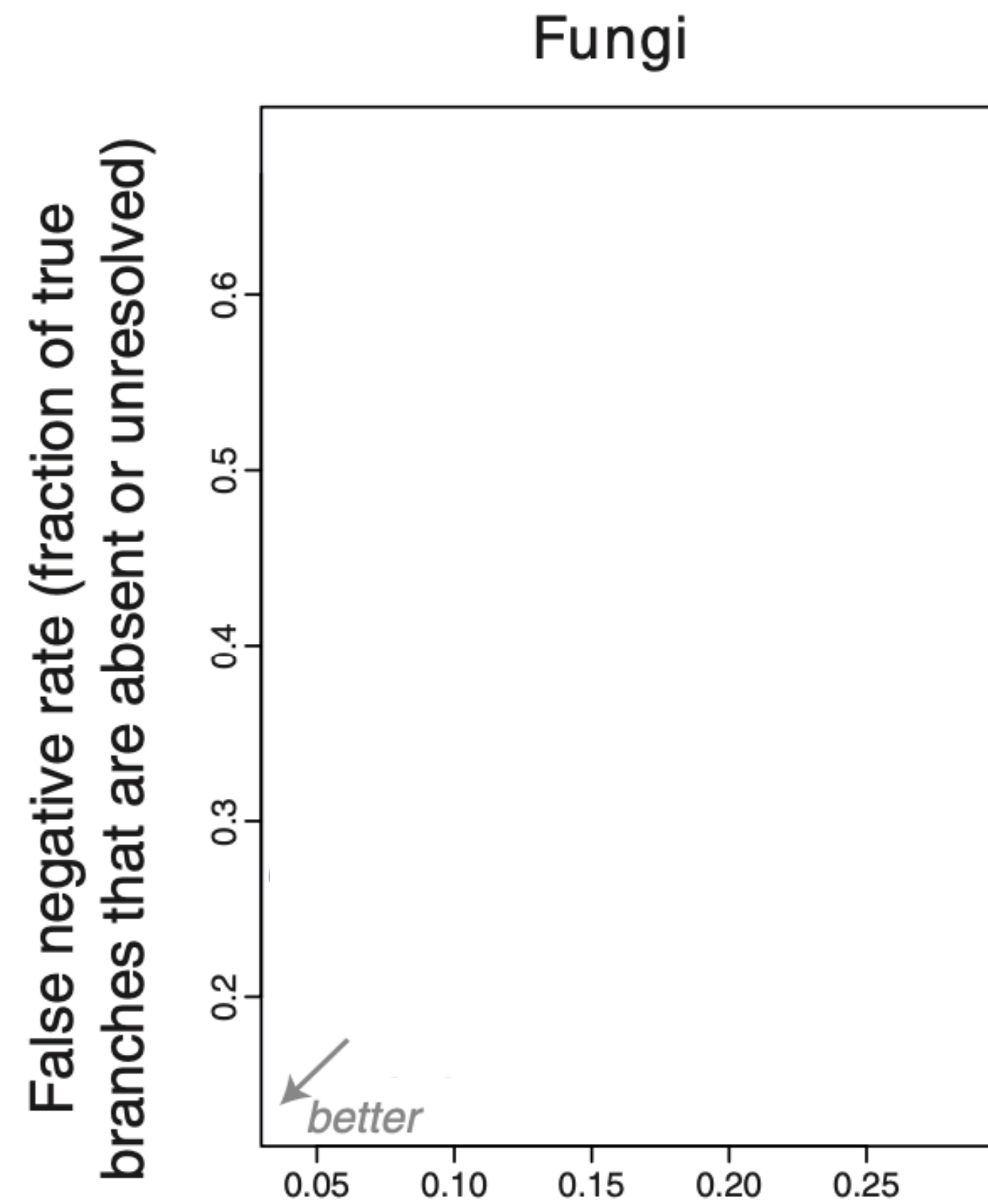
Comparing inferred tree to “incontestable” tree



Testing the impact of trimming

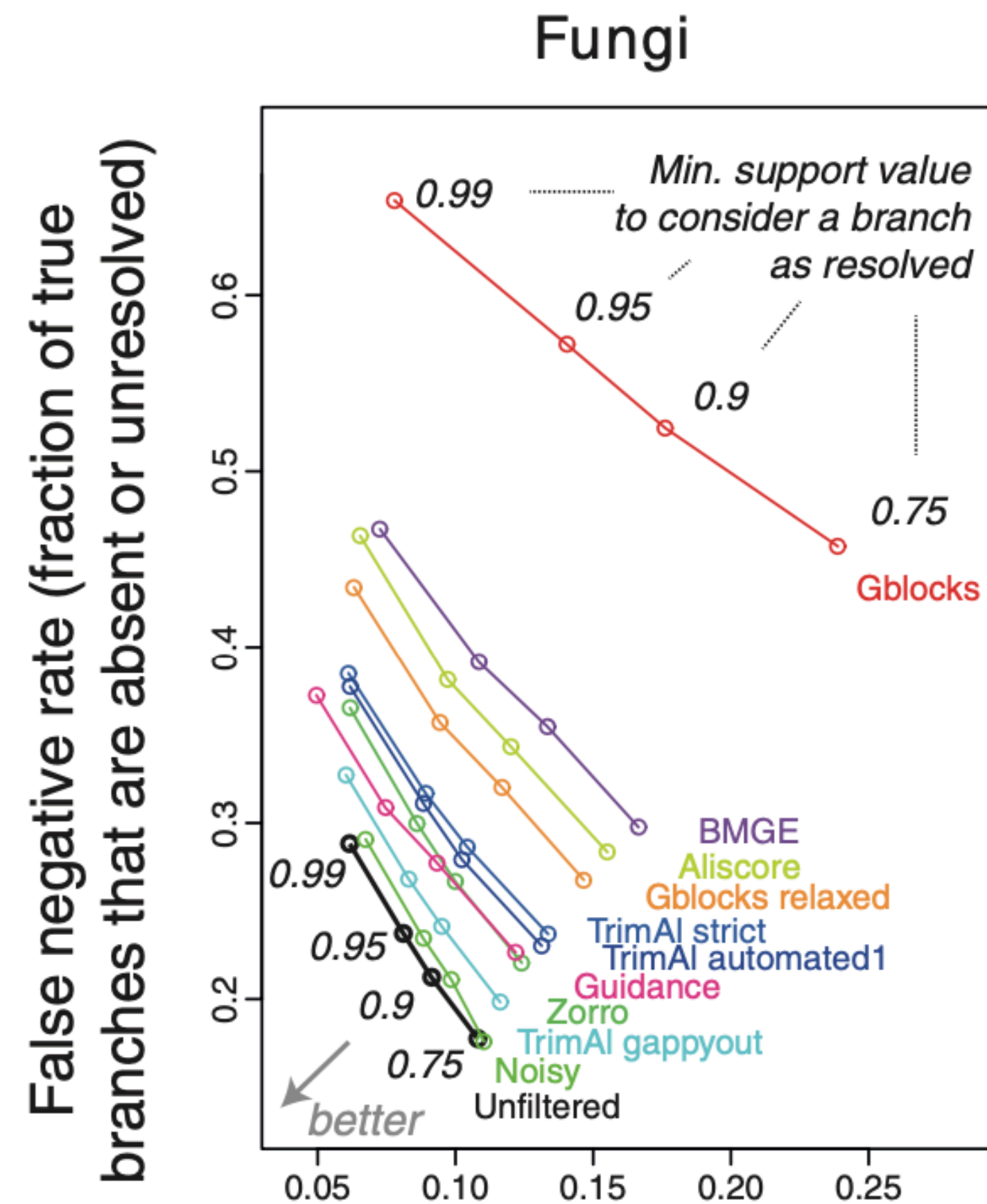


Testing the impact of trimming



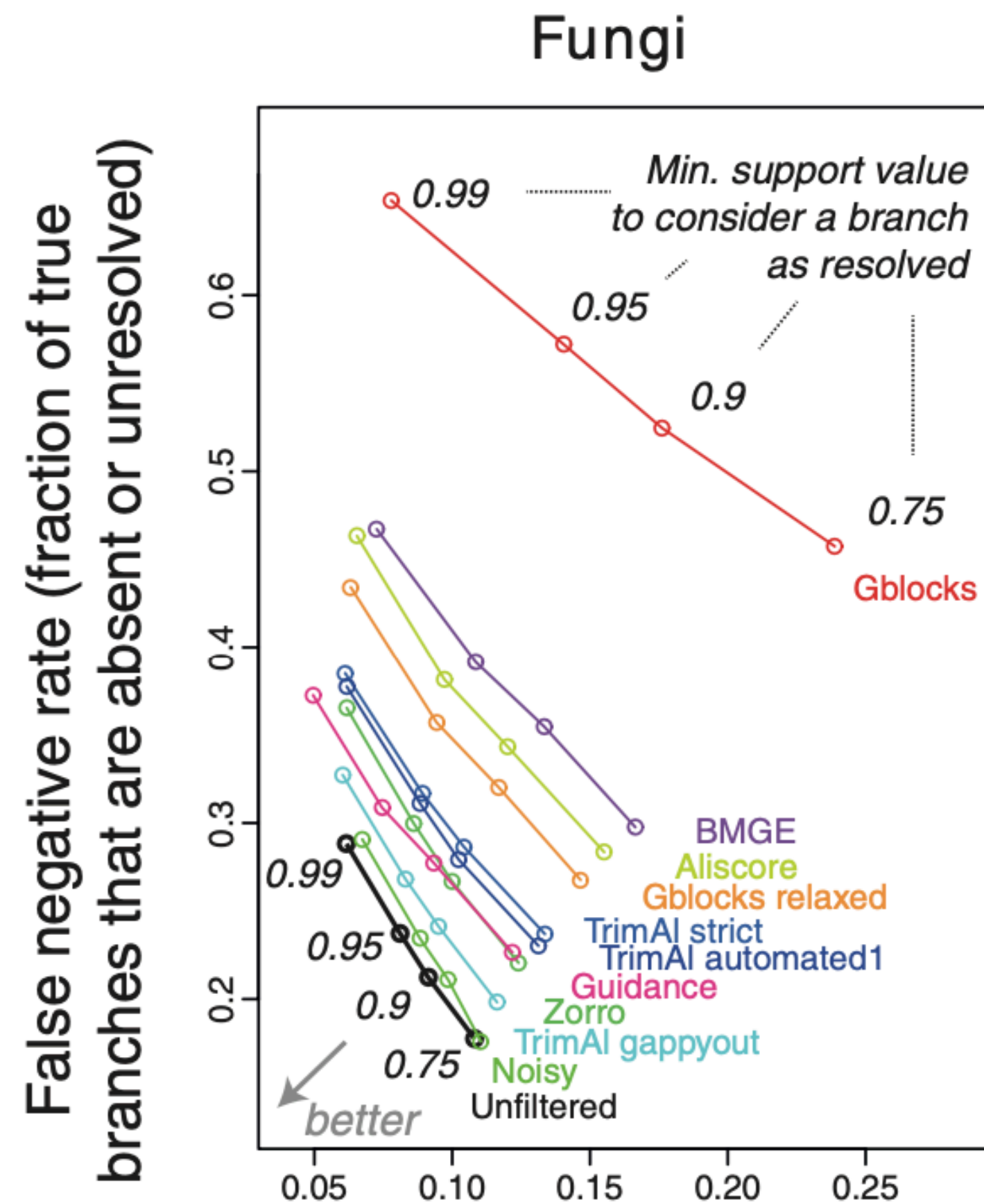
False positive rate (fraction of resolved branches that are incorrect)

Testing the impact of trimming



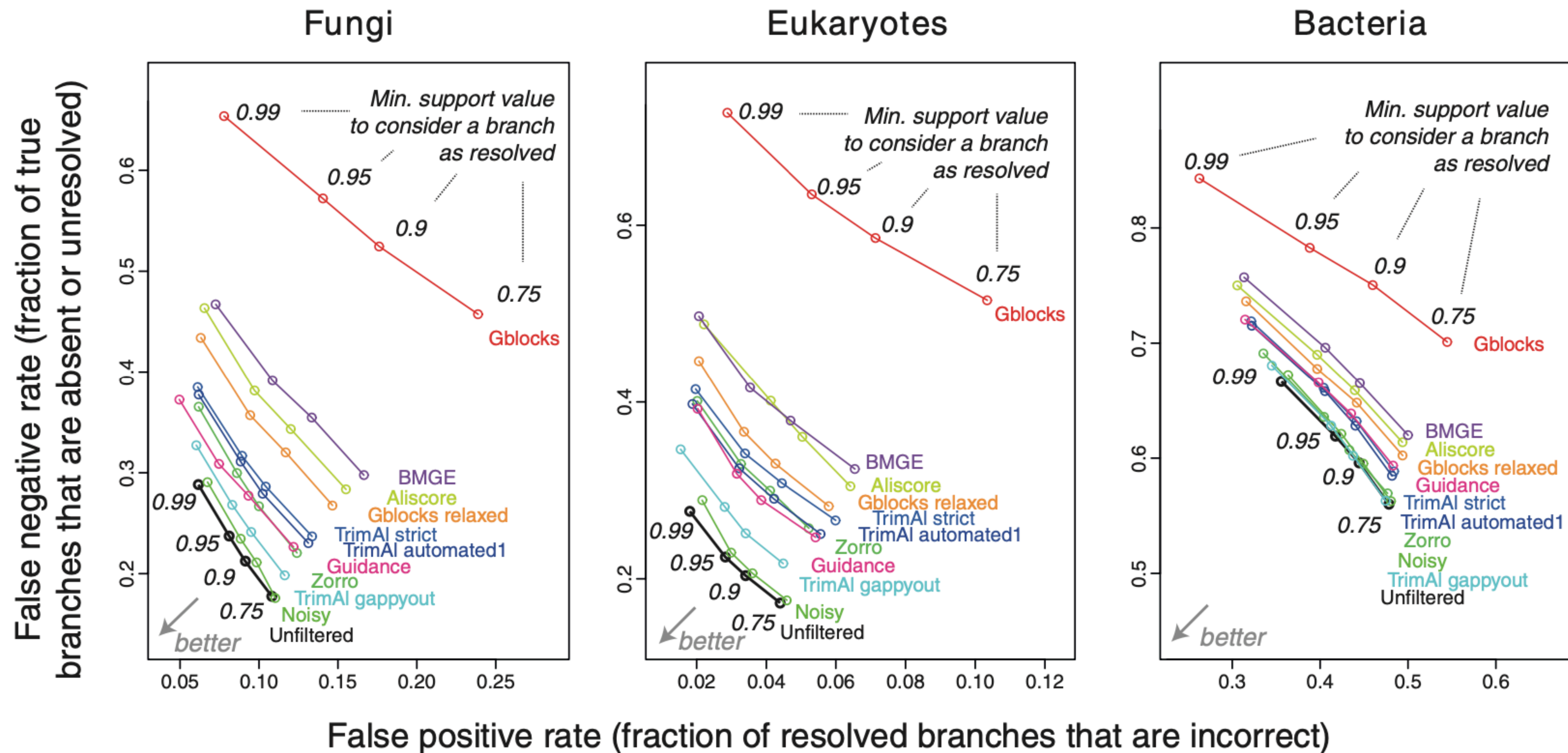
False positive rate (fraction of resolved branches that are incorrect)

Testing the impact of trimming



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

Testing the impact of trimming



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”

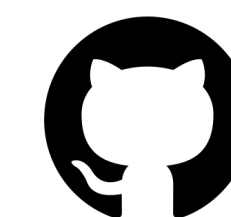
This suggest current methods remove sites with phylogenetic signal

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

>1

A-GTAT

>2

A-G-AT

>3

A-G-TA

>4

AGA-TA

>5

ACa-T-

kpi

>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

>1

A-GTAT

>2

A-G-AT

>3

A-G-TA

>4

AGA-TA

>5

ACa-T-

kpi

>1

A-GTAT

>2

A-G-AT

>3

A-G-TA

>4

AGA-TA

>5

ACa-T-

kpik

>1

A-GTAT

>2

A-G-AT

>3

A-G-TA

>4

AGA-TA

>5

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)

ClipKIT has several modes

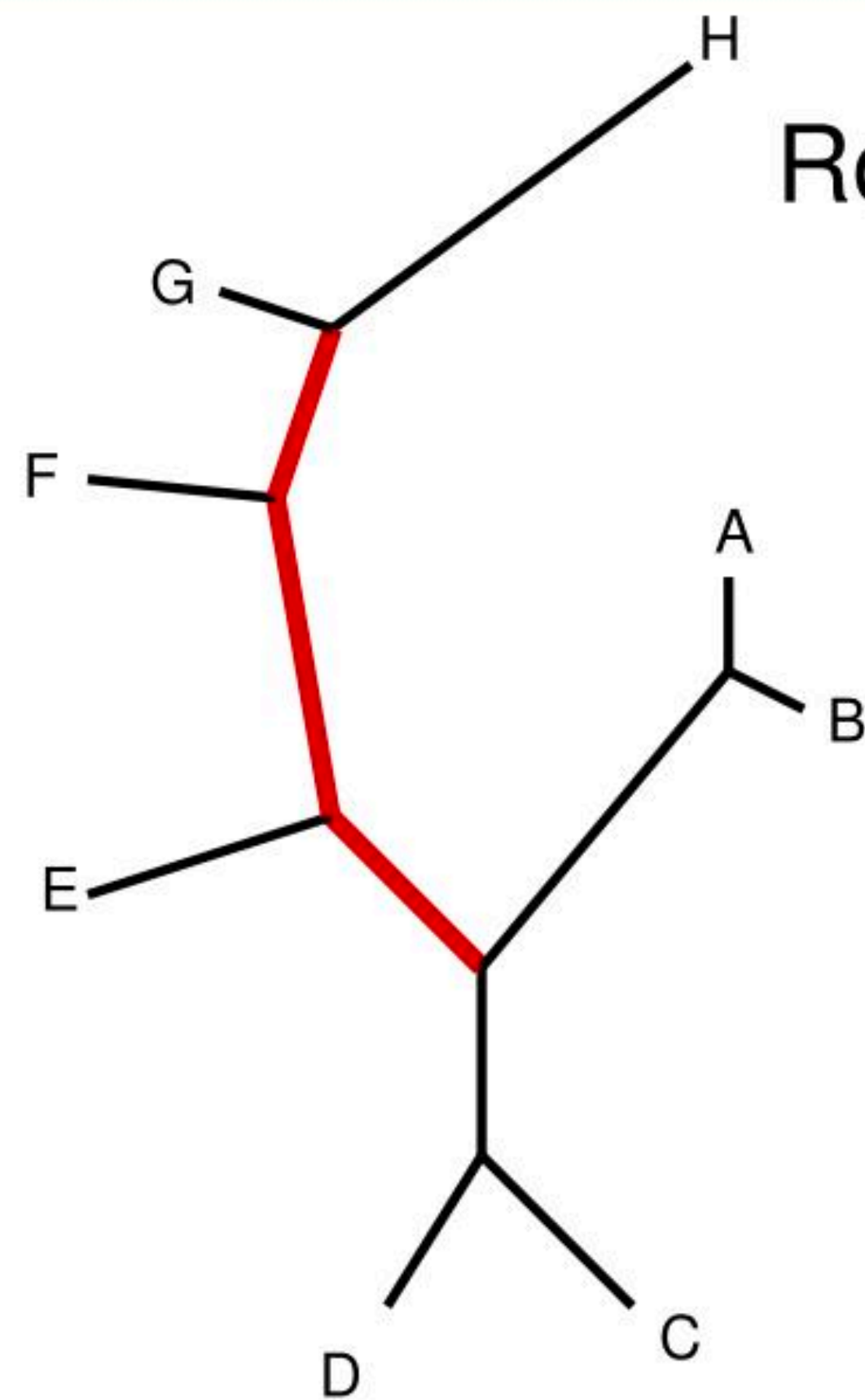
- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
 - kpi (keeps only parsimony informative sites)
 - kplic (keeps parsimony informative sites and constant sites)
 - smart-gap (dynamic gappyness threshold determination)
 - gappy (removes sites with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

ClipKIT has several modes

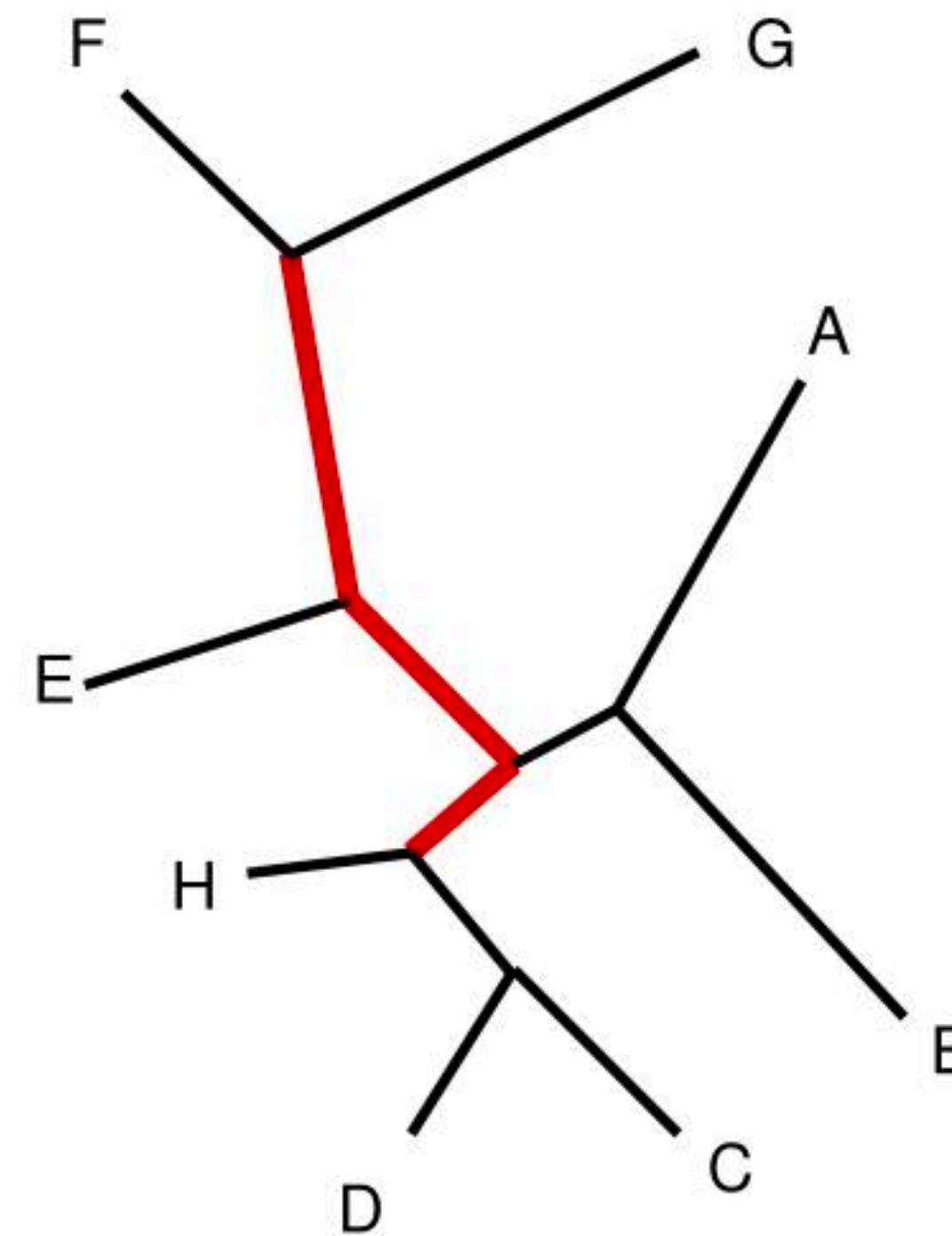
- based on keeping parsimony informative sites and sites that aren't gappy-rich
- ClipKIT has five modes:
 - kpi (keeps only parsimony informative sites)
 - kpic (keeps parsimony informative sites and constant sites)
 - smart-gap (dynamic gappyness threshold determination)
 - gappy (removes sites with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

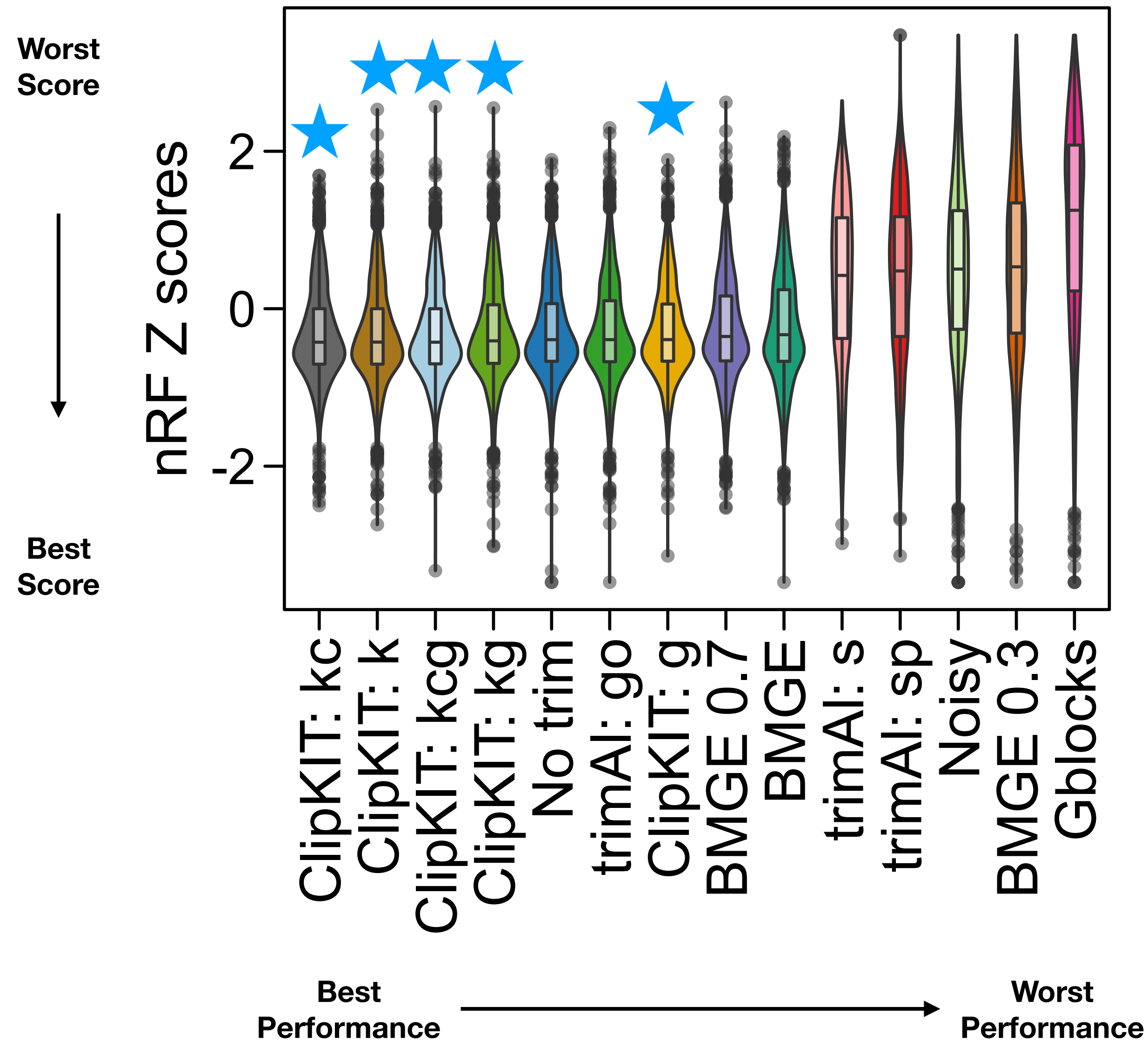
Internal branches exist in one tree but not in the other



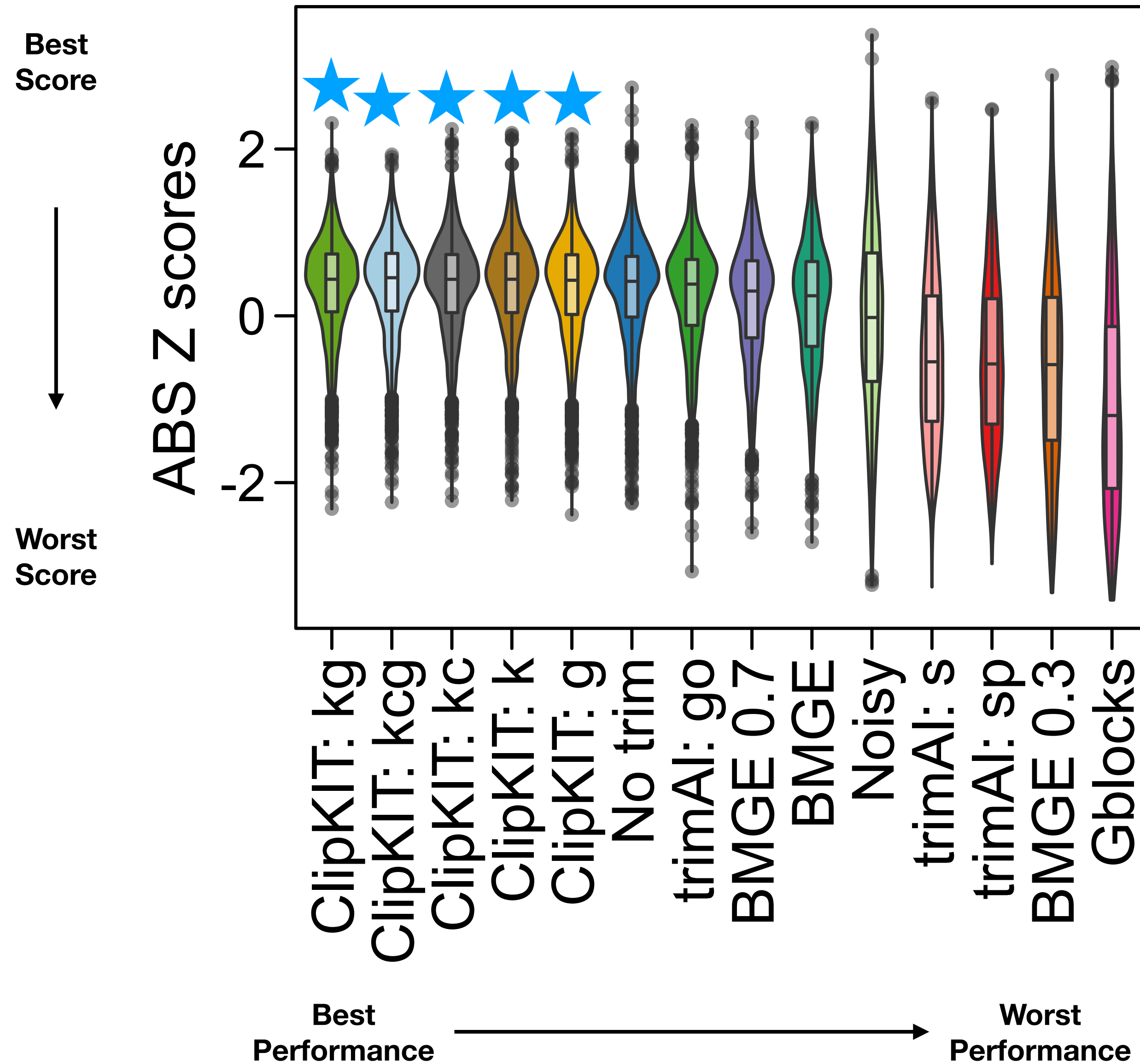
Robinson-Foulds distance = 6



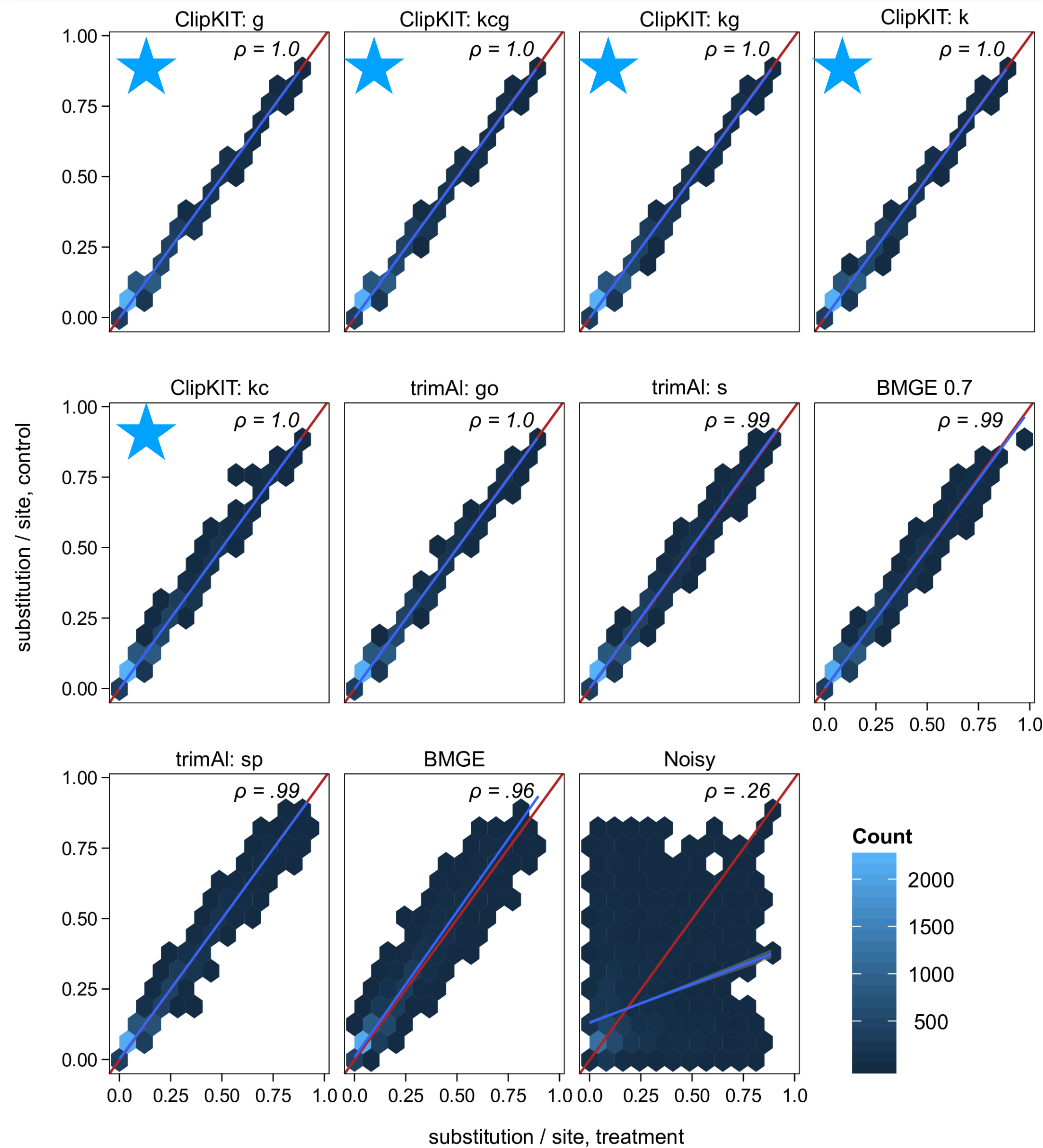
Trees inferred using ClipKIT are accurate



Trees inferred using ClipKIT are well supported



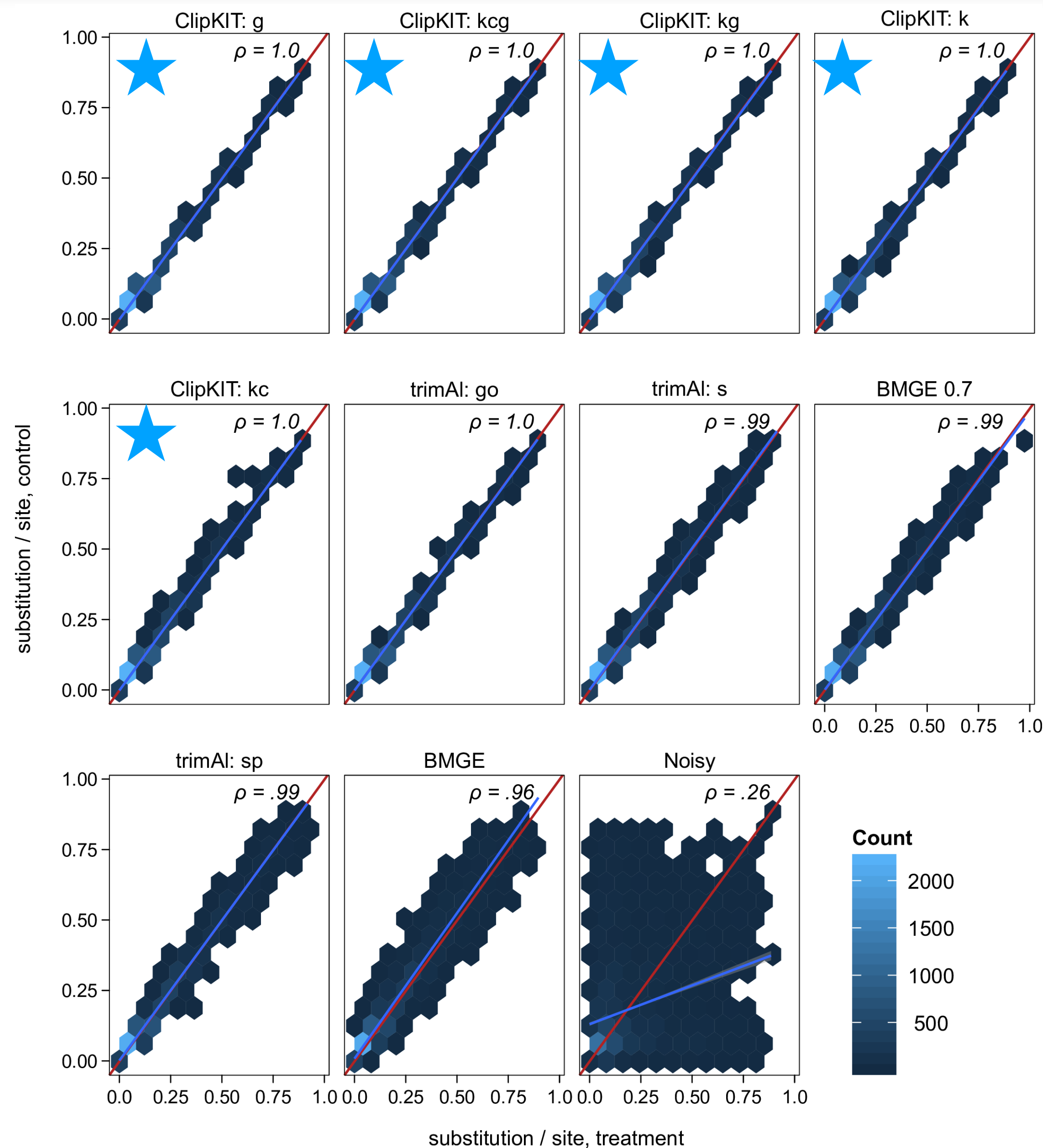
Branch lengths estimates after trimming are typically accurate



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

Choose FASTA(s)

Mode

smart-gap (default) ▼

Trim FASTA(s)

Sequence Type

Auto detect ▼

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



Buida, J



O'Meara, T



Li, Y



Verbruggen, H



King, N



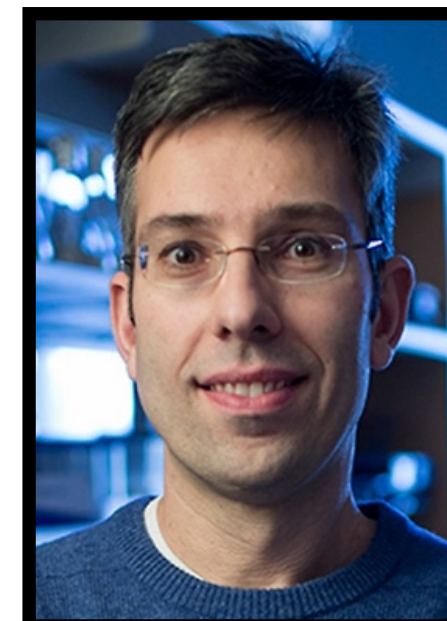
Coyle, M



Geiser, D



Goldman, G



Rokas, A



Hittinger, C



Berman, J



Shen, X

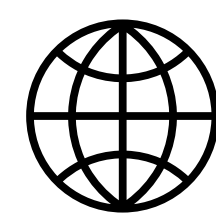


Stay tuned for a silly quiz in the last 10 minutes!

Trimming MSAs



@JLSteenwyk

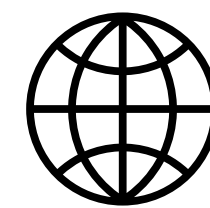


<https://jlsteenwyk.com/>

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



@JLSteenwyk



<https://jlsteenwyk.com/>