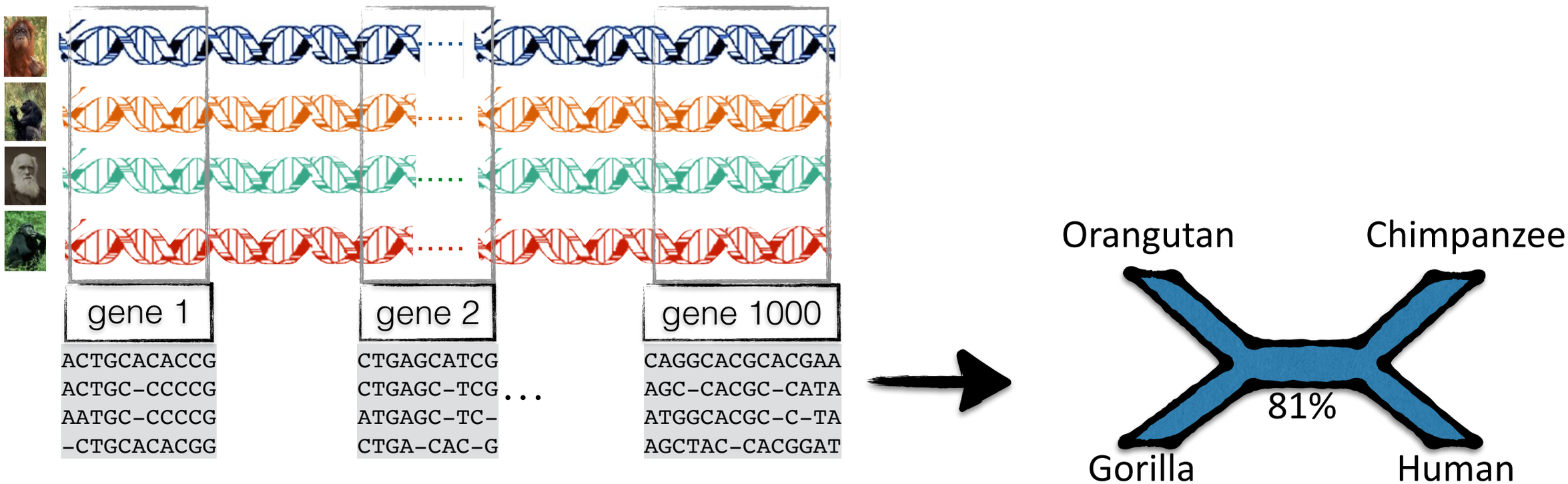
Site-based quartet-based estimation of species trees (CASTER)

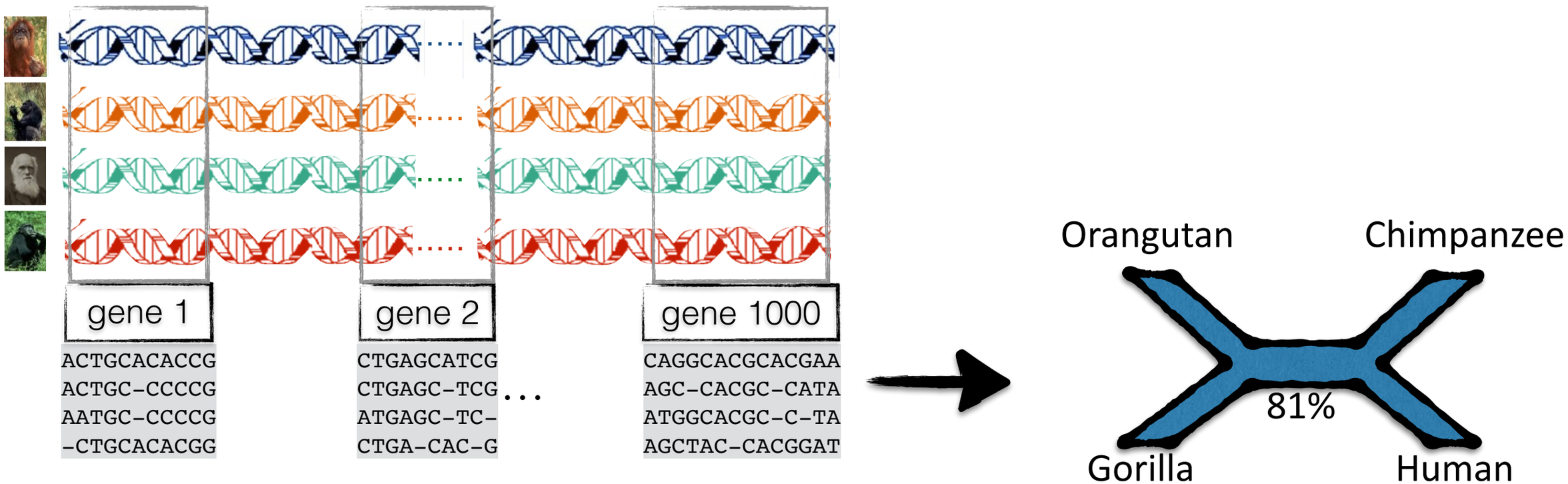
Chao Zhang, Rasmus Nielsen, Siavash Mirarab

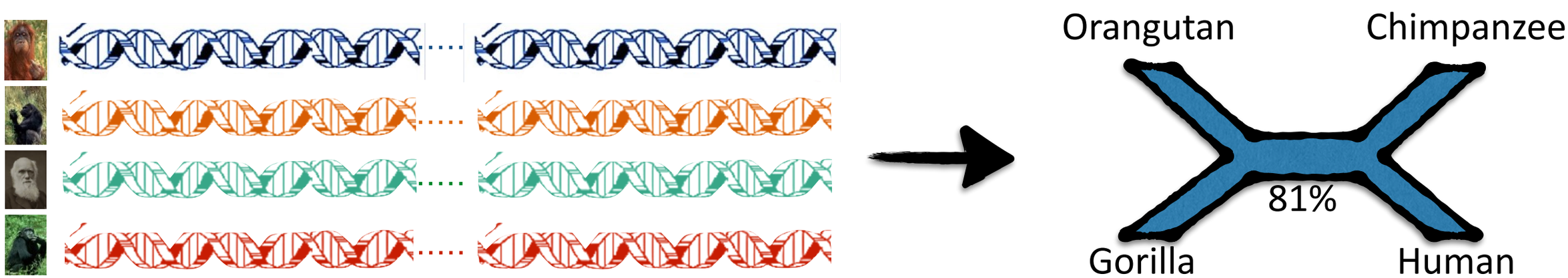
Chao Zhang

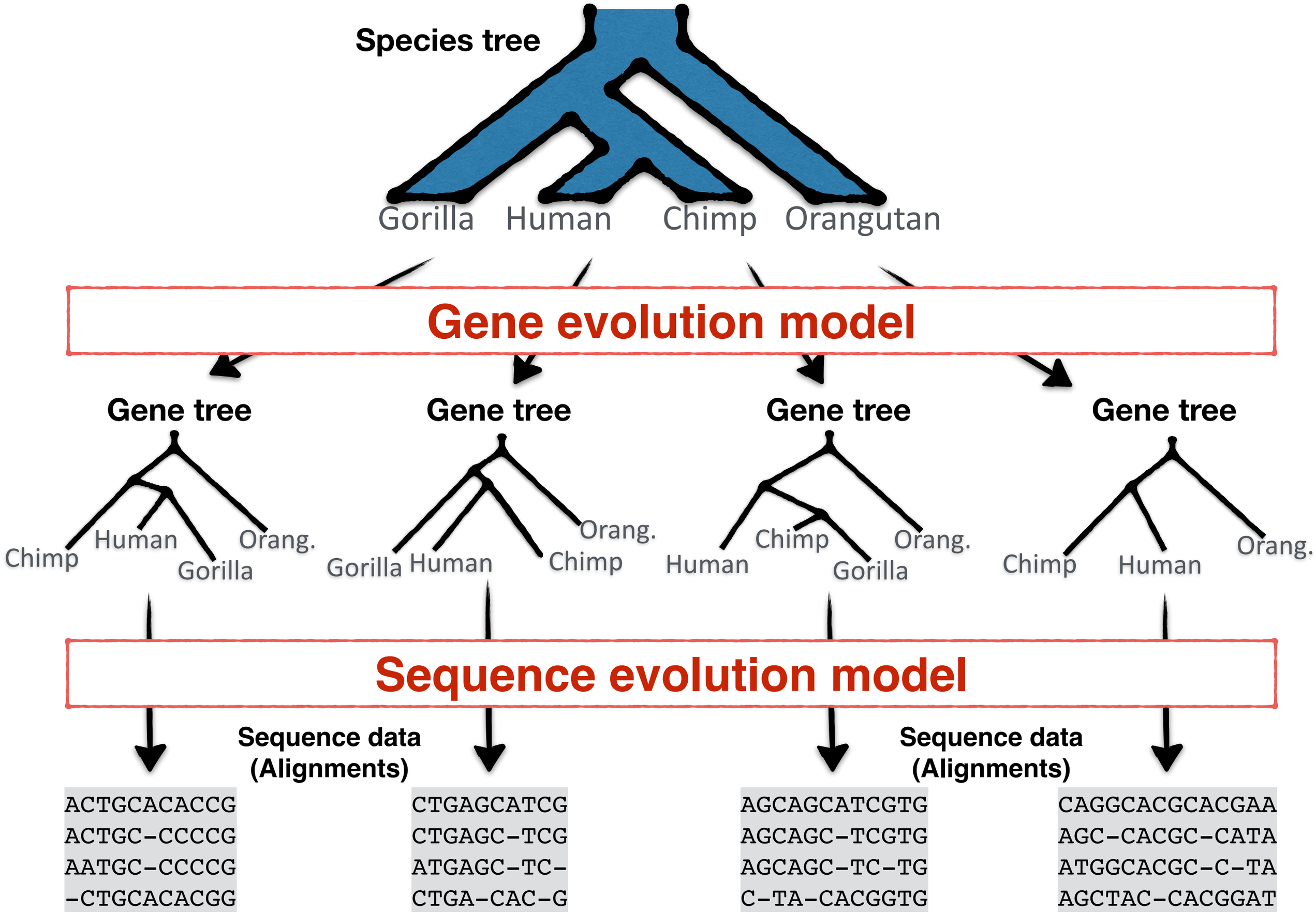
Truly genome-wide phylogenomics!

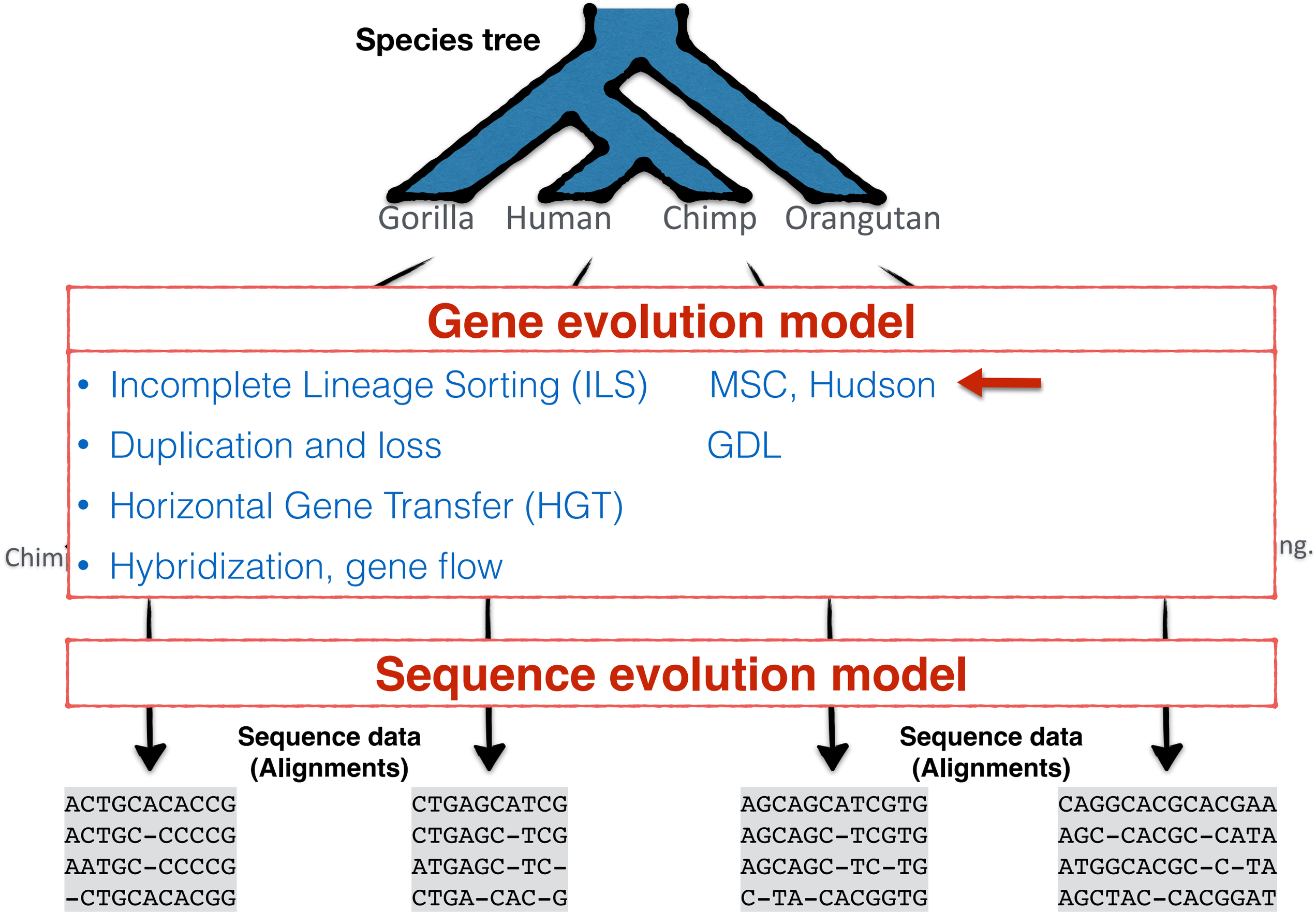


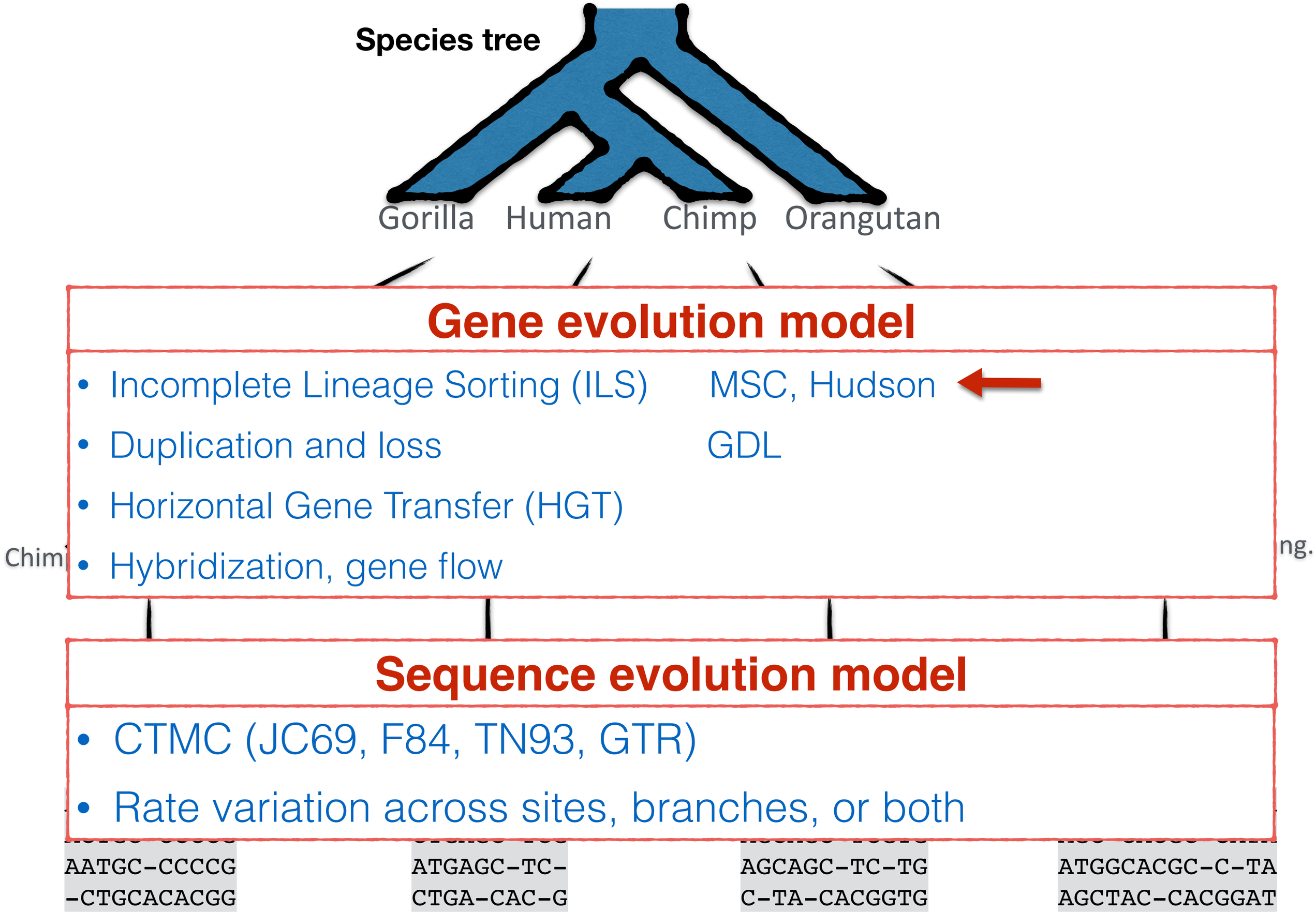
Truly genome-wide phylogenomics!






Species tree
Gorilla
Human
Chimp
Orangutan
Gene evolution model
Gene tree
Chimp
Human
Orang.
Gorilla
Gene tree
Gorilla
Human
Orang.
Chimp
Gene tree
Human
Chimp
Orang.
Gorilla
Gene tree
Chimp
Human
Orang.
Sequence evolution model
Sequence data
(Alignments)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
Sequence data
(Alignments)
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

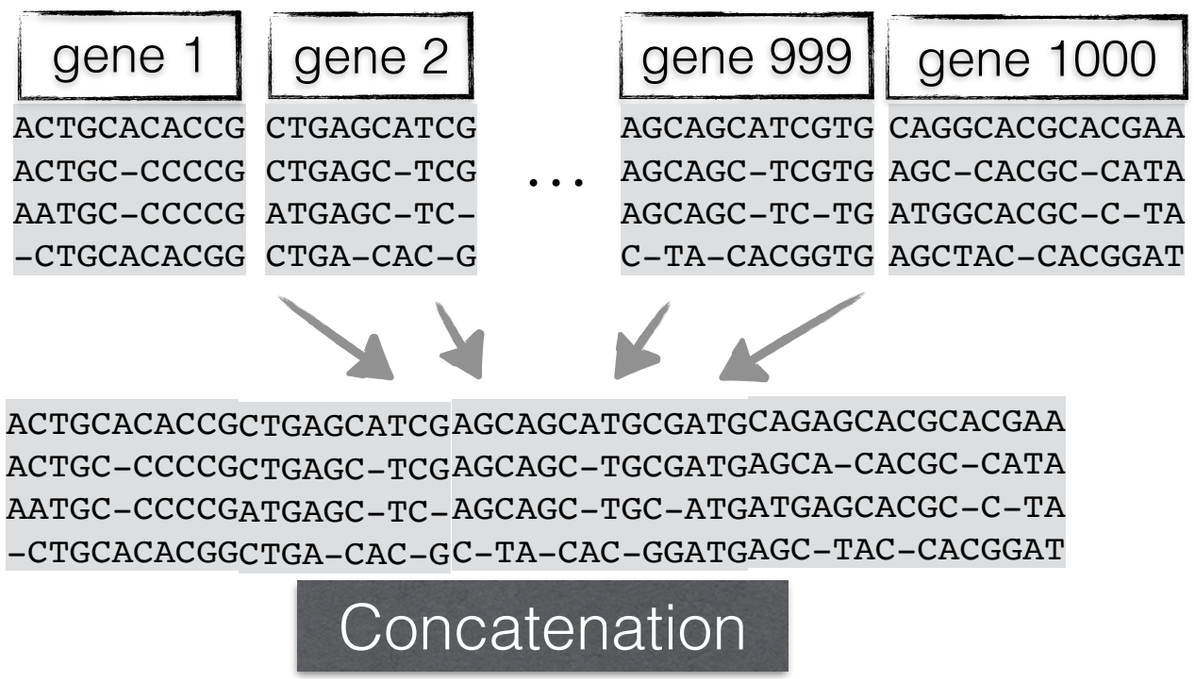

Species tree
Gorilla
Human
Chimp
Orangutan
Gene evolution model
Incomplete Lineage Sorting (ILS)
MSC, Hudson
Duplication and loss
GDL
Horizontal Gene Transfer (HGT)
Hybridization, gene flow
Sequence evolution model
Sequence data (Alignments)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
Sequence data (Alignments)
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT


Species tree
Gorilla
Human
Chimp
Orangutan
Gene evolution model
Incomplete Lineage Sorting (ILS)
MSC, Hudson
Duplication and loss
GDL
Horizontal Gene Transfer (HGT)
Hybridization, gene flow
Sequence evolution model
CTMC (JC69, F84, TN93, GTR)
Rate variation across sites, branches, or both
AATGC-CCCCG
-CTGCACACGG
ATGAGC-TC-
CTGA-CAC-G
AGCAGC-TC-TG
C-TA-CACGGTG
ATGGCACGC-C-TA
AGCTAC-CACGGAT

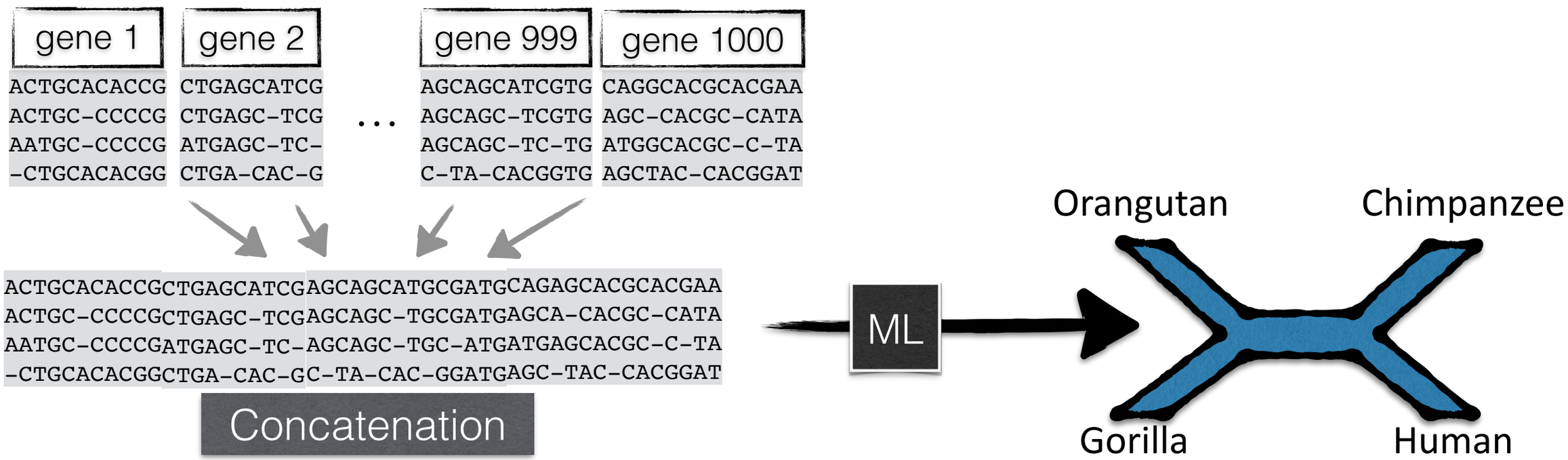
Concatenation is not enough!

gene 1	gene 2		gene 999	gene 1000
ACTGCACACCG	CTGAGCATCG		AGCAGCATCGTG	CAGGCACGCACGAA
ACTGC-CCCCG	CTGAGC-TCG	...	AGCAGC-TCGTG	AGC-CACGC-CATA
AATGC-CCCCG	ATGAGC-TC-		AGCAGC-TC-TG	ATGGCACGC-C-TA
-CTGCACACGG	CTGA-CAC-G		C-TA-CACGGTG	AGCTAC-CACGGAT

Concatenation is not enough!



Concatenation is not enough!



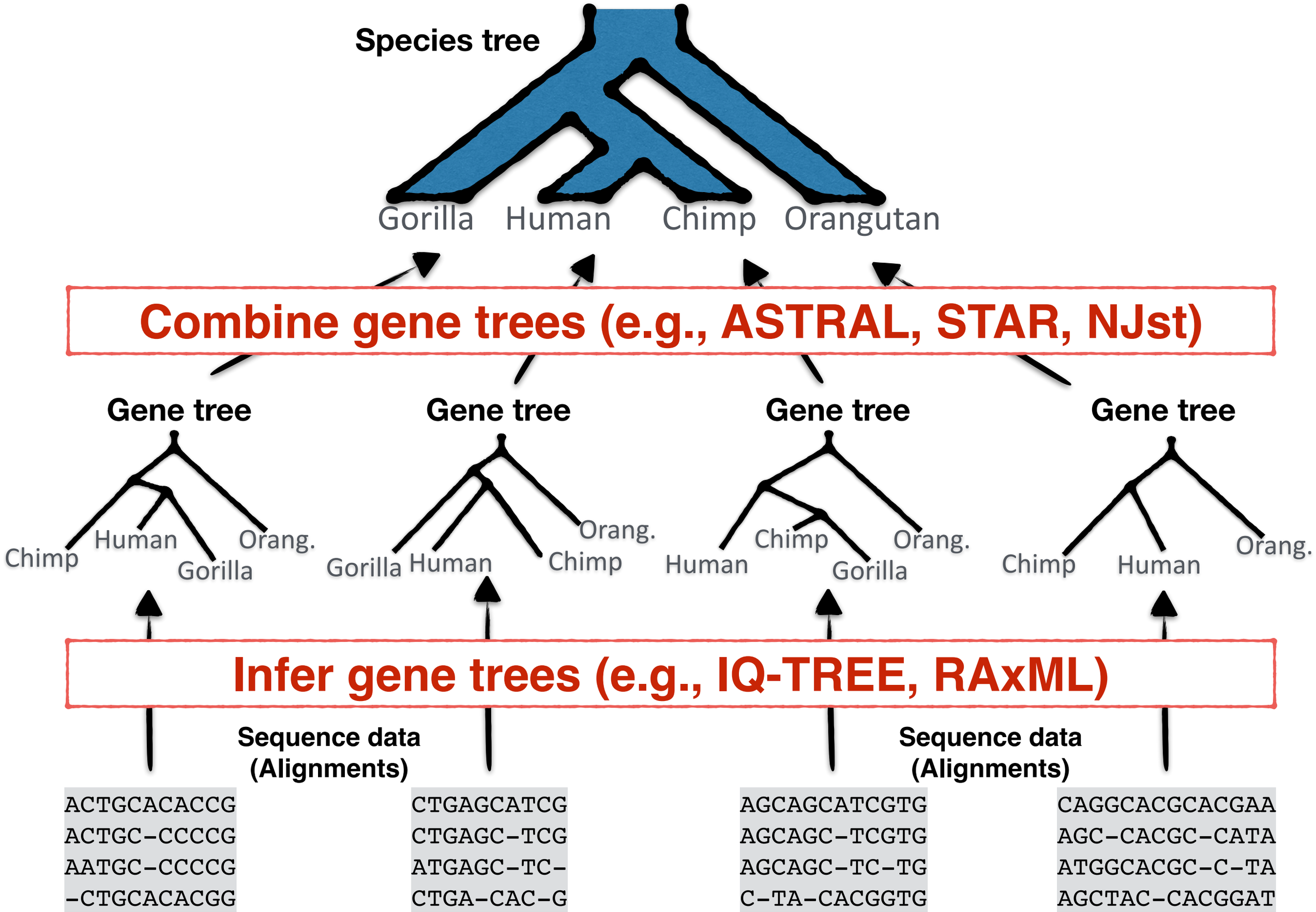
Statistically <u>inconsistent</u> & positively misleading

(Roch and Steel, Theo. Pop. Gen., 2014)

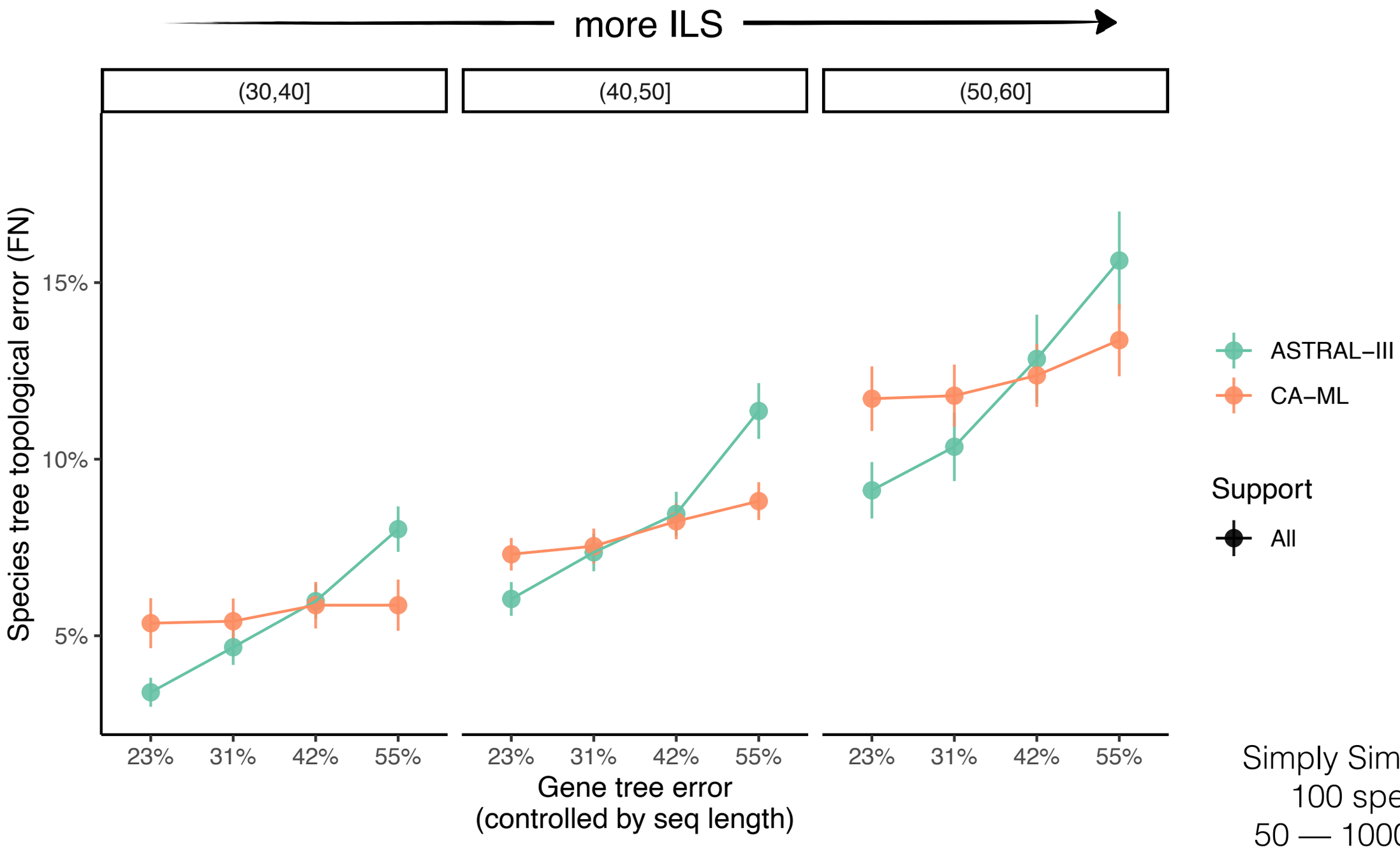
Mixed accuracy in simulations

(Kubatko and Degnan, Systematic Biology, 2007)
(Mirarab, et al., Systematic Biology, 2014)

Memory can become a bottleneck


Species tree
Gorilla
Human
Chimp
Orangutan
Combine gene trees (e.g., ASTRAL, STAR, NJst)
Gene tree
Gene tree
Gene tree
Gene tree
Chimp
Human
Orang.
Gorilla
Gorilla
Human
Orang.
Chimp
Human
Chimp
Orang.
Gorilla
Chimp
Human
Orang.
Infer gene trees (e.g., IQ-TREE, RAxML)
Sequence data
(Alignments)
Sequence data
(Alignments)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

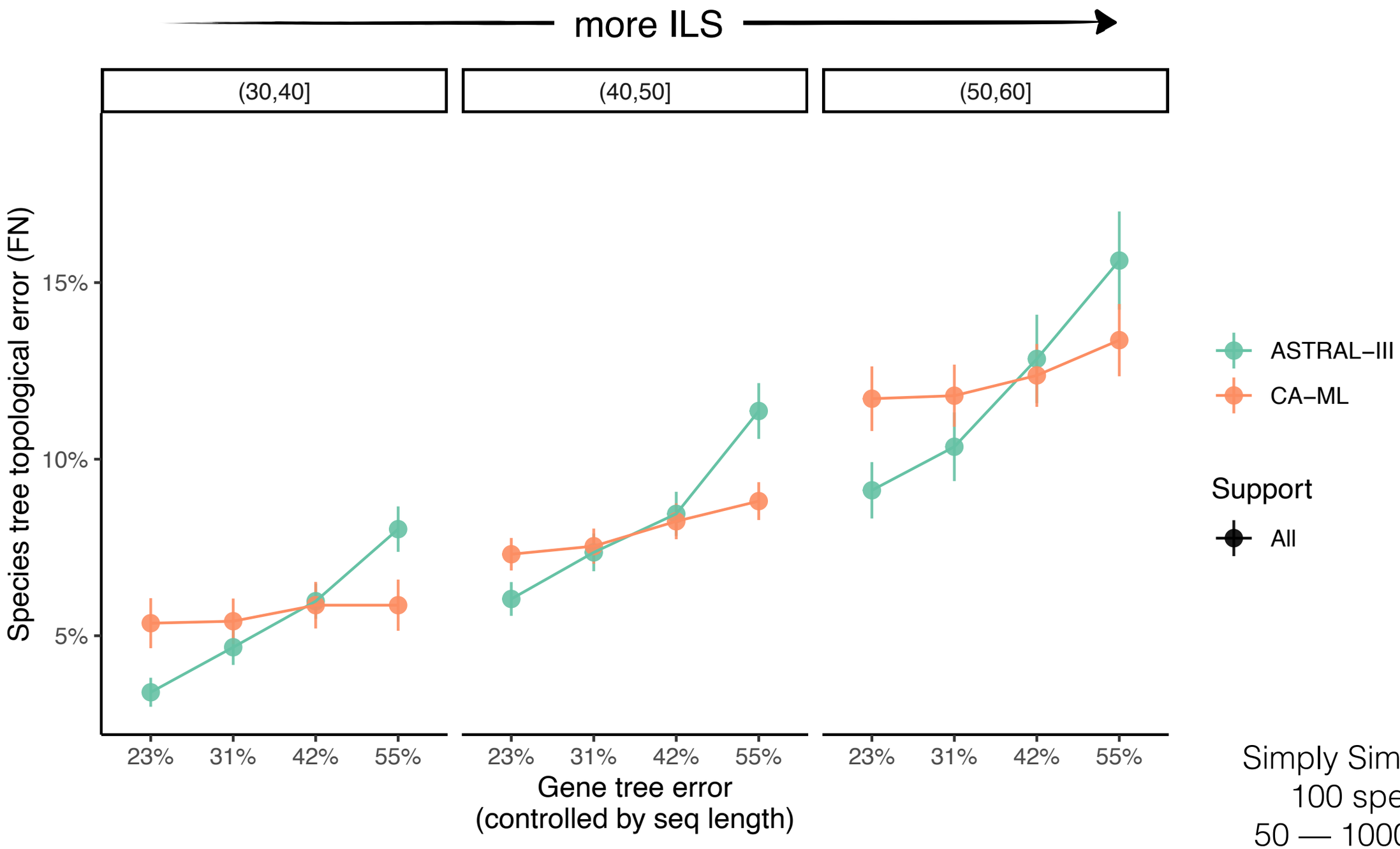
Sensitive to high levels of gene estimation tree error



Simply Simulations,
100 species,
50 — 1000 genes

Zhang, C. & Mirarab, S. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. MBE (2022). doi.org/10.1093/molbev/msac215

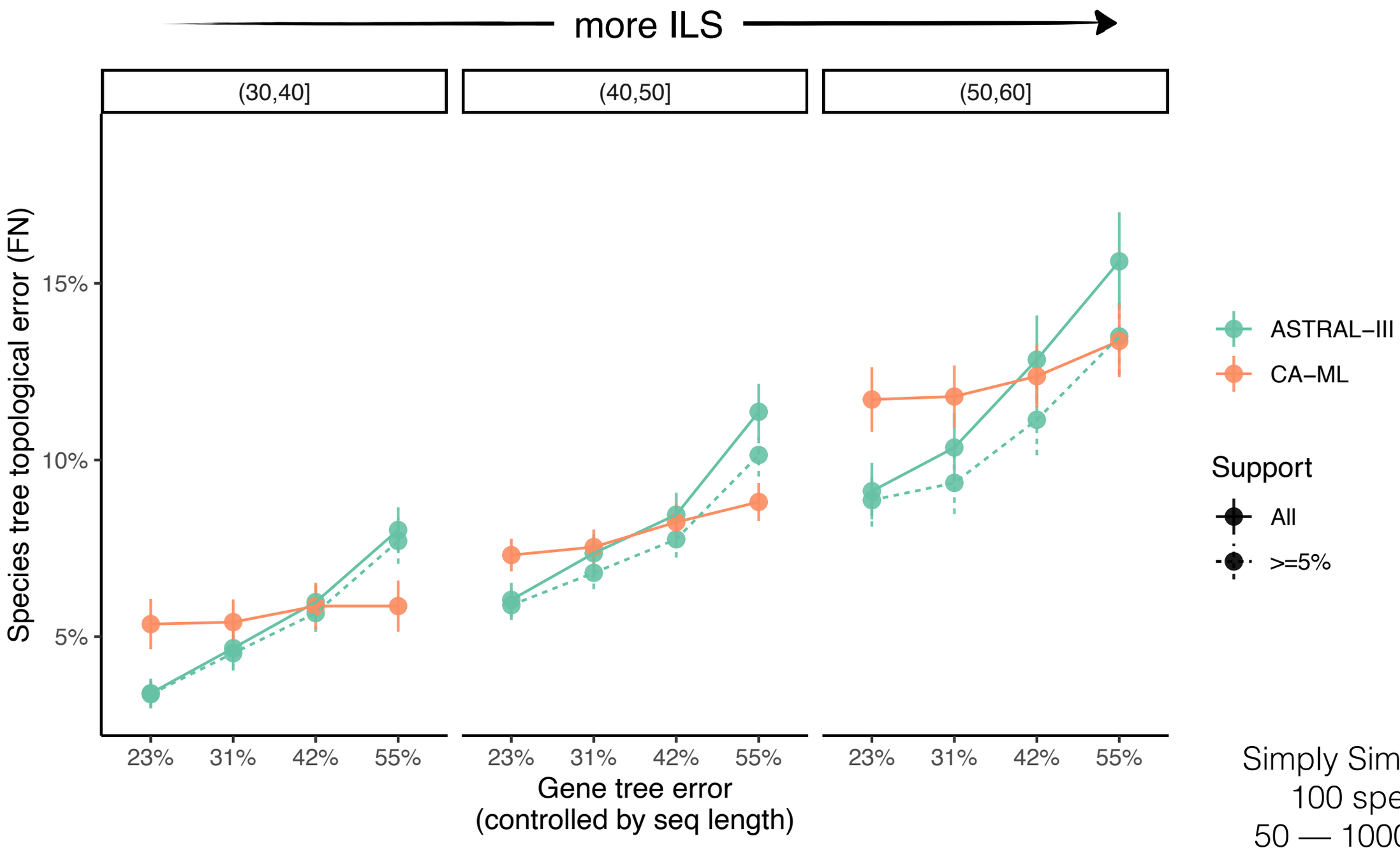
Sensitive to high levels of gene estimation tree error



Simply Simulations,
100 species,
50 — 1000 genes

Zhang, C. & Mirarab, S. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. MBE (2022). doi.org/10.1093/molbev/msac215

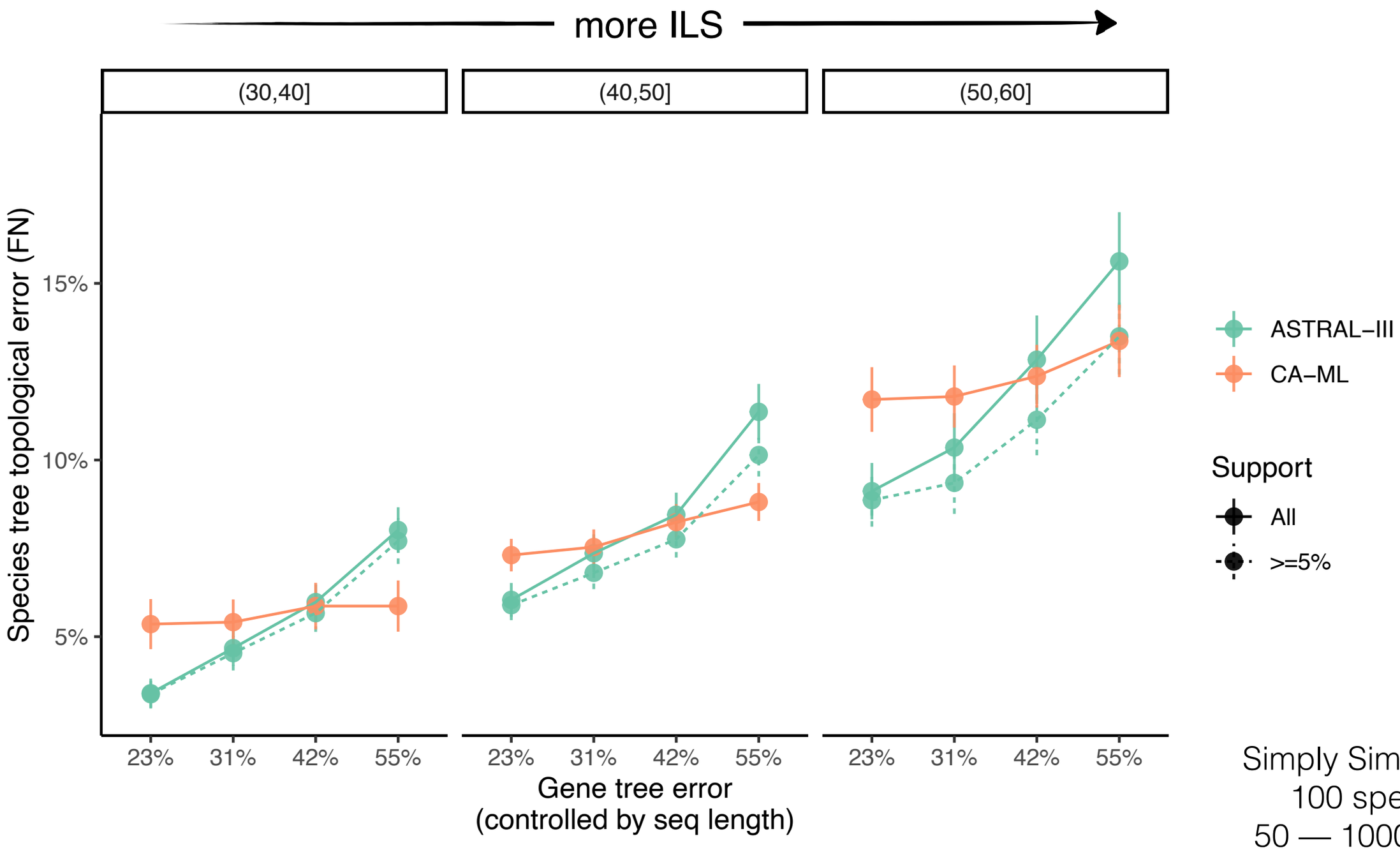
Contracting low support helps but …



Simply Simulations,
100 species,
50 — 1000 genes

Zhang, C. & Mirarab, S. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. MBE (2022). doi.org/10.1093/molbev/msac215

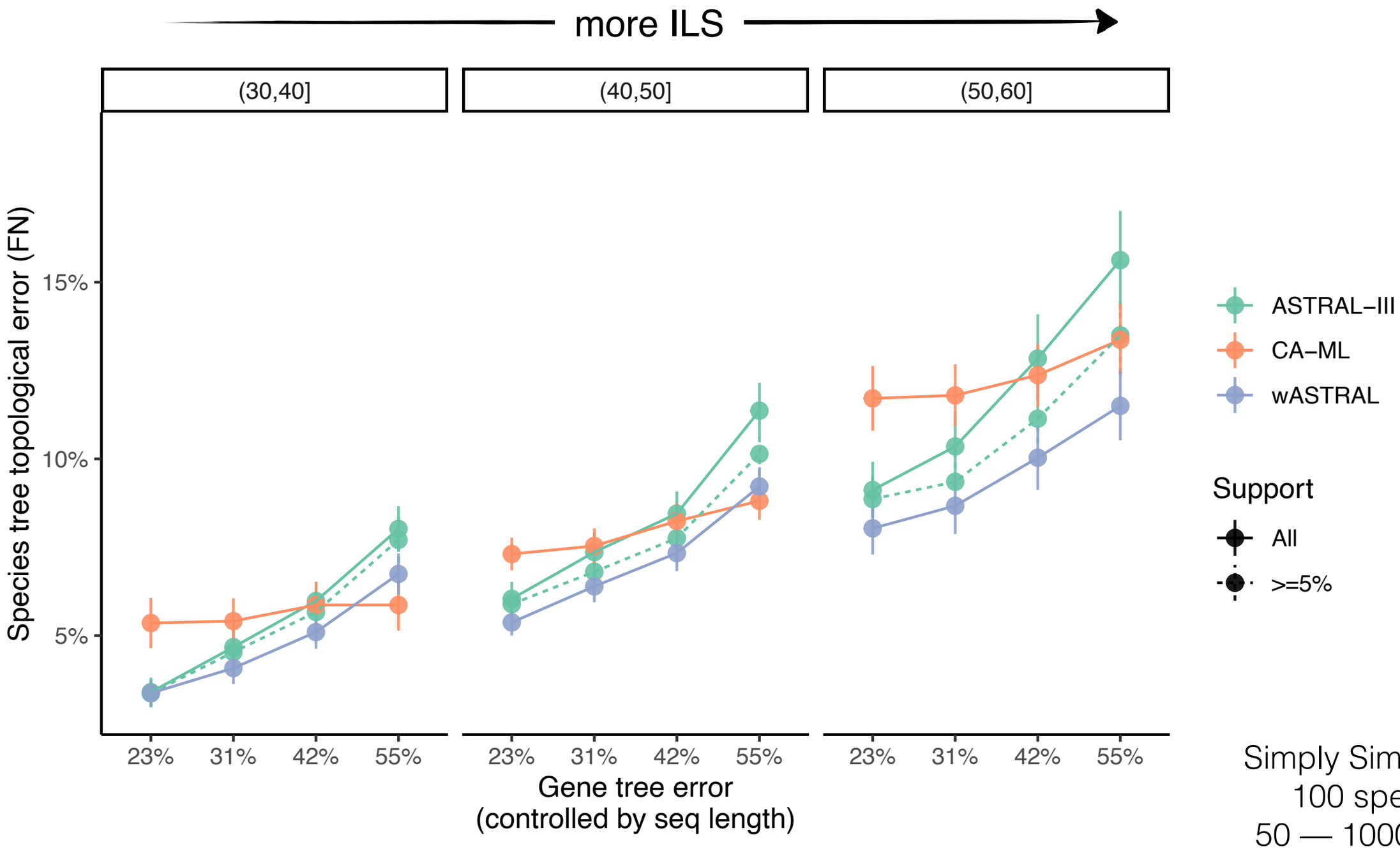
1. Not clear <u>what threshold</u> to use (data dependent)
2. Even after contraction, can be worse than CA-ML



Simply Simulations,
100 species,
50 — 1000 genes

Zhang, C. & Mirarab, S. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. MBE (2022). doi.org/10.1093/molbev/msac215

Weighting (almost) closes the gap to concatenation



Simply Simulations,
100 species,
50 — 1000 genes

Zhang, C. & Mirarab, S. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. MBE (2022). doi.org/10.1093/molbev/msac215

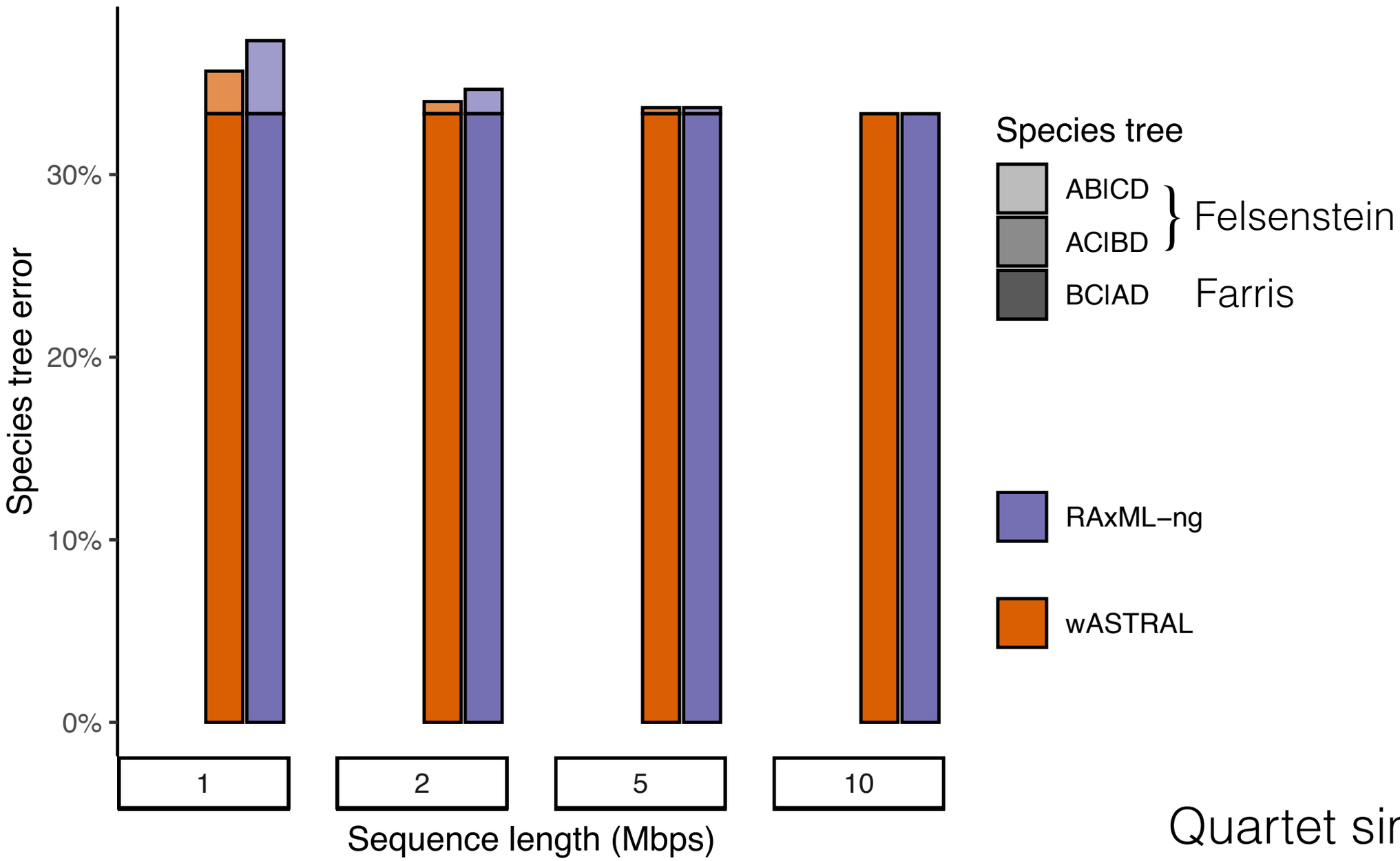
How about systematic error? (Long Branch Attraction/Revulsion)

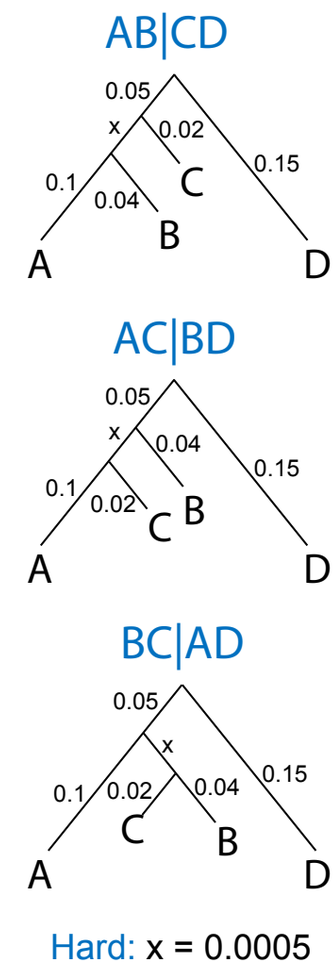
Quartet simulations

Recombination+ILS simulations
(msprime; Hudson model)
4 species, 10Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished

How about systematic error? (Long Branch Attraction/Revulsion)





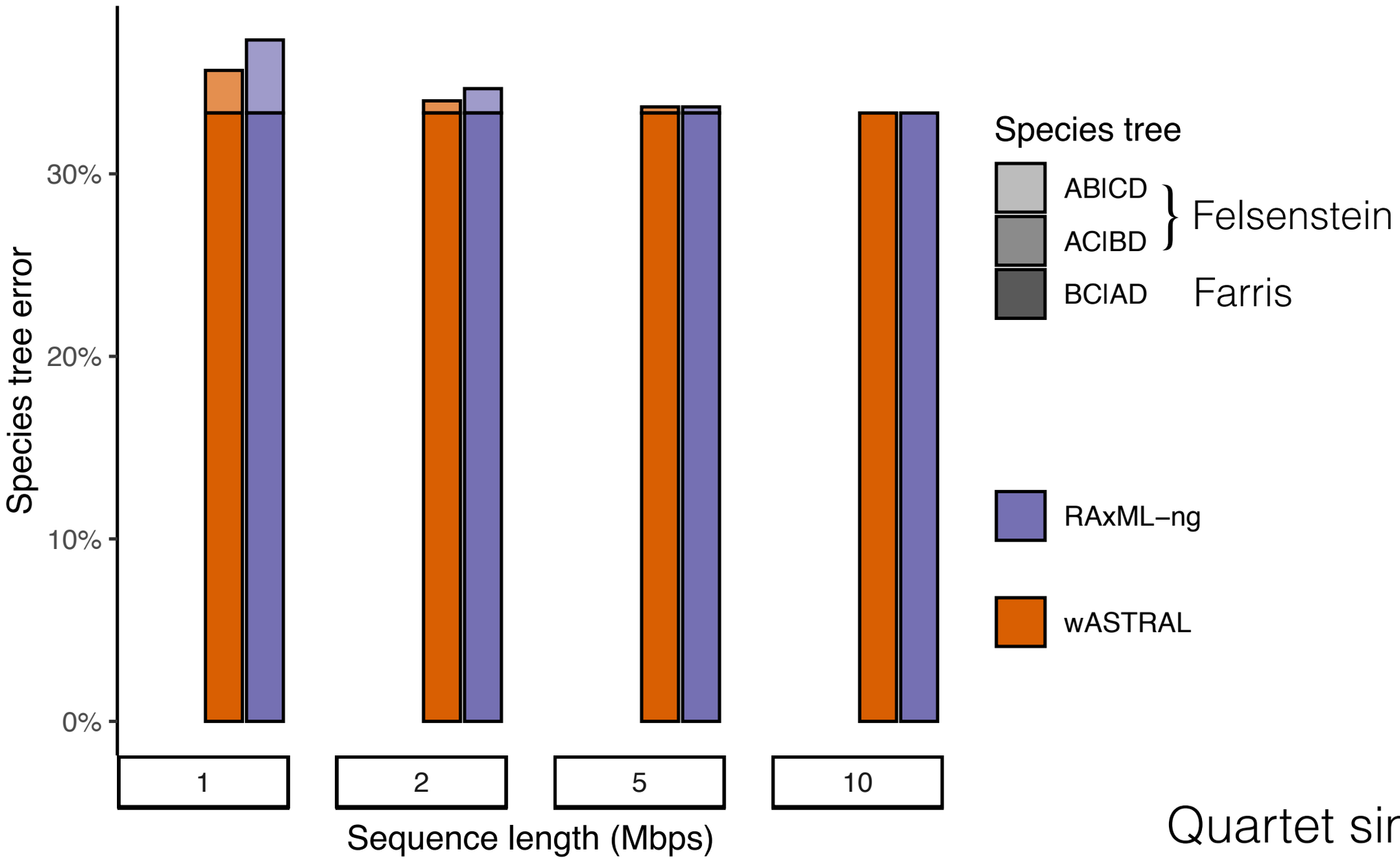
Quartet simulations

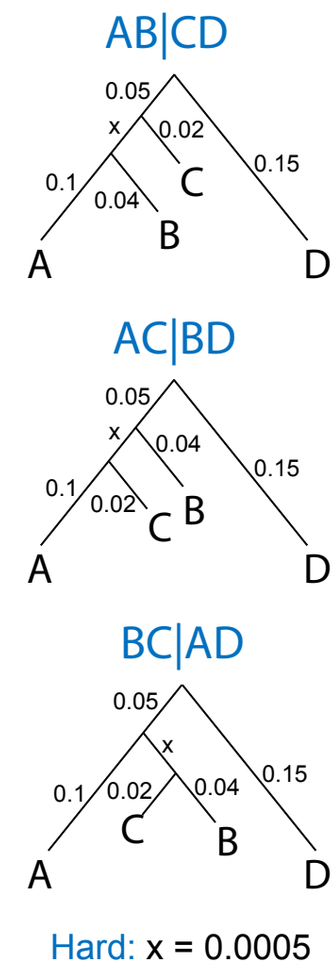
Recombination+ILS simulations
(msprime; Hudson model)
4 species, 10Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished

Challenges with wASTRAL+ML gene trees:

A. Long branch attraction can prevent good gene trees





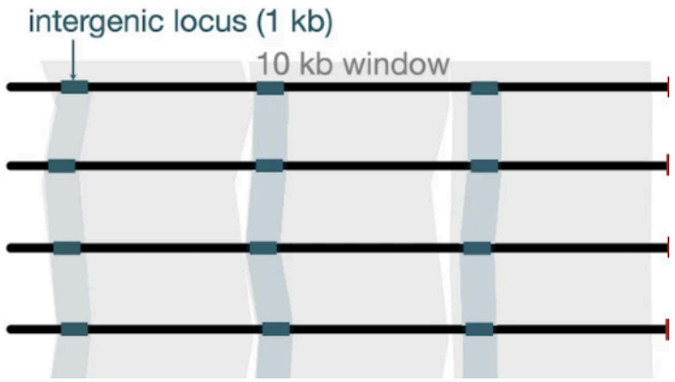
Quartet simulations

Recombination+ILS simulations
(msprime; Hudson model)
4 species, 10Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished

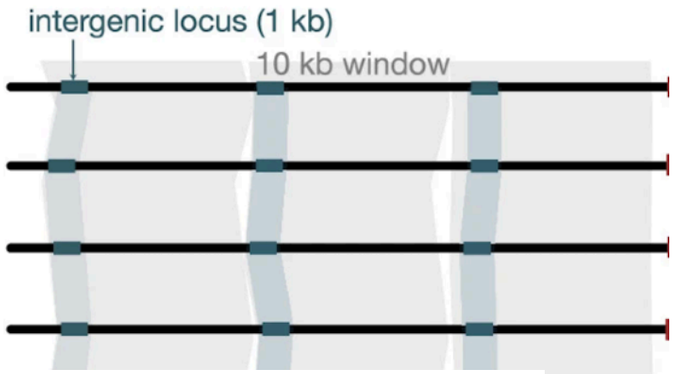
How about recombination?

- Choose short(ish) genes (e.g., 1000bp)
  Leave gaps between genes (e.g., 9000bp)

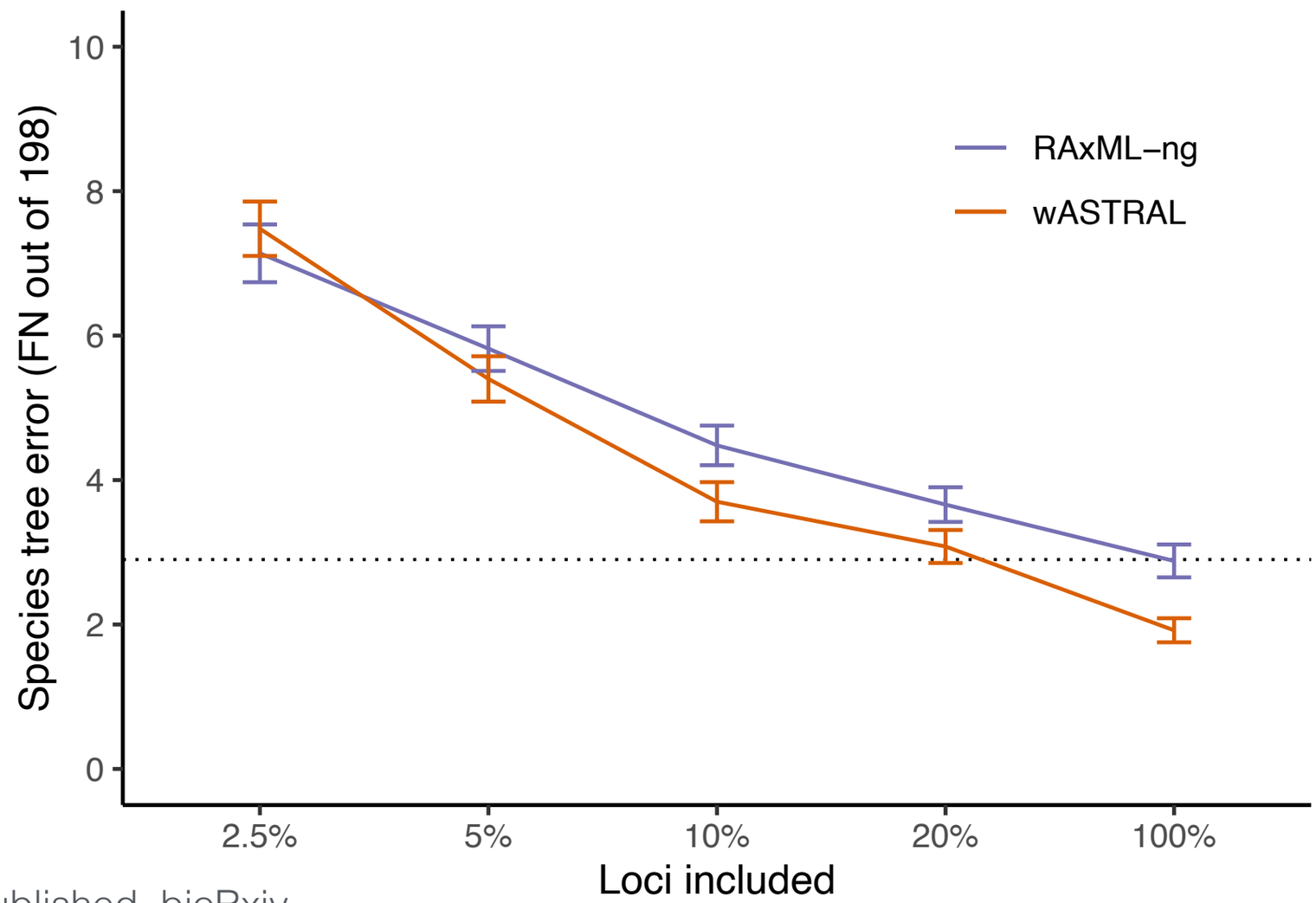


How about recombination?

- Choose short(ish) genes (e.g., 1000bp)
  Leave gaps between genes (e.g., 9000bp)



Recombination simulations
(msprime; Hudson model)
201 species, 5Mbp,
recombination rate =
substitution rate,
non-ultrametric,
rate heterogeneity

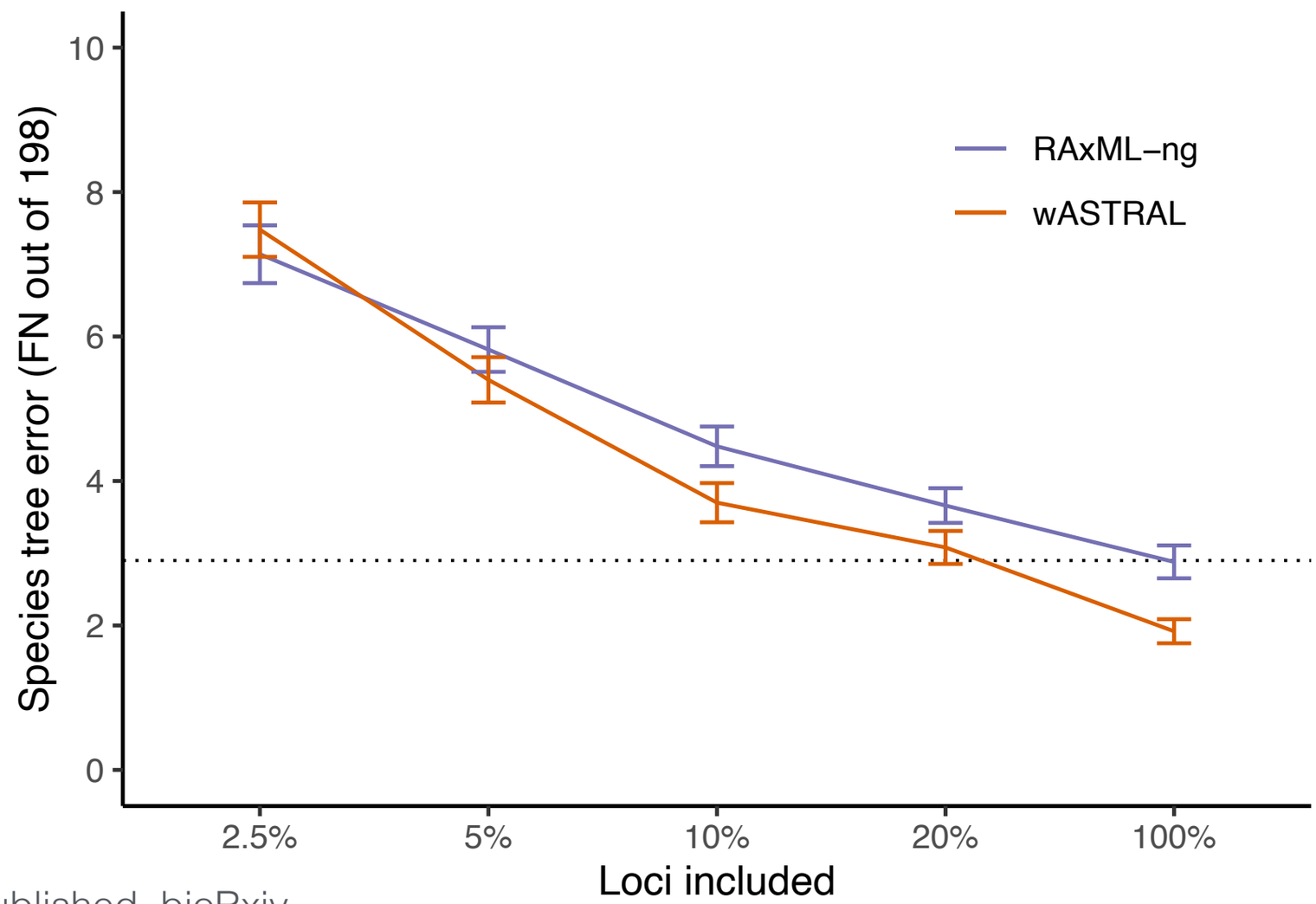


Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

Challenges with wASTRAL+ML gene trees:

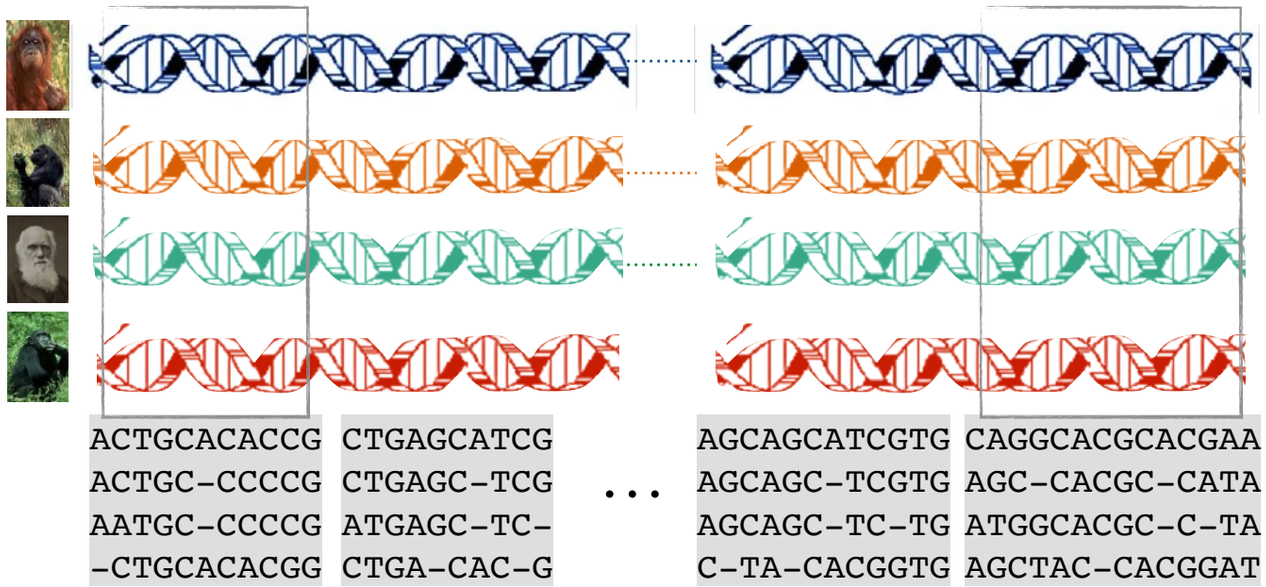
A. Long branch attraction can prevent good gene trees
B. Looses theoretical guarantees when used with all loci (though in practice it may not be an issue)

Recombination simulations (msprime; Hudson model) 201 species, 5Mbp, recombination rate = substitution rate, non-ultrametric, rate heterogeneity

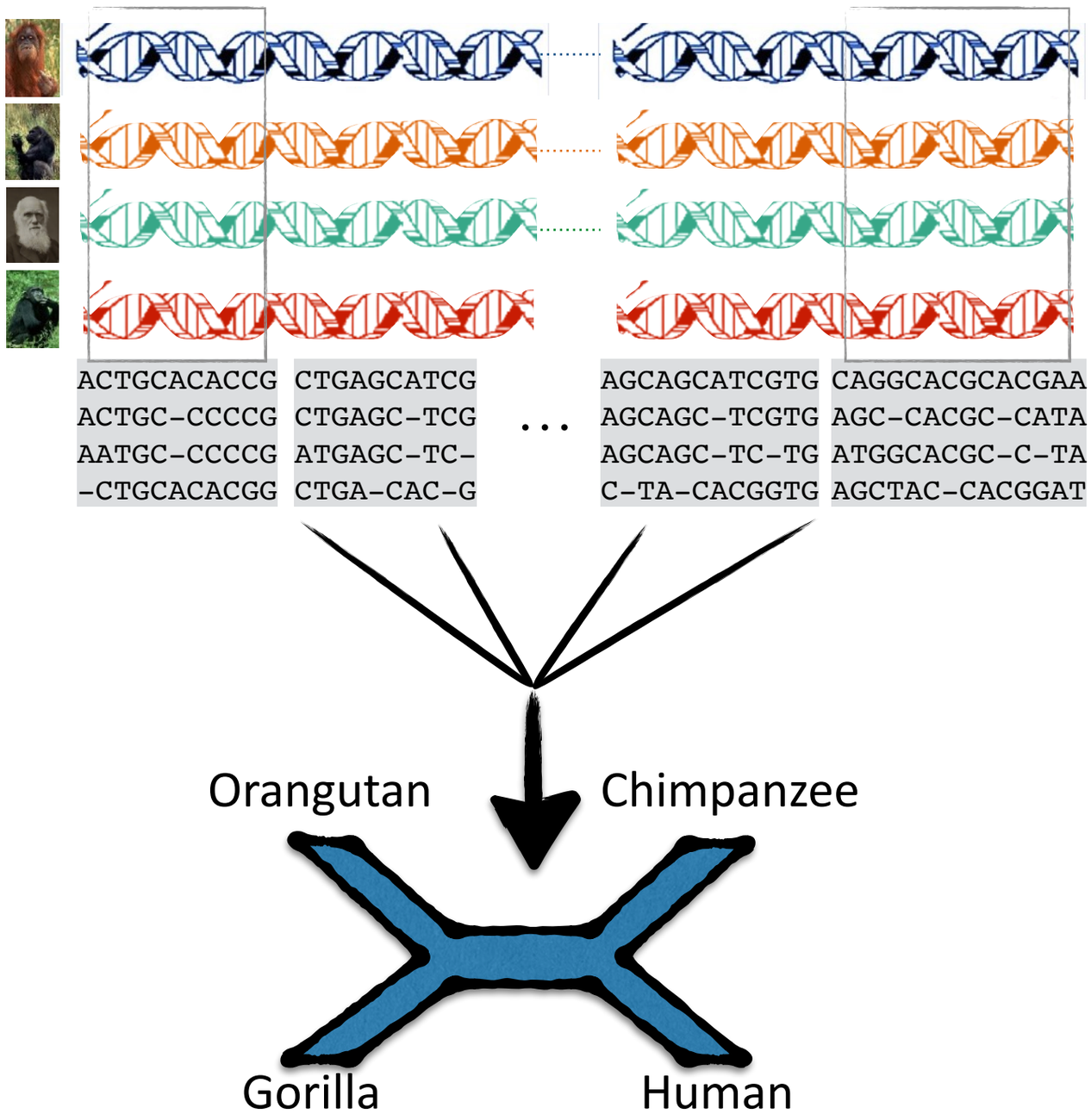


Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv, doi.org/10.1101/2023.10.04.560884

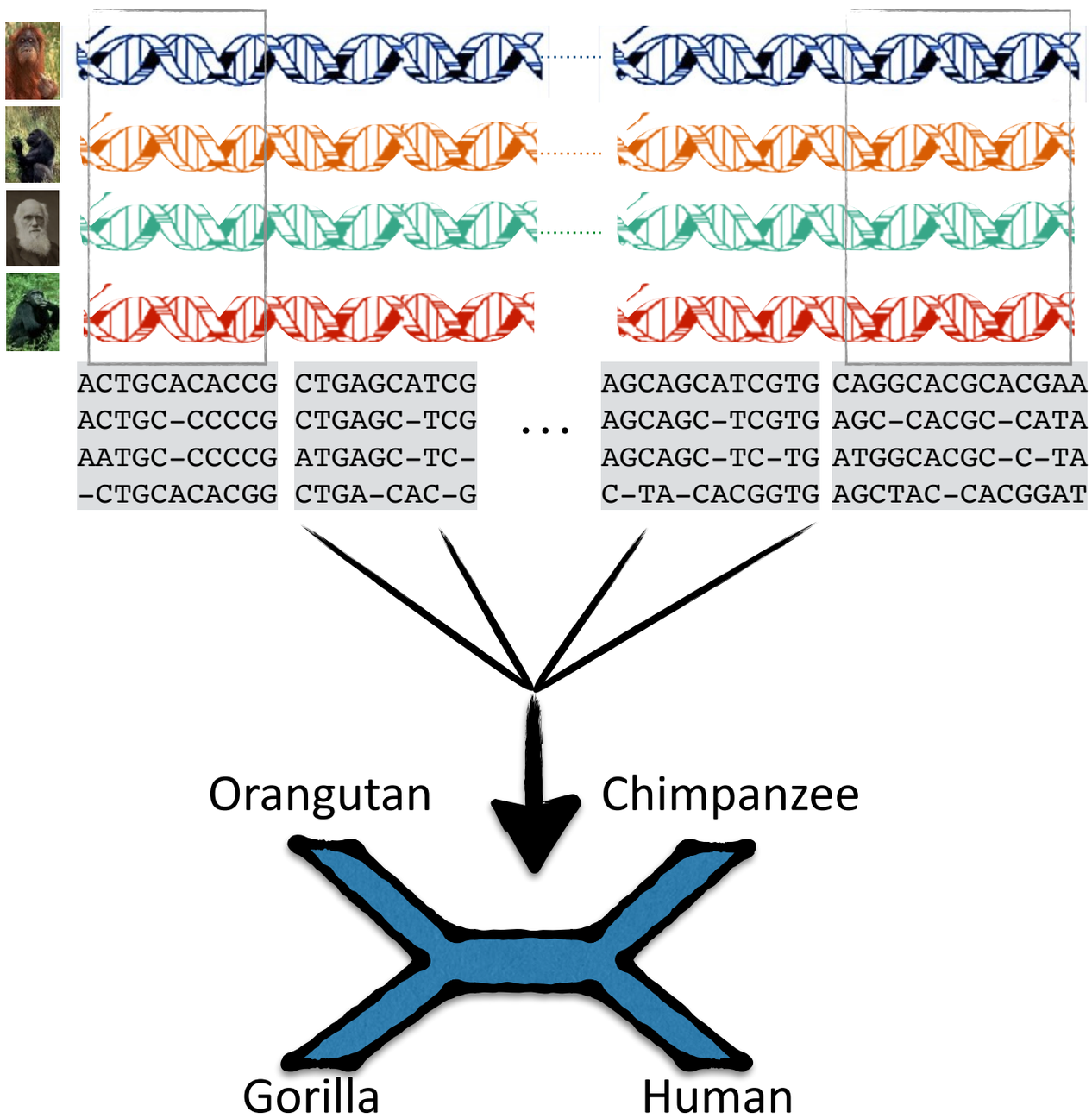
Site-based methods



Site-based methods

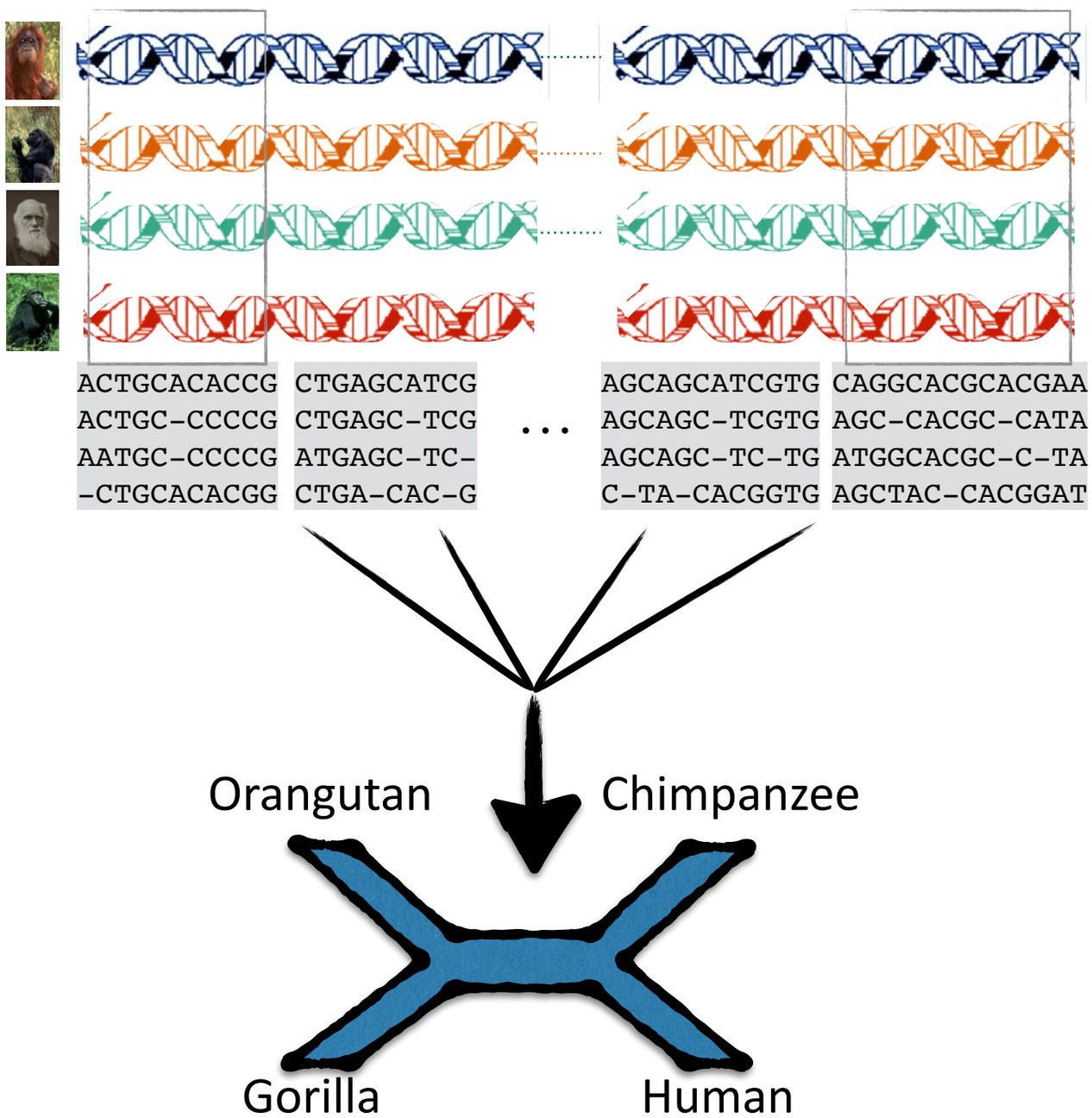


Site-based methods



— Bryant, Bouckaert, Felsenstein, Rosenberg, Roychoudhury, (2012).
— Chifman, L. Kubatko,(2014).
— Dasarathy, Nowak, Roch (2015).
— Vachaspati, Warnow. (2018).
— Allman, Long, Rhodes (2019).
— Stoltz, Baeumer, Bouckaert, Fox, Hiscott, Bryant (2021).

Site-based methods



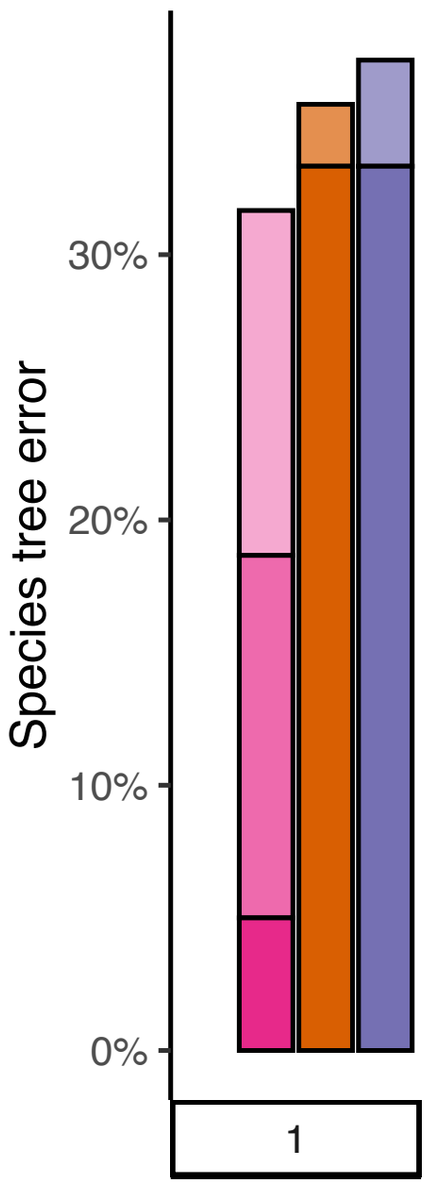
— Bryant, Bouckaert, Felsenstein, Rosenberg, Roychoudhury, (2012).
— Chifman, L. Kubatko,(2014).
— Dasarathy, Nowak, Roch (2015).
— Vachaspati, Warnow. (2018).
— Allman, Long, Rhodes (2019).
— Stoltz, Baeumer, Bouckaert, Fox, Hiscott, Bryant (2021).

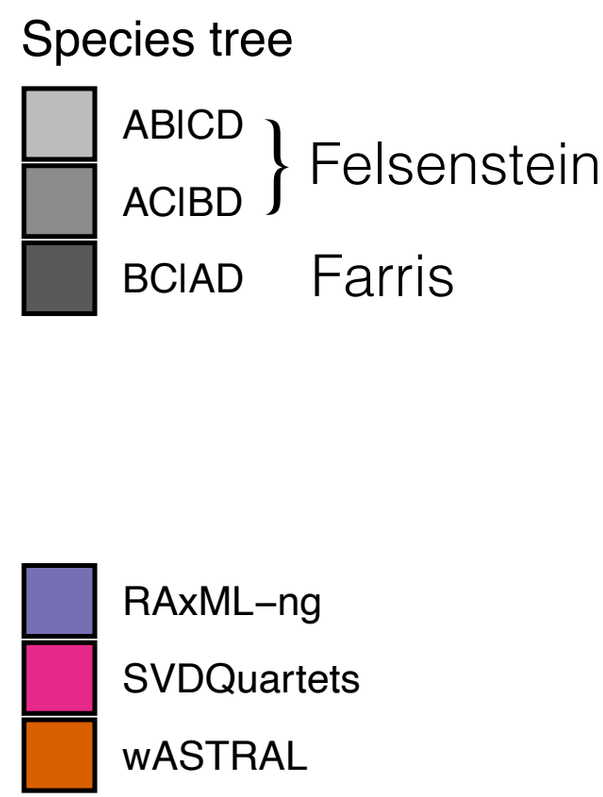
SVDQuartet:

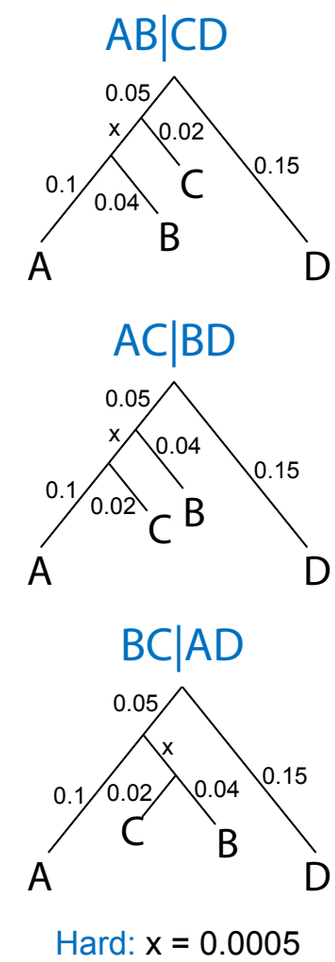
Quartet Inference from SNP Data Under the Coalescent Model FREE

Julia Chifman, Laura Kubatko ✉ Author Notes

Does a leading site-based method work better?





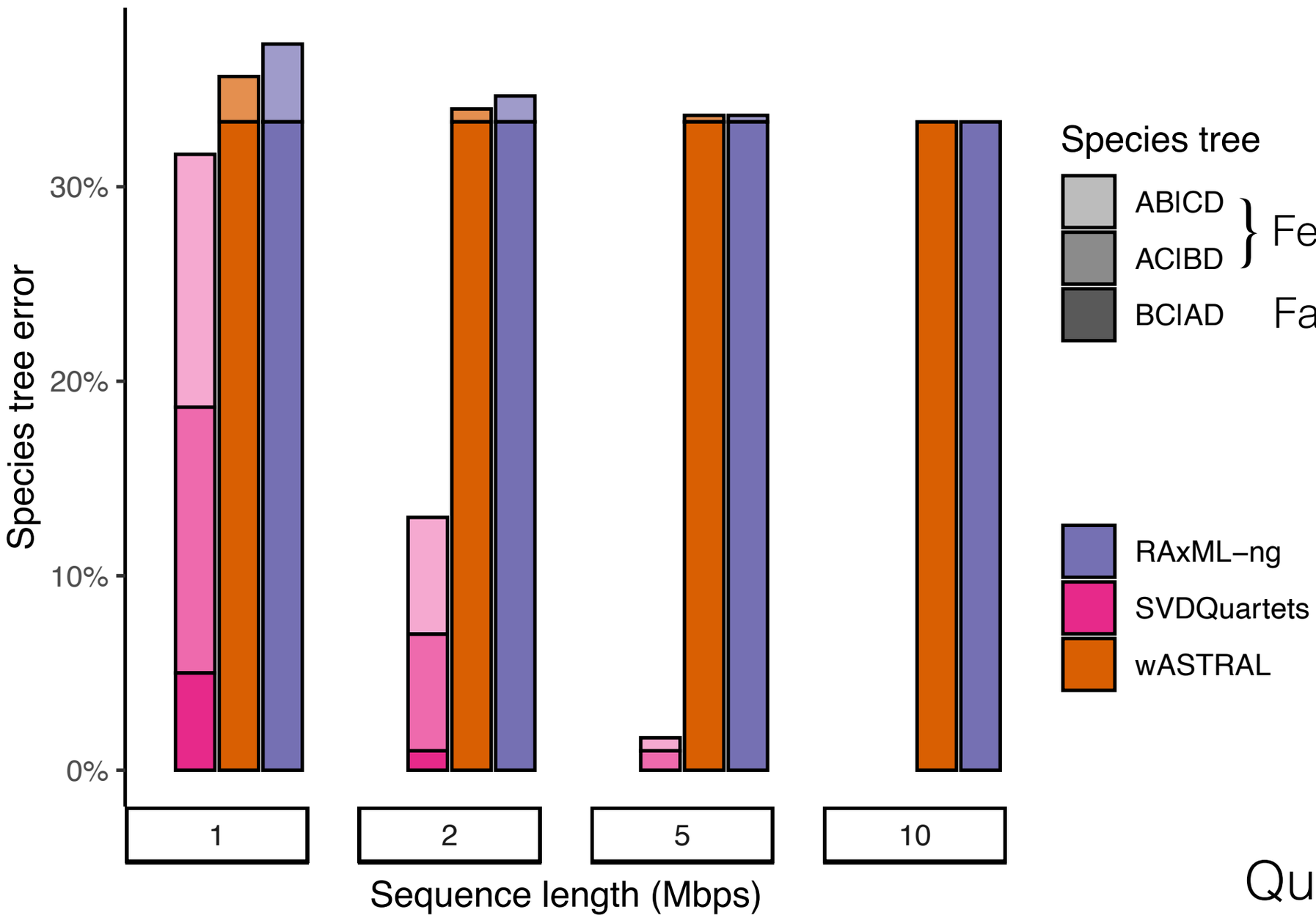


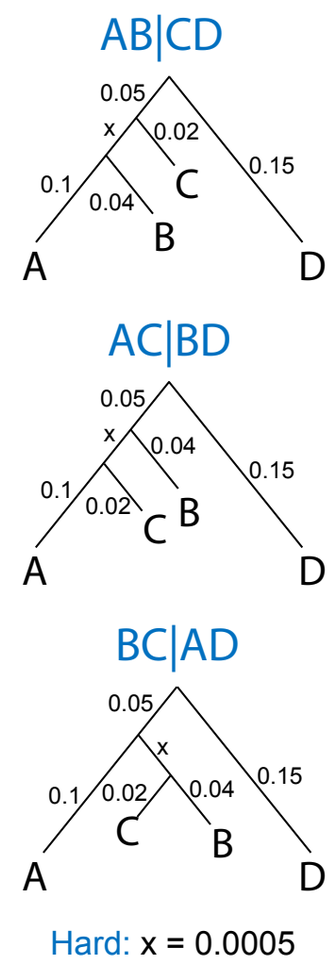
Quartet simulations

Recombination simulations
(msprime; Hudson model)
4 species, 10Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished

Does a leading site-based method work better? It can!



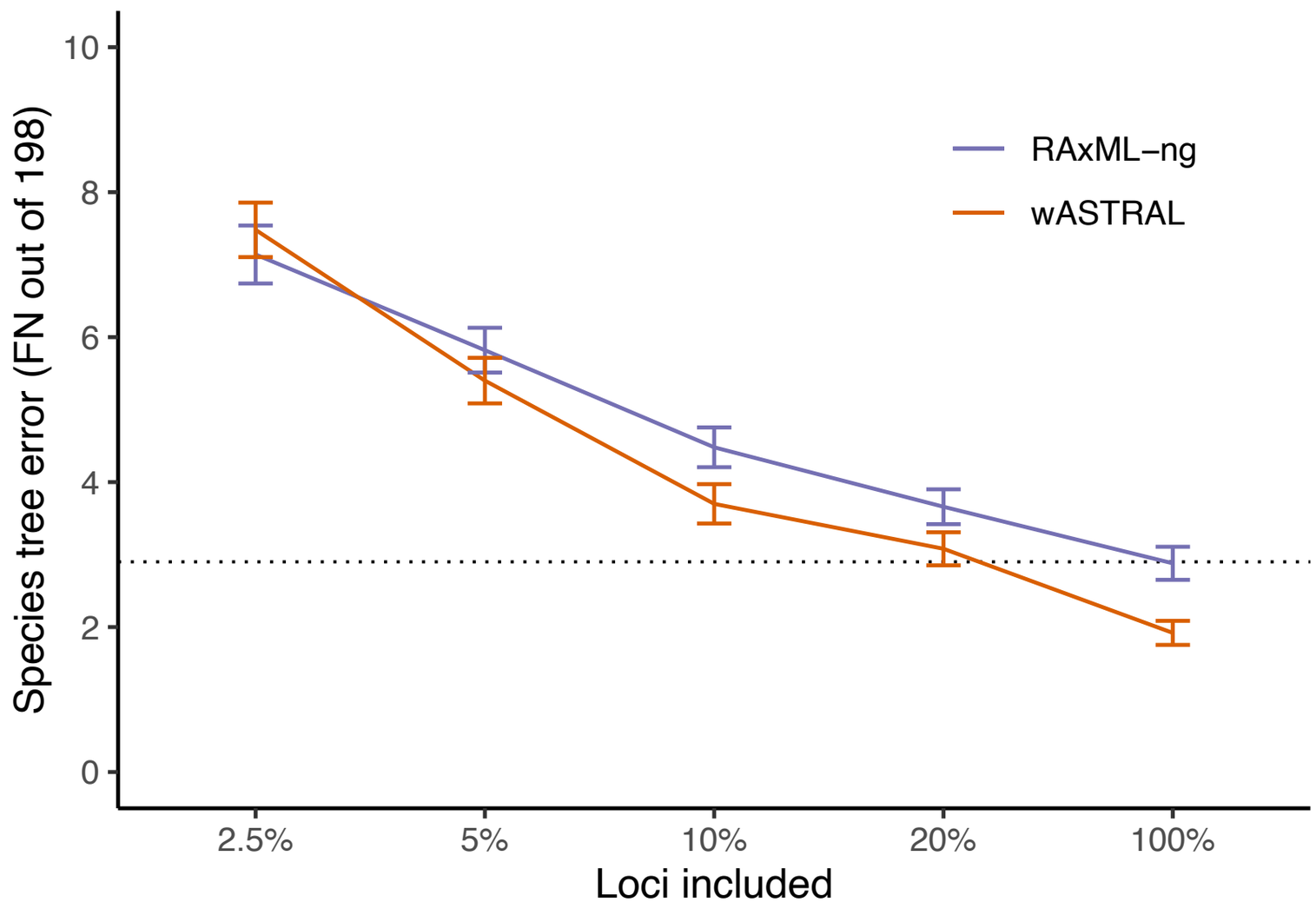


Quartet simulations

Recombination simulations
(msprime; Hudson model)
4 species, 10Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished

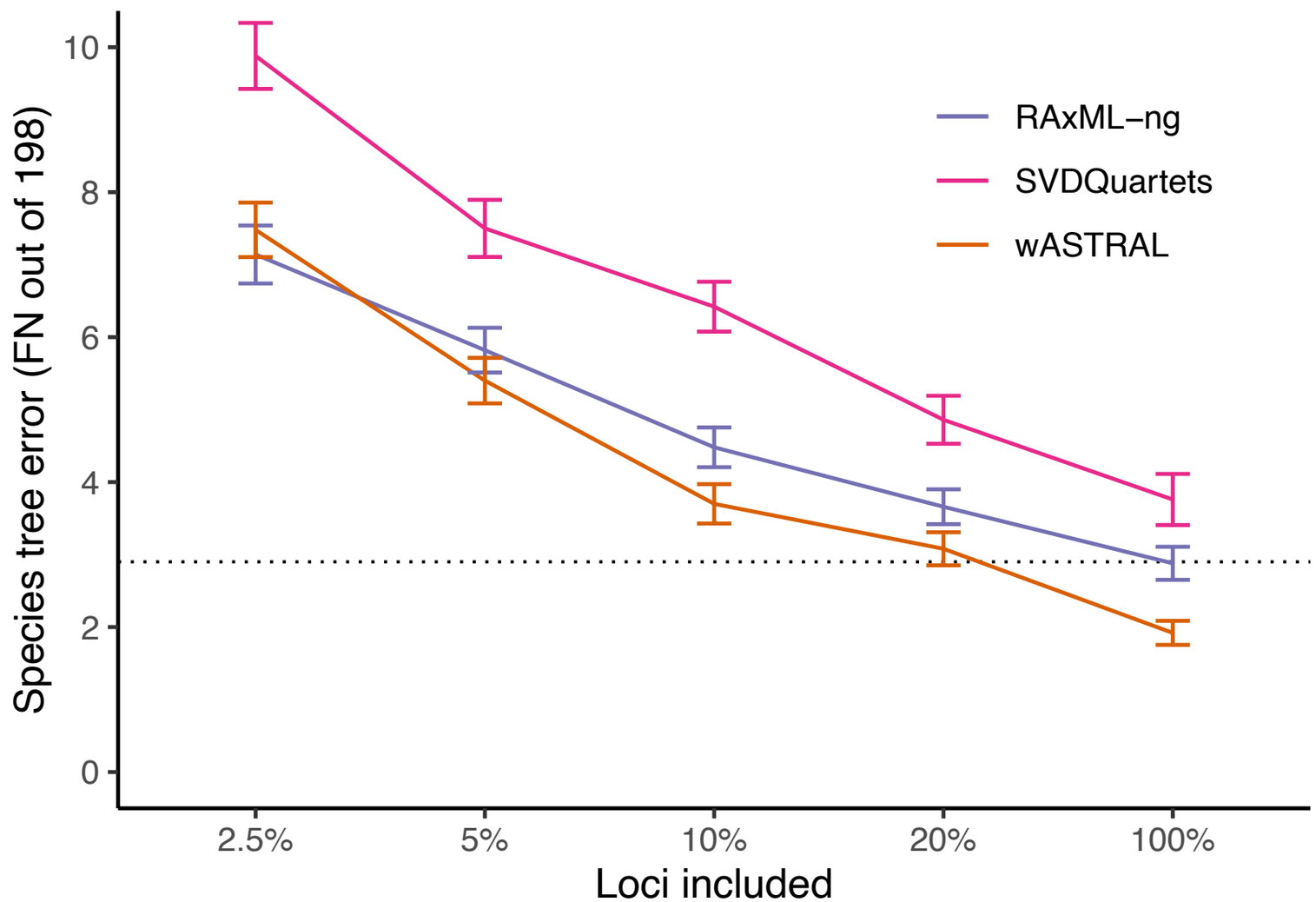
Does a leading site-based method work better?



Recombination simulations
(msprime; Hudson model)
201 species, 5Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

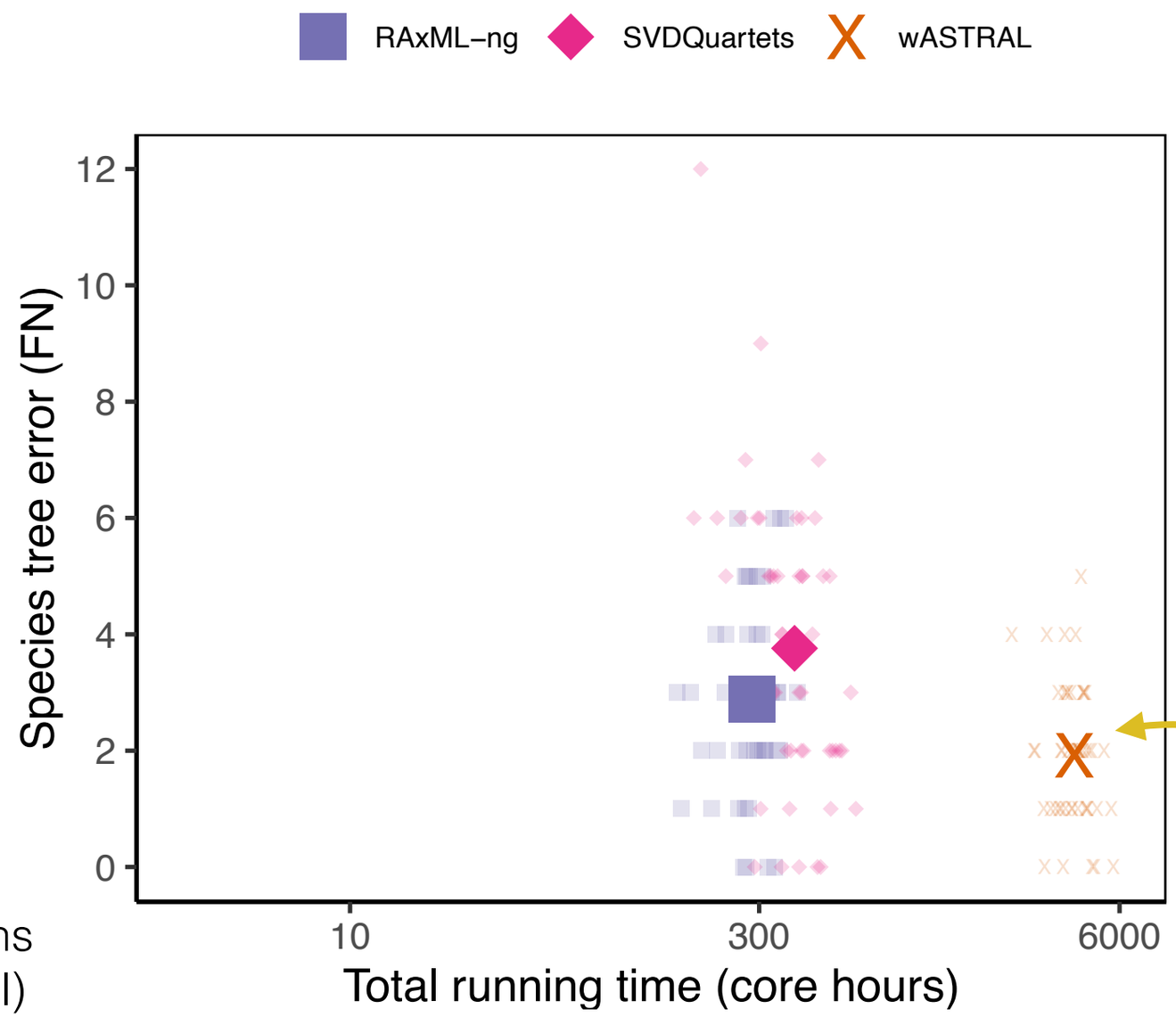
Does a leading site-based method work better? Not always!



Recombination simulations
(msprime; Hudson model)
201 species, 5Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

And how about running time?



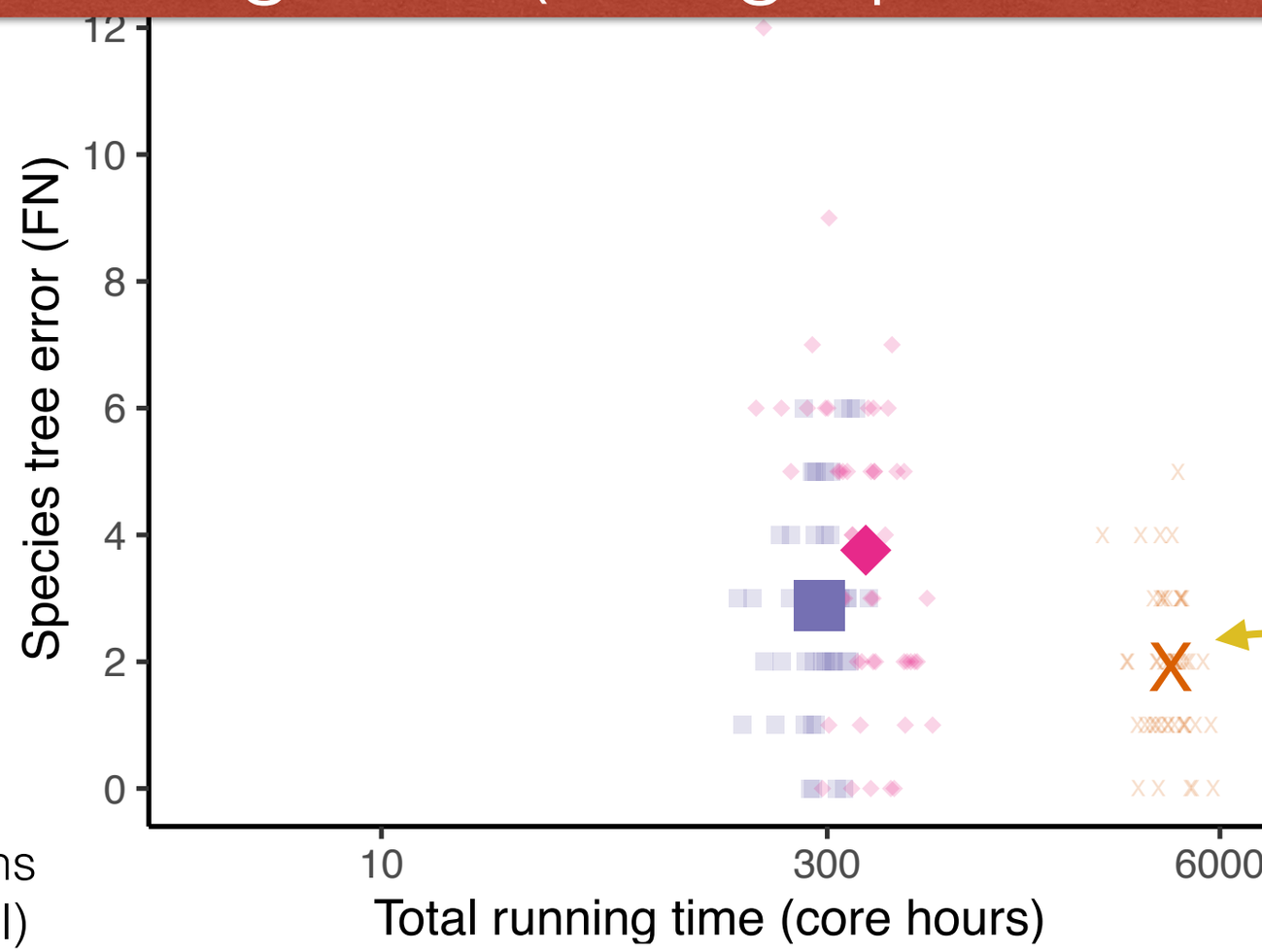
Recombination simulations
(msprime; Hudson model)
201 species, 5Mbp

"Only" 10,000 gene trees, 500bp each

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv, doi.org/10.1101/2023.10.04.560884

Challenges with wASTRAL+ML gene trees:

A. Long branch attraction can prevent good gene trees
B. Looses theoretical guarantees when used with all loci (though in practice it may not be an issue)
C. High total running time (though parallelizable)



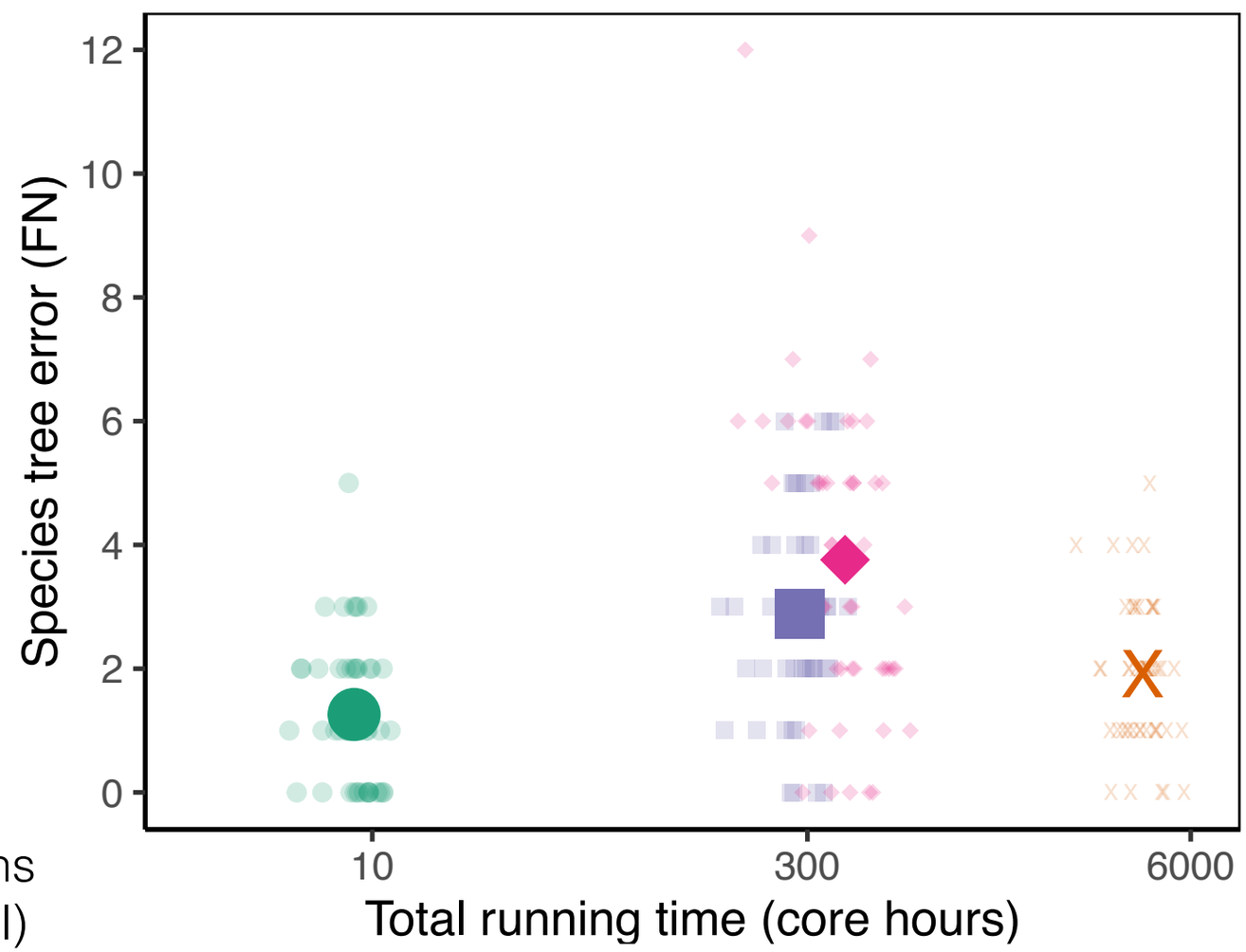
Recombination simulations
(msprime; Hudson model)
201 species, 5Mbp

"Only" 10,000 gene trees, 500bp each

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv, doi.org/10.1101/2023.10.04.560884

We can do better: **CASTER**

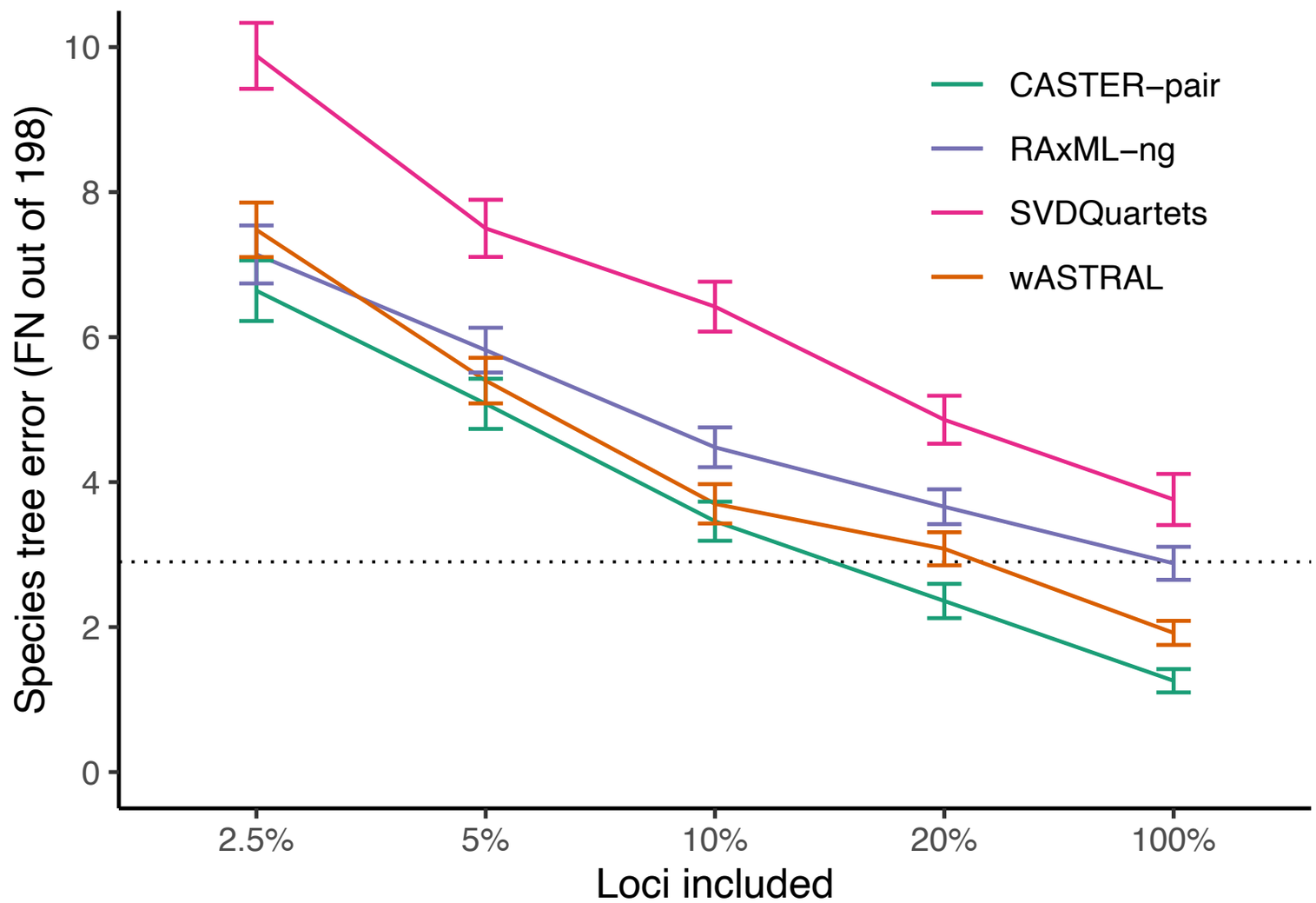
A. Site-based, thus rescues theoretical guarantees
B. Total running time a fraction of all other methods
C. Slightly more accurate



Recombination simulations
(msprime; Hudson model)
201 species, 5Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

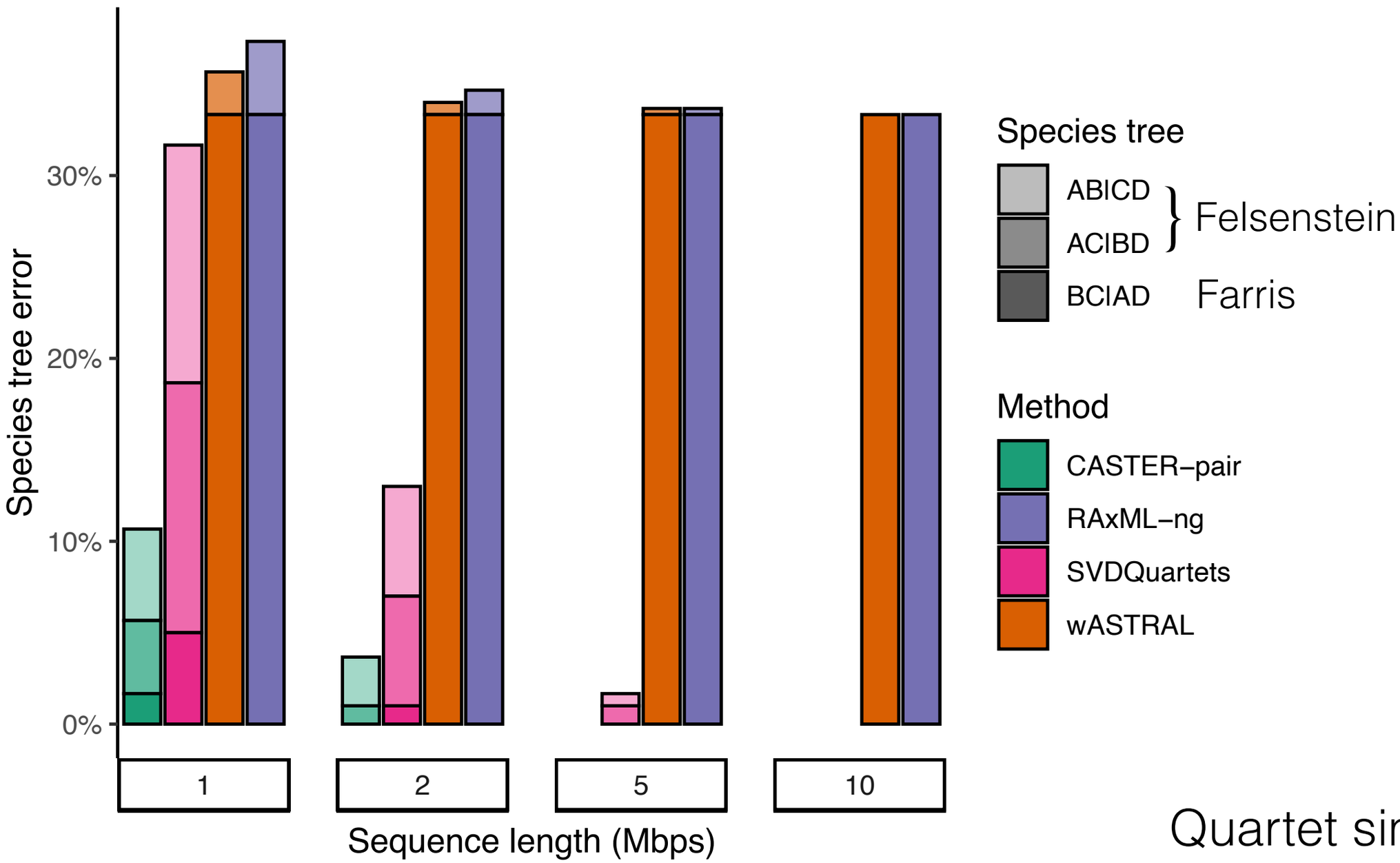
And works well with fewer loci too

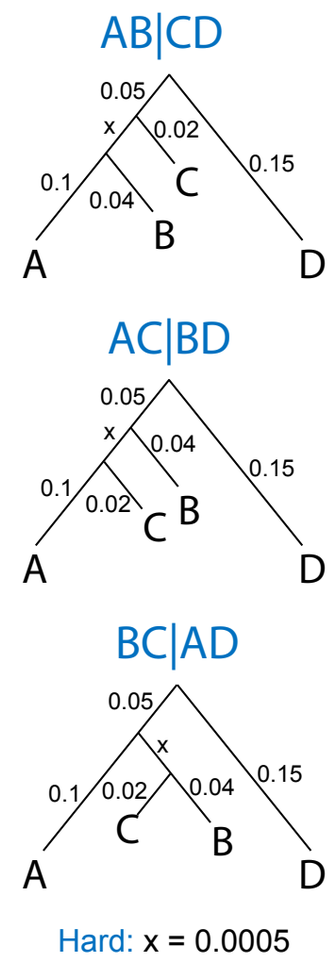


Recombination simulations
(msprime; Hudson model)
200 species, 5Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

CASTER works with non-ultra metric trees too!





Quartet simulations

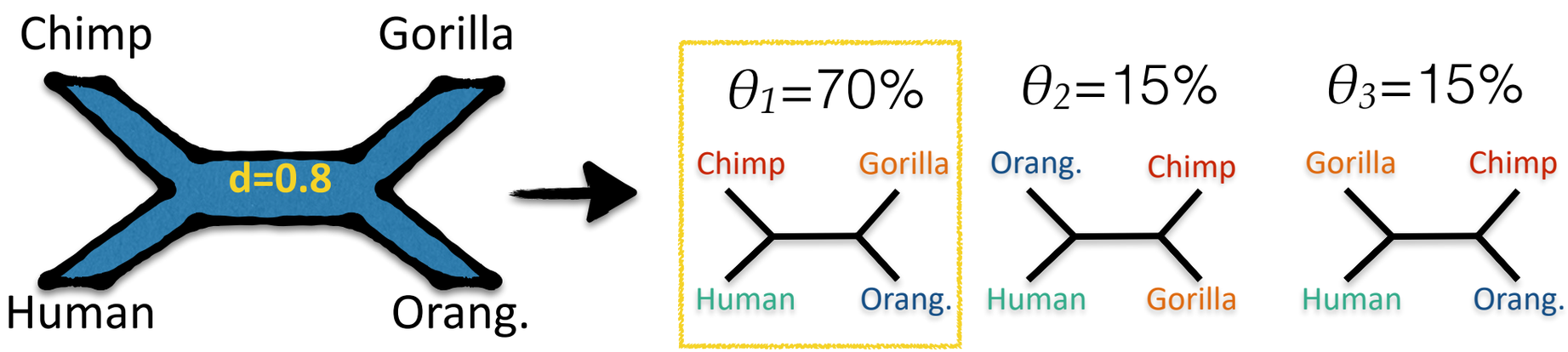
Recombination simulations
(msprime; Hudson model)
4 species, 10Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished

How does CASTER work?

Unrooted quartets under MSC model

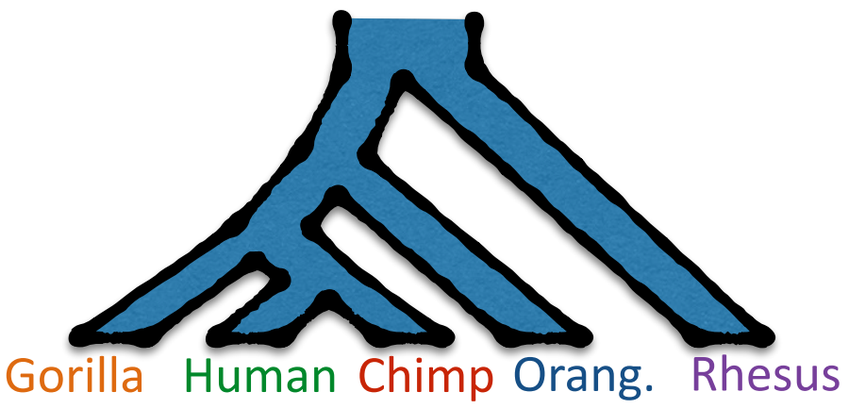
For a quartet (4 species), the most probable unrooted quartet tree among the gene trees is the unrooted species tree topology
(Allman, Degnan, Rhodes, J. Theo. Bio., 2011)



The most frequent gene tree
=
The most likely species tree

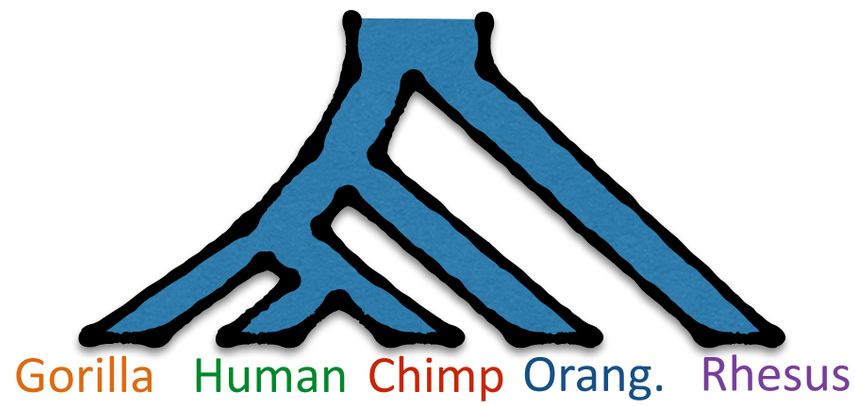
More than 4 species

For 5 or more species, the unrooted species tree topology can be different from the most probable gene tree ("anomaly zone")
(Degnan and Rosenberg, 2006) (Degnan, 2013) (Rosenberg, 2013)



More than 4 species

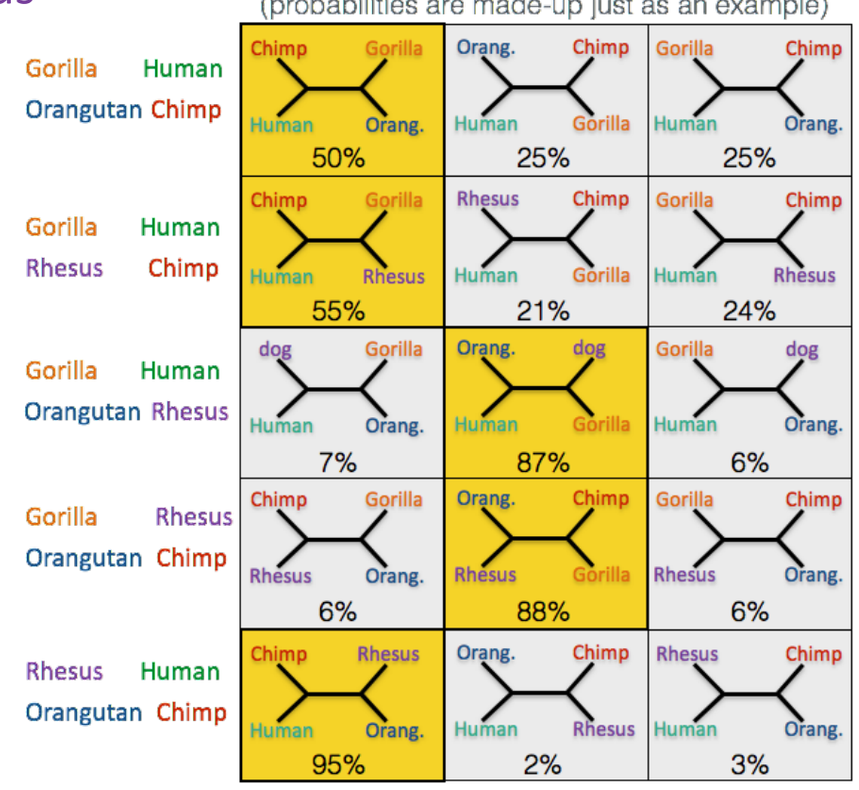
For 5 or more species, the unrooted species tree topology can be different from the most probable gene tree ("anomaly zone")
(Degnan and Rosenberg, 2006) (Degnan, 2013) (Rosenberg, 2013)



1. Break gene trees into $\binom{n}{4}$ quartets of species
2. Find the dominant tree for all quartets of taxa
3. Combine quartet trees

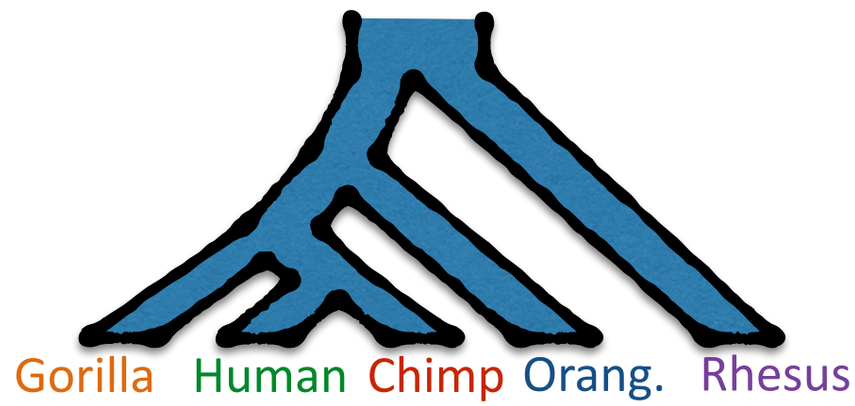
   (e.g., SVDQuartet, BUCKy-p (Larget, et al., 2010))

(probabilities are made-up just as an example)



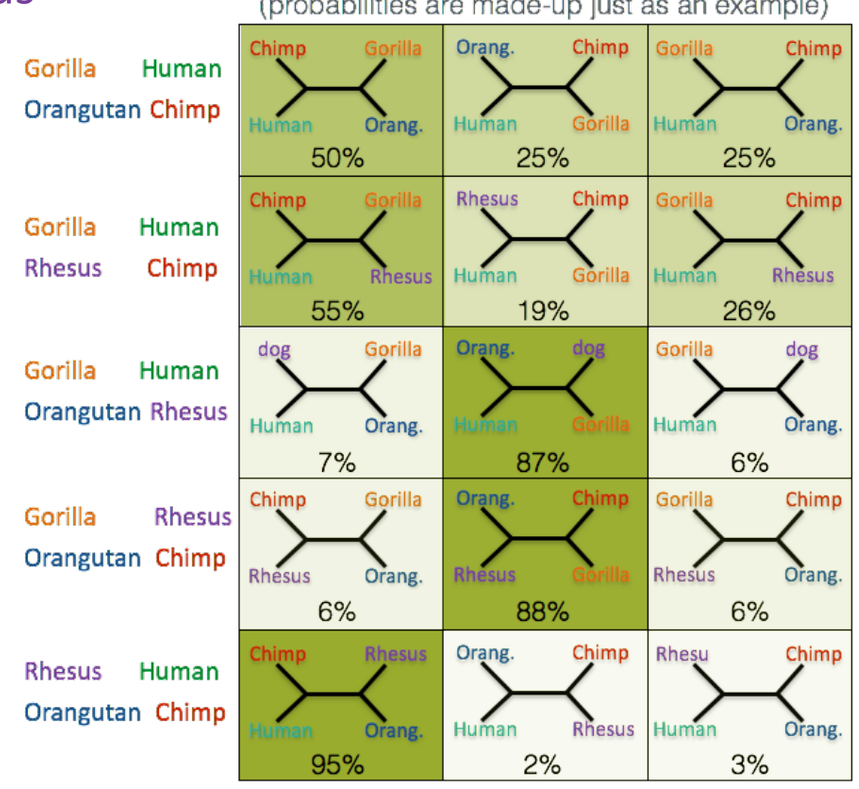
More than 4 species

For 5 or more species, the unrooted species tree topology can be different from the most probable gene tree ("anomaly zone")
(Degnan and Rosenberg, 2006) (Degnan, 2013) (Rosenberg, 2013)



Alternative:
Give each of $3\binom{n}{4}$ quartet topologies a score and find the tree with the maximum total score

(probabilities are made-up just as an example)



ASTRAL: Maximum Quartet Support

$$S(T) = \sum_{i=1}^{k} \sum_{j=1}^{\binom{n}{4}} \mathbf{I}(T \,|\, q_j = t_i \,|\, q_j)$$

a gene tree

a quartet of taxa

$$= \sum_{i=1}^{k} |\, Q(T) \cap Q(t_i) \,|$$

the set of $\binom{n}{4}$ quartet trees induced by T

Find the species tree with <u>the maximum</u> number of induced <u>quartet trees</u> <u>shared</u> with the input gene trees

$$\mathrm{argmax}_T \; S(T)$$

- **Guarantee**: Statistically consistent under the MSC

CASTER: a site-based method inspired by ASTRAL

- Each site with site pattern s votes for (or against) each quartet topology q with weight $w(q, s)$

$$\text{argmax}_T \sum_{j=1}^{\binom{n}{4}} \sum_{i=1}^{L} w(T \mid q_j, s_i \mid q_j)$$

Each site

CASTER: a site-based method inspired by ASTRAL

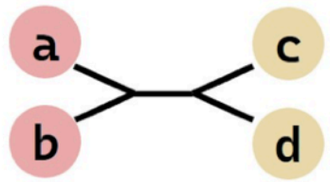
- Each site with site pattern s votes for (or against) each quartet topology q with weight $w(q, s)$

$$\text{argmax}_T \sum_{j=1}^{\binom{n}{4}} \sum_{i=1}^{L} w(T \mid q_j, s_i \mid q_j)$$

Each site

- What voting scheme $w(q, s)$ leads to a statistically consistent estimator?

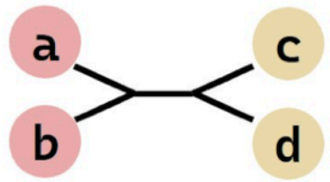
Jukes-Cantor



JC69

a	X
b	X
c	Y
d	Y
W	+1

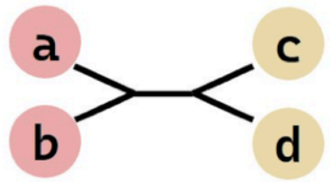
Jukes-Cantor



JC69

a	X	X	Y	Others
b	X	X	Z	
c	Y	Y	X	
d	Y	Z	X	
W	+1	?	?	0

Jukes-Cantor



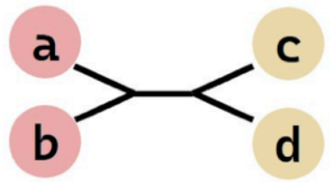
JC69

a	X	X	Y	Others
b	X	X	Z	
c	Y	Y	X	
d	Y	Z	X	
W	+1	?	?	0

Does it work for a gene tree?

$$\Delta = \mathbb{E}[w(ab\,|\,cd)] - \mathbb{E}[w(ac\,|\,bd)] > 0$$

Jukes-Cantor

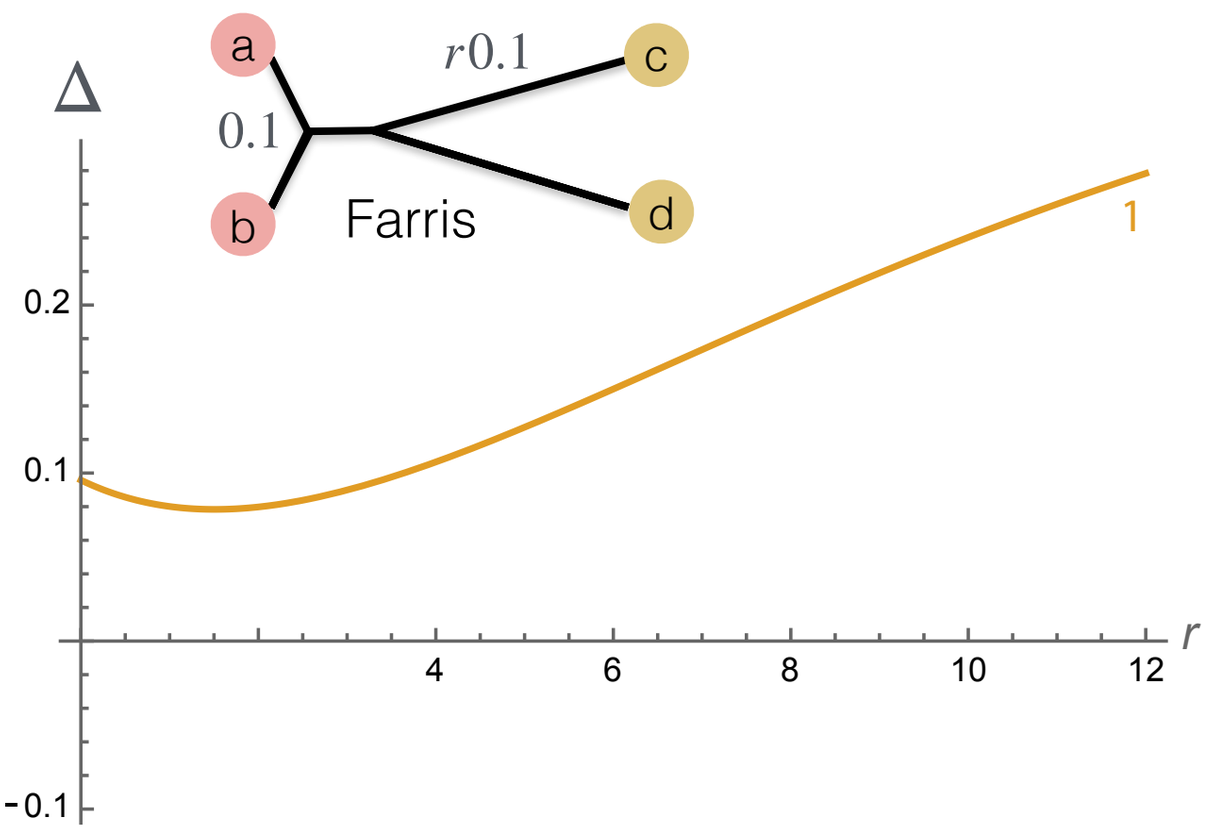


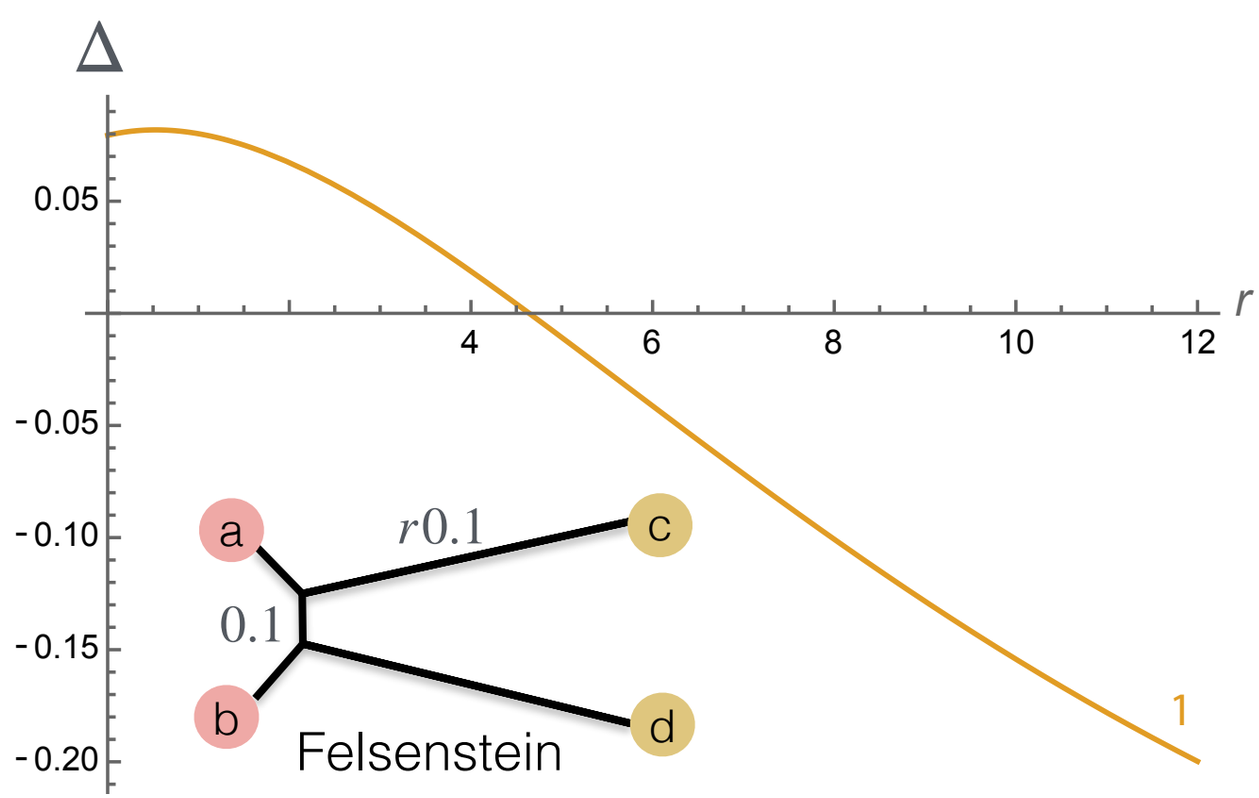
JC69

a	X	X	Y	Others
b	X	X	Z	
c	Y	Y	X	
d	Y	Z	X	
W	+1	?	?	0

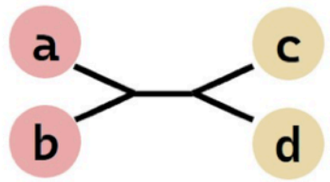
Does it work for a gene tree?

$$\Delta = \mathbb{E}[w(ab\,|\,cd)] - \mathbb{E}[w(ac\,|\,bd)] > 0$$





Jukes-Cantor

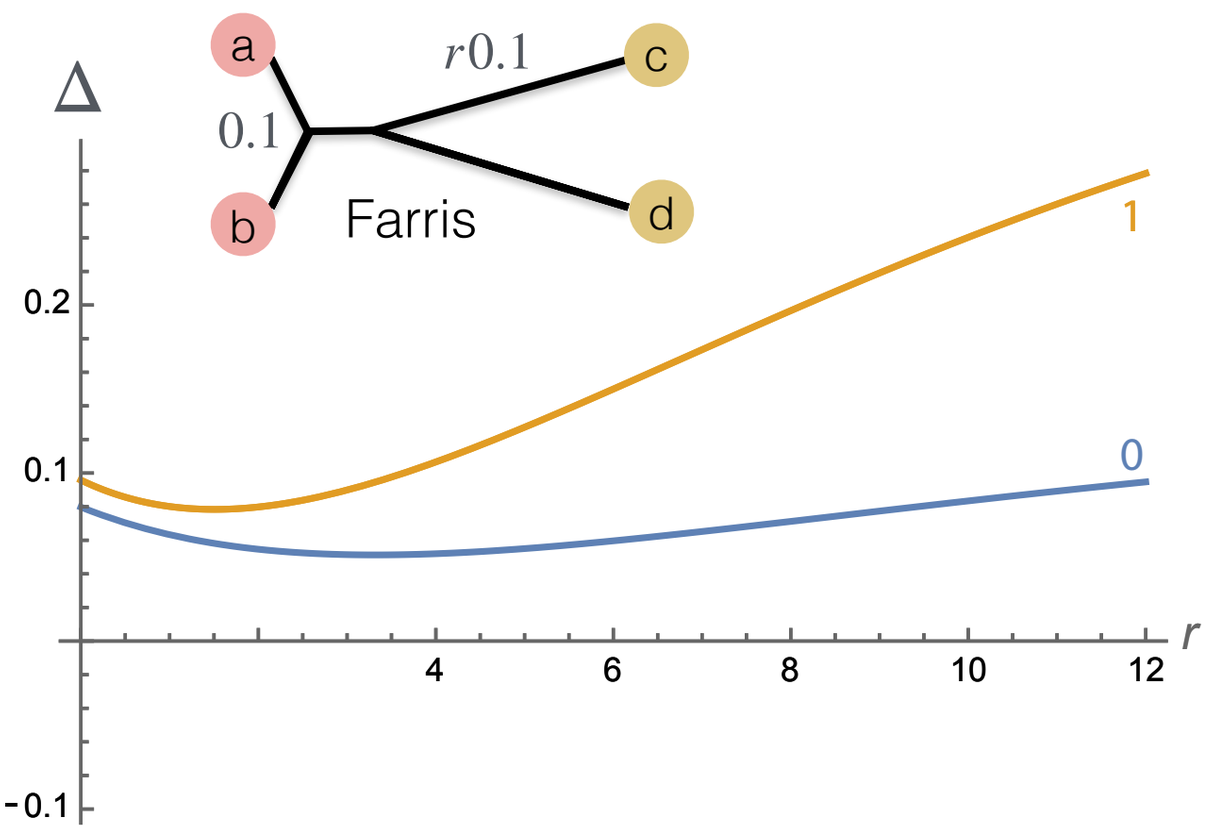


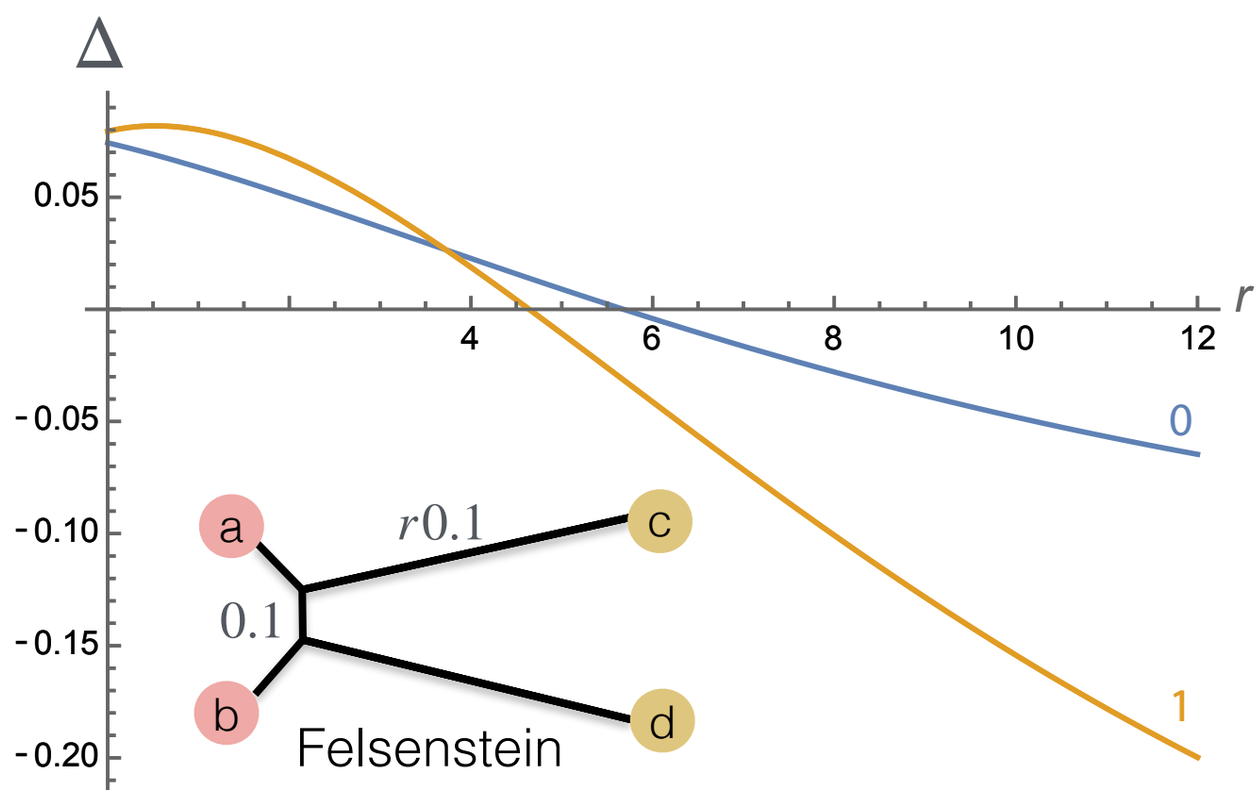
JC69

a	X	X	Y	Others
b	X	X	Z	
c	Y	Y	X	
d	Y	Z	X	
W	+1	?	?	0

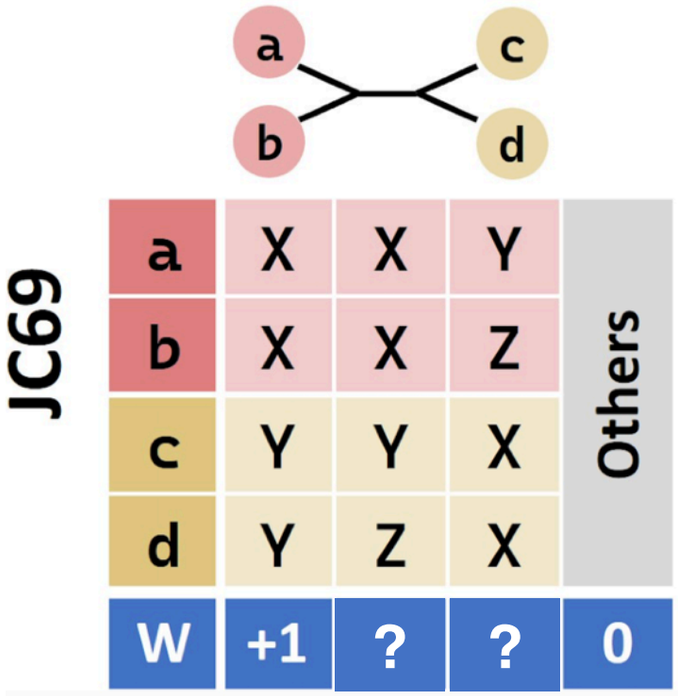
Does it work for a gene tree?

$$\Delta = \mathbb{E}[w(ab\,|\,cd)] - \mathbb{E}[w(ac\,|\,bd)] > 0$$



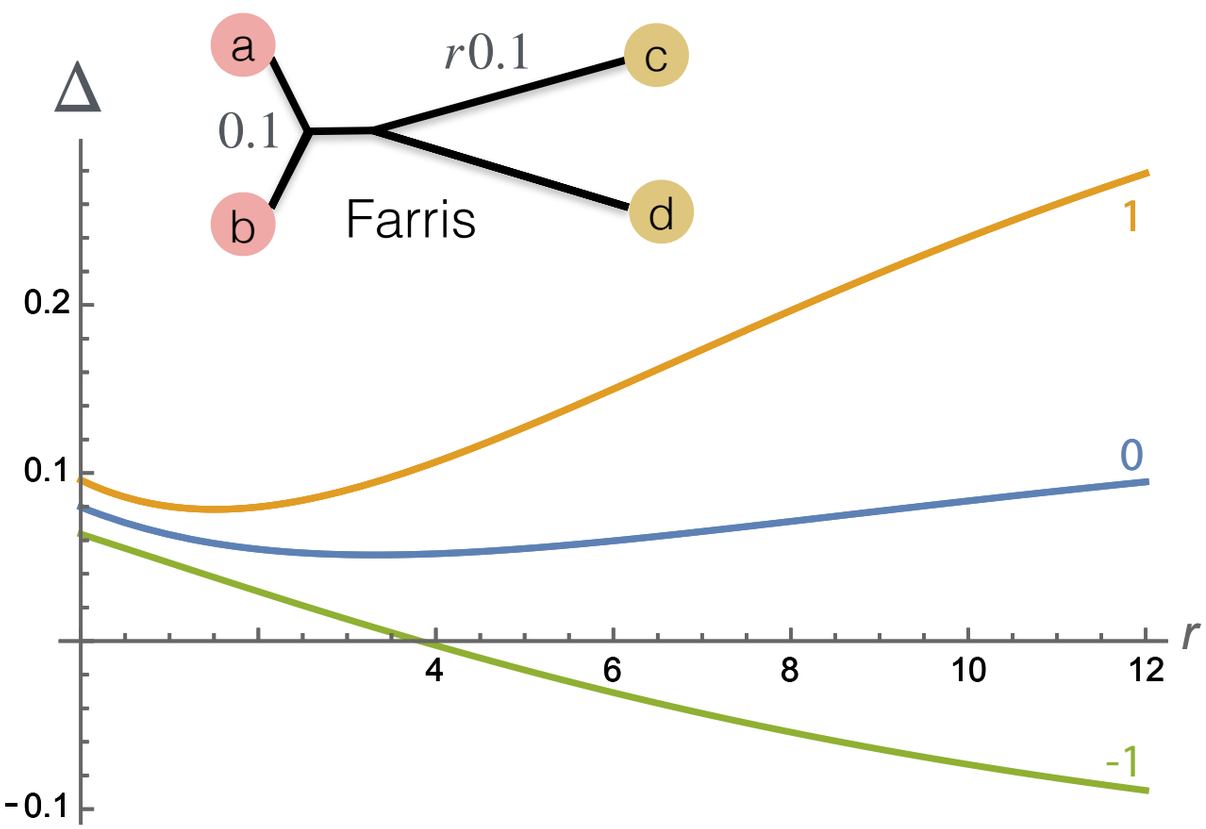


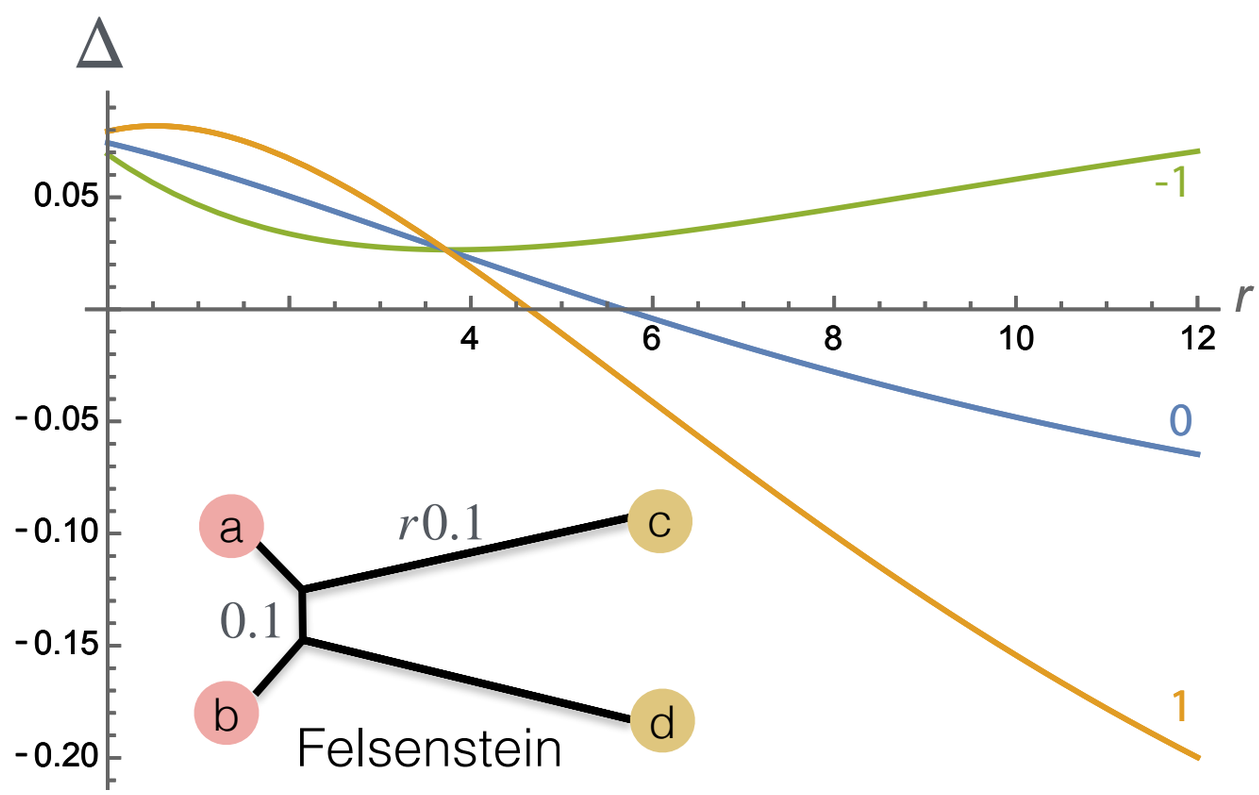
Jukes-Cantor



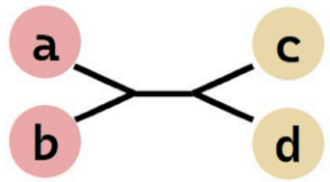
Does it work for a gene tree?

$$\Delta = \mathbb{E}[w(ab\,|\,cd)] - \mathbb{E}[w(ac\,|\,bd)] > 0$$





Jukes-Cantor

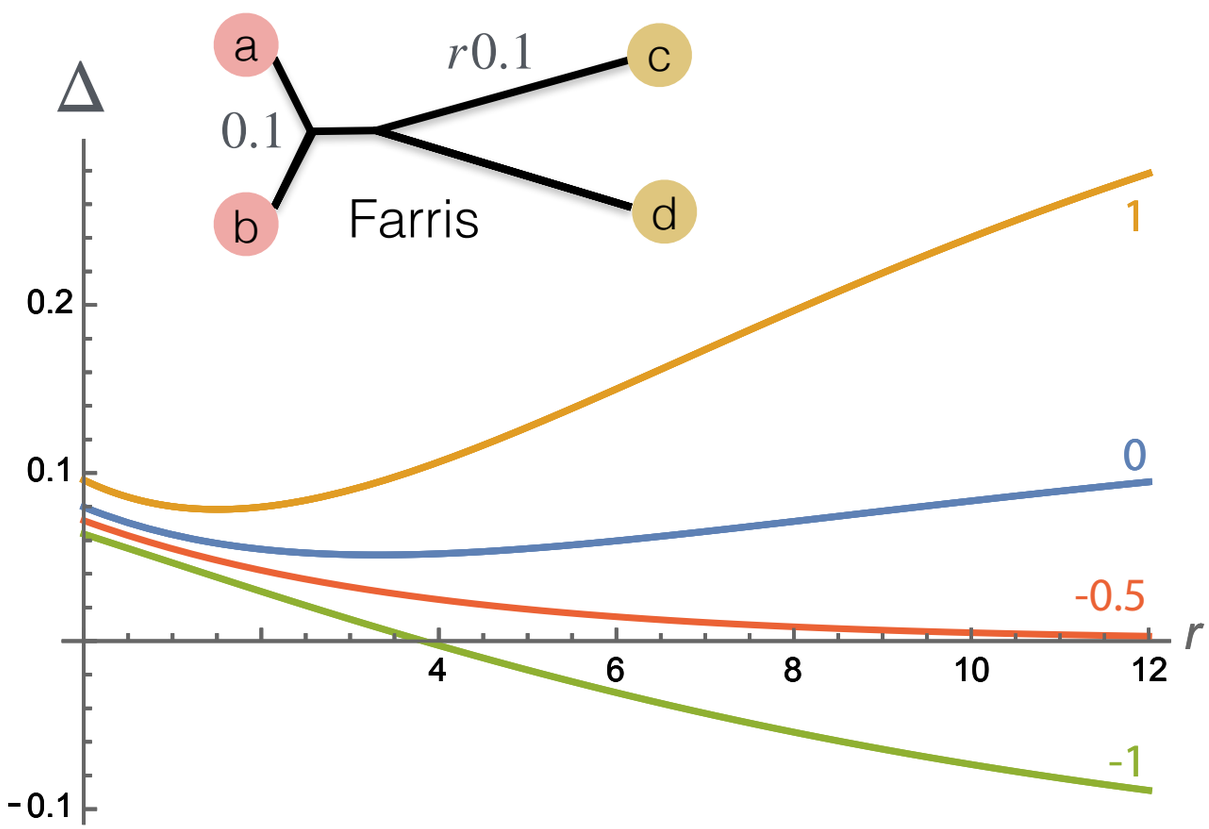


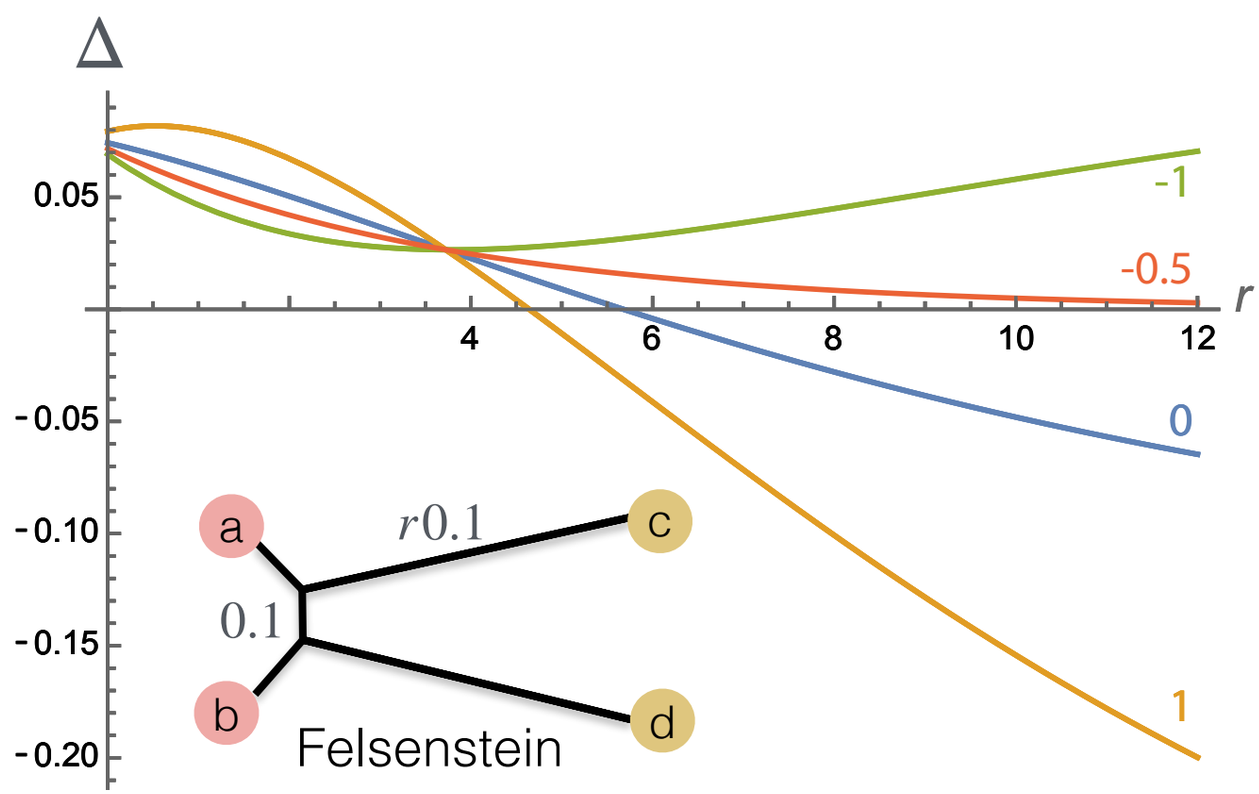
JC69

a	X	X	Y	Others
b	X	X	Z	
c	Y	Y	X	
d	Y	Z	X	
W	+1	?	?	0

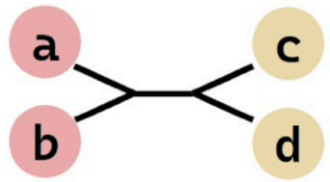
Does it work for a gene tree?

$$\Delta = \mathbb{E}[w(ab\,|\,cd)] - \mathbb{E}[w(ac\,|\,bd)] > 0$$





Minor miracle

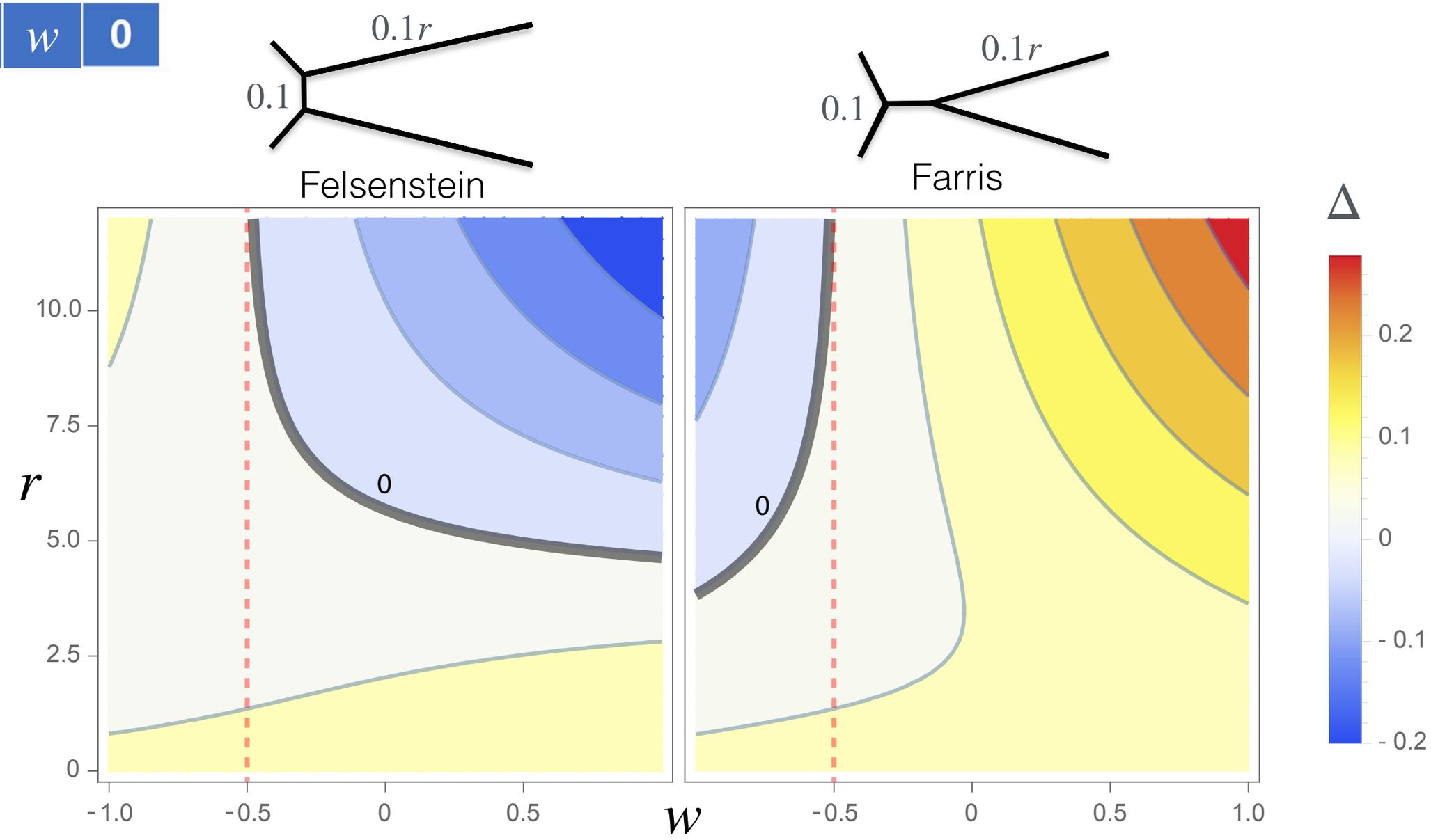


JC69

a	X	X	Y	Others
b	X	X	Z	
c	Y	Y	X	
d	Y	Z	X	
W	+1	w	w	0

$$w = -1/2:$$

$$\Delta = \mathbb{E}[w(ab\,|\,cd)] - \mathbb{E}[w(ac\,|\,bd)] > 0$$



Works for every gene tree quartet!

For a quartet gene tree G_i with topology $xy\,|\,zw$, terminal branch lengths l_x, l_y, l_z, l_w and internal branch length t in substitution units, for some $\alpha > 0,\ \beta > 0,\ \gamma > 0$:

$$\mathbb{E}\left[w_i(xy\,|\,zw)\right] - \gamma e^{-\alpha(l_x+l_y+l_w+l_z)}\left(1 - e^{-\beta t}\right) = \mathbb{E}\left[w_i(xz\,|\,yw)\right]$$

Works for every gene tree quartet!

For a quartet gene tree G_i with topology $xy\,|\,zw$, terminal branch lengths l_x, l_y, l_z, l_w and internal branch length t in substitution units, for some $\alpha > 0,\ \beta > 0,\ \gamma > 0$:

$$\mathbb{E}\left[w_i(xy\,|\,zw)\right] - \gamma e^{-\alpha(l_x+l_y+l_w+l_z)}\left(1 - e^{-\beta t}\right) = \mathbb{E}\left[w_i(xz\,|\,yw)\right]$$

Works for a species tree quartet!

For the true species tree S of four leaves with topology $ab\,|\,cd$, for each gene tree i:

$$\mathbb{E}\left[w_i(ab\,|\,cd)\right] > \mathbb{E}\left[w_i(ac\,|\,bd)\right] = \mathbb{E}\left[w_i(ac\,|\,bd)\right]$$

Works for every gene tree quartet!

For a quartet gene tree G_i with topology $xy\,|\,zw$, terminal branch lengths l_x, l_y, l_z, l_w and internal branch length t in substitution units, for some $\alpha > 0$, $\beta > 0$, $\gamma > 0$:

$$\mathbb{E}\left[w_i(xy\,|\,zw)\right] - \gamma e^{-\alpha(l_x+l_y+l_w+l_z)}\left(1 - e^{-\beta t}\right) = \mathbb{E}\left[w_i(xz\,|\,yw)\right]$$

Works for a species tree quartet!

For the true species tree S of four leaves with topology $ab\,|\,cd$, for each gene tree i:

$$\mathbb{E}\left[w_i(ab\,|\,cd)\right] > \mathbb{E}\left[w_i(ac\,|\,bd)\right] = \mathbb{E}\left[w_i(ac\,|\,bd)\right]$$

Hence:

CASTER is statistically consistent for JC69!

Extends to F84 (CASTER-site)

a	R_1	Y_1	R_1	Y_1	R_1	Y_1	R_1	Y_1	Others
b	R_1	Y_1	R_2	Y_1	R_1	Y_2	R_2	Y_2	
c	Y_1	R_1	Y_1	R_1	Y_1	R_1	Y_1	R_1	
d	Y_1	R_1	Y_1	R_2	Y_2	R_1	Y_2	R_2	
Weight	$4\pi_A\pi_G$ $\pi_C\pi_T$		$-2\pi_A\pi_G$ $(\pi_C^2+\pi_T^2)$		$-2\pi_C\pi_T$ $(\pi_A^2+\pi_G^2)$		$(\pi_A^2+\pi_G^2)$ $(\pi_C^2+\pi_T^2)$		0

Beyond F84? CASTER-Pair

- We could not extend these scores beyond F84! ¯_(ツ)_/¯

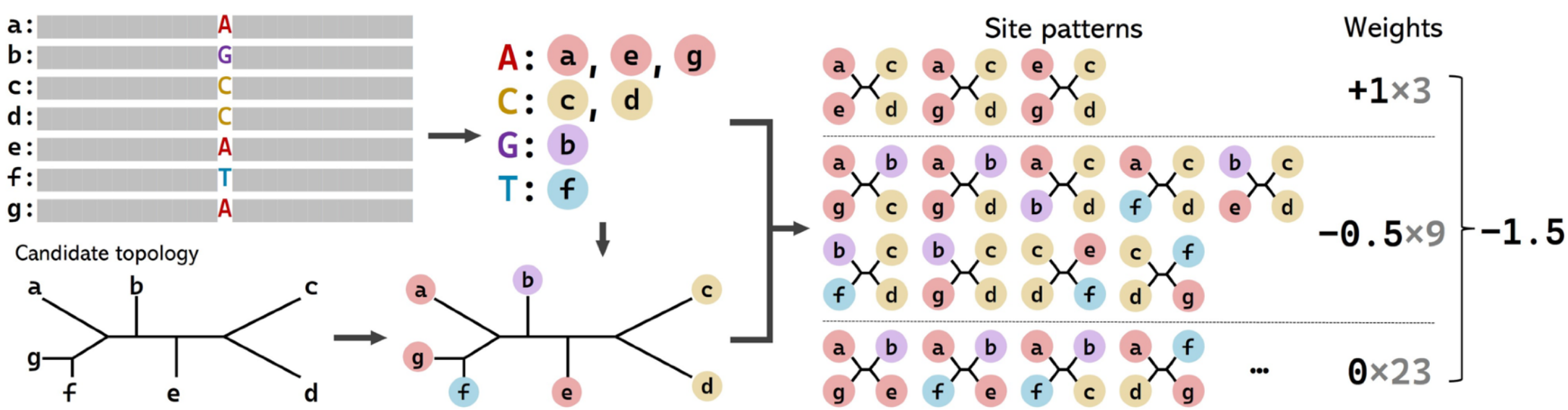
Beyond F84? CASTER-Pair

- We could not extend these scores beyond F84! ¯_(ツ)_/¯
- What if we use a pair of sites?
  - We can use RY coding
  - Consistent under Markovian Reducible (MR) models
    — MR1: a 7-parameter submodule of GTR (2 fewer parameters, a generalization of TN93) that remain a CTMC after RY recoding

a	RN	RN	YN	YN	NR	NR	NY	NY	RN	YN	NN	NN
b	YN	YN	RN	RN	NY	NY	NR	NR	YN	RN	NN	NN
c	NR	NY	NR	NY	RN	YN	RN	YN	NN	NN	RN	YN
d	NY	NR	NY	NR	YN	RN	YN	RN	NN	NN	YN	RN
W	$+1$								$-4\pi_R\pi_Y$			

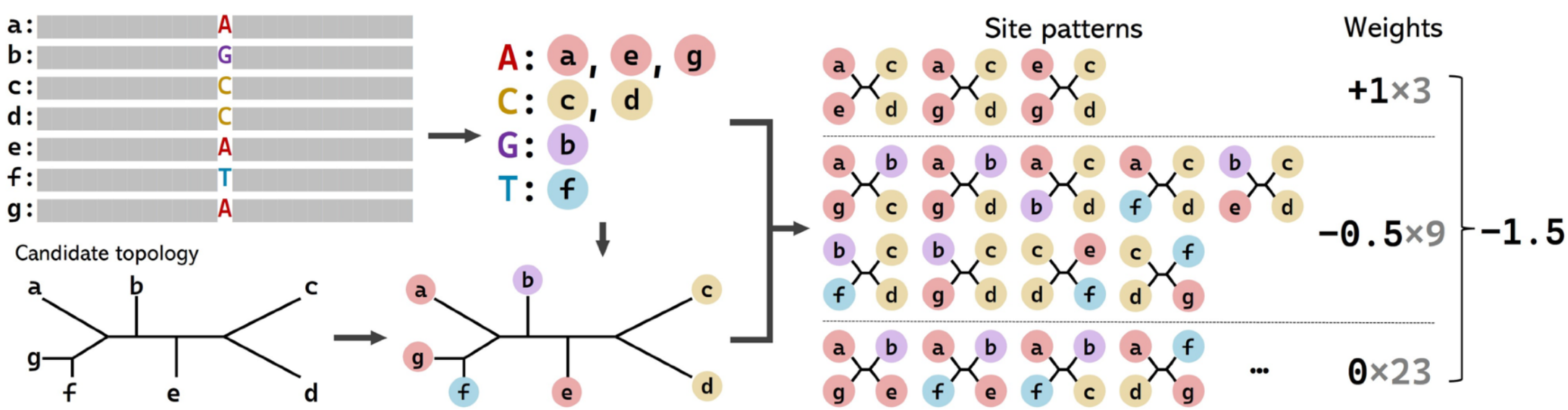
Optimization

- We can sum $w_i(T \mid q_j)$ over all $\binom{n}{4}$ quartets without listing all of them (no subsampling needed)



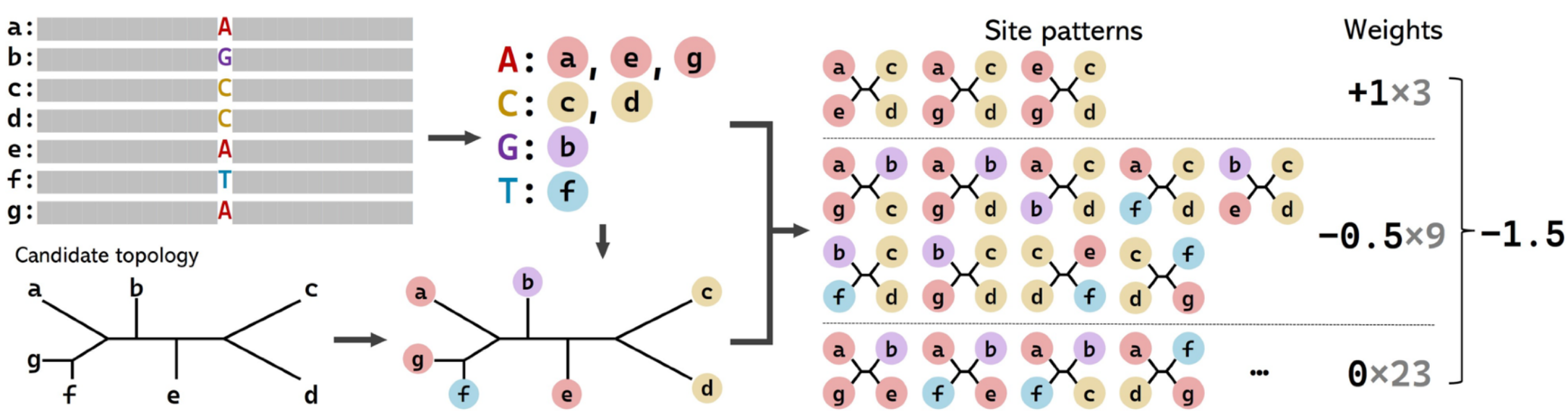
Optimization

- We can sum $w_i(T \mid q_j)$ over all $\binom{n}{4}$ quartets without listing all of them (no subsampling needed)
- Dynamic programming, similar to ASTRAL (identical to wASTRAL)



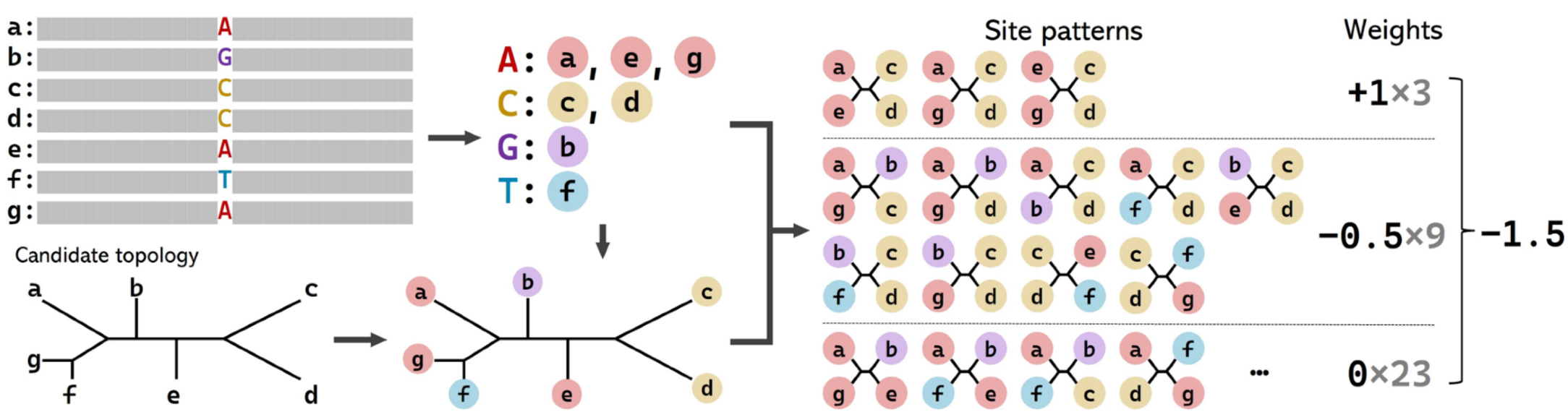
Optimization

- We can sum $w_i(T \mid q_j)$ over all $\binom{n}{4}$ quartets without listing all of them (no subsampling needed)
- Dynamic programming, similar to ASTRAL (identical to wASTRAL)
- Roughly $\mathcal{O}(kn^2 \log(n))$



Optimization

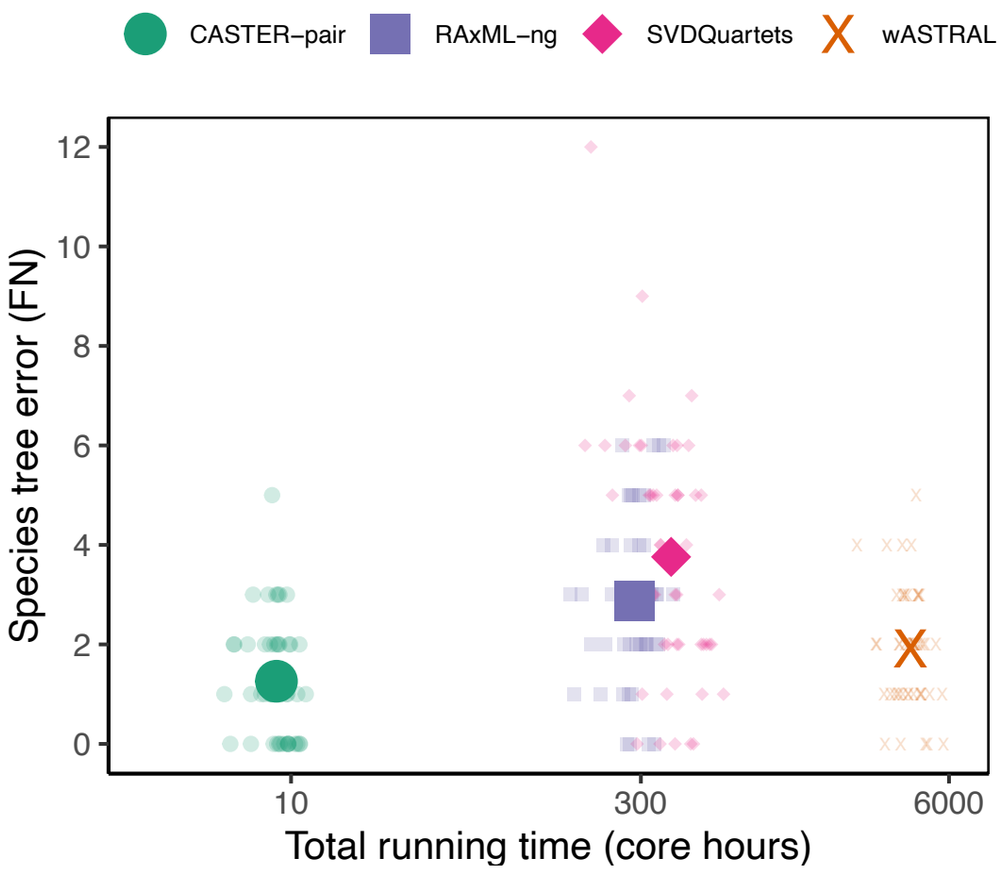
- We can sum $w_i(T \mid q_j)$ over all $\binom{n}{4}$ quartets without listing all of them (no subsampling needed)
- Dynamic programming, similar to ASTRAL (identical to wASTRAL)
- Roughly $\mathcal{O}(kn^2 \log(n))$
- This algorithm keeps statistical consistency guarantees



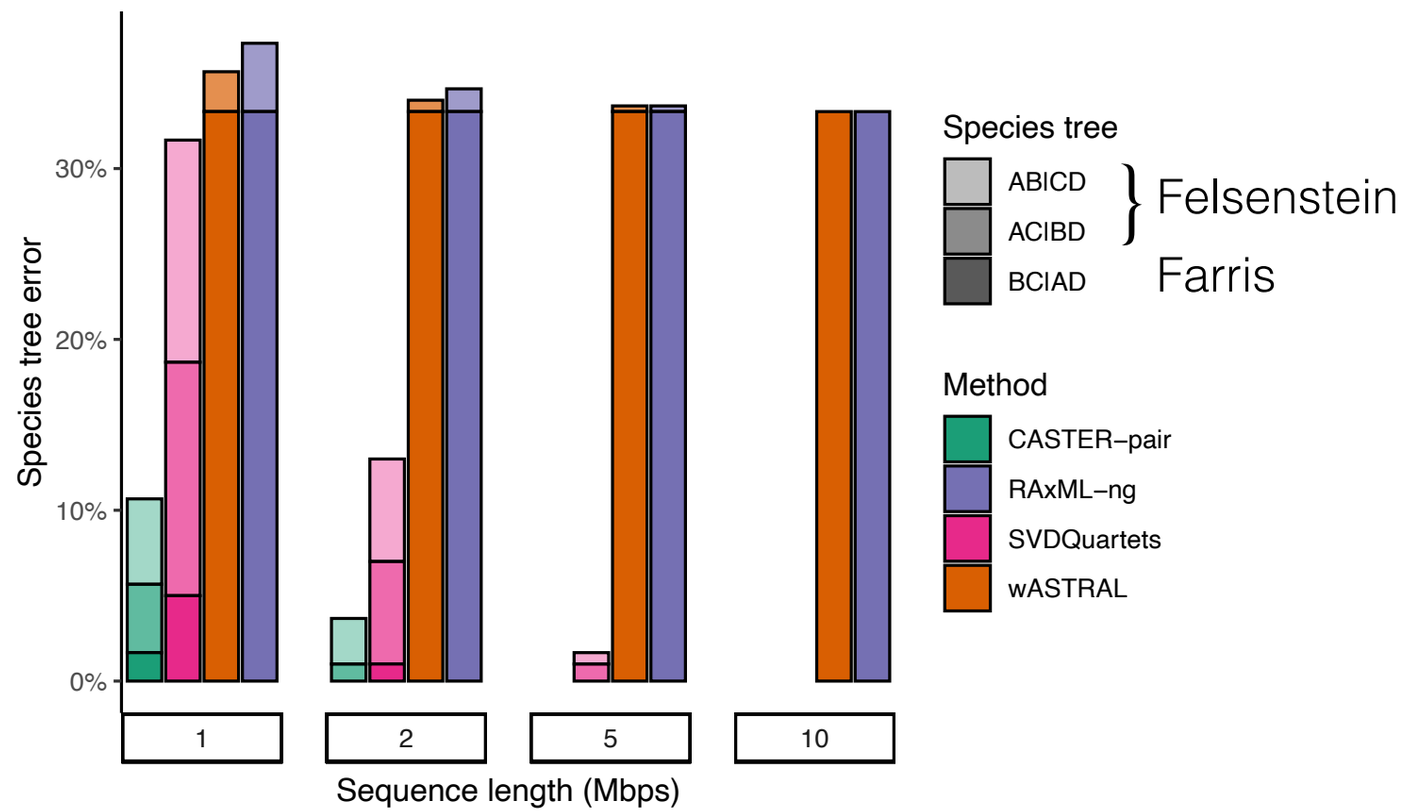
How well does it work?

Simulations with recombinations!

Recombination simulations
(msprime; Hudson model)
201 species, 5Mbp

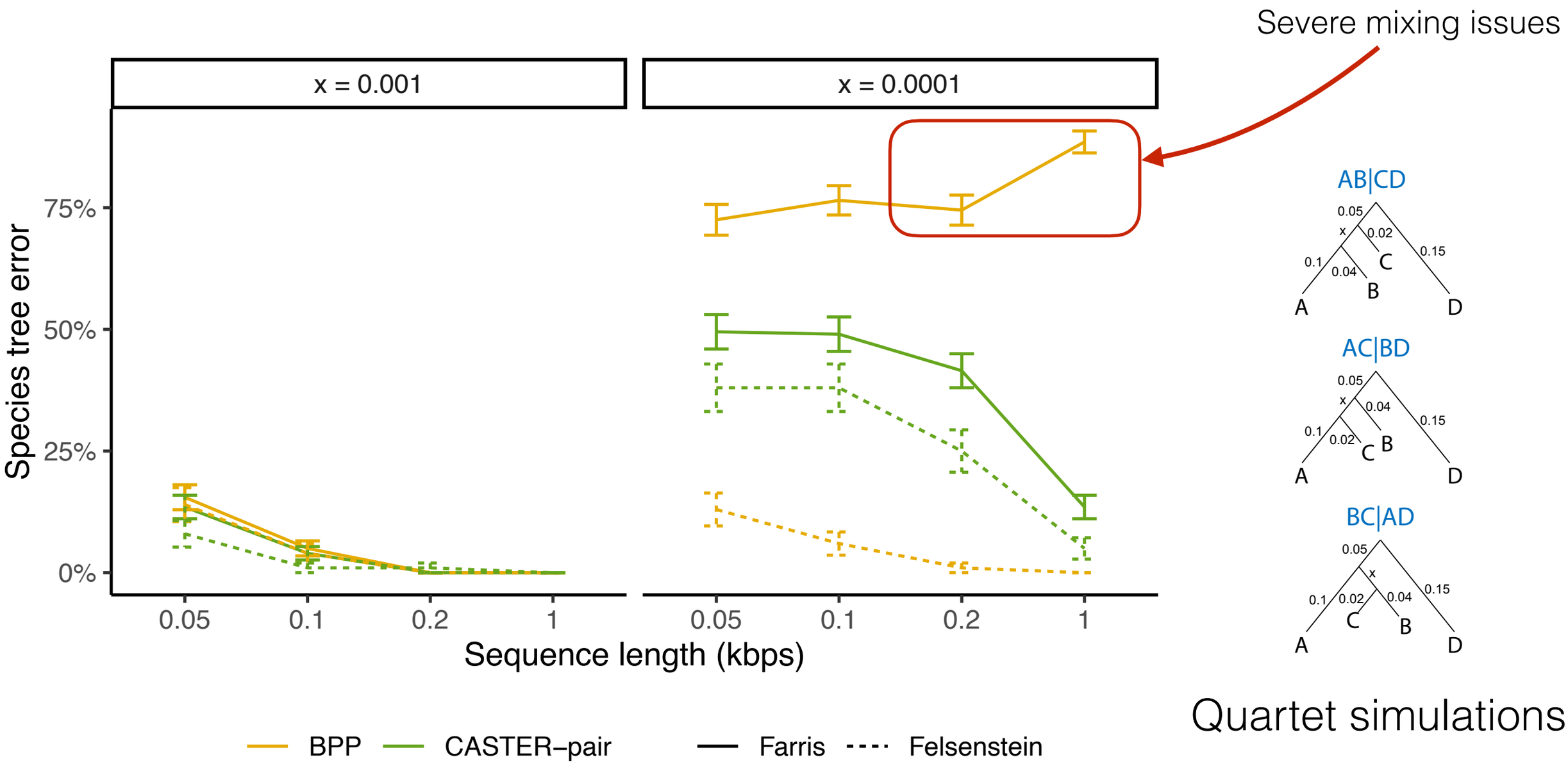


Recombination simulations
(msprime; Hudson model)
4 species, 10Mbp



Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

How about Bayesian co-estimation

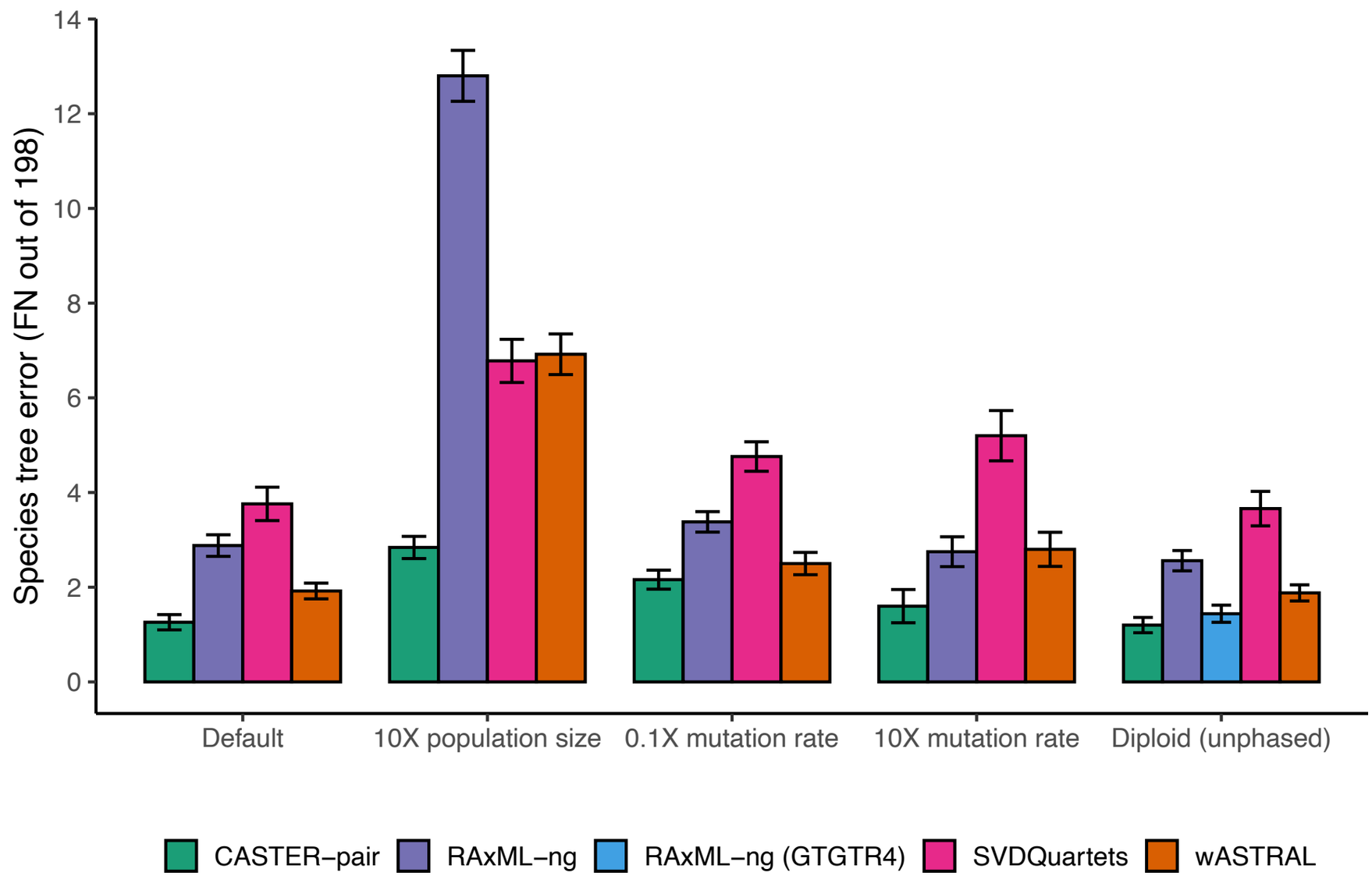


Quartet simulations

Recombination+ILS simulations
(msprime; Hudson model)
4 species, 10Mbp

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

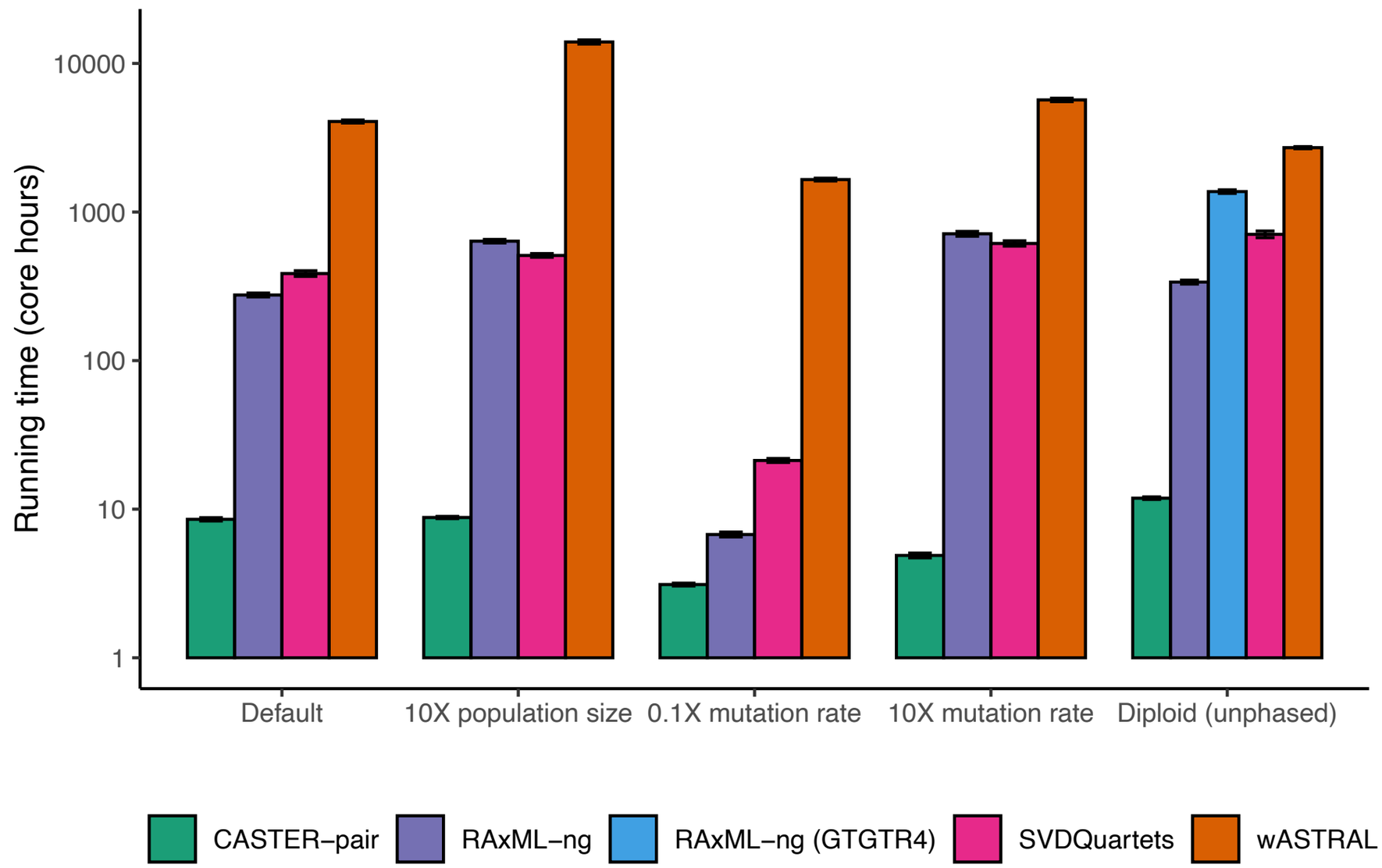
Same across rates of evolution, population size, and ploidy



Simulations recombination
(Hudson model)
201 species, 5Mbp,
recombination rate =
substitution rate,
non-ultra metric,
rate heterogeneity

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

Much better running time …



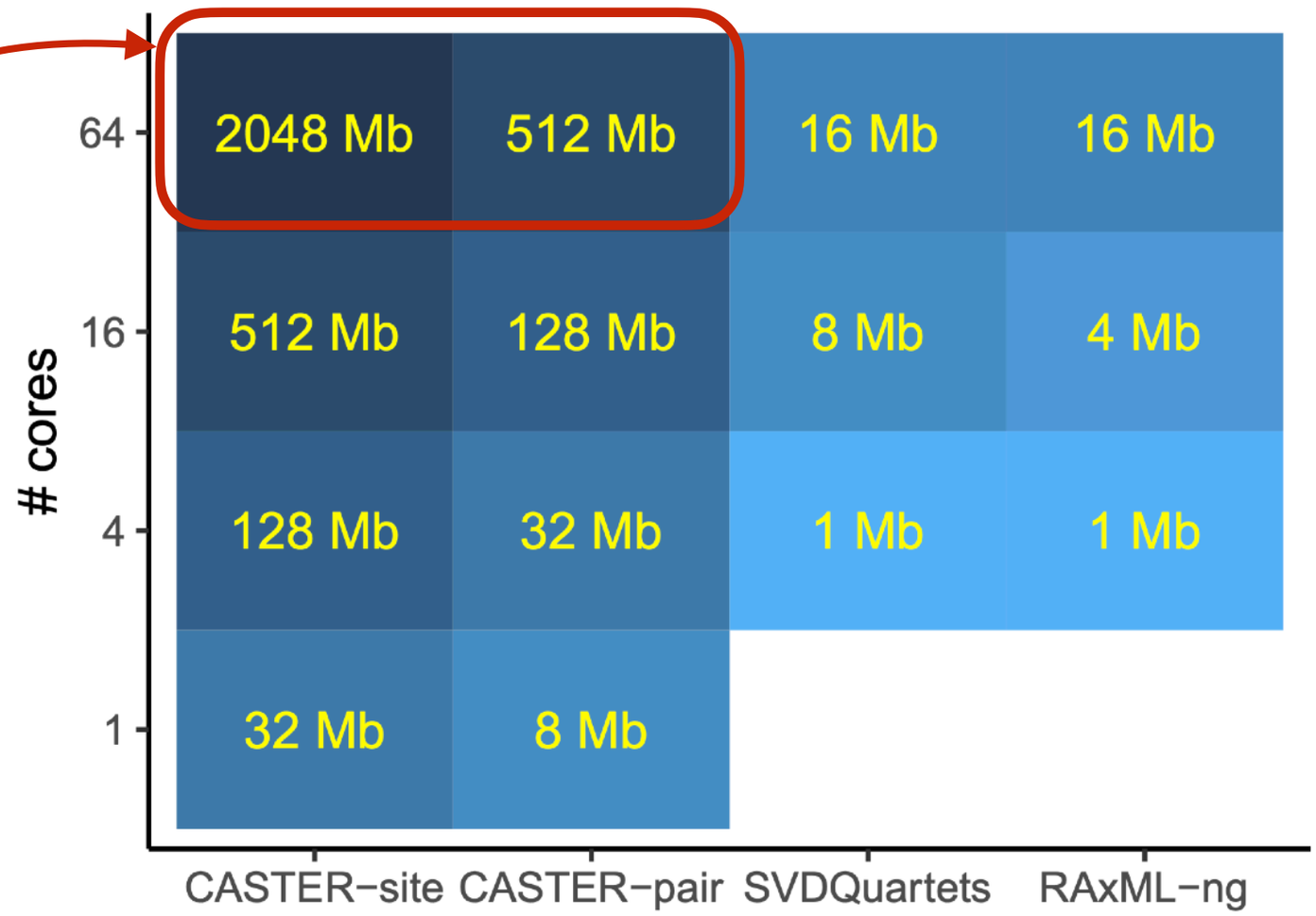
Simulations recombination
(Hudson model)
201 species, 5Mbp,
recombination rate =
substitution rate,
non-ultra metric,
rate heterogeneity

Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv,
doi.org/10.1101/2023.10.04.560884

Scalability

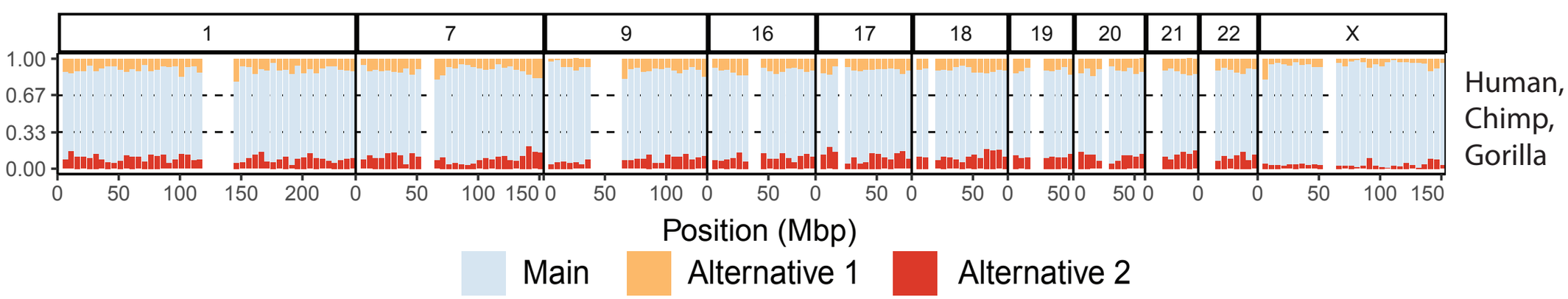
sites analyzed in 2 days (2GB per core)

Real genome-scale data



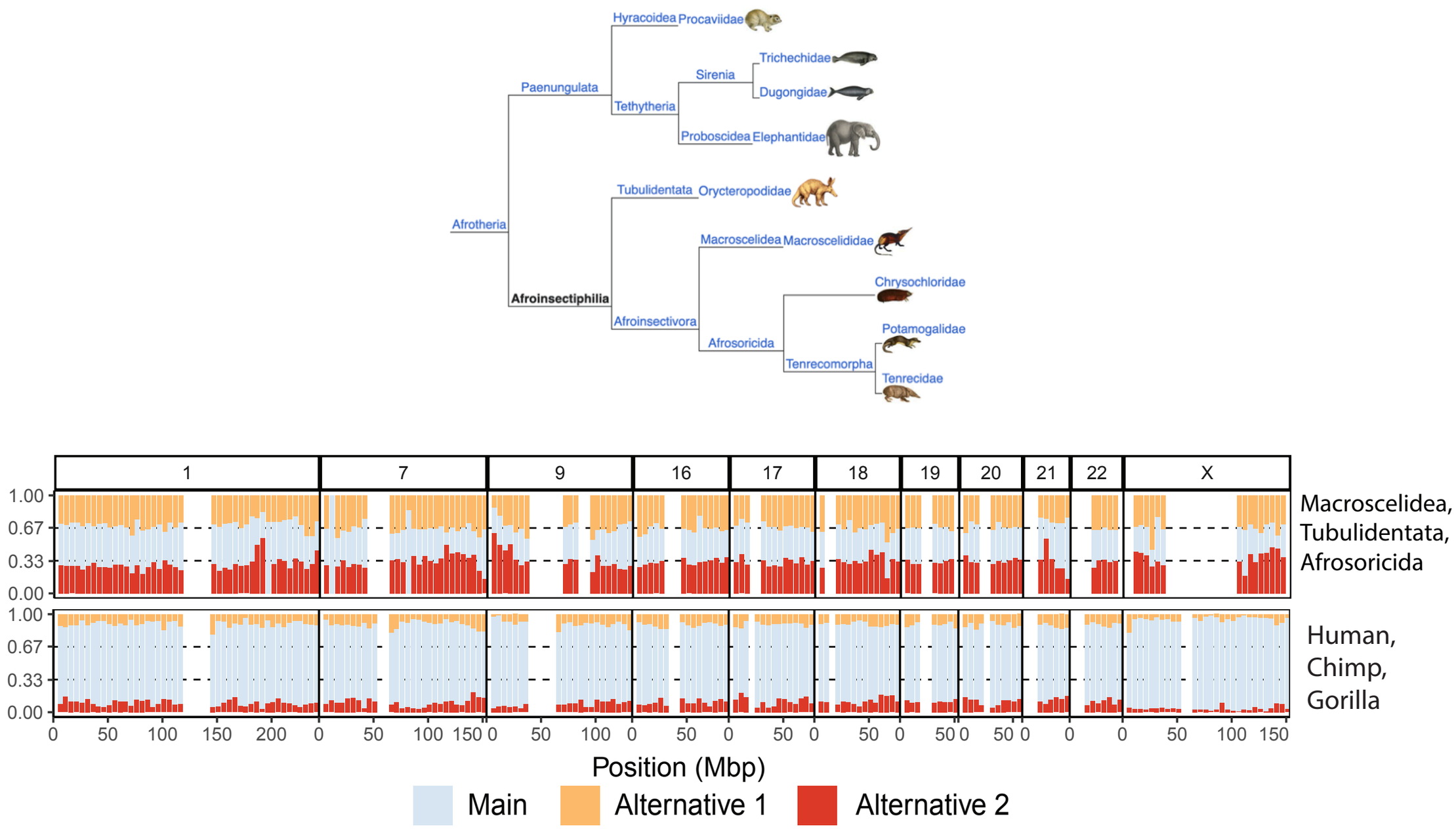
Zhang, Nielsen, Mirarab, 2024, Unpublished, bioRxiv, doi.org/10.1101/2023.10.04.560884

Applied to full mammalian genomes



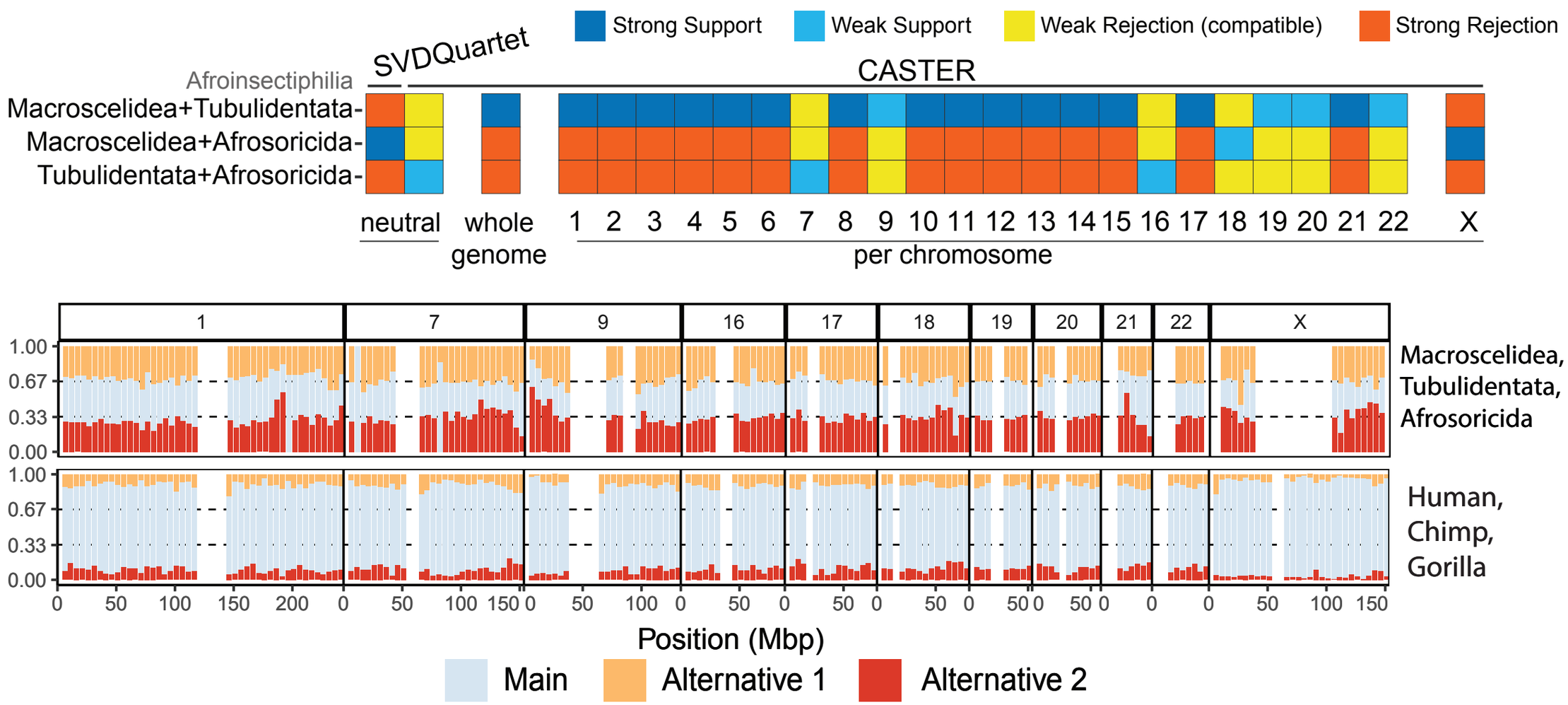
- Data from Foley et al, 2023, Science.

Applied to full mammalian genomes



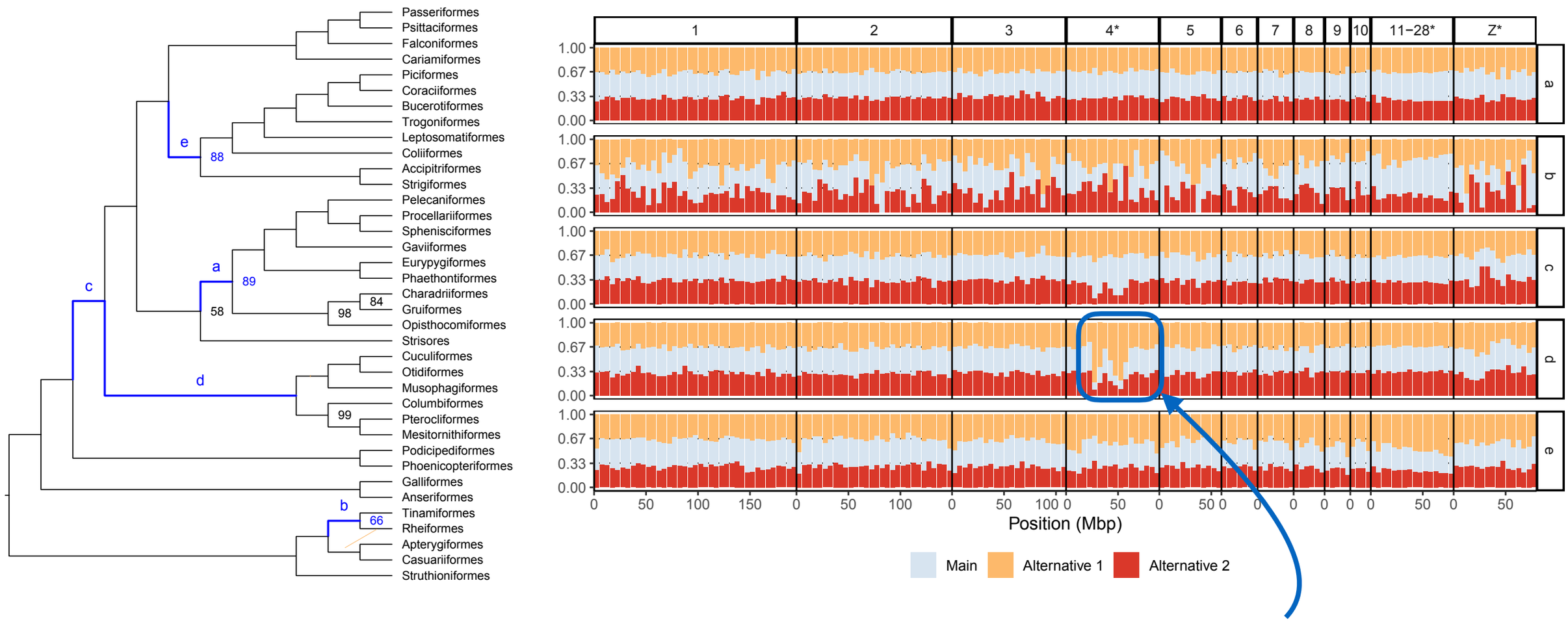
- Data from Foley et al, 2023, Science.

Applied to full mammalian genomes



- Data from Foley et al, 2023, Science.

Works even for birds!



The subject of a whole separate paper

A region of suppressed recombination misleads neoavian phylogenomics

Siavash Mirarab[a,1], Iker Rivas-González[b], Shaohong Feng[c,d], Josefin Stiller[e], Qi Fang[f], Uyen Mai[a], Glenn Hickey[g], Guangji Chen[c,d], Nadolina Brajuka[a], Olivier Fedrigo[h], Giulio Formenti[h], Jochen B. W. Wolf, Kerstin Howe[i], Agostinho Antunes[k], Mikkel H. Schierup[b], Benedict Paten[g], Erich D. Jarvis[h], Guojie Zhang[c], and Edward L. Braun[m]

- Stiller et al, 2024, 64000 intergenic regions, bird genomes

Open questions

Theoretical questions

A. Why $-1/2$ for JC69? We can prove it, but is there a more elegant explanation?

Theoretical questions

A. Why $-1/2$ for JC69? We can prove it, but is there a more elegant explanation?

B. Using a single site, can we design weights for any model more complex than F84?

Theoretical questions

A. Why $-1/2$ for JC69? We can prove it, but is there a more elegant explanation?

B. Using a single site, can we design weights for any model more complex than F84?

C. Using a pair of sites, we could only prove consistency for (three) submodels of GTR with two fewer parameters:

   1. Is the pair approach consistent for GTR?

   2. Are there other weight schemes based on pairs that are?

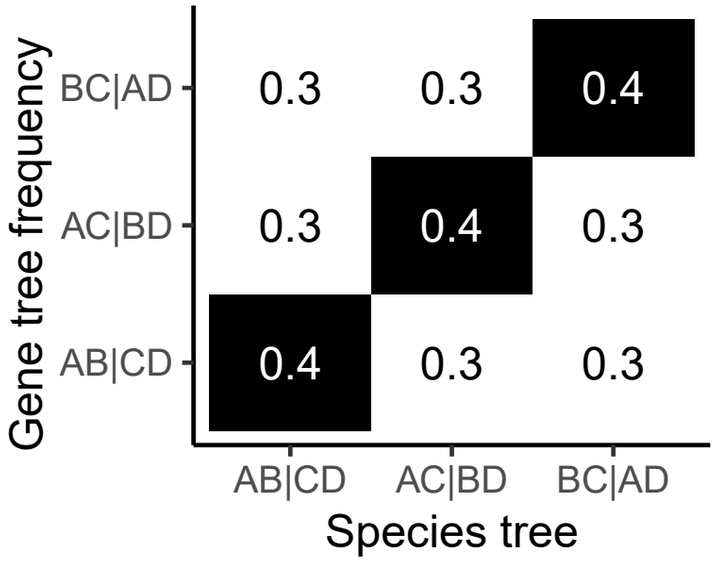
   3. How about if we use more than two sites?

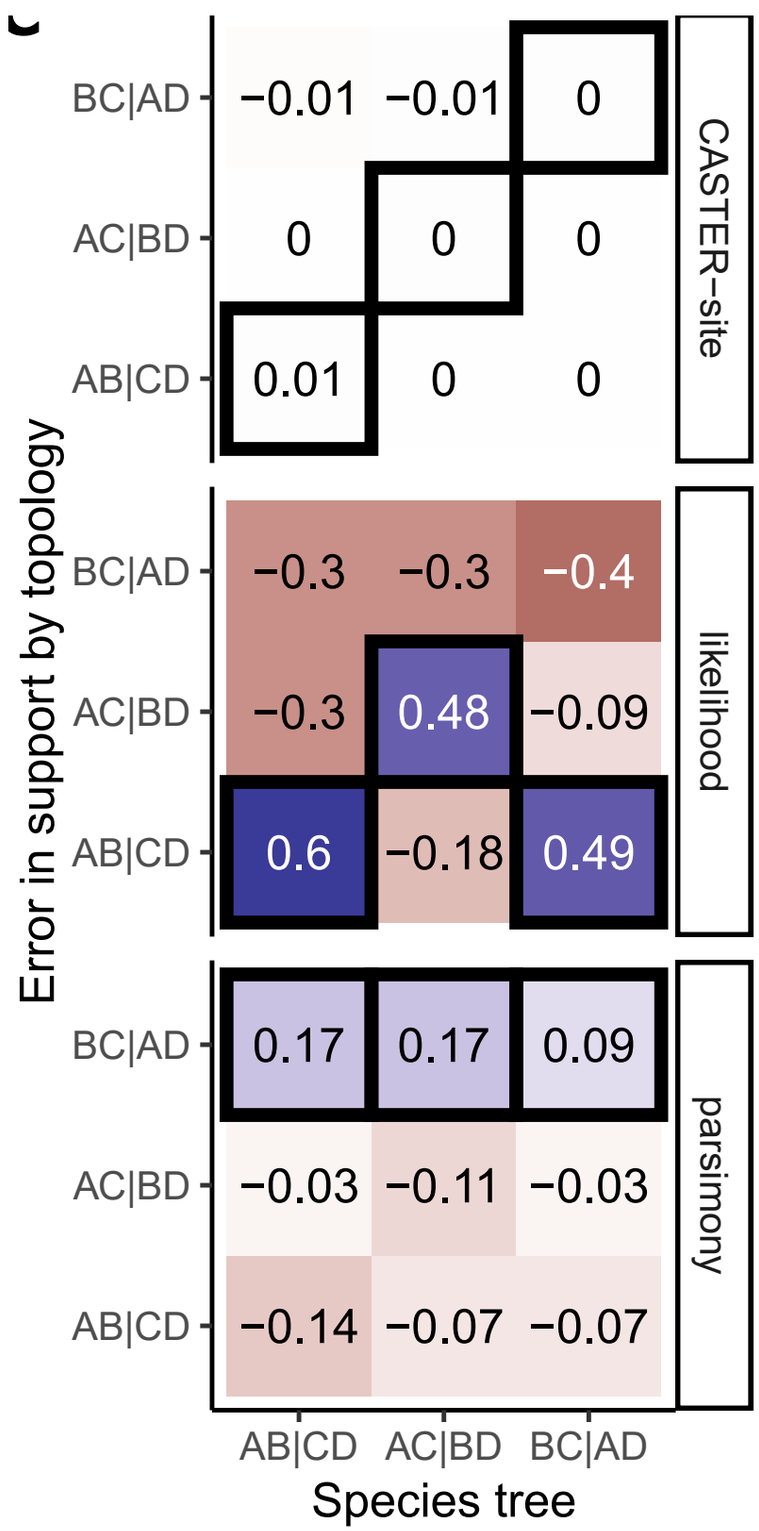
Future work

- No branch lengths!
- No duplication and loss!
- Scores used in moving window analysis; more elegant ways
  - Can we use it to detect introgression?
  - Can we downright poorly aligned sites?
- Comparing to other site-based methods (METAL, etc.), both in simulation but also in sample complexity
- Amino acid and binary data.

Chao Zhang

Rasmus Nielsen


Gene tree frequency
BC|AD
AC|BD
AB|CD
0.3
0.3
0.4
0.3
0.4
0.3
0.4
0.3
0.3
AB|CD
AC|BD
BC|AD
Species tree


Error in support by topology
CASTER−site
likelihood
parsimony
BC|AD
AC|BD
AB|CD
−0.01
−0.01
0
0
0
0
0.01
0
0
−0.3
−0.3
−0.4
−0.3
0.48
−0.09
0.6
−0.18
0.49
0.17
0.17
0.09
−0.03
−0.11
−0.03
−0.14
−0.07
−0.07
AB|CD
AC|BD
BC|AD
Species tree