

Phlag: Scalable detection of genomics regions with unexplained phylogenetic heterogeneity

Ali Osman Berk Şapcı, Shayesteh Arasti, Edward L. Braun and Siavash Mirarab

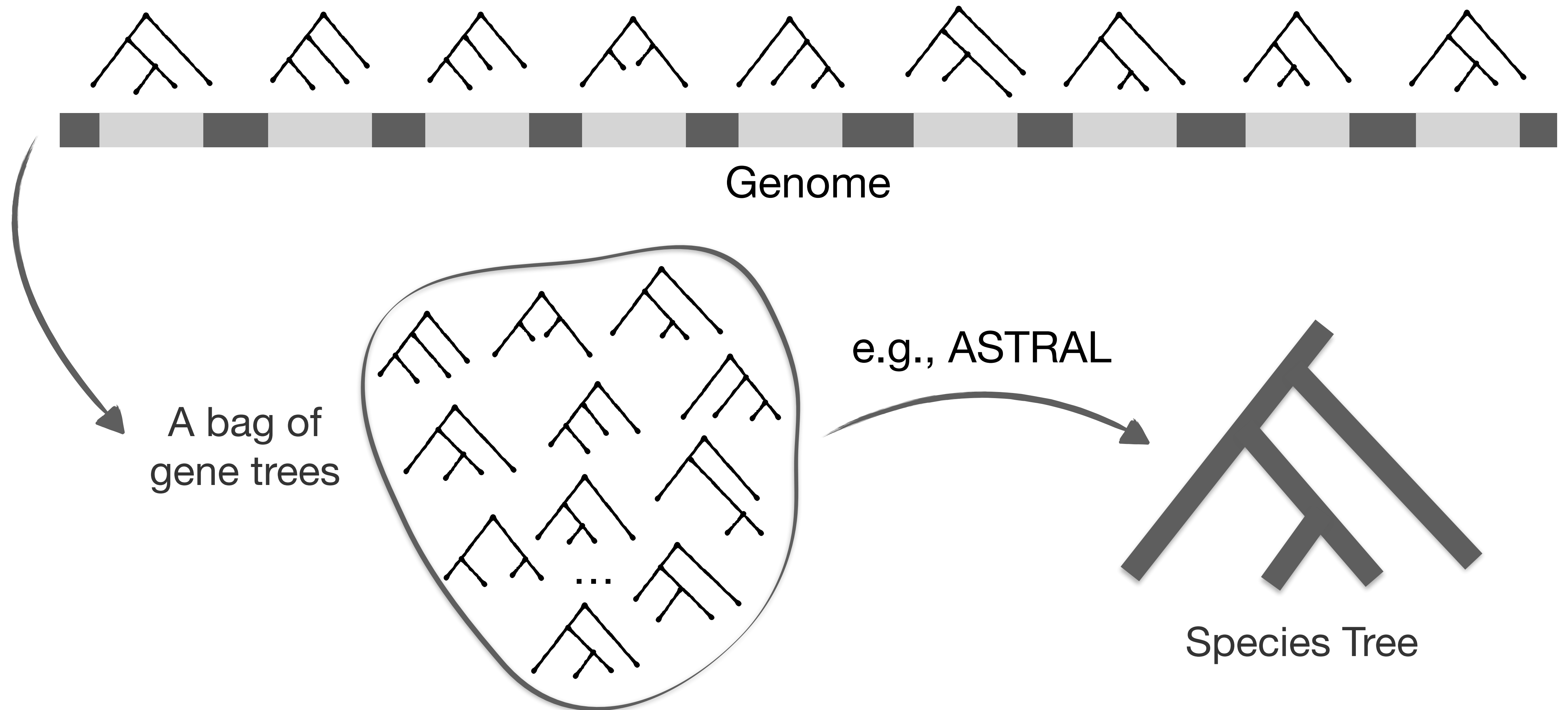


Phlag: Scalable detection of genomics regions with **unexplained phylogenetic heterogeneity**

Ali Osman Berk Şapcı, Shayesteh Arasti, Edward L. Braun and Siavash Mirarab

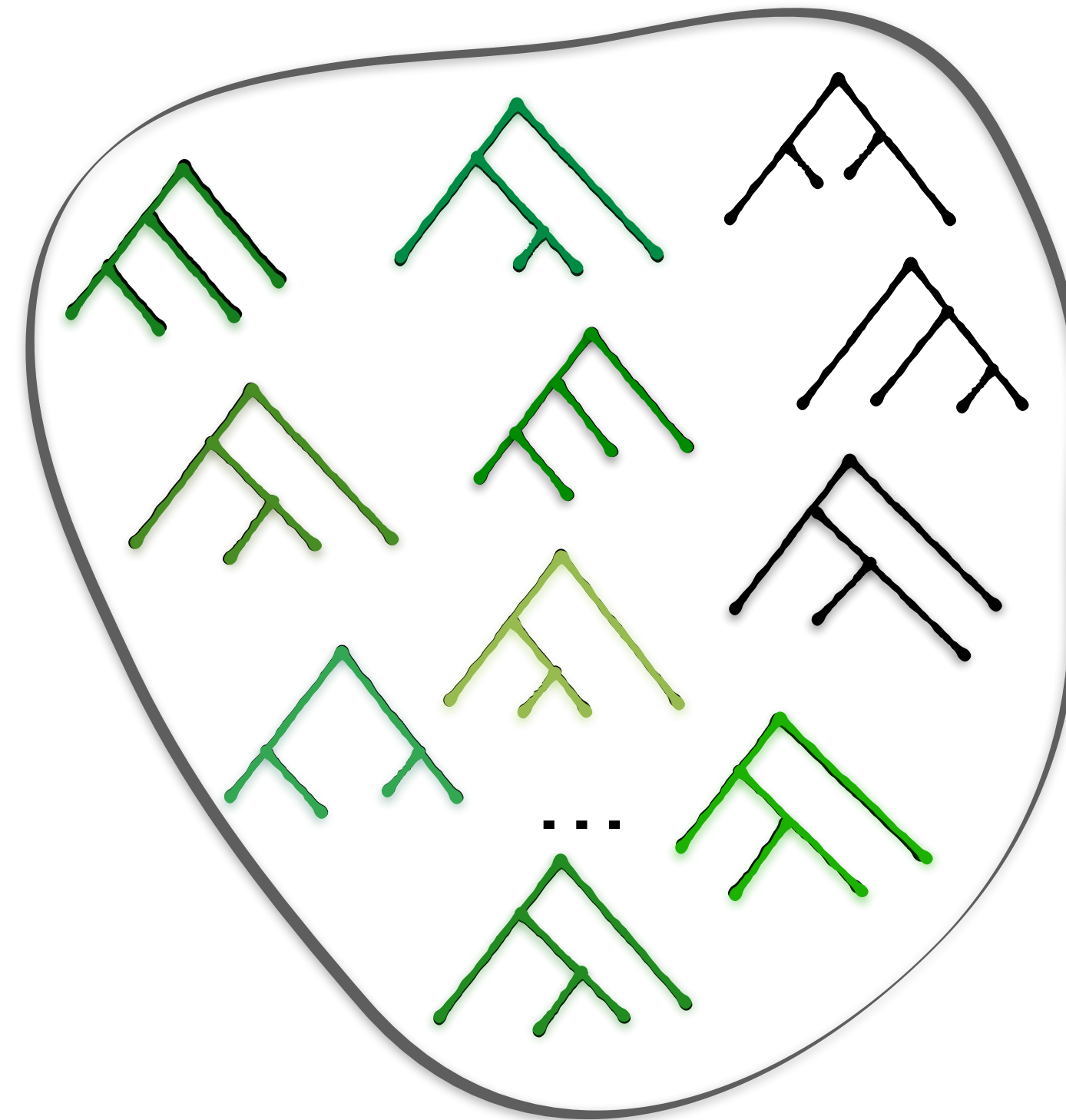


Phylogenomics through a “bag of genes” model

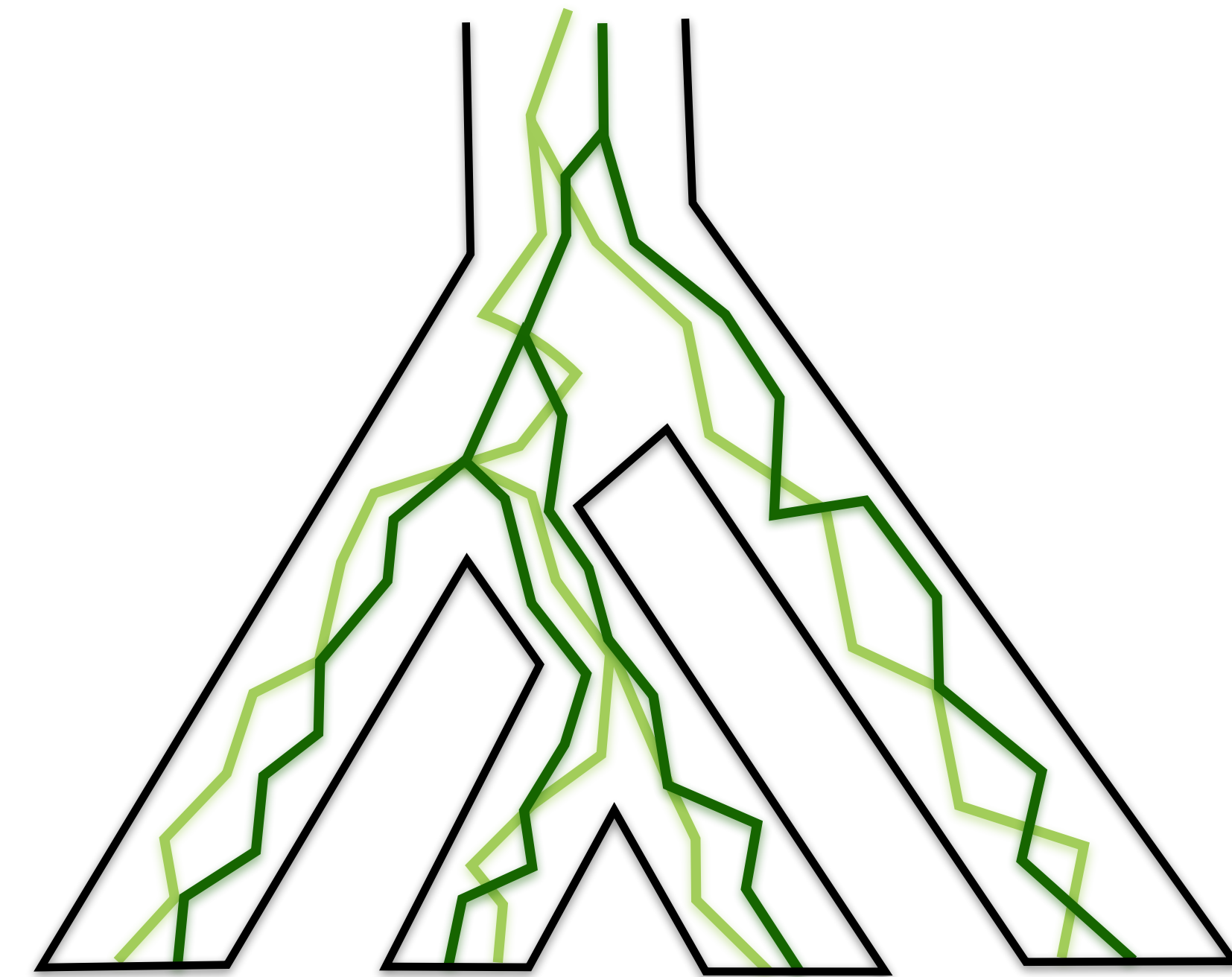


Multi-species coalescent model and violations

Expected discordance:
incomplete lineage sorting
► **Multi-species coalescent**



Gene Trees



Species Tree

Multi-species coalescent model and violations

Expected discordance:

incomplete lineage sorting

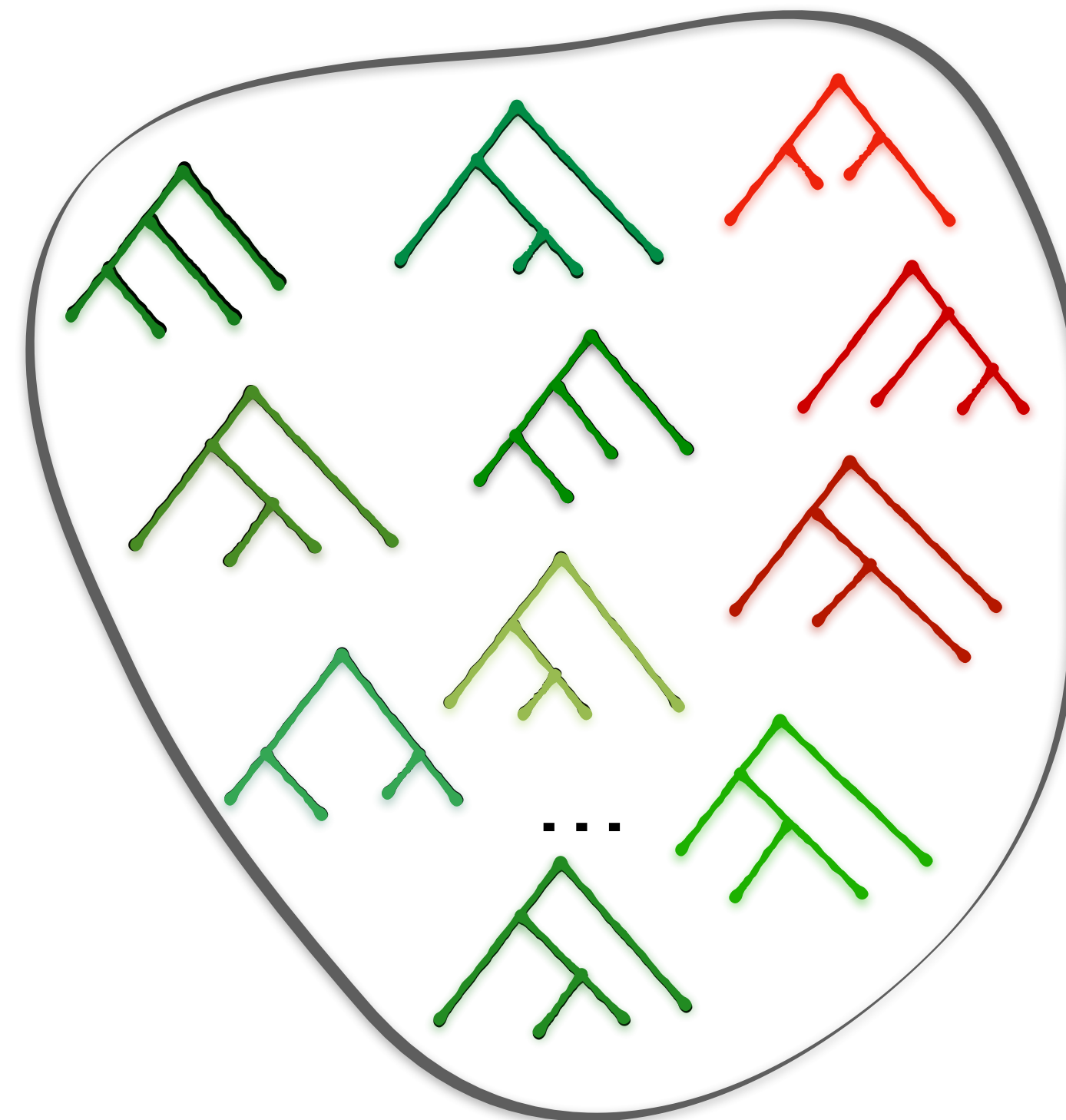
► Multi-species coalescent

Errors and artifacts:

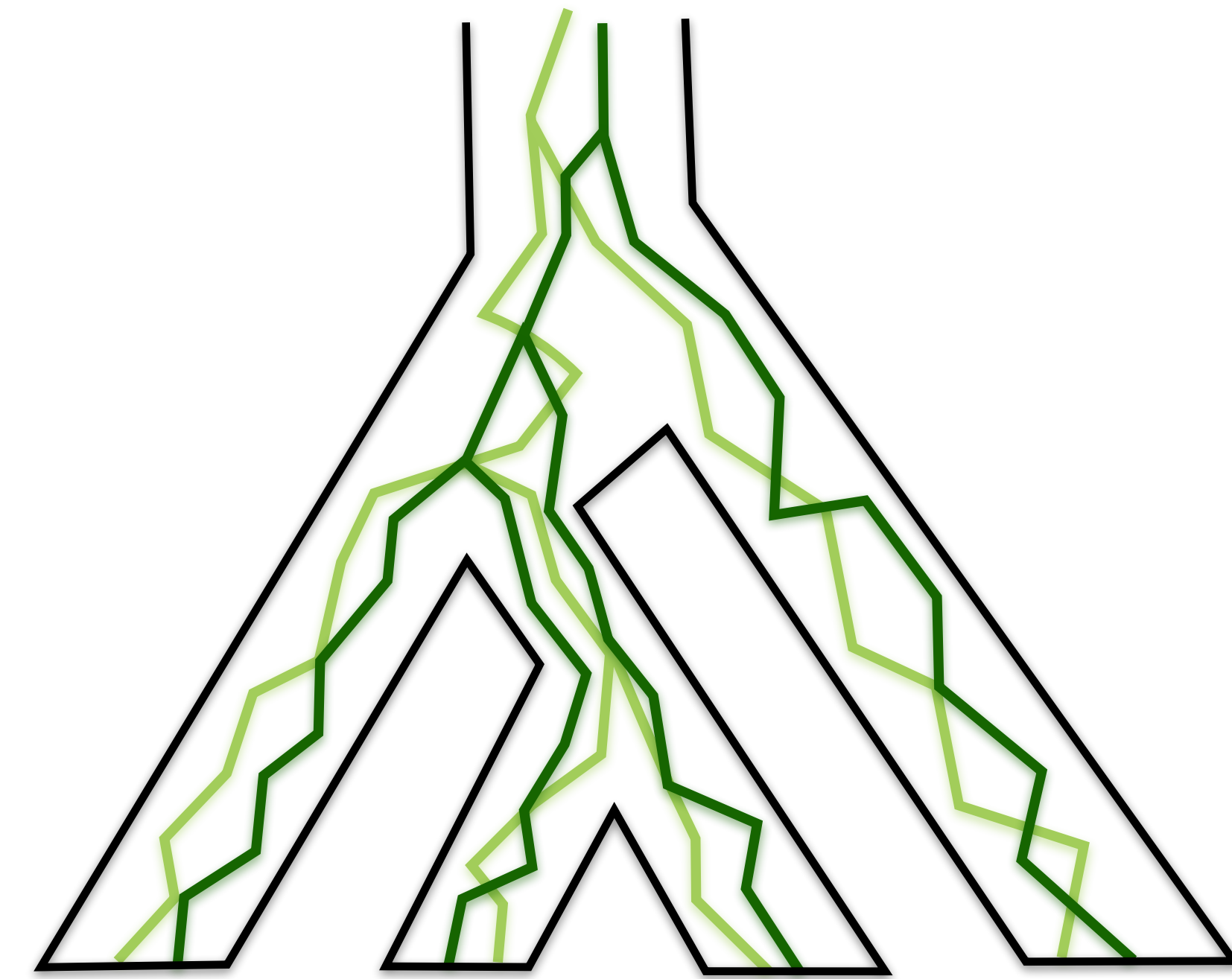
- Alignment, paralogy/homology

Biological causes:

- Hybridization & introgression
- Dramatic change in parameters (pop. size & recombination rate)

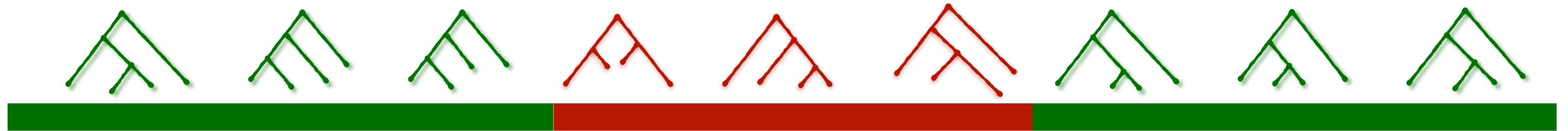


Gene Trees



Species Tree

Multi-species coalescent model and violations



Expected discordance:

incomplete lineage sorting

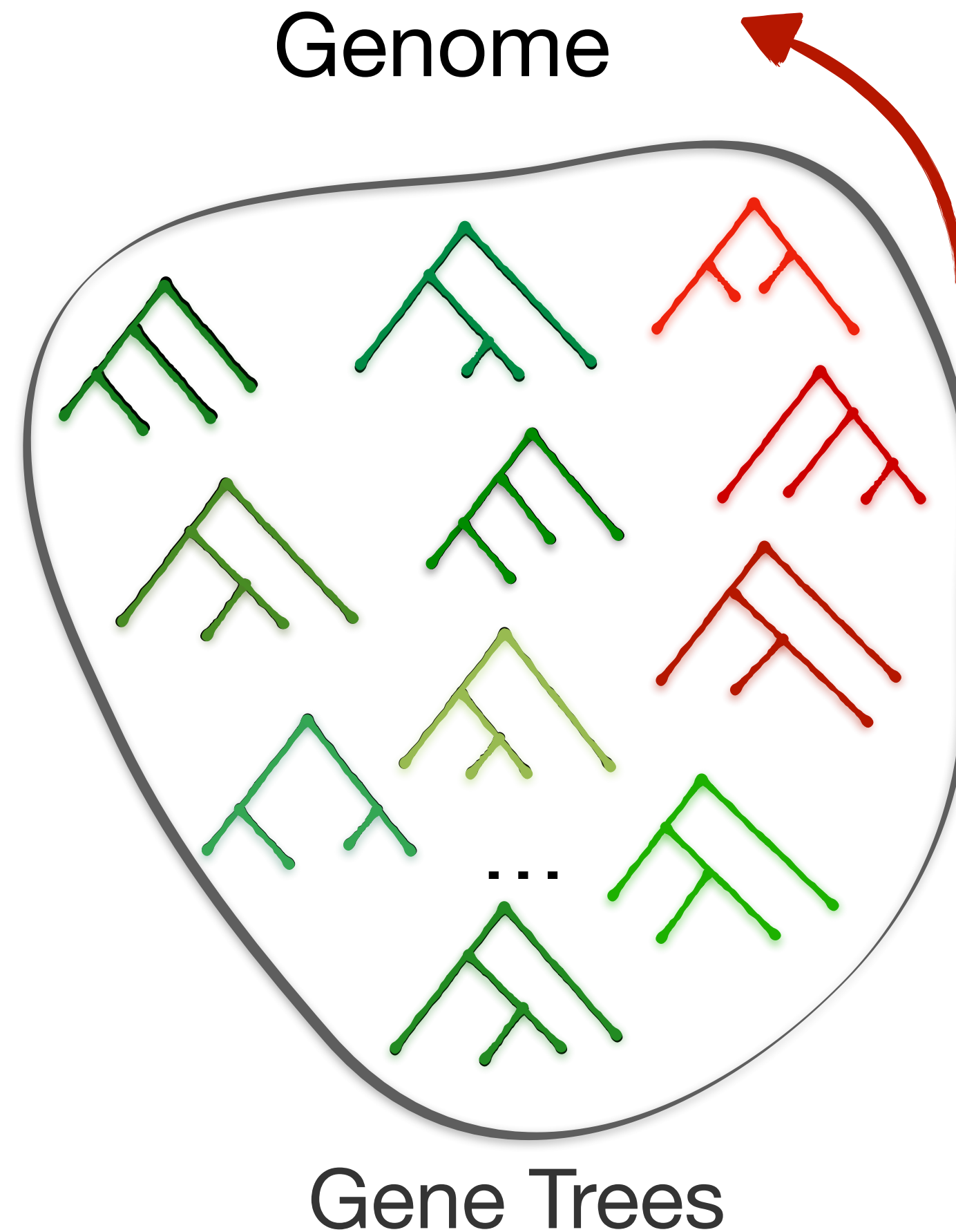
► Multi-species coalescent

Errors and artifacts:

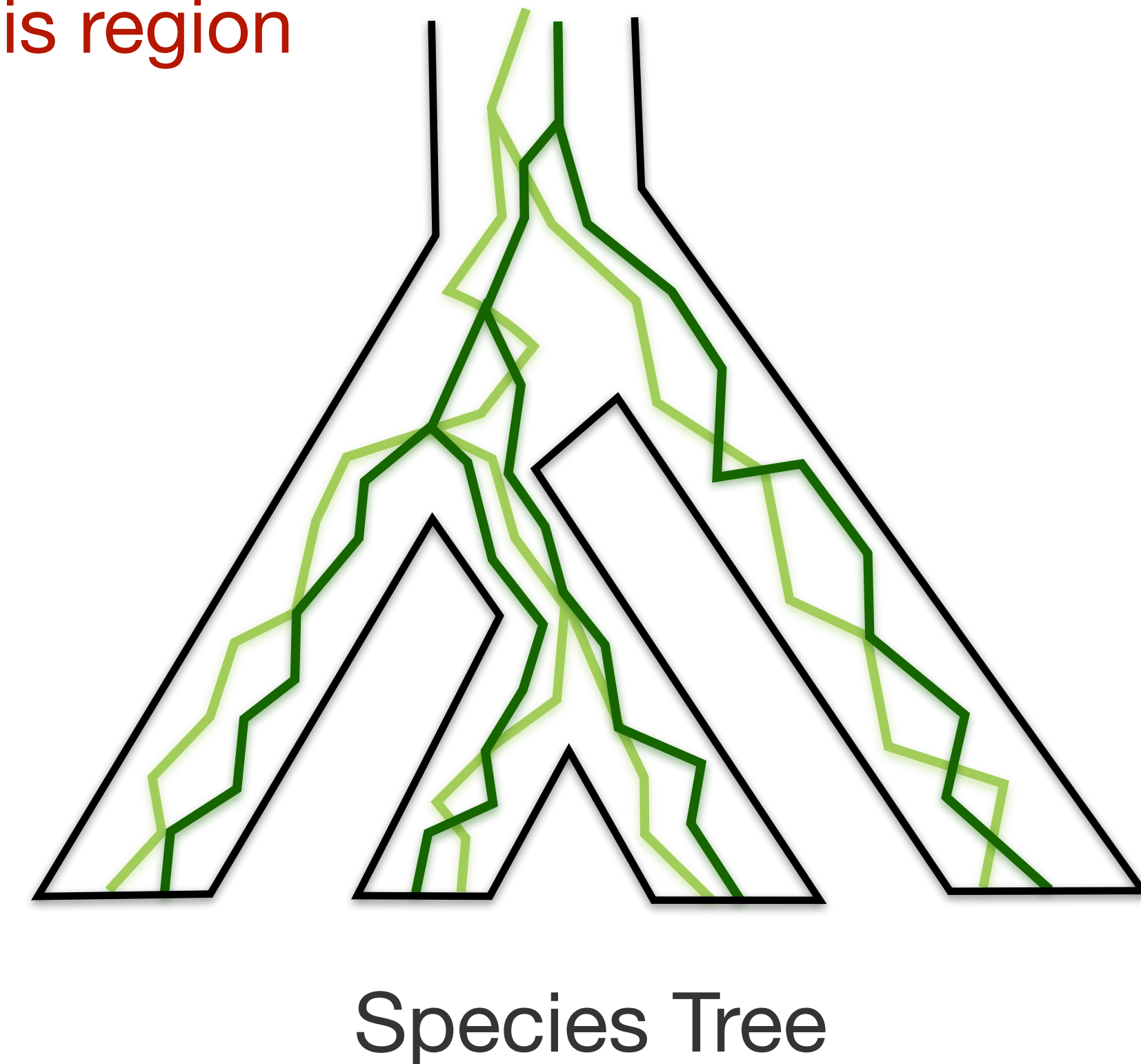
- Alignment, paralogy/homology

Biological causes:

- Hybridization & introgression
- Dramatic change in parameters (pop. size & recombination rate)



resulted from
this region



Deviations from the species tree are often localized

**Signal from non-MSC processes
concentrates in certain regions:**

Deviations from the species tree are often **localized**

Signal from non-MSA processes concentrates in certain regions:

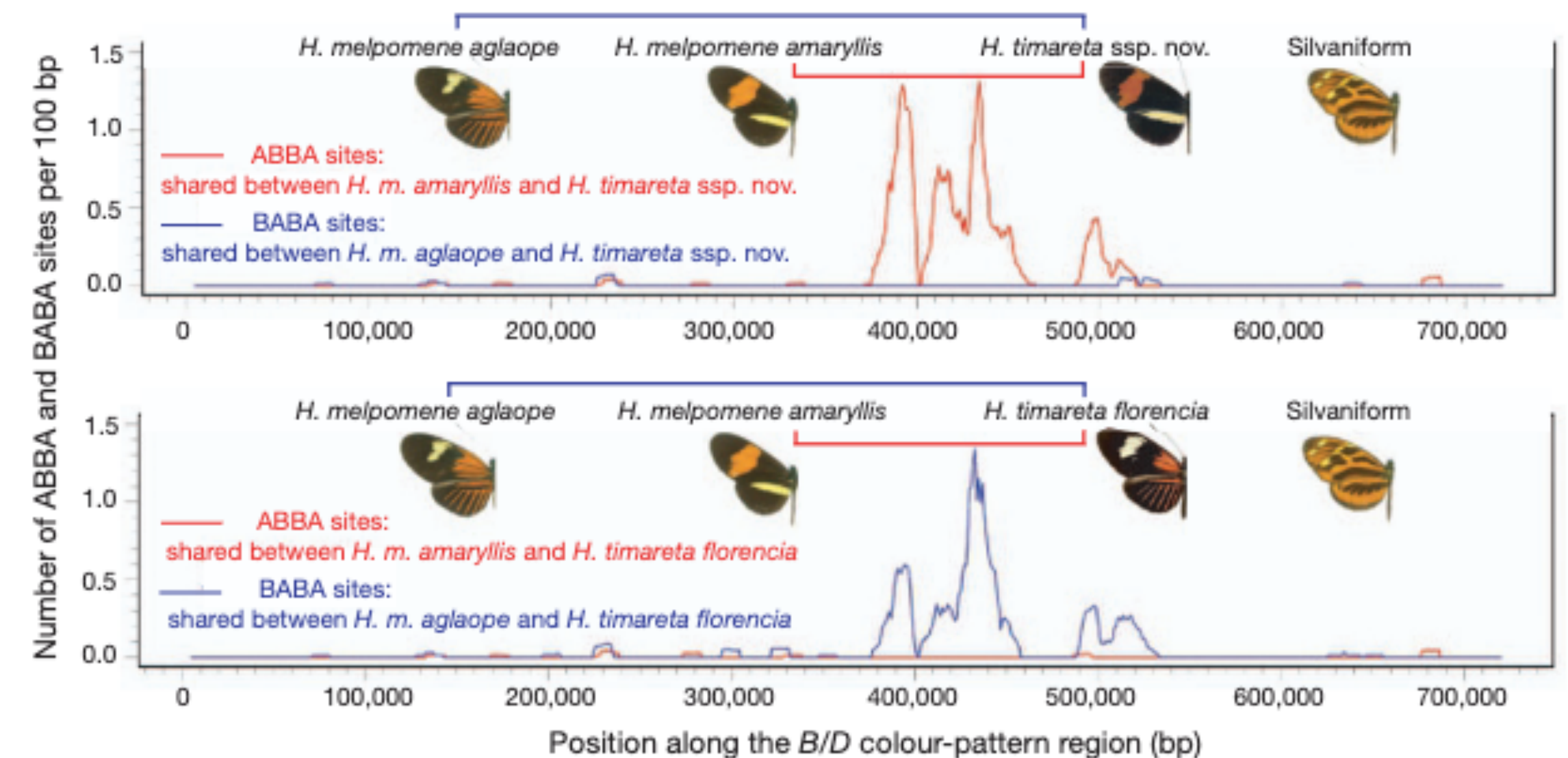
- **Introgression & hybridization:**
 - ▶ Butterflies (chr15, chr18)
 - ▶ Mouse & other rodents (chr1)
 - ▶ ...

Major patterns in the introgression history of *Heliconius* butterflies

Yuttapong Thawornwattana^{1*}, Fernando Seixas¹, Ziheng Yang^{2*}, James Mallet^{1*}

Butterfly genome reveals promiscuous exchange of mimicry adaptations among species

The *Heliconius* Genome Consortium*



Müllerian
mimicry in



Deviations from the species tree are often **localized**

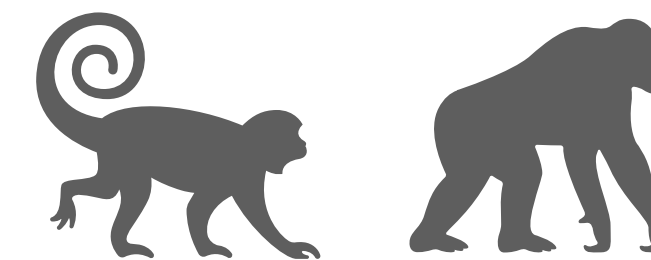
Signal from non-MS processes concentrates in certain regions:

- **Introgression & hybridization:**
 - Butterflies (chr15, chr18)
 - Mouse & other rodents (chr1)
 - ...
- **Selective sweep & unusual ILS** in primates (chr3)

PRIMATE GENOMES

Pervasive incomplete lineage sorting illuminates speciation and selection in primates

Iker Rivas-González[†], Marjolaine Rousselle[†], Fang Li[†], Long Zhou, Julien Y. Dutheil, Kasper Munch, Yong Shao, Dongdong Wu, Mikkel H. Schierup*, Guojie Zhang*



Deviations from the species tree are often **localized**

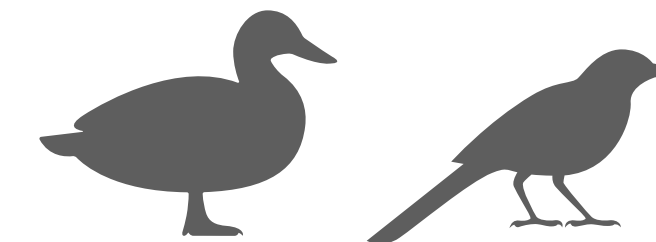
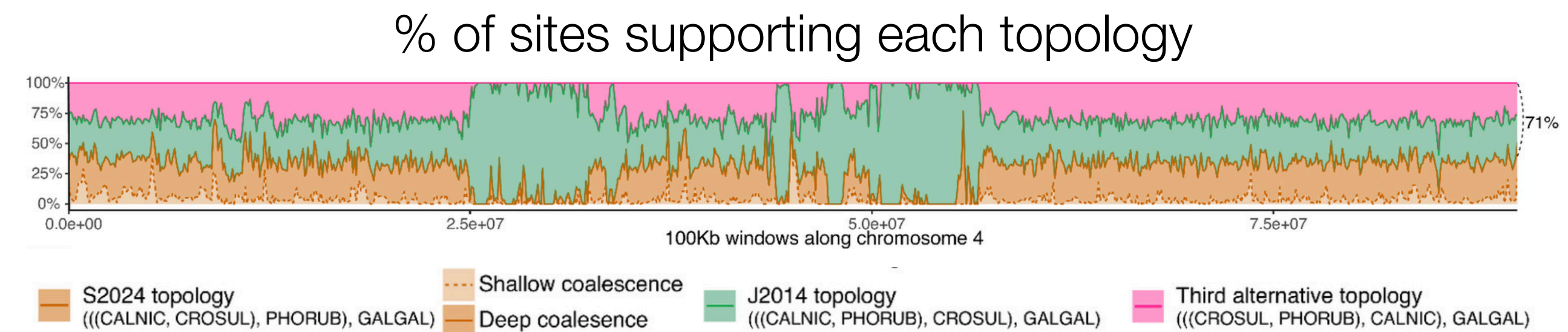
Signal from non-MS processes concentrates in certain regions:

- **Introgression & hybridization:**
 - Butterflies (chr15, chr18)
 - Mouse & other rodents (chr1)
 - ...
- **Selective sweep & unusual ILS** in primates (chr3)
- **Recombination suppression** in Neoaves (chr4)

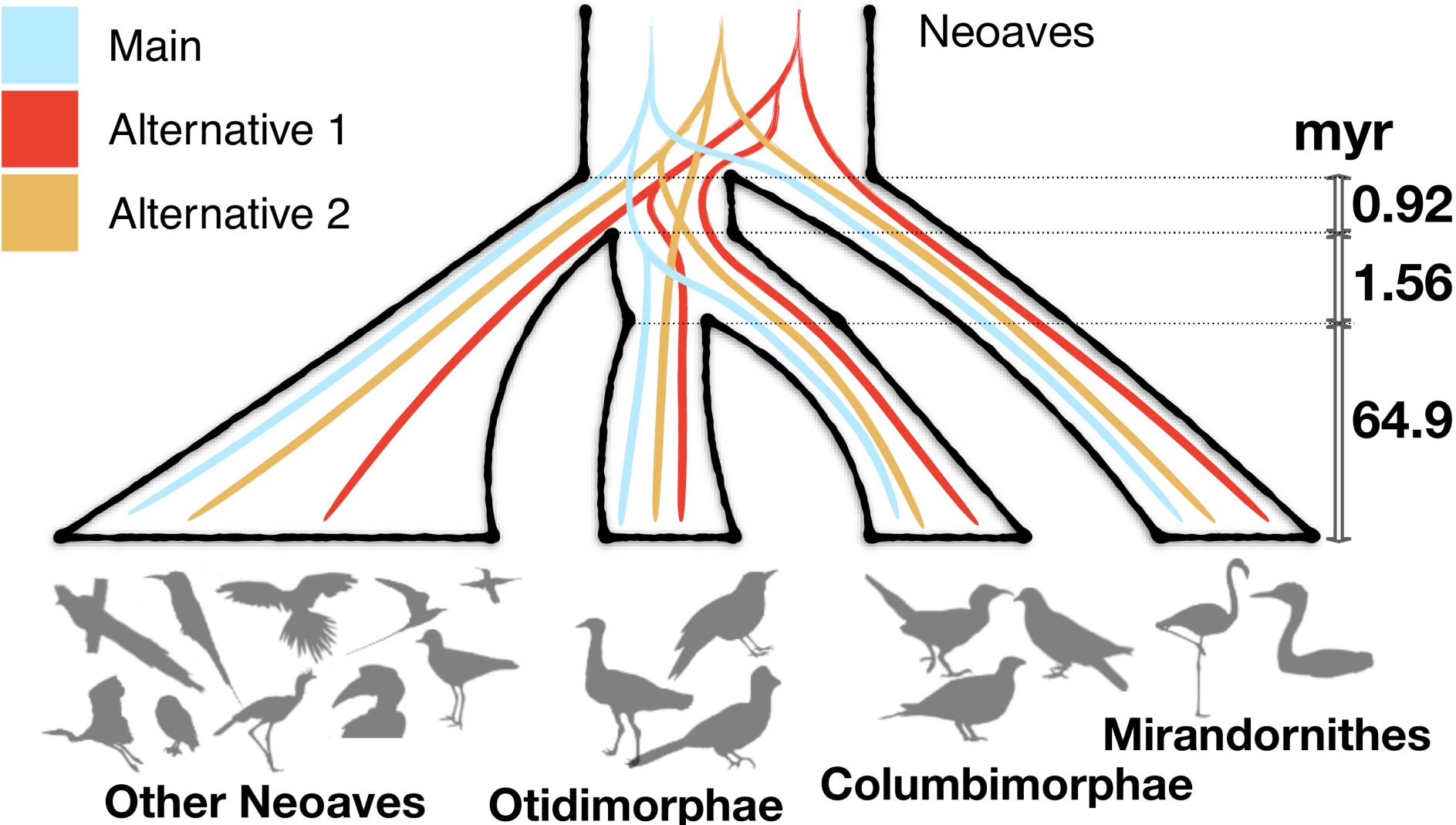
A region of suppressed recombination misleads neoavian phylogenomics

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^h, Olivier Fedrigo^h, Giulio Formenti^h, Jochen B. W. Wolfⁱ, Kerstin Howe^j, Agostinho Antunes^{k,l}, Mikkel H. Schierup^b, Benedict Paten^e, Erich D. Jarvis^h, Guojie Zhang^c, and Edward L. Braun^m

Edited by David Hillis, The University of Texas at Austin, Austin, TX; received November 14, 2023; accepted February 7, 2024



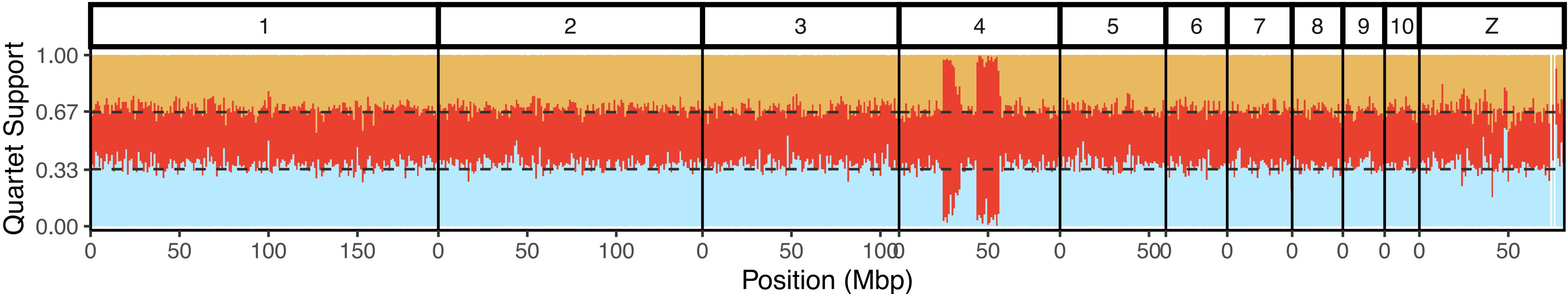
Suppressed recombination misleads avian phylogeny



Short branch lengths; difficult to resolve:

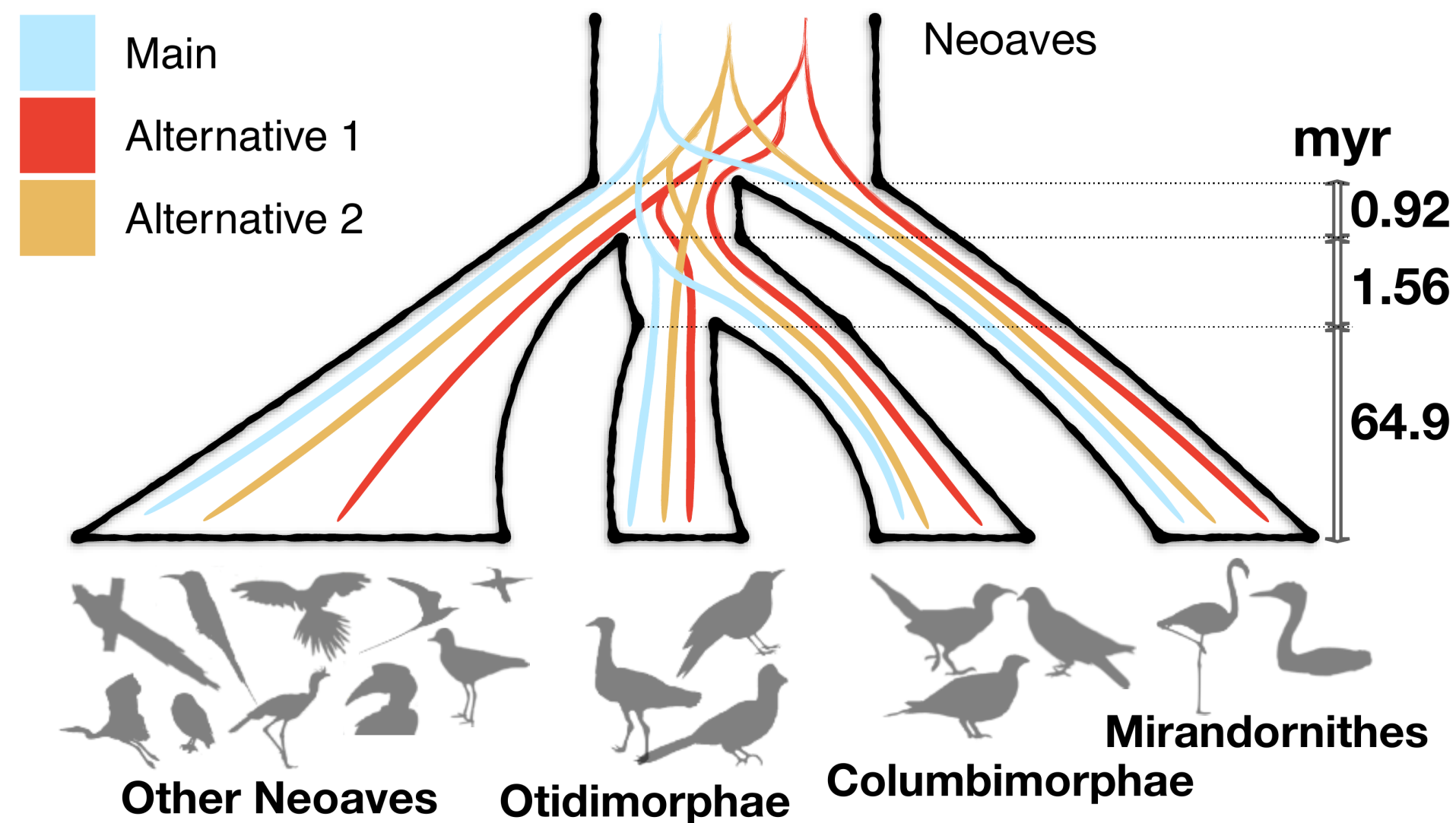
Jarvis et al., 2014 → Columbea (alternative)

Stiller et al., 2024 → Main topology



[Rivas-González et al., 2026]

Suppressed recombination misleads avian phylogeny

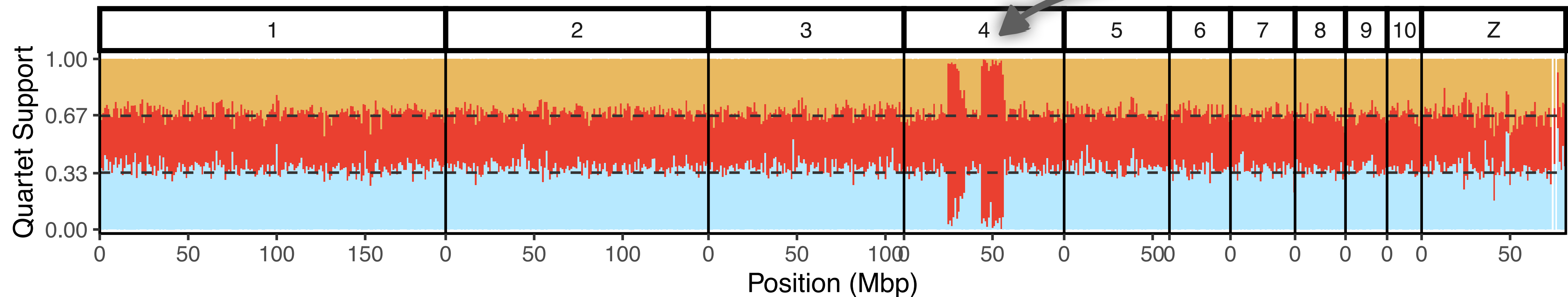


Short branch lengths; difficult to resolve:

Jarvis et al., 2014 → Columbea (alternative)

Stiller et al., 2024 → Main topology

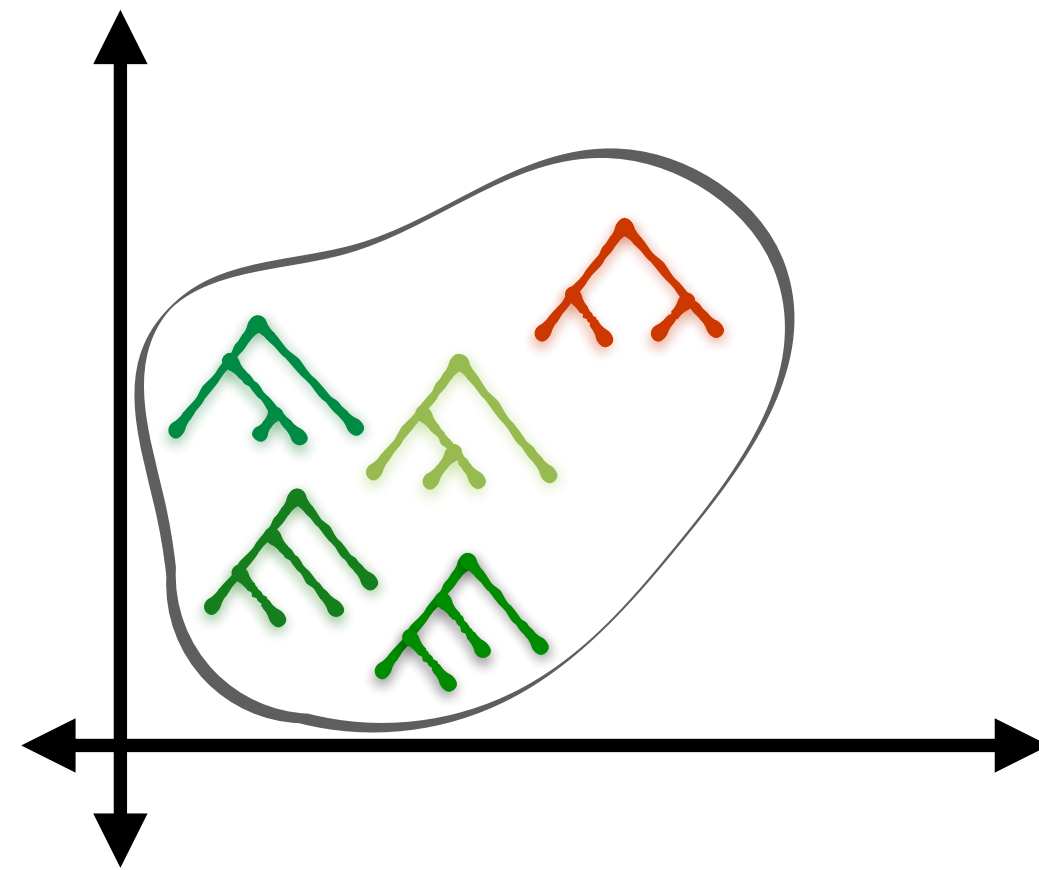
How to identify such gene trees systematically?
Even when the signal is subtle...



[Rivas-González et al., 2026]

Challenge: Distinguishing ILS from real alternative processes

Compare gene trees to each other?
e.g., [PhylteR - Devienne et al. 2023]



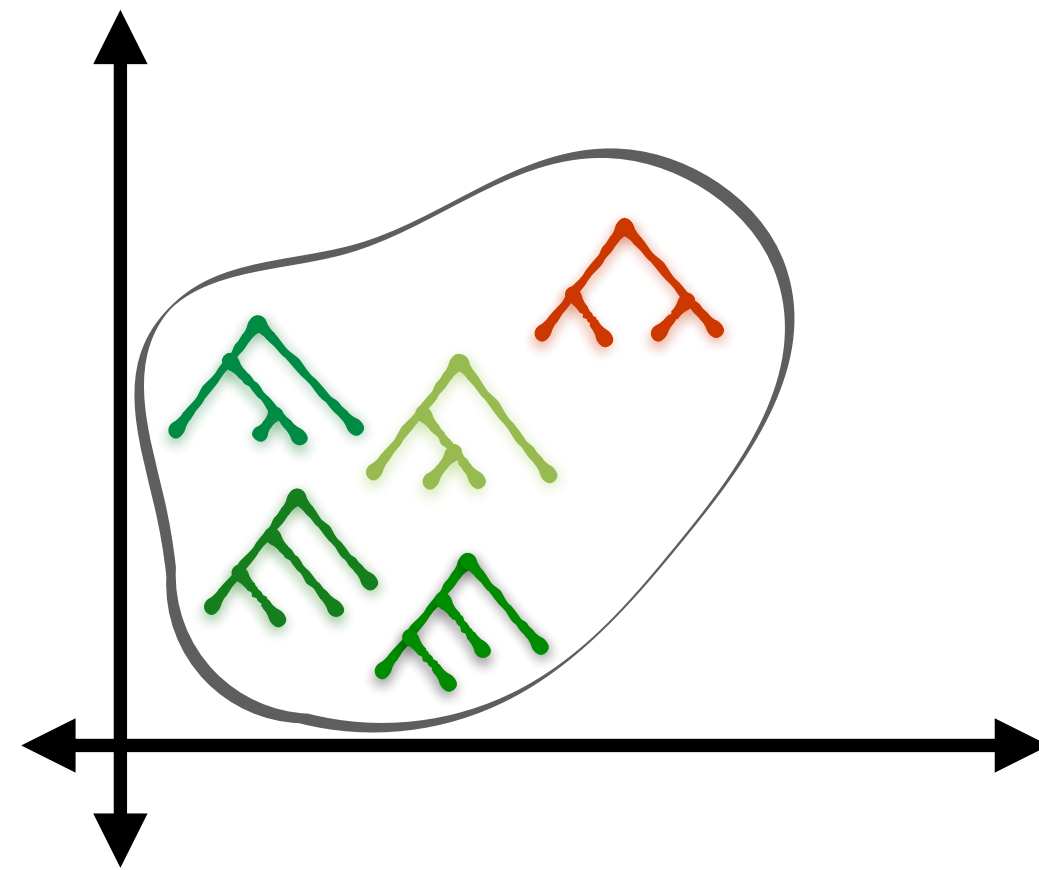
Difficult to make a conclusion if you look at a single gene tree at a time...

Likelihoods of gene trees under $MSC(T)$?



Challenge: Distinguishing ILS from real alternative processes

Compare gene trees to each other?
e.g., [PhylteR - Devienne et al. 2023]



Likelihoods of gene trees under $\text{MSC}(T)$?



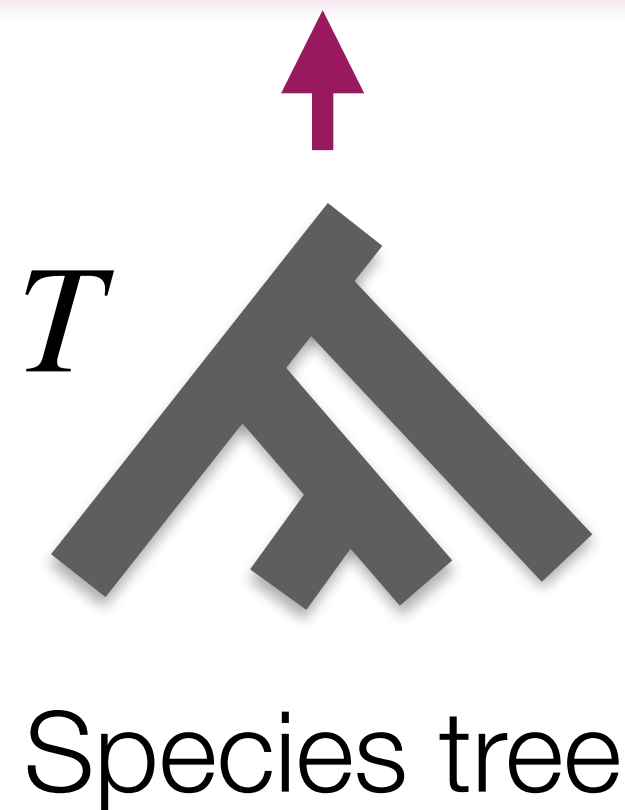
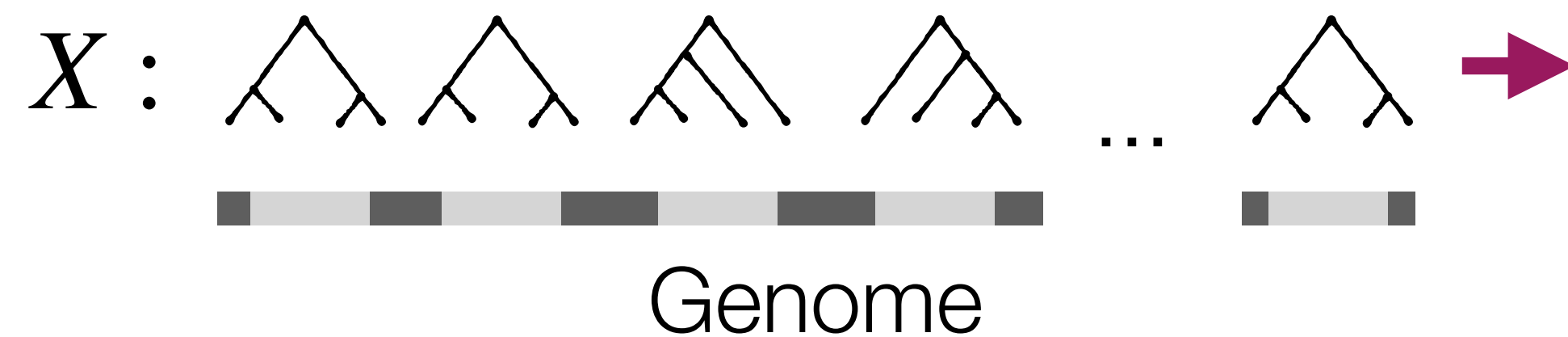
Difficult to make a conclusion if you look at a single gene tree at a time...

Idea: treat gene trees as a sequence & incorporate positions!

Phlag: the problem & the framework

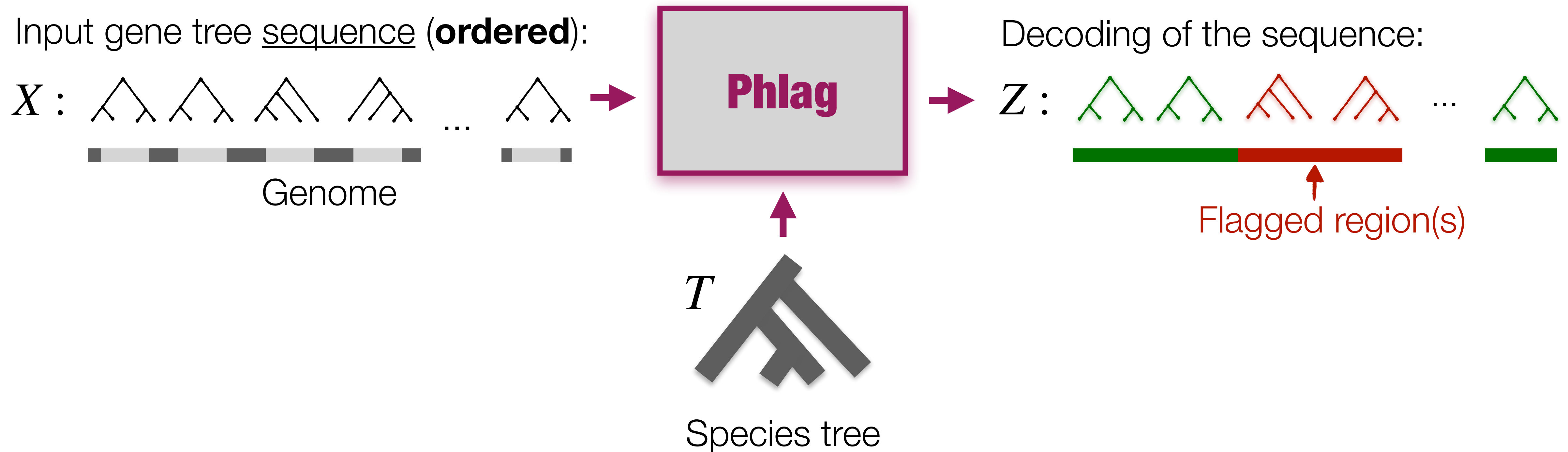
Automatic segmentation of regions generated under $\text{MSC}(T)$ or an **alternative process**

Input gene tree sequence (**ordered**):

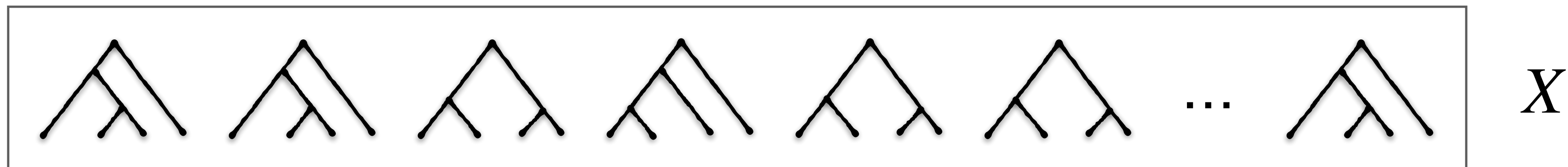


Phlag: the problem & the framework

Automatic segmentation of regions generated under $\text{MSC}(T)$ or an **alternative process**



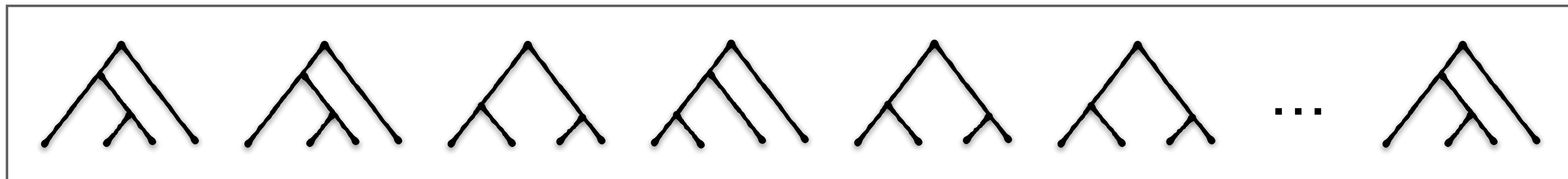
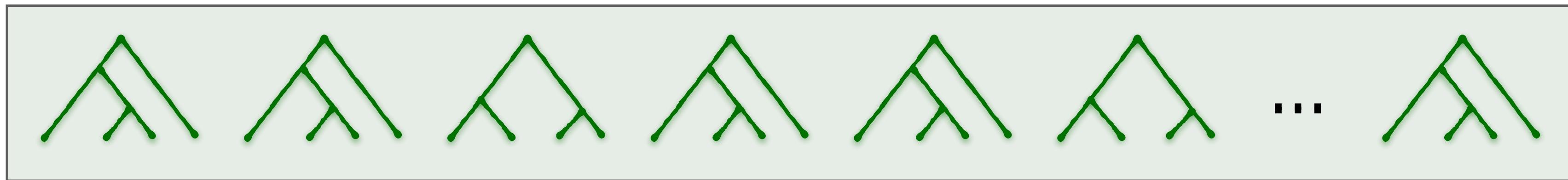
Detecting deviations through a mixture process



Detecting deviations through a mixture process

Marginal distribution: MSC parameterized by the species tree

MSC(T):

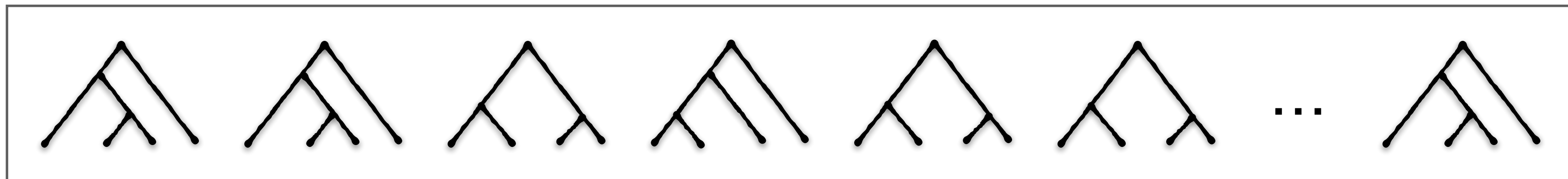
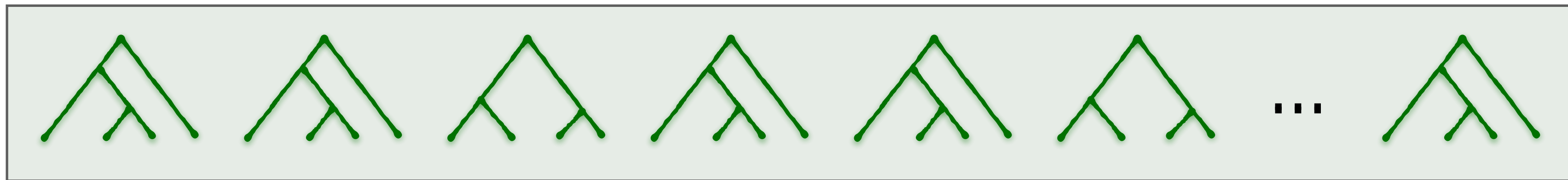


X

Detecting deviations through a mixture process

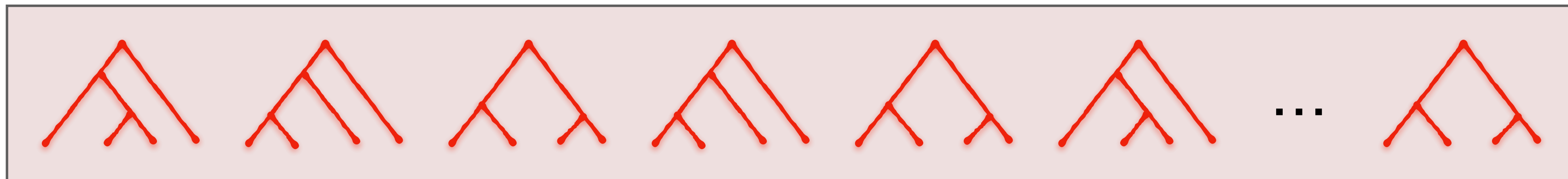
Marginal distribution: MSC parameterized by the species tree

MSC(T):



X

**Alternative
process:**



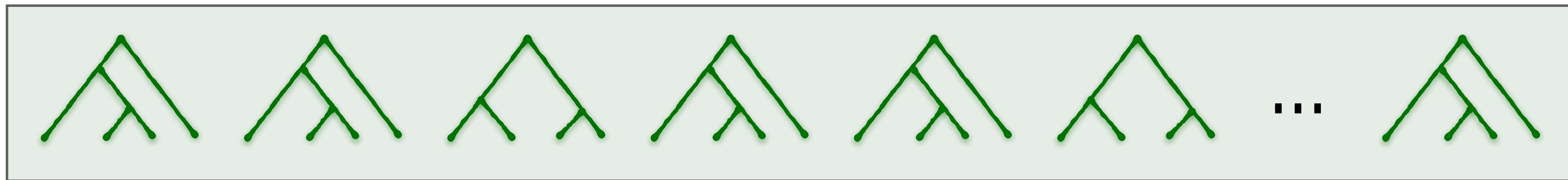
Marginal distribution: unknown model (e.g., gene flow, different pop. size)

Detecting deviations through a mixture process

Inferring latent variables $Z_1, \dots, Z_N \in \{0,1\}$ for gene trees $X_1, \dots, X_n$

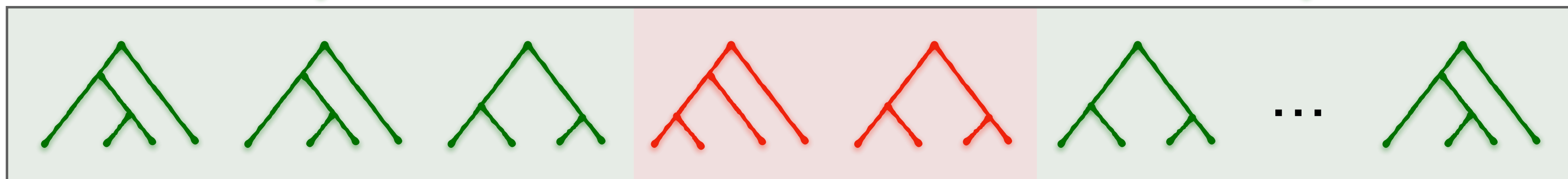
Marginal distribution: MSC parameterized by the species tree

MSC(T):



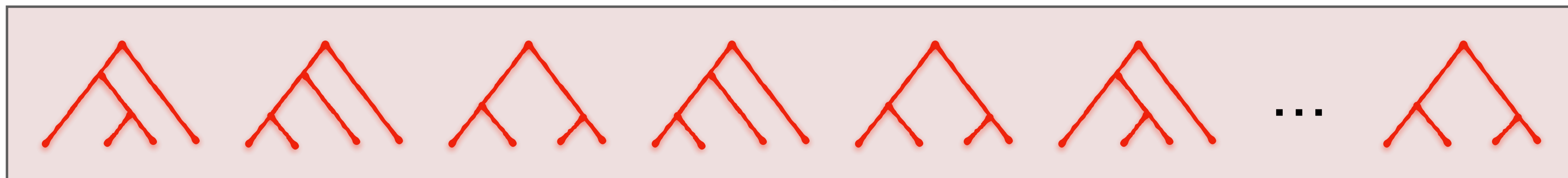
$Z_i = 0$
 $P(X_i | \theta_0)$

Mixture process:



X

Alternative process:

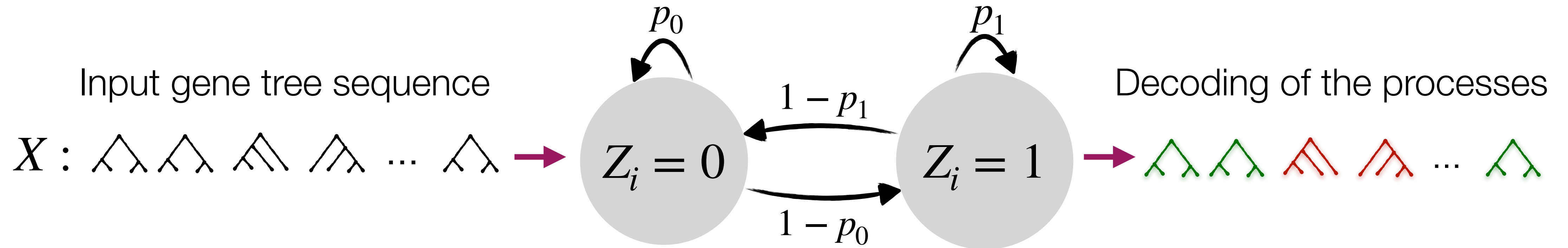


$Z_i = 1$
 $P(X_i | \theta_1)$

Marginal distribution: unknown model (e.g., gene flow, different pop. size)

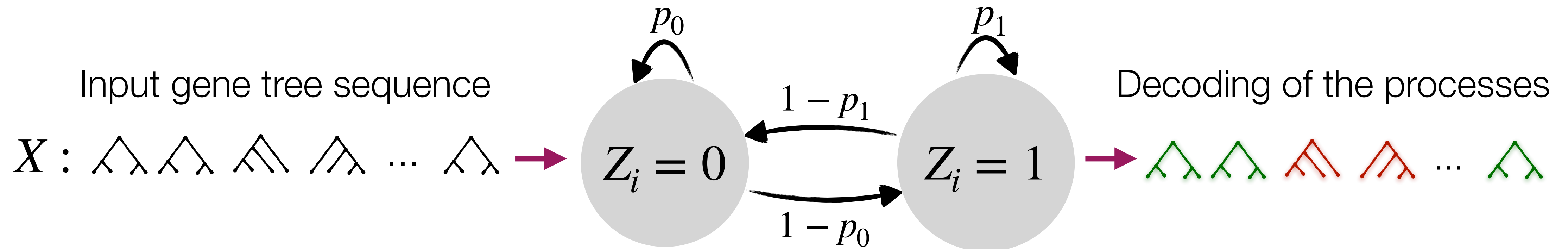
Phlag: a hidden Markov model

A two-state HMM if we reasonably assume Z is to be Markovian:



Phlag: a hidden Markov model

A two-state HMM if we reasonably assume Z is to be Markovian:

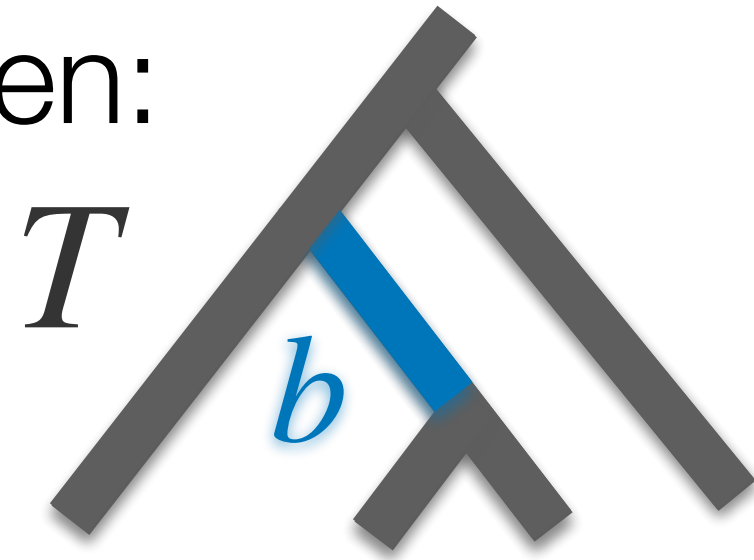


Should we directly use gene trees as our emissions?

It is possible [Wu 2012], but **not scalable for MSC** & **unknown for the alternative...**

An alternative to gene tree likelihoods

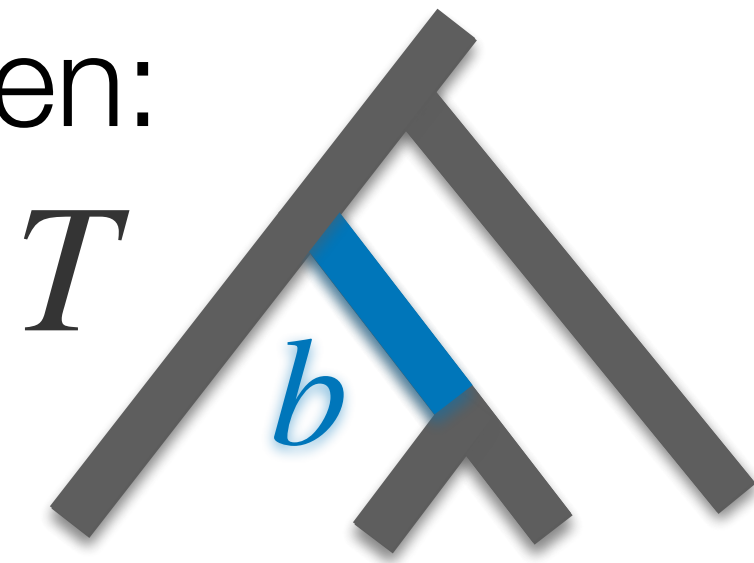
Given:



Focus on some **focal branches**: $(b_1, \dots, b_m)$
where the impact is potentially more palpable

An alternative to gene tree likelihoods

Given:

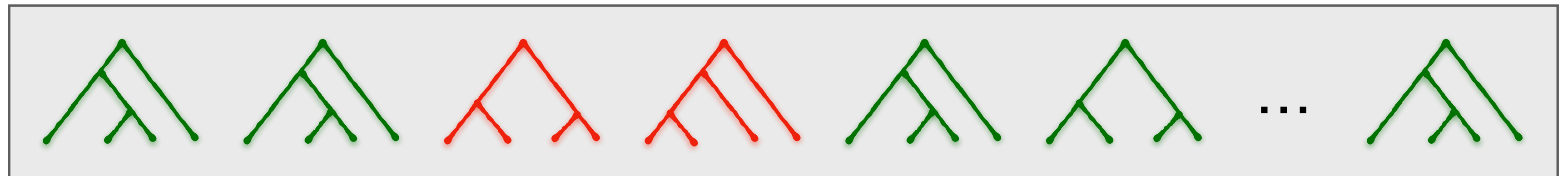


Focus on some **focal branches**: $(b_1, \dots, b_m)$
where the impact is potentially more palpable

Use a **scalable test statistic** $F(X_i, b)$ instead of gene tree likelihoods $P(X_i | \theta_{Z_i})$:

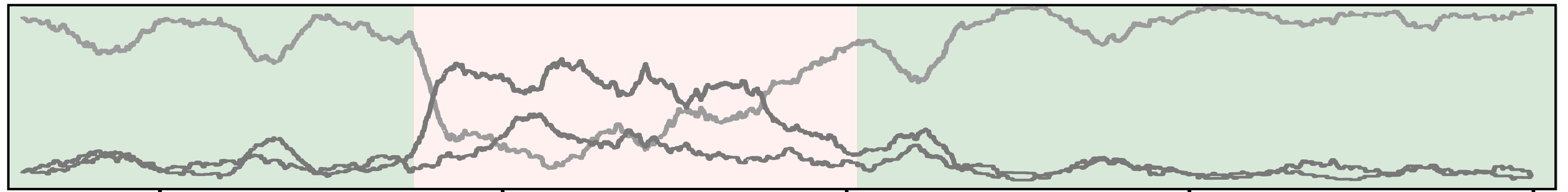
Mixture process:

(X_i, Z_i)



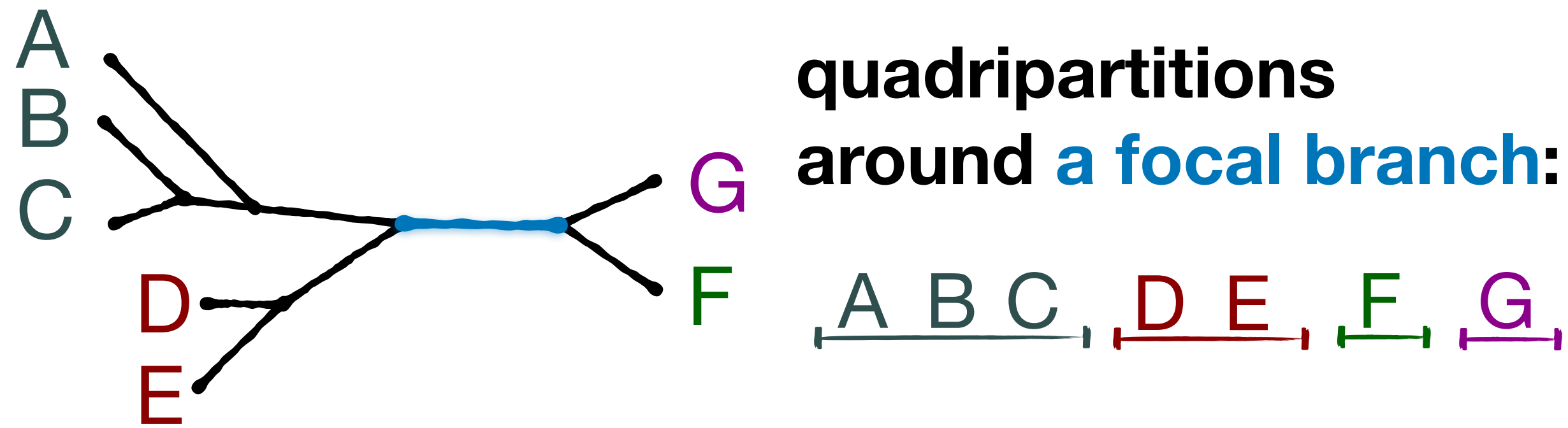
Emissions:

$$F(X_i, b_j) = Y_i^{(j)}$$

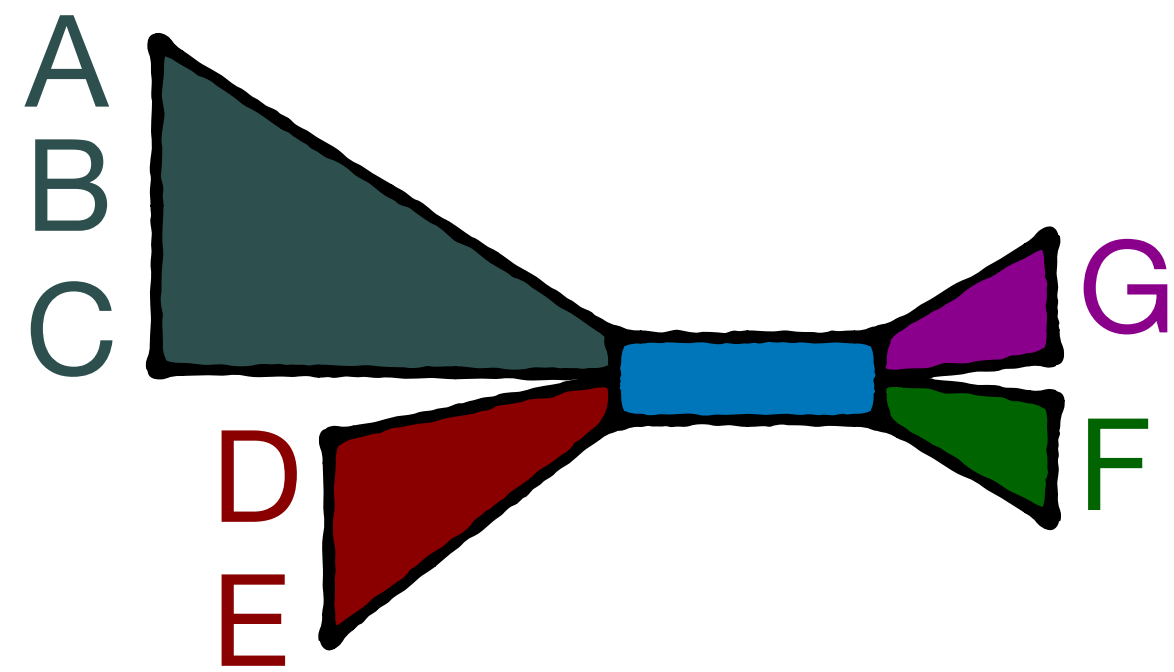


Position on chromosome

Quadripartition quartet scores as emissions



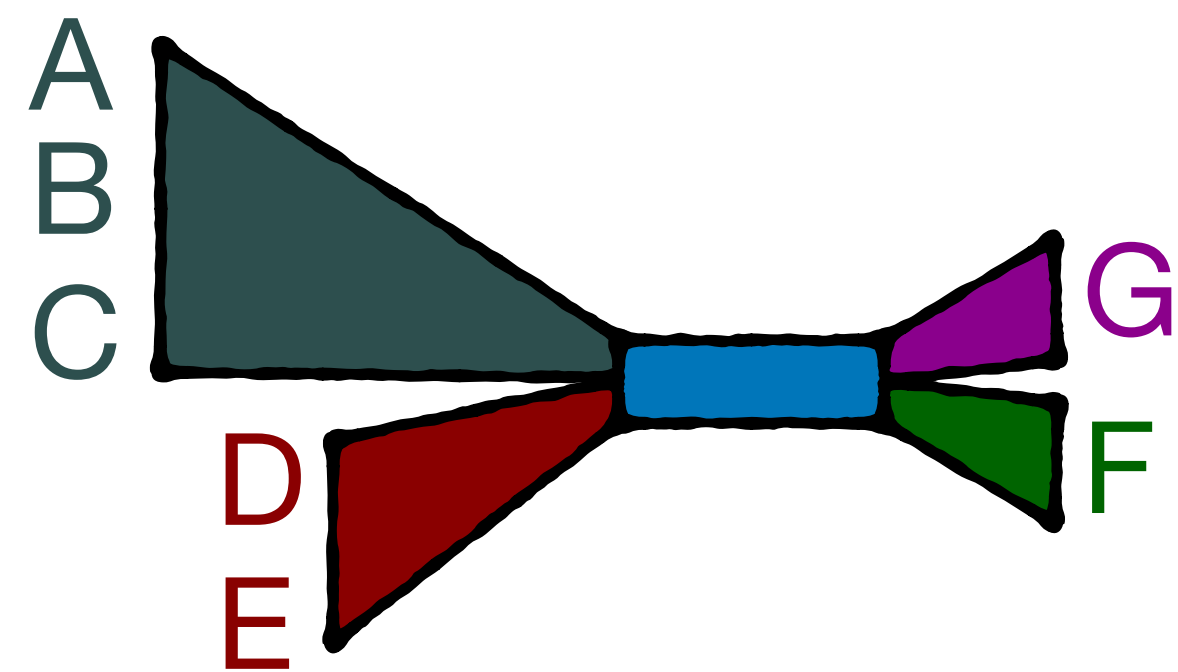
Quadripartition quartet scores as emissions



quadripartitions
around **a focal branch**:

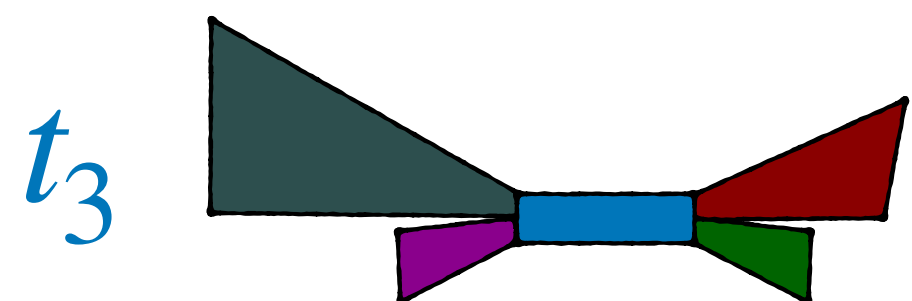
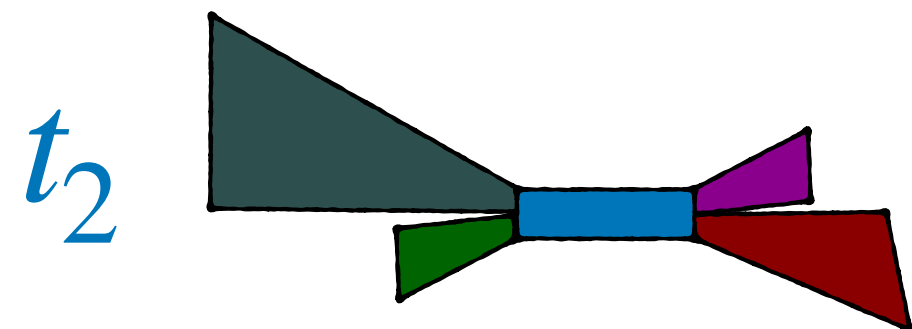
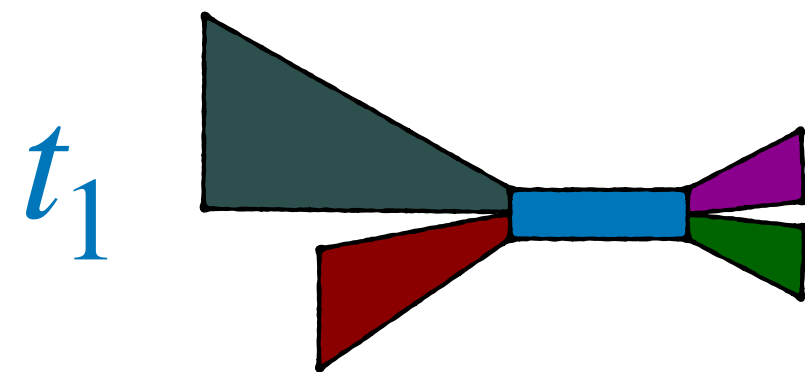
A B C, D E, F, G

Quadripartition quartet scores as emissions

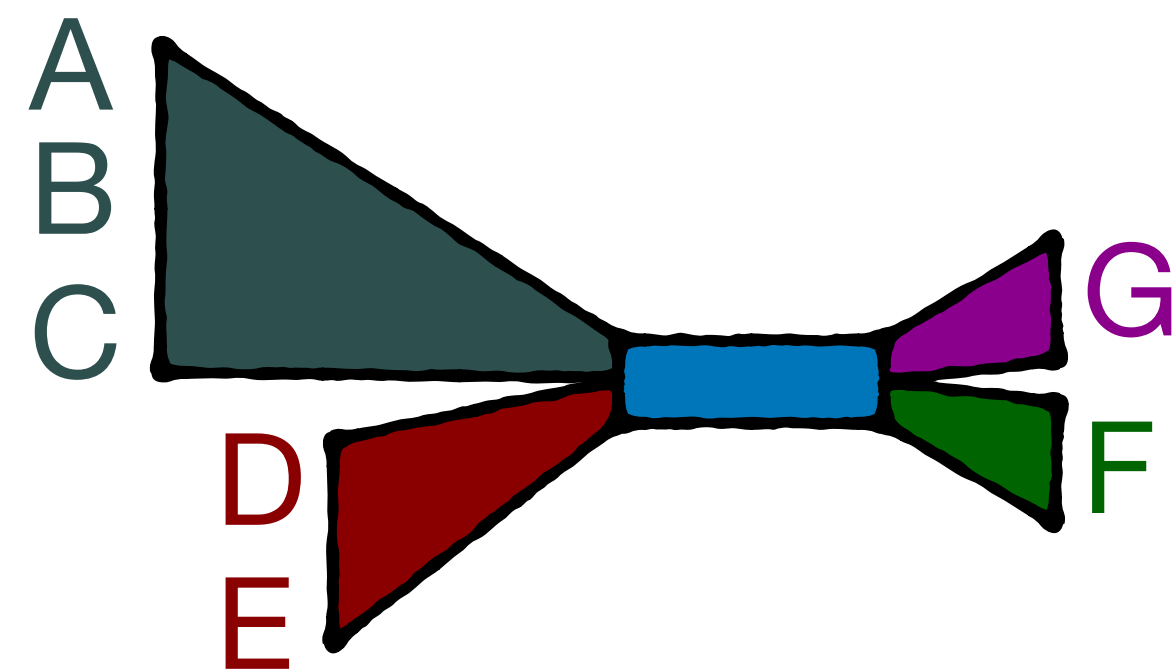


quadripartitions
around **a focal branch**:

A B C, D E, F, G



Quadripartition quartet scores as emissions



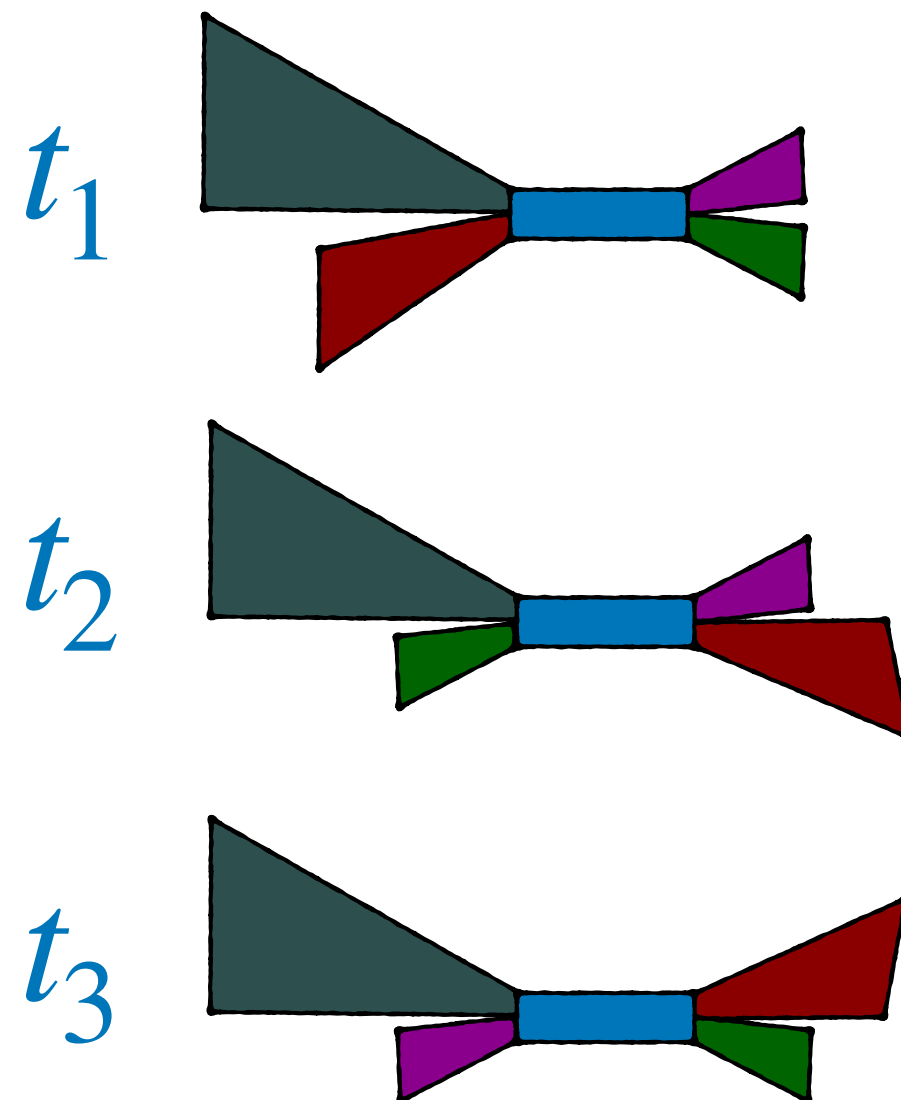
quadripartitions
around **a focal branch**:

A B C, D E, F, G

normalized quartet
frequencies



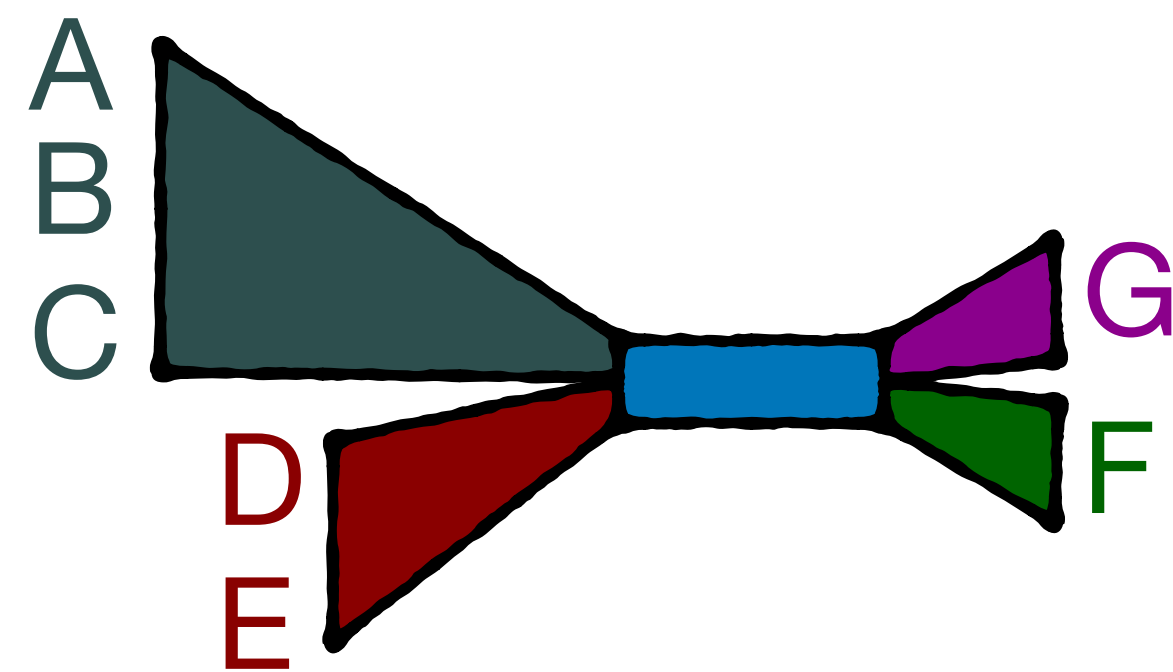
	X_1	X_2	...	X_N
t_1	0.66	0.66	...	0.22
t_2	0.16	0.16	...	0.39
t_3	0.16	0.16	...	0.39



QQS values:

- add up to one — **compositional**
- can be calculated in $\mathcal{O}(n \log^2(n))$
- Under MSC, related to **branch lengths**

Quadripartition quartet scores as emissions



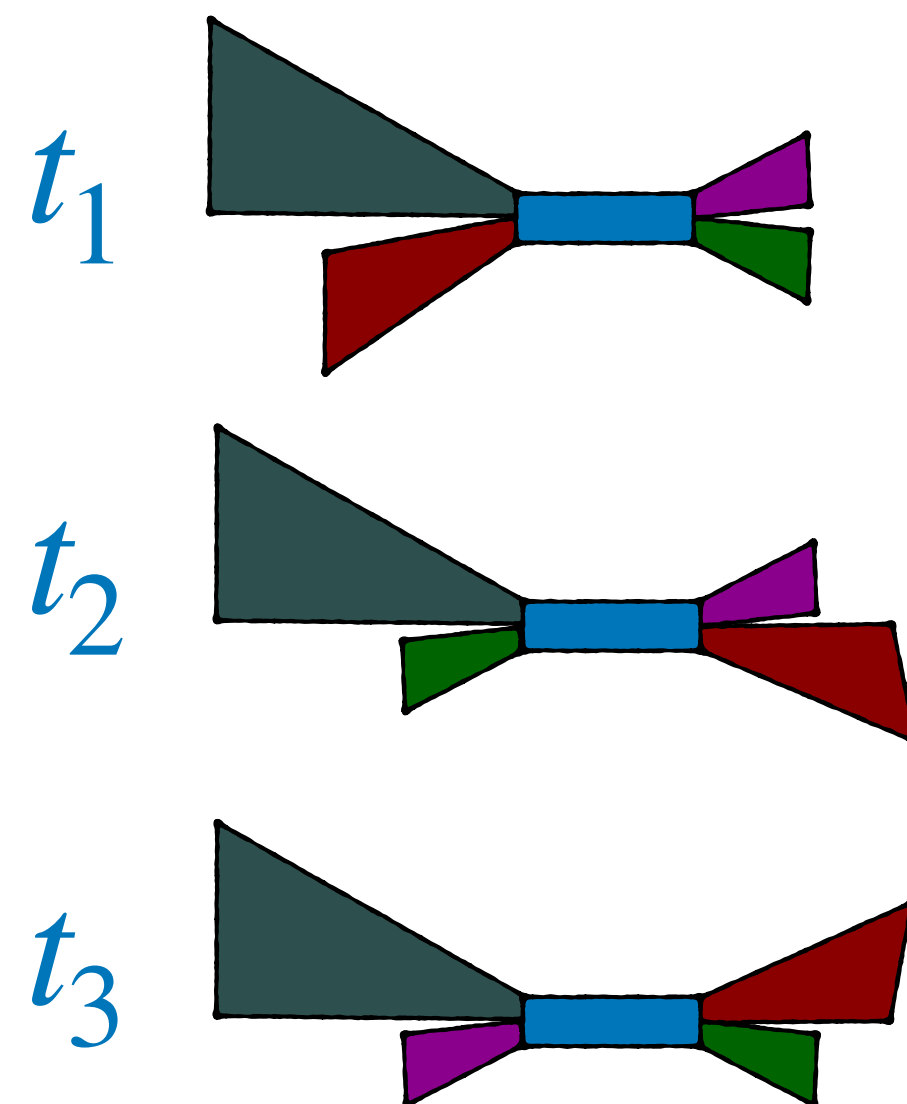
quadripartitions
around **a focal branch**:

A B C, D E, F, G

normalized quartet
frequencies



	X_1	X_2	...	X_N
t_1	0.66	0.66	...	0.22
t_2	0.16	0.16	...	0.39
t_3	0.16	0.16	...	0.39



QQS values:

- add up to one — **compositional**
- can be calculated in $\mathcal{O}(n \log^2(n))$
- Under MSC, related to **branch lengths**

**What about their
distribution?**

*We parameterized
and approximated:*

$$t_1 > t_2 > t_3$$

$$t_1 > t_3 > t_2$$

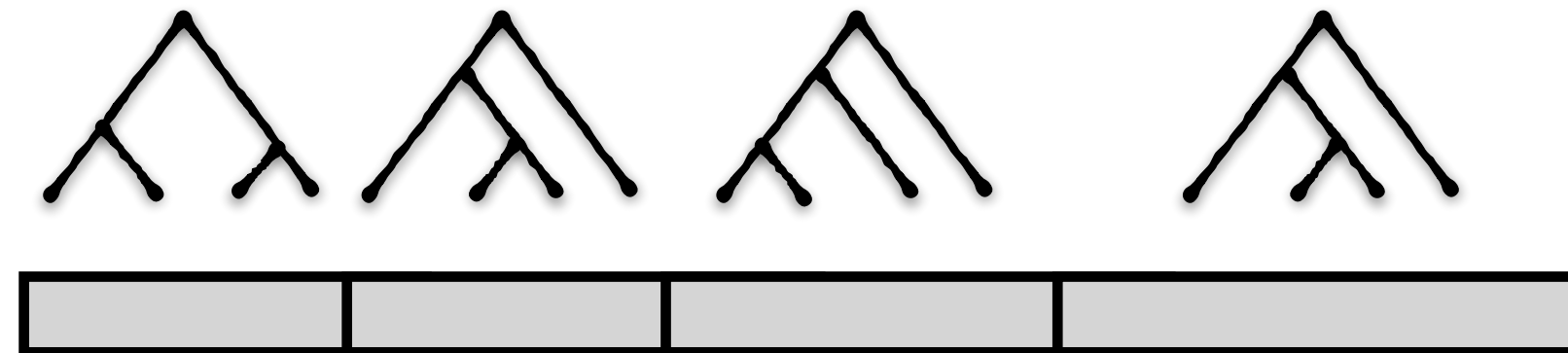
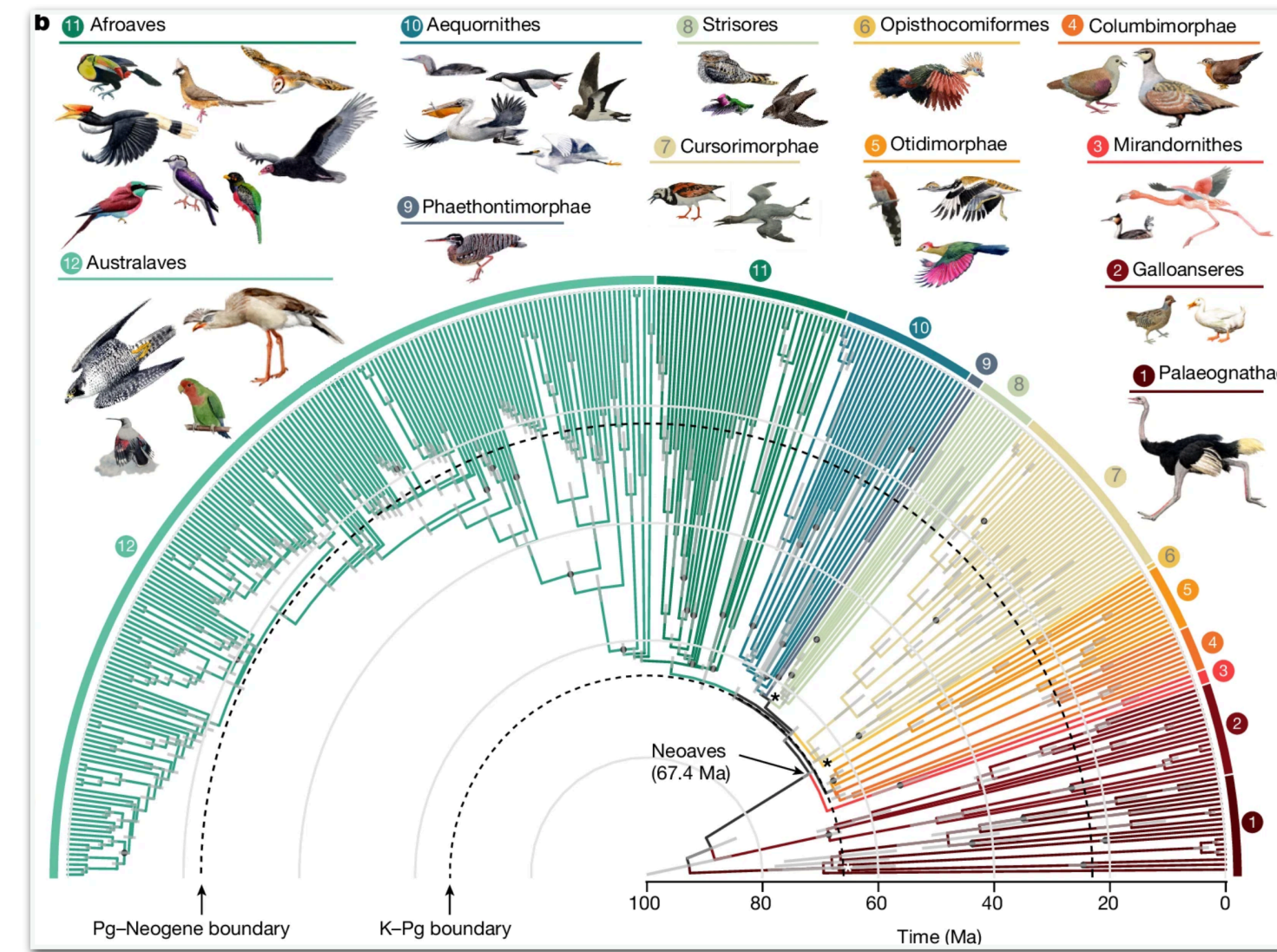
$$t_2 > t_1 > t_3$$

...

Simulations with recombination and rate variation

Demographic model:
Stiller et al. 2024 (363 taxa)

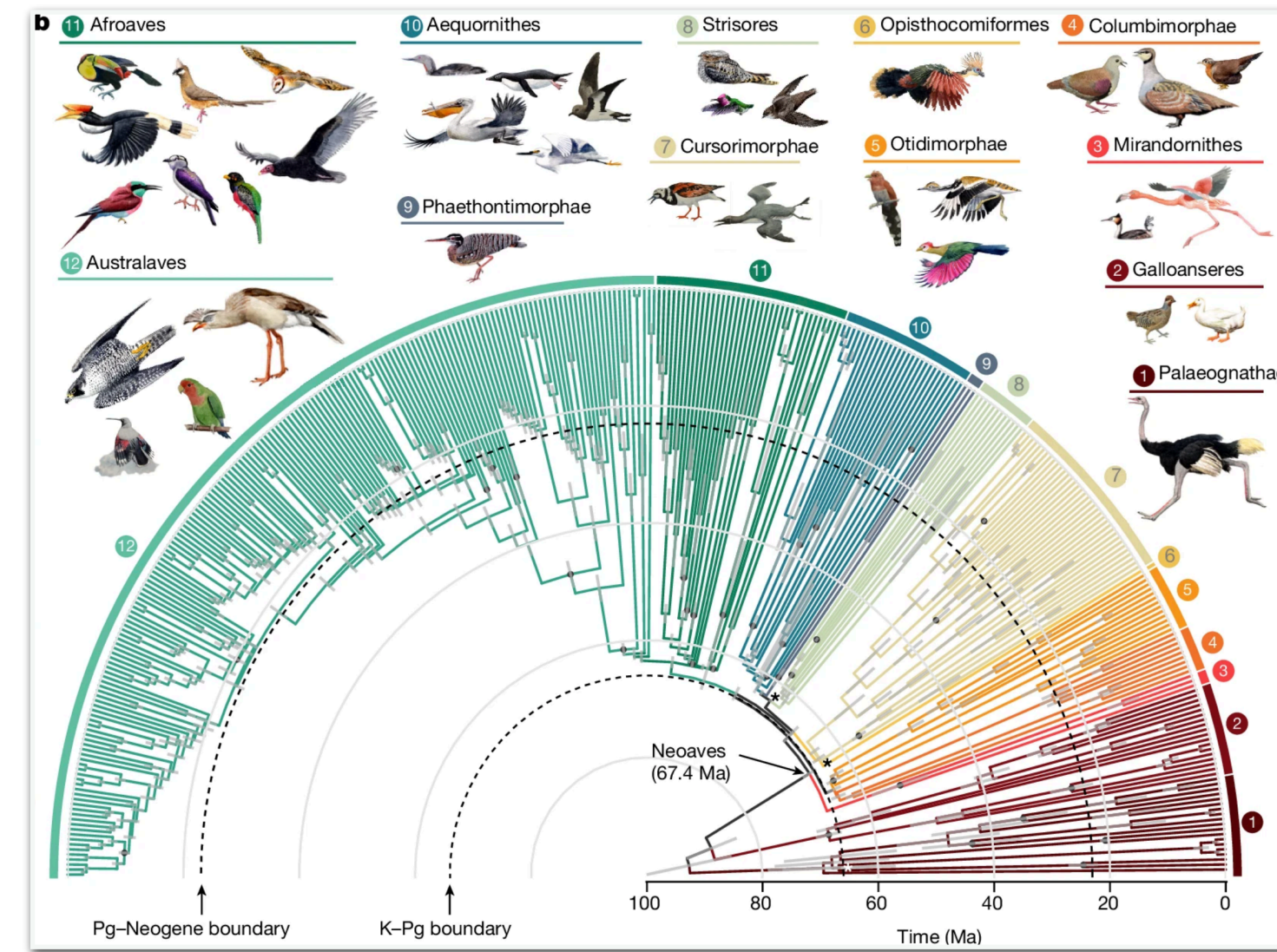
Genealogies under
coalescent & recombination
(msprime)



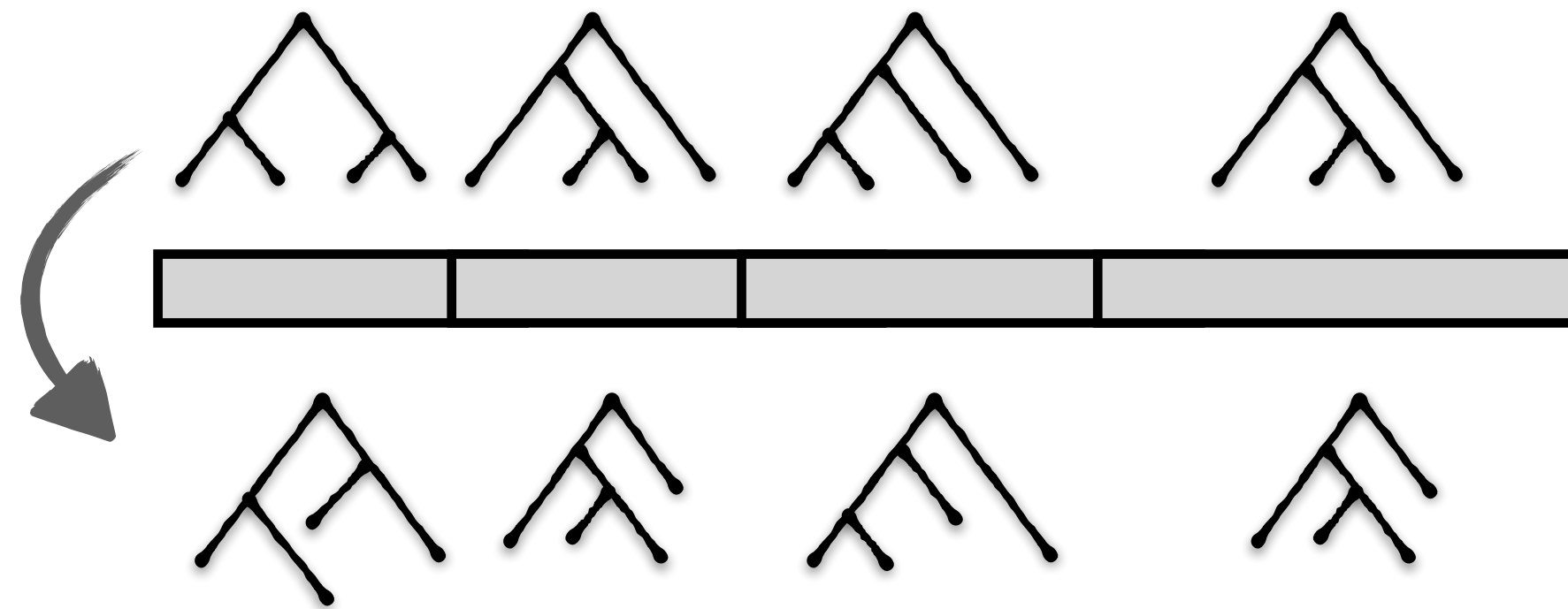
Simulations with recombination and rate variation

Demographic model:
Stiller et al. 2024 (363 taxa)

Genealogies under
coalescent & recombination
(msprime)



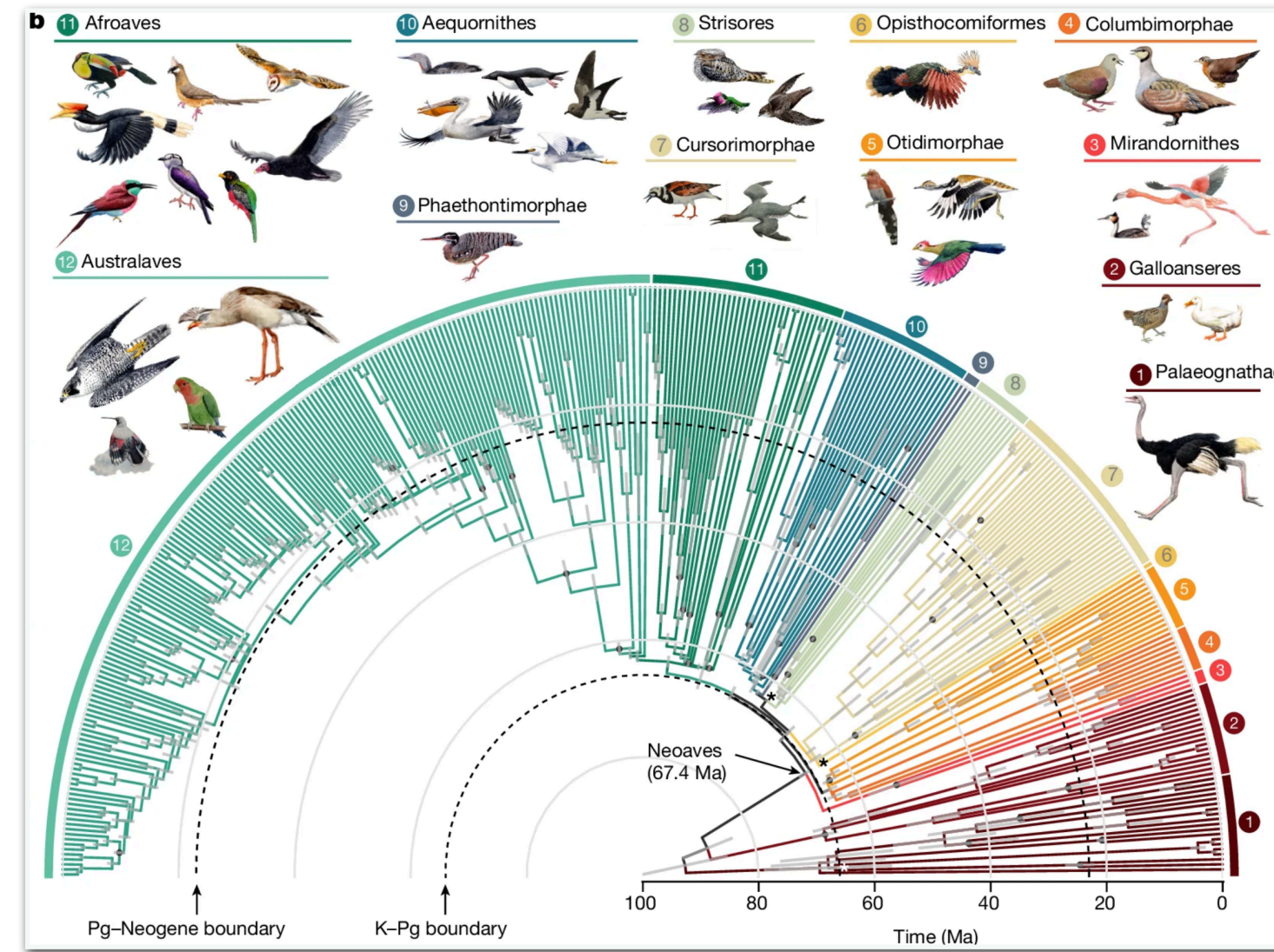
Add rate
heterogeneity



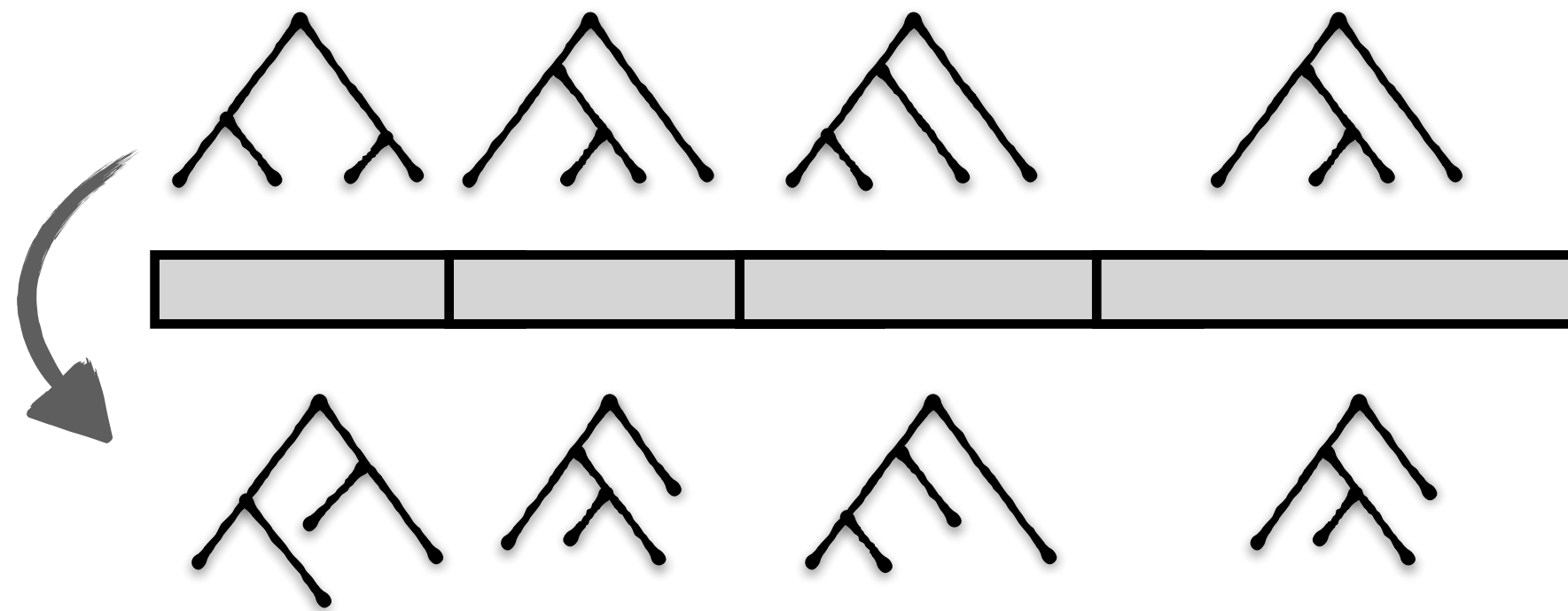
Simulations with recombination and rate variation

Demographic model:
Stiller et al. 2024 (363 taxa)

Genealogies under
coalescent & recombination
(msprime)



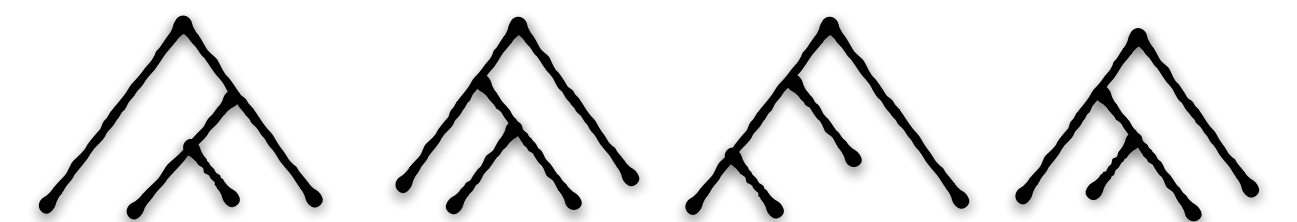
Add rate
heterogeneity



Simulate **6Mbp**
alignments

Estimate
gene trees

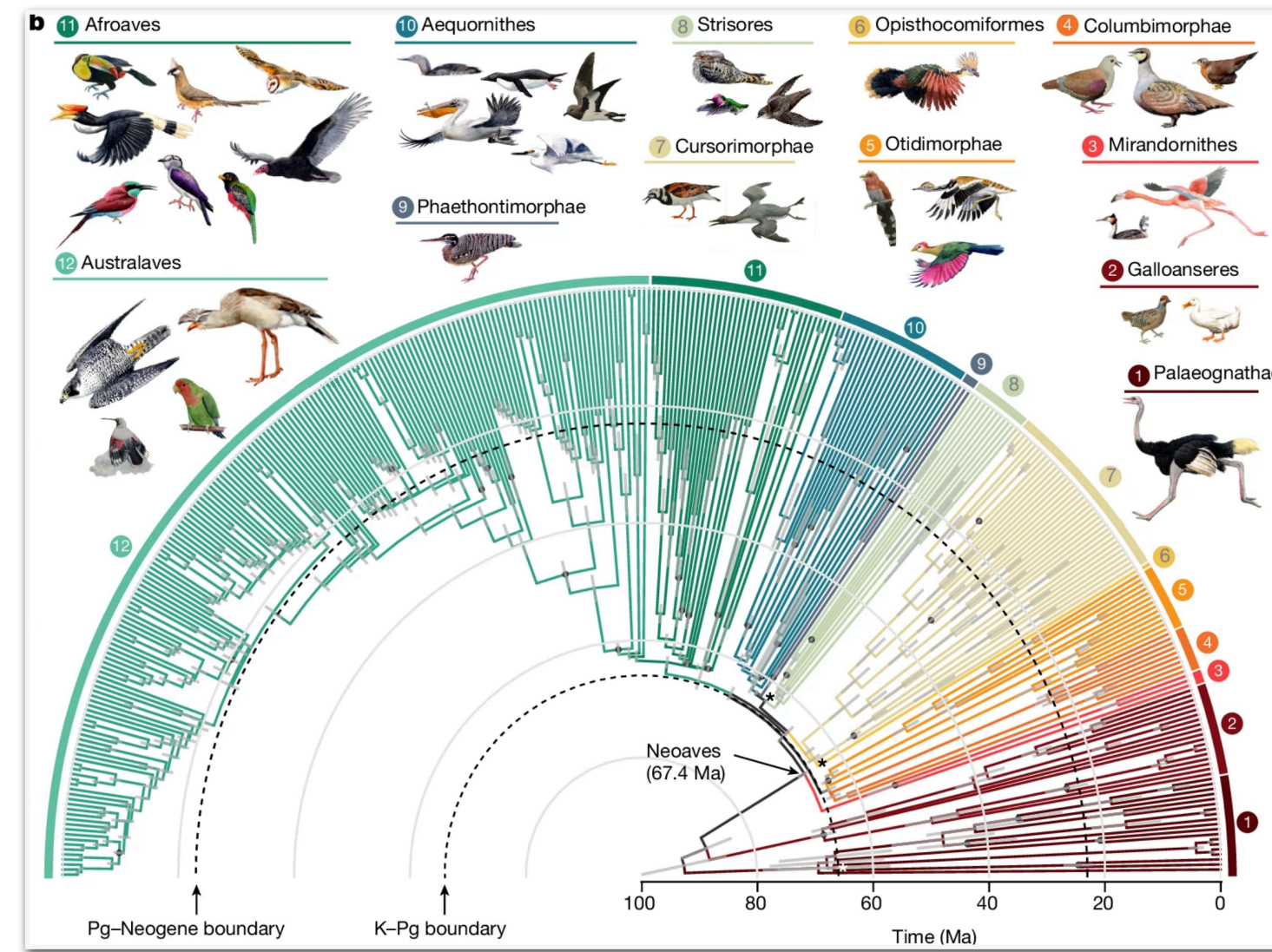
```
ATTCGTGC...
ATCCGTCG...
TTCCGTCG...
ATTCGTGC...
```



Simulations with recombination and rate variation

Demographic model:
Stiller et al. 2024 (363 taxa)

Genealogies under
coalescent & recombination
(`msprime`)



Alternative process:

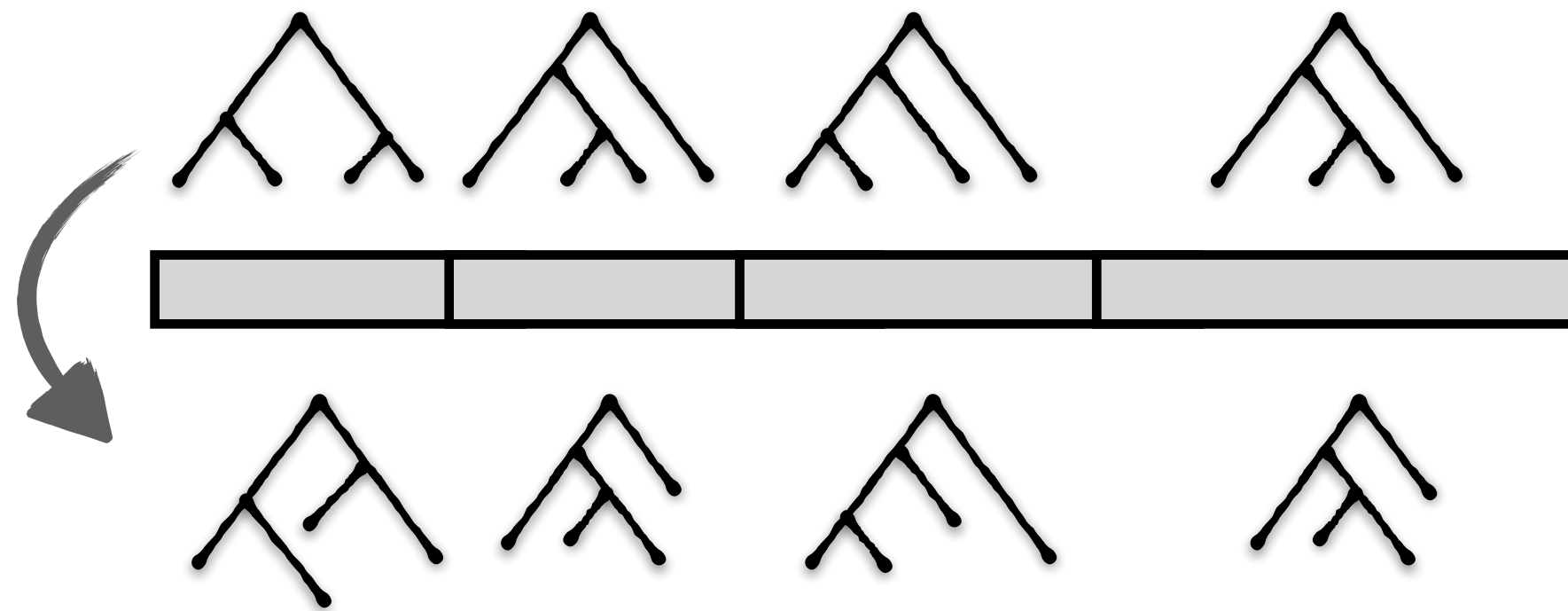
An admixture event

Change in parameters ($\uparrow$ or $\downarrow$)

► Population size: 10X

► Recombination rate: 1000X

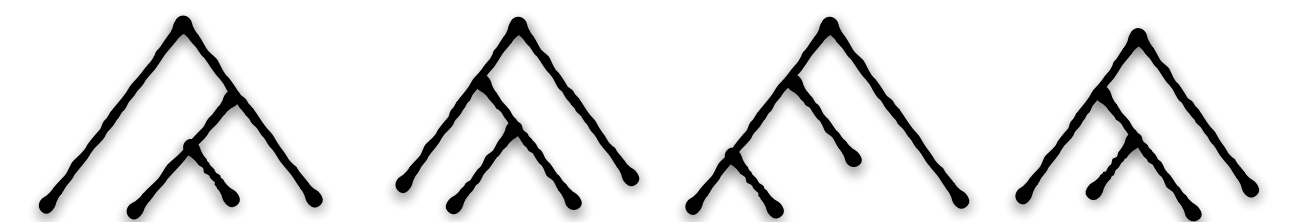
Add rate
heterogeneity



Simulate **6Mbp**
alignments

Estimate
gene trees

```
ATTCGTGC...
ATCCGTCG...
TTCCGTCG...
ATTCGTGC...
```



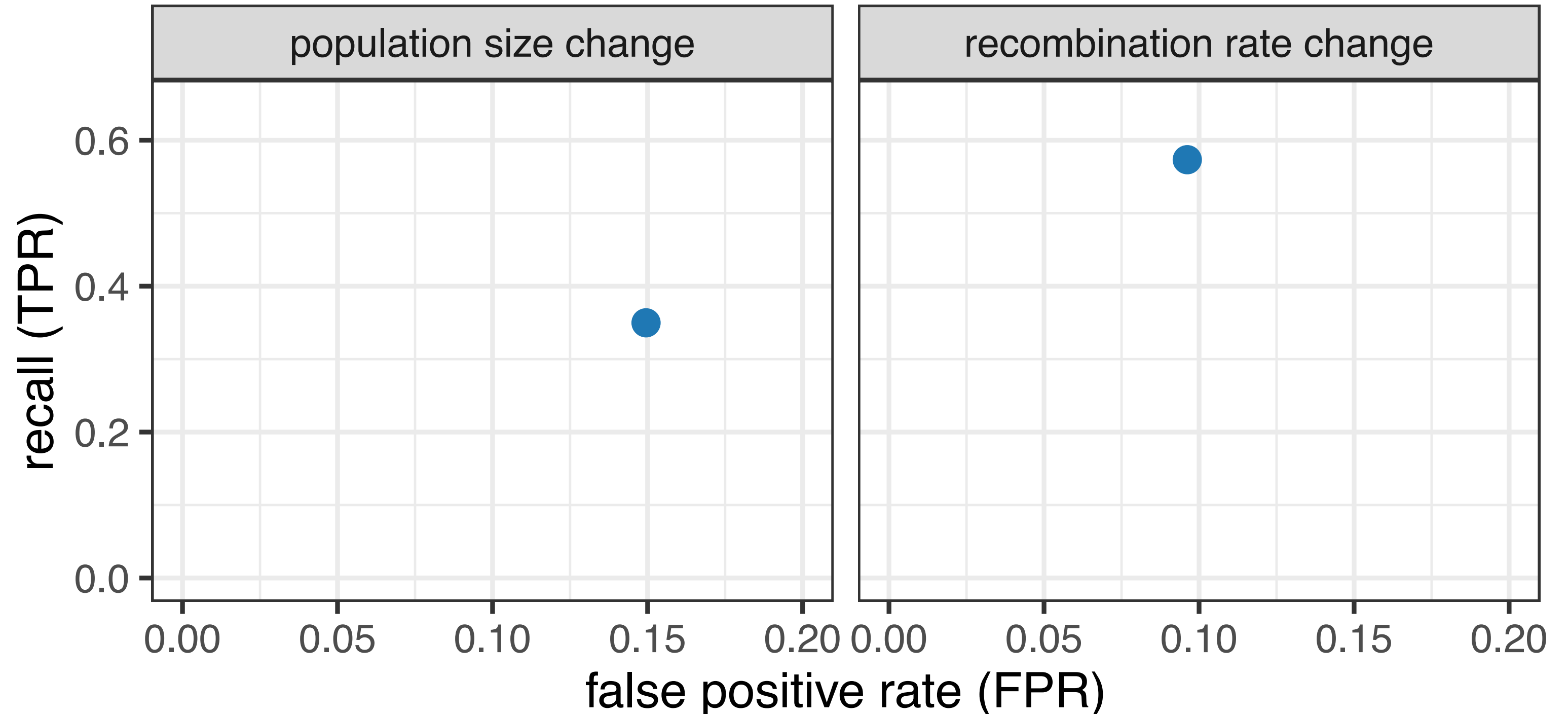
Segmentation using QQS-based emissions & a vanilla HMM

one region spanning 5%-25% of the data

10× ↑ or ↓

1000× ↑ or ↓

- Parameters inferred from data
- No explicit modeling of MSC
- T is not incorporated



Segmentation using QQS-based emissions & a vanilla HMM

one region spanning 5%-25% of the data

10× ↑ or ↓

1000× ↑ or ↓

- Parameters inferred from data
- No explicit modeling of MSC
- T is not incorporated



Incorporating priors for emissions and transitions

Preventing rapid state switches & favoring continuity:

A ***sticky HMM*** with significantly higher probabilities for the self-transitions!

Incorporating priors for emissions and transitions

Preventing rapid state switches & favoring continuity:

A *sticky HMM* with significantly higher probabilities for the self-transitions!

Incorporating the species tree and MSC:

Given correct topology
but not branch lengths:



Incorporating priors for emissions and transitions

Preventing rapid state switches & favoring continuity:

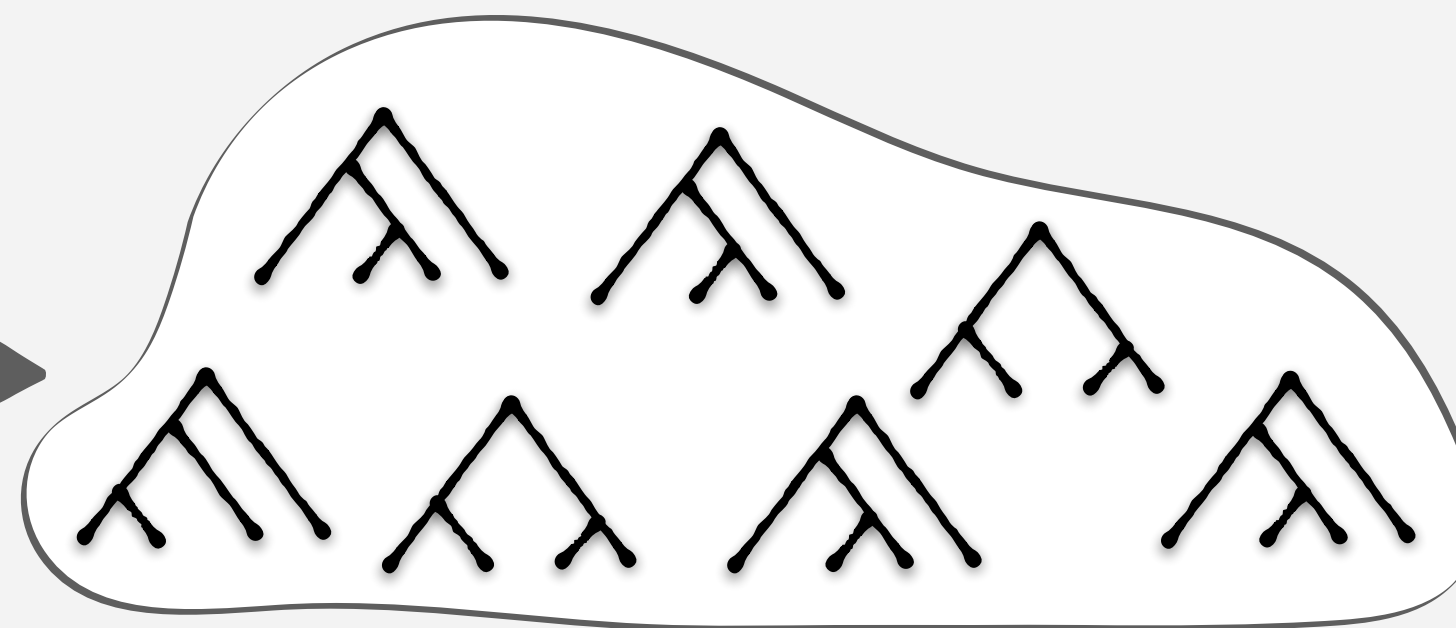
A **sticky HMM** with significantly higher probabilities for the self-transitions!

Incorporating the species tree and MSC:

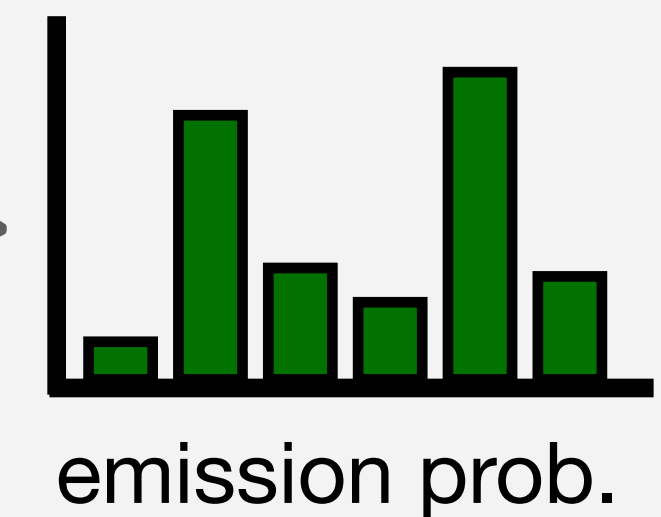
Given correct topology
but not branch lengths:



simulations



A prior for the emission
prob. of the MSC state

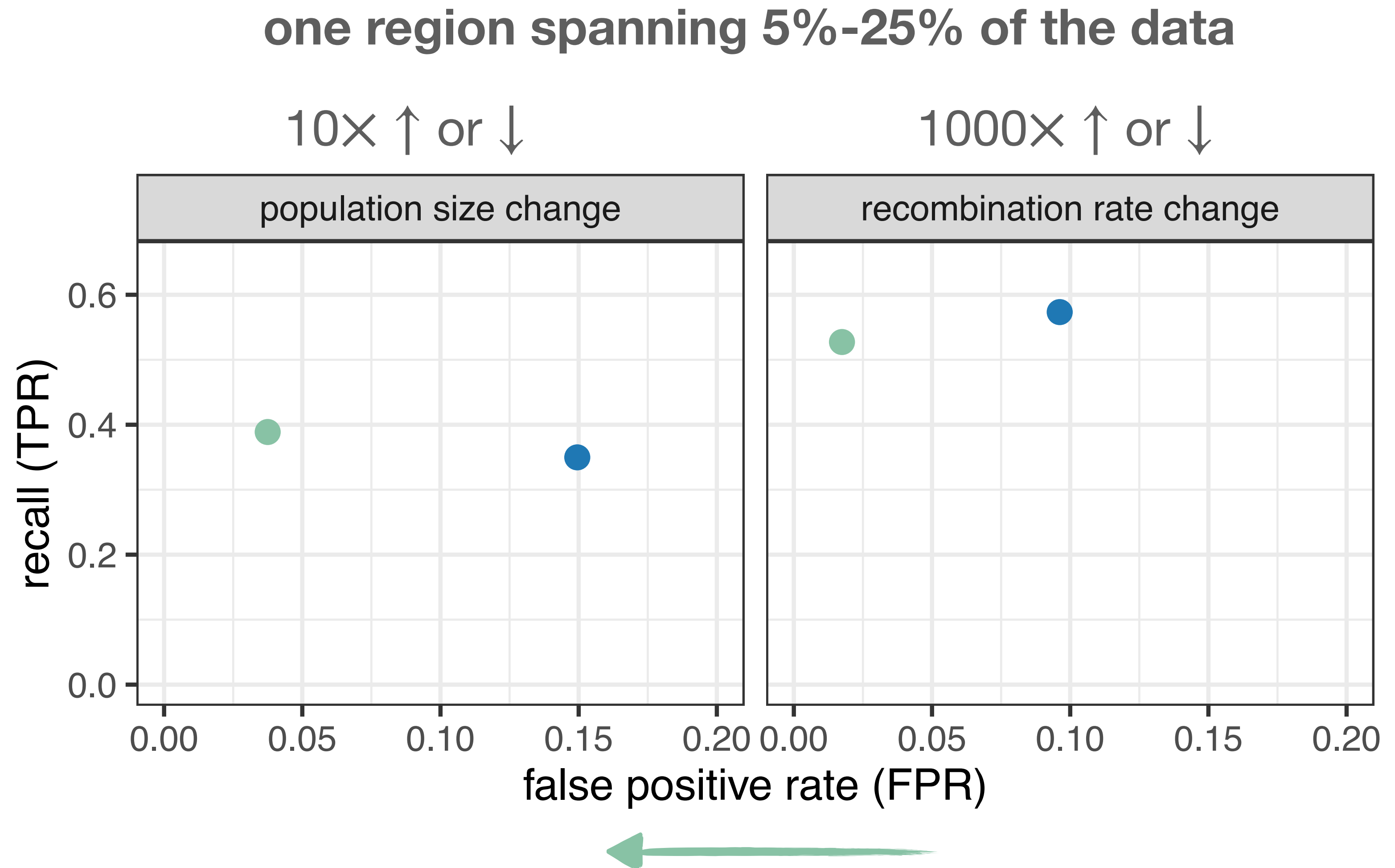


Empirical Bayes: update branch lengths during EM

Priors help improve recall and reduce false positives

Sticky transitions: lower FP

- Vanilla HMM
- Sticky HMM

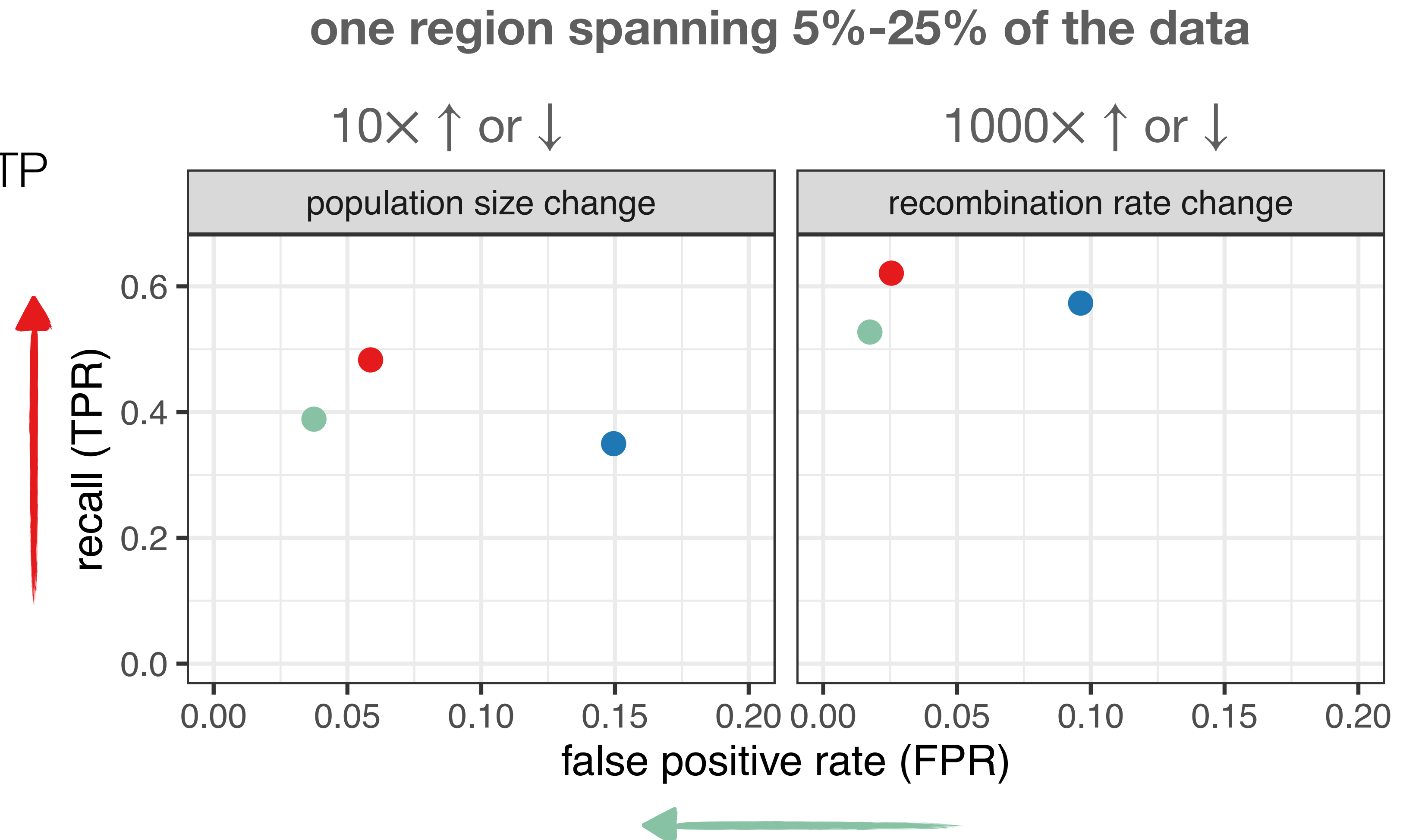


Priors help improve recall and reduce false positives

Sticky transitions: lower FP

MSC-informed emissions: higher TP

- Vanilla HMM
- Sticky HMM
- Sticky HMM w/ MSC-priors



Priors help improve recall and reduce false positives

Sticky transitions: lower FP

MSC-informed emissions: higher TP

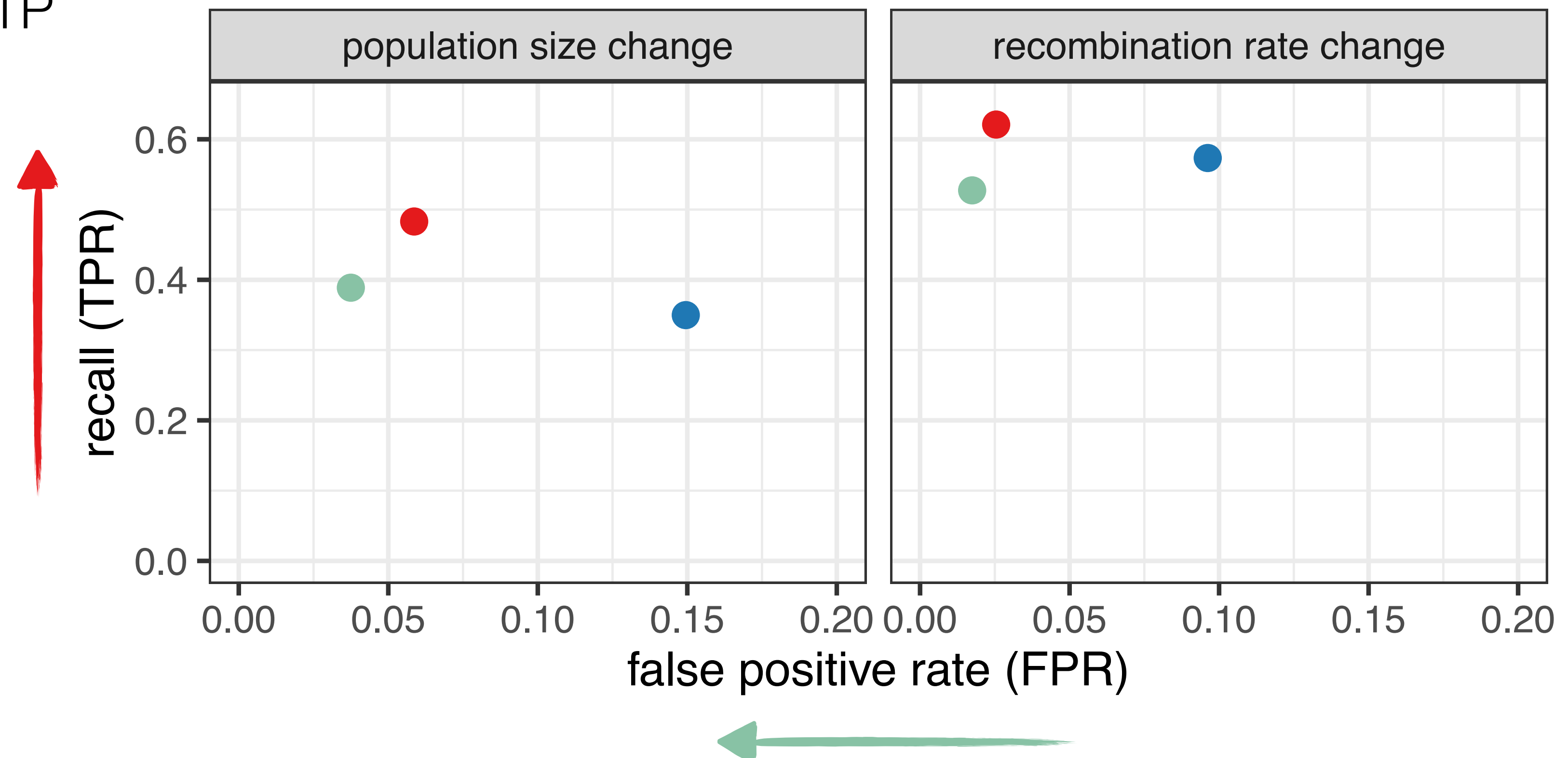
- Vanilla HMM
- Sticky HMM
- Sticky HMM w/ MSC-priors

Phlag!

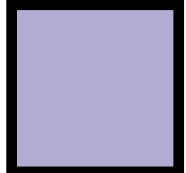
one region spanning 5%-25% of the data

10× ↑ or ↓

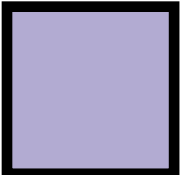
1000× ↑ or ↓



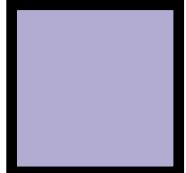
Comparing methods on different types of alternative processes

Incorporating positions of the gene trees:  

Treating gene trees independently: 

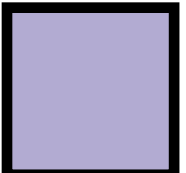
-  **PhylteR:** deviations from gene-specific distance matrices
-  **Moving average of QQS & z -score based outlier detection**
-  **Phlag**

Comparing methods on different types of alternative processes

Incorporating positions of the gene trees:  

Treating gene trees independently: 

★  **PhylteR**: deviations from gene-specific distance matrices

★  **Moving average of QQS & z -score based outlier detection**

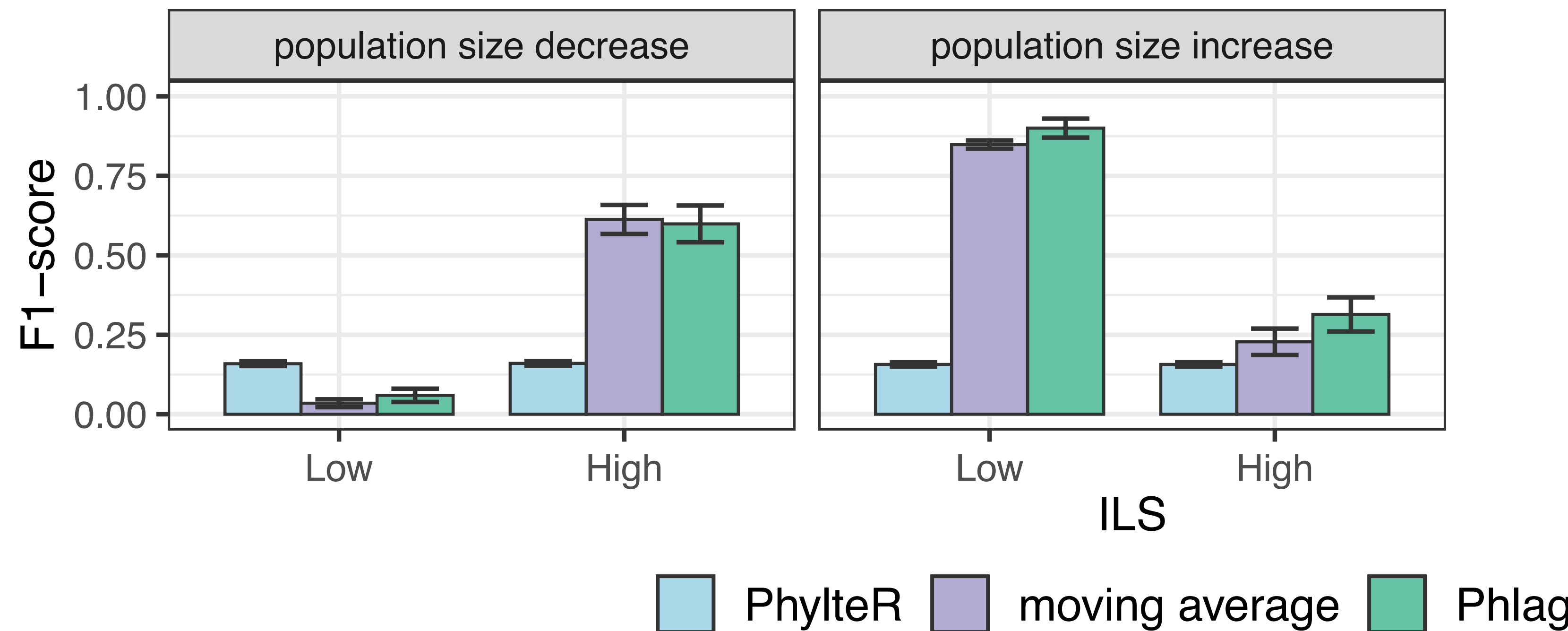
 **Phlag**

★ : tested multiple parameters, reporting best

Dramatic changes model parameters

- **High ILS:** difficult for pop. size increase
- **Low ILS:** difficult for pop. size decrease

Incorporating positions
makes a difference!

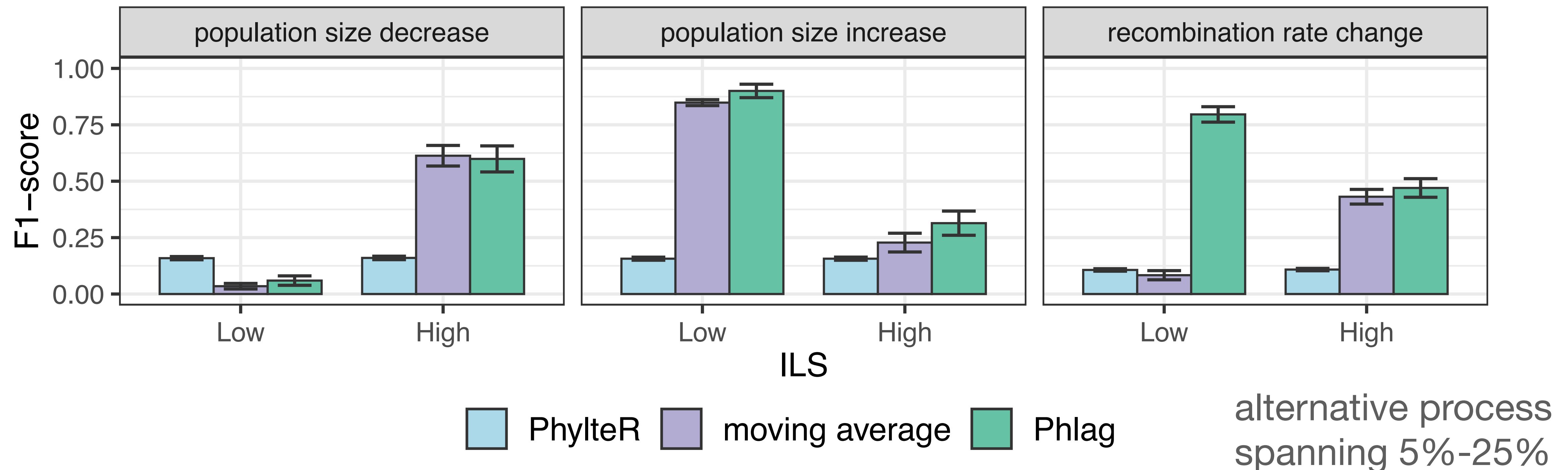


alternative process
spanning 5%-25%

Dramatic changes model parameters

- **High ILS:** difficult for pop. size increase
- **Low ILS:** difficult for pop. size decrease

Incorporating positions
makes a difference!



Detecting deep admixture events

10 events 16-70 Mya

► Gene flow rate: 90%

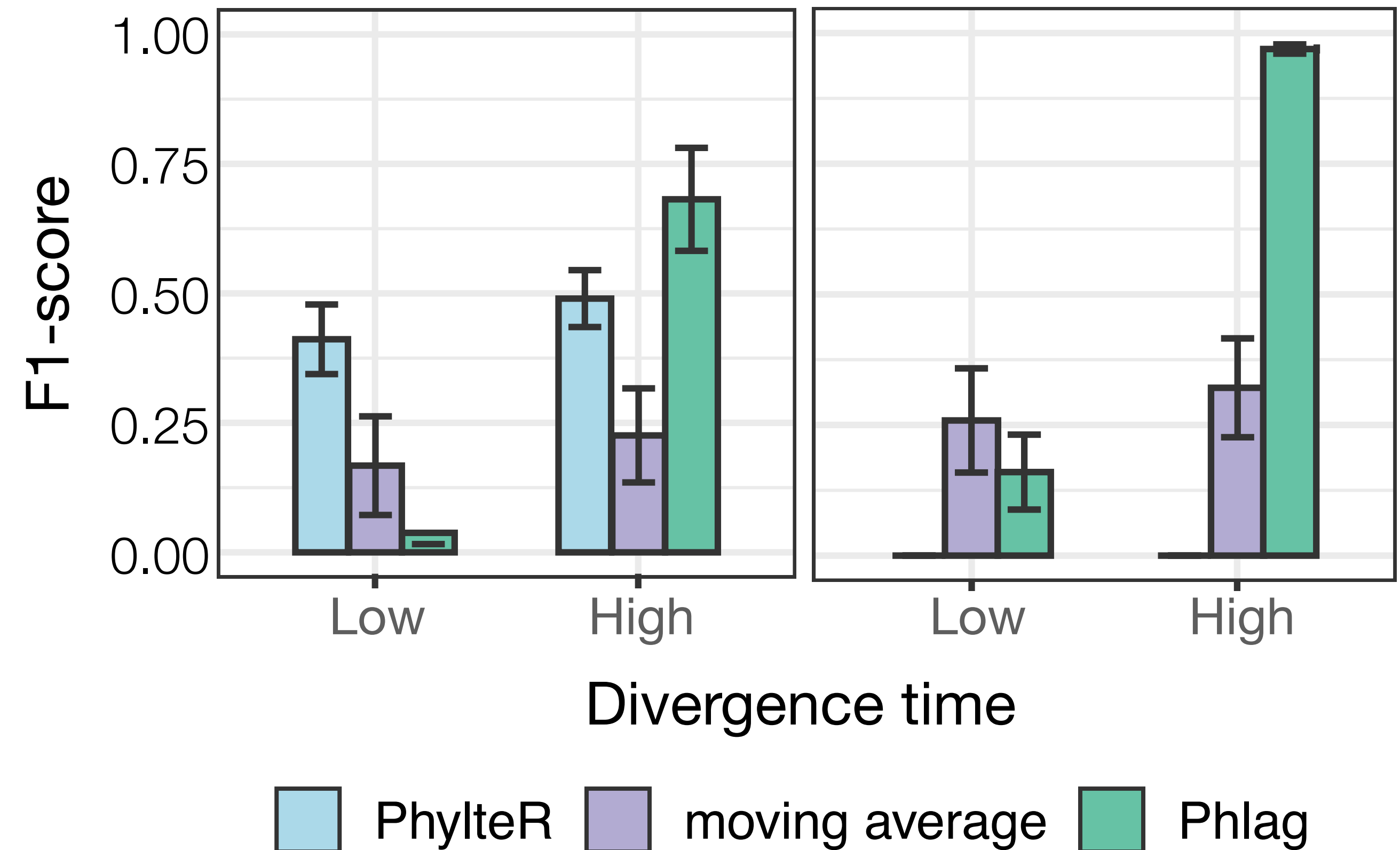
Phlag performs well when admixed lineages diverge at least several Myrs.

PhylteR degrades if the region is >10%!

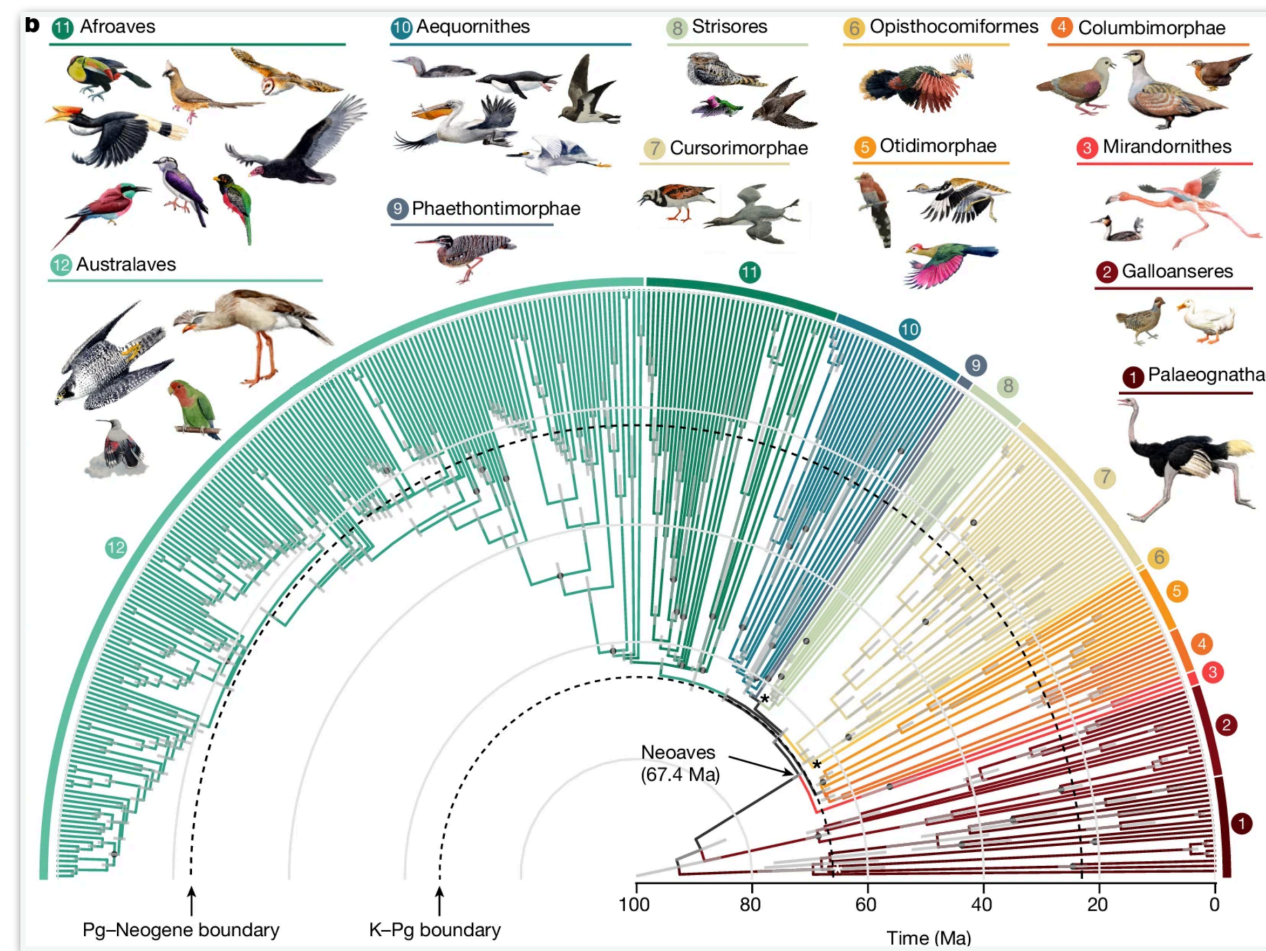
one region spanning:

1-10%

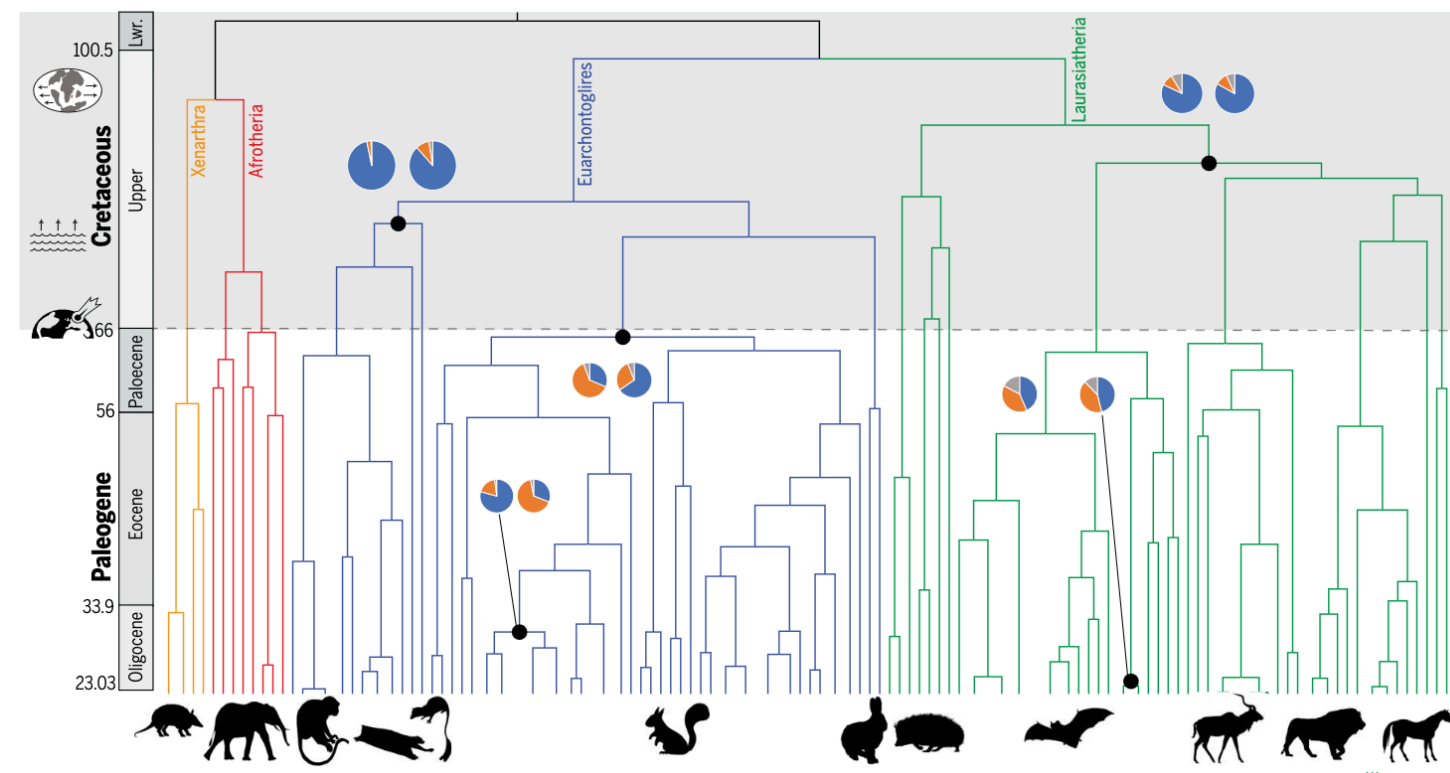
10-25%



Applying Phlag on two large empirical phylogenies



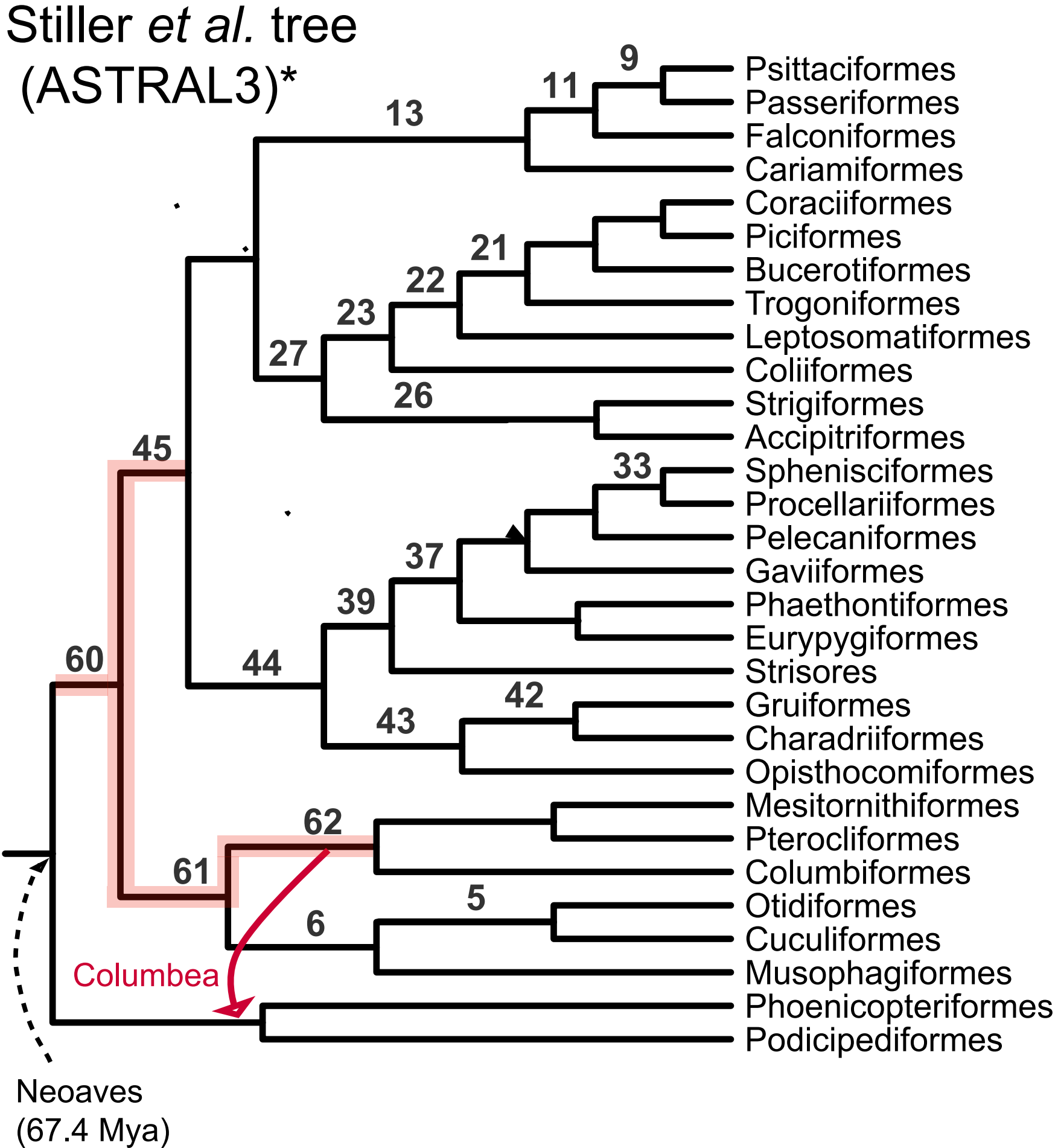
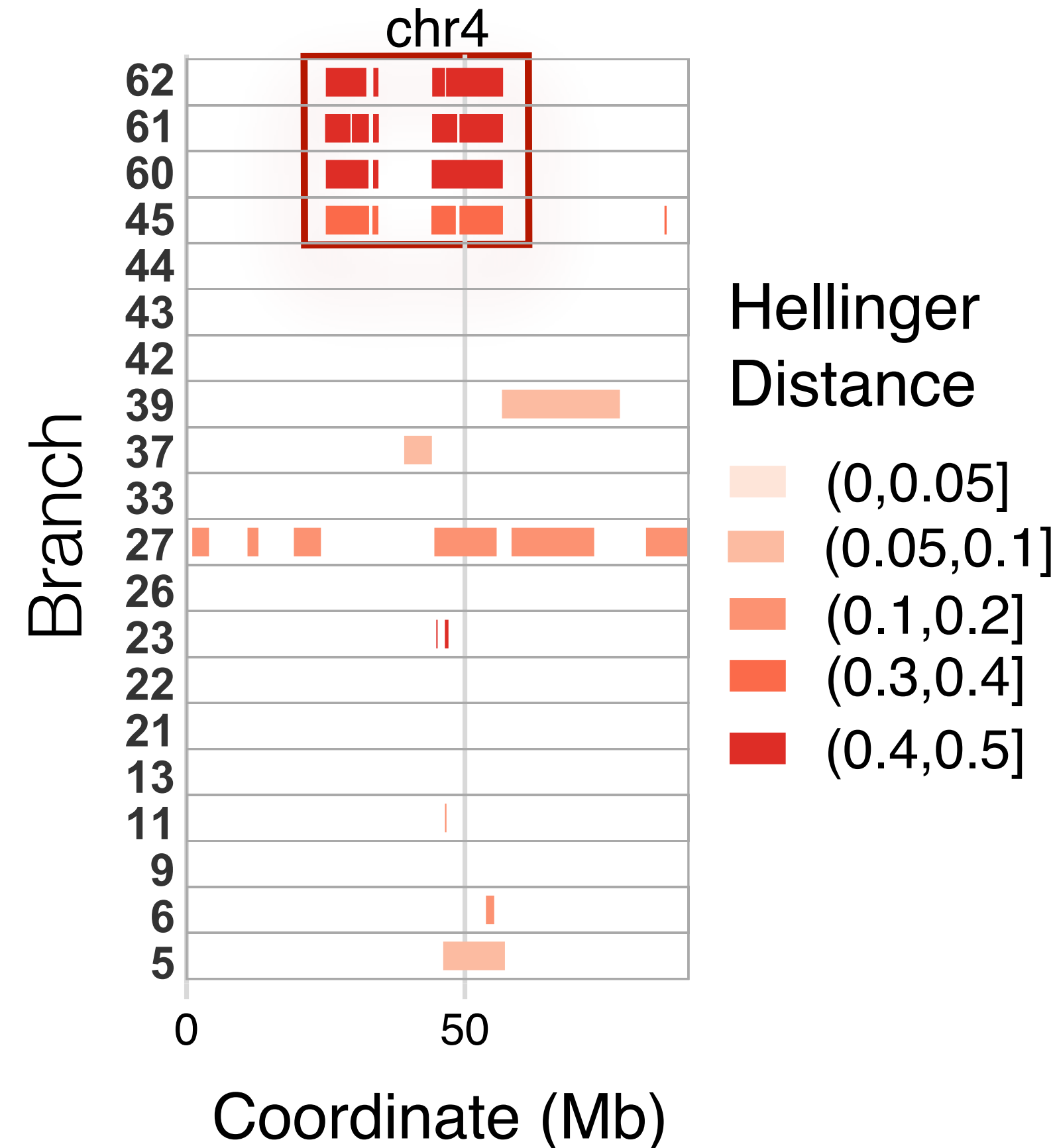
- [Stiller et al. 2024] 363-taxon avian dataset
- 39,849 gene trees from five macrochromosomes
- Analyzed 20 branches under Neoaves w/ support < 0.5



- [Foley et al. 2023] 241-taxon mammalian dataset
- 19,465 gene trees from chromosome 3 (unusual ILS)
- Applied to all internal branches — one at a time

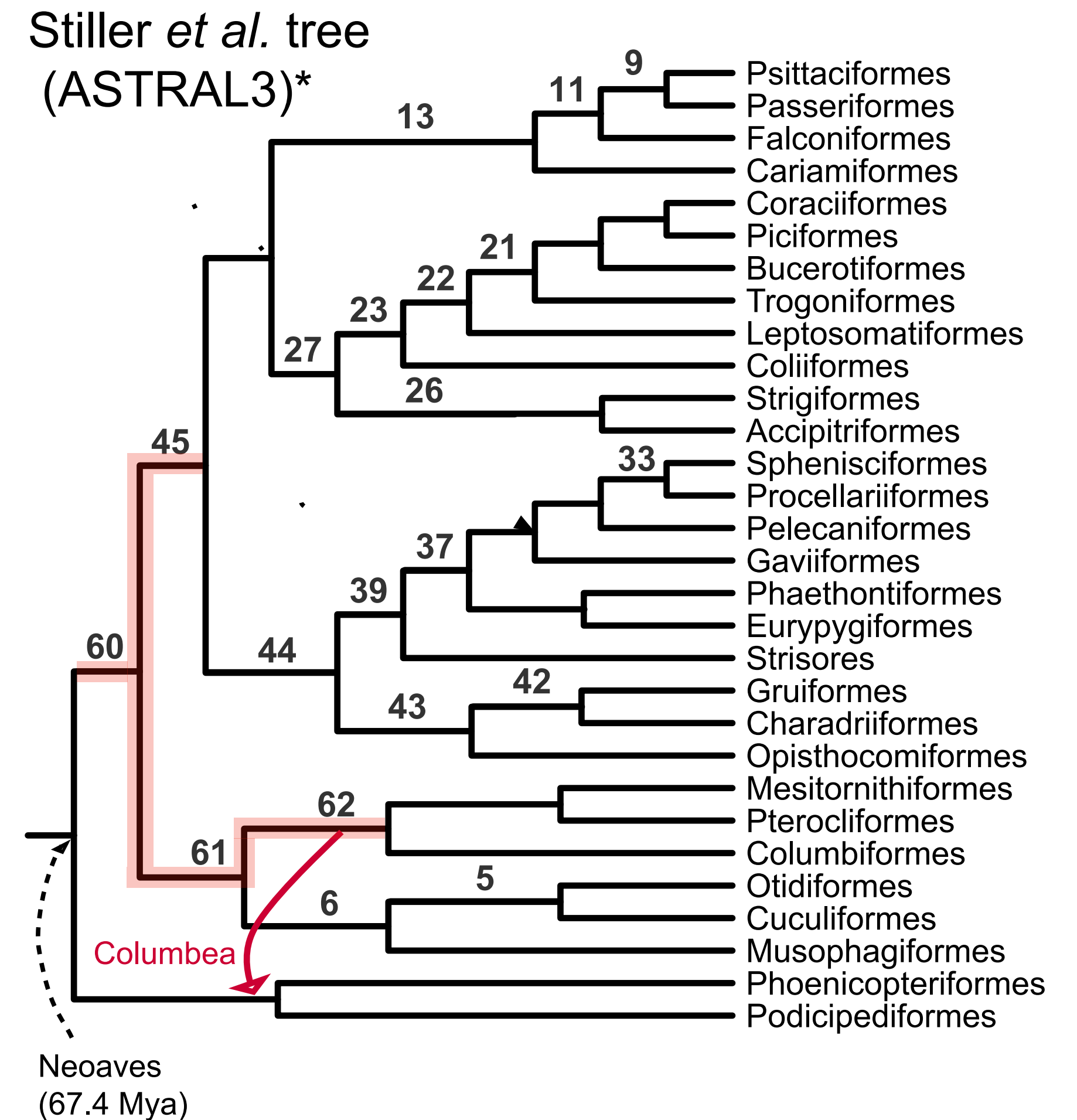
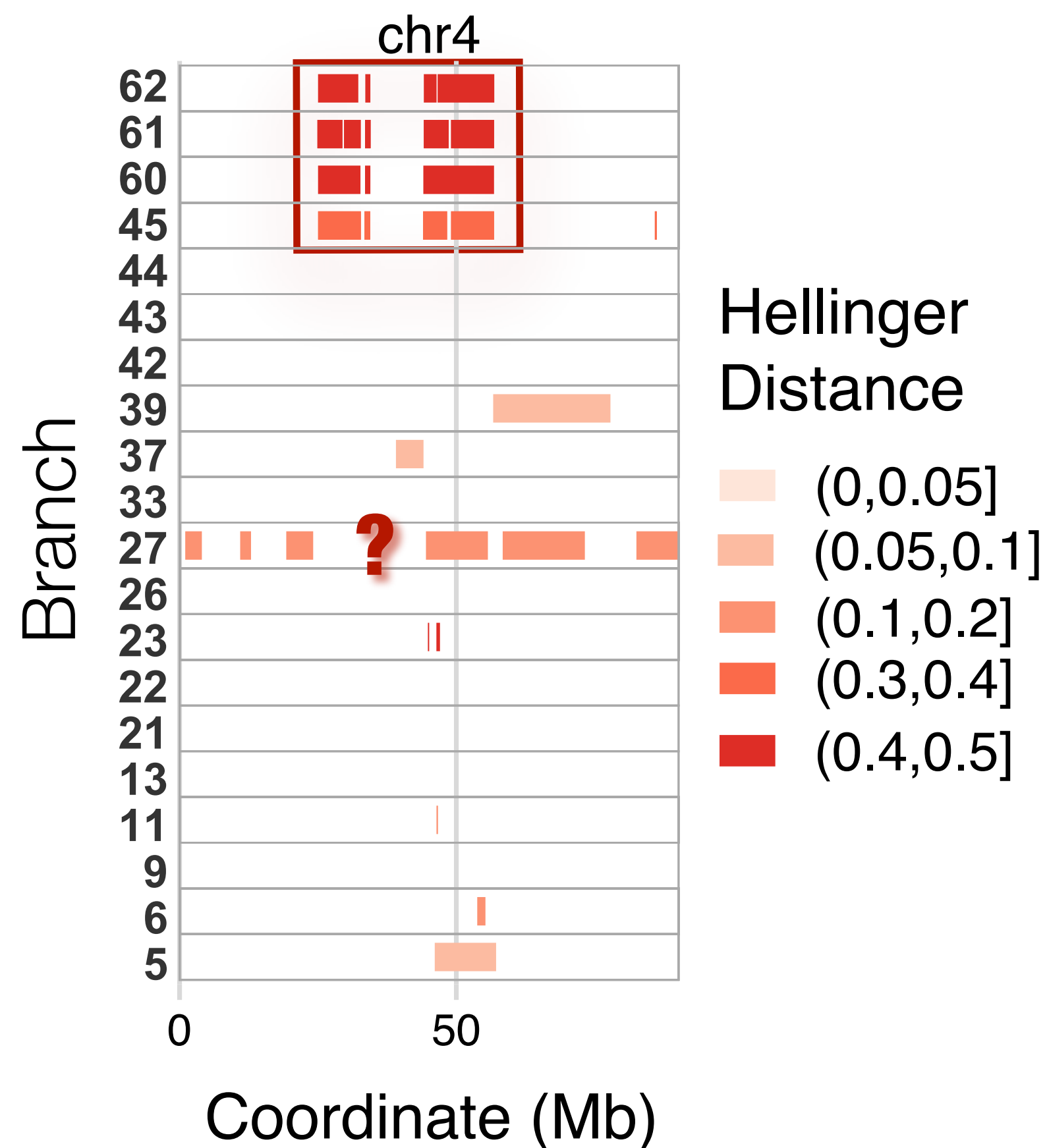
“Columbea” rearrangement & recombination suppression

Consistent support for
an alternative topology:



“Columbea” rearrangement & recombination suppression

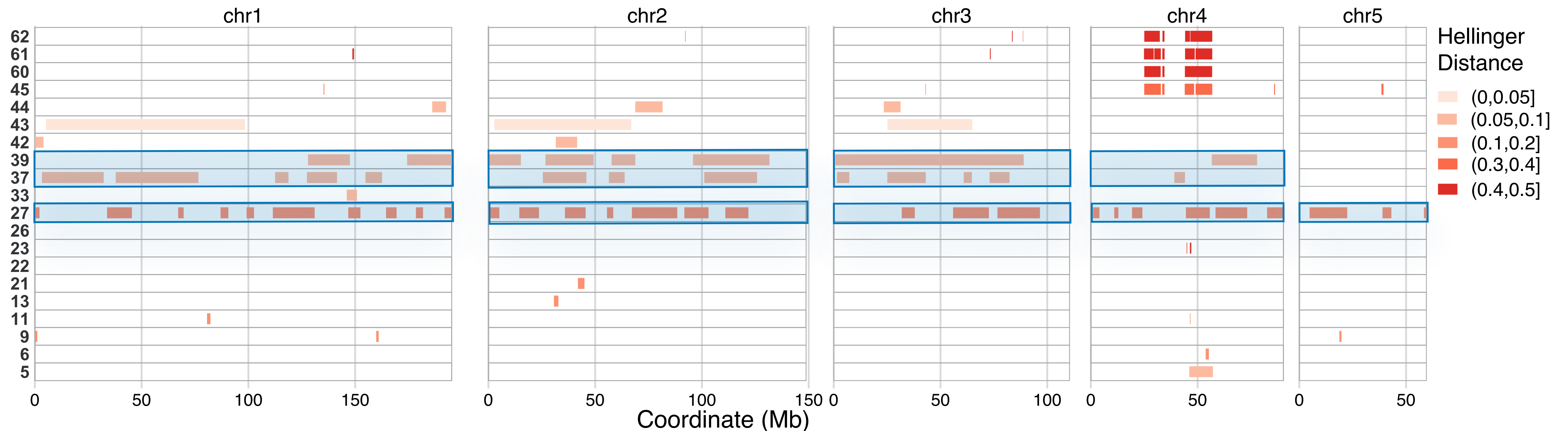
Consistent support for
an alternative topology:



Phlag flags branches that are difficult to resolve

Branches that were **inconclusive in Stiller et al.** — potentially reticulate origin:

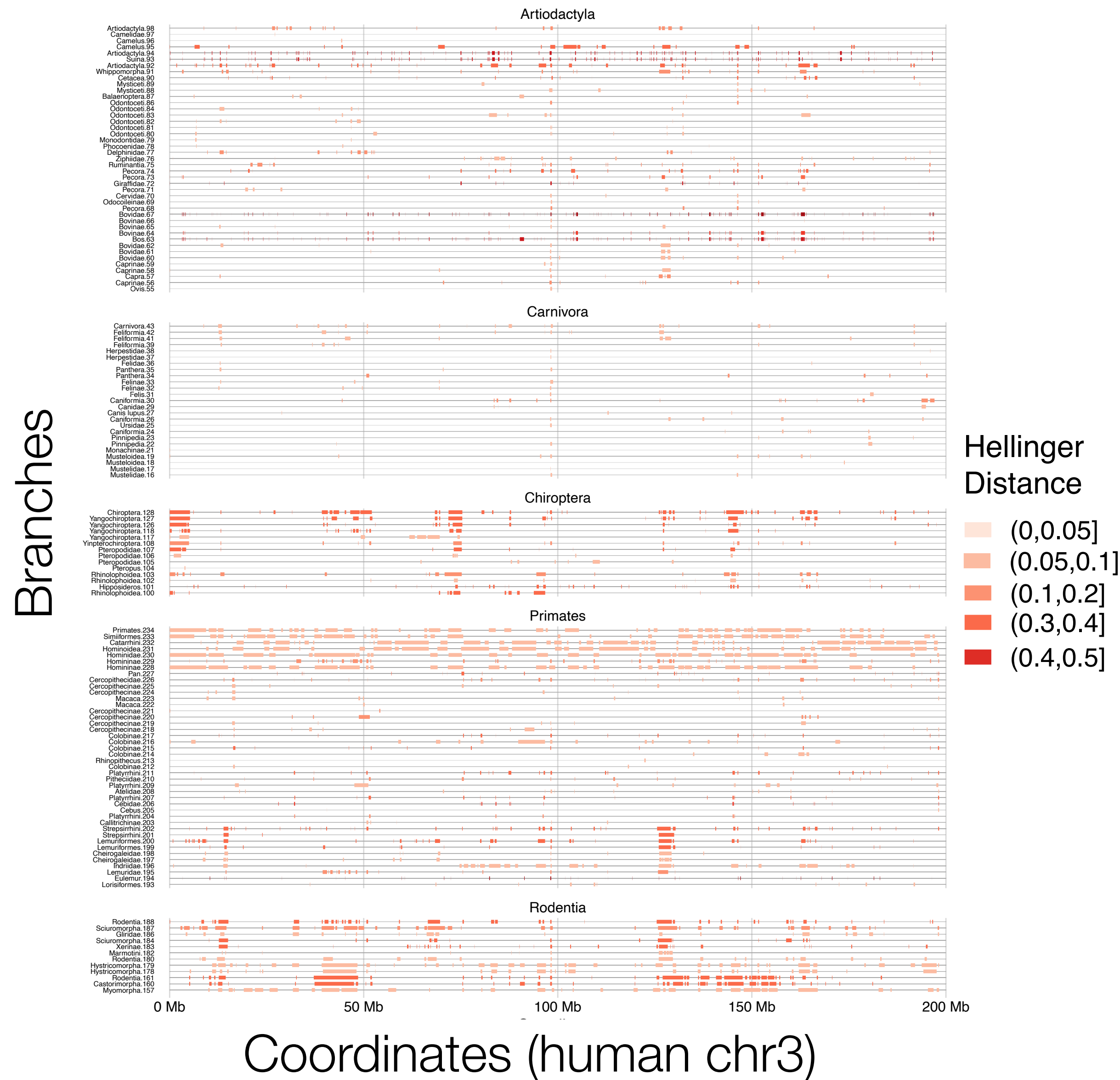
- ▶ Phaethontimorphae & Strisores (branches **37** and **39**)
- ▶ Accipitriformes & Strigiformes (branch **27**)



Explorative results on mammalian phylogeny & chr3

Many regions across various branches:

Some clades seem to have more “phlagged” regions than others



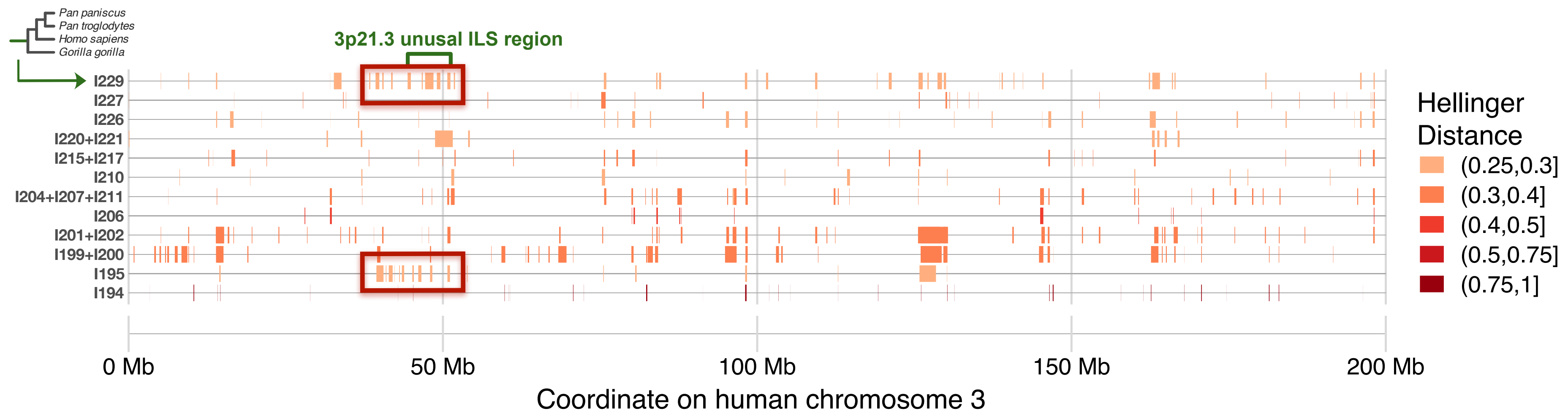
Explorative results on mammalian phylogeny & chr3

Many regions across various branches:

Some clades seem to have more
“phlagged” regions than others

Unusual ILS patterns, selection?

(e.g., primates around 50Mb on chr3)



Explorative results on mammalian phylogeny & chr3

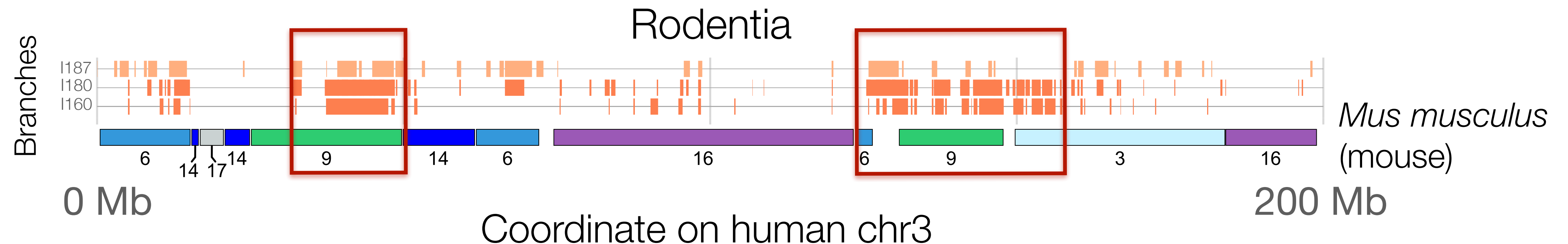
Many regions across various branches:

Some clades seem to have more “phlagged” regions than others

Signs of introgression?

(e.g., Rodentia & Cercopithecinae)

Unusual ILS patterns, selection?
(e.g., primates around 50Mb on chr3)



Explorative results on mammalian phylogeny & chr3

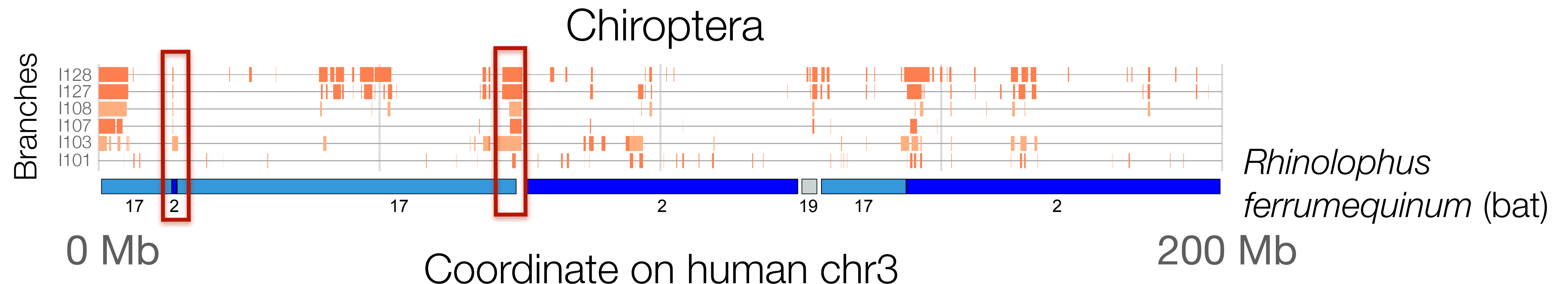
Many regions across various branches:

Some clades seem to have more “phlagged” regions than others

Signs of introgression?
(e.g., Rodentia & Cercopithecinae)

Unusual ILS patterns, selection?
(e.g., primates around 50Mb on chr3)

Detected regions and boundaries coincide with rearrangements:
(e.g., chr2 of horseshoe bat)



Summary

Phlag

- a general framework for detecting deviations from the species tree and MSC
- could help refine the species tree & further inspect alternative processes
 - scalable & flexible (interpretable hyperparameters)
 - cannot distinguish between different types of alternatives

Future work:

incorporating synteny, allowing multiple states, exploring different emission types

Thank you!

Funding



ISMB Travel Fellowship
ISCB

software



paper



Shayesteh Arasti



Edward L. Braun



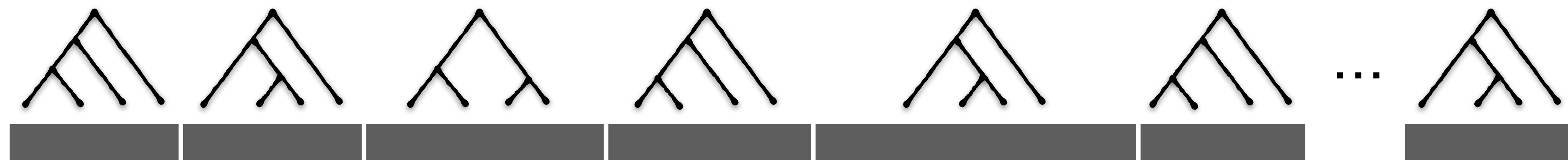
Siavash Mirarab

Extra Slides

Finding regions with strong deviations from the species tree model

Not a new notion: Finding changes in locus tree and non-tree like scenarios
e.g., PhyloNet-HMM, CoalHMM, ERICA, TRAILS, Twisst...

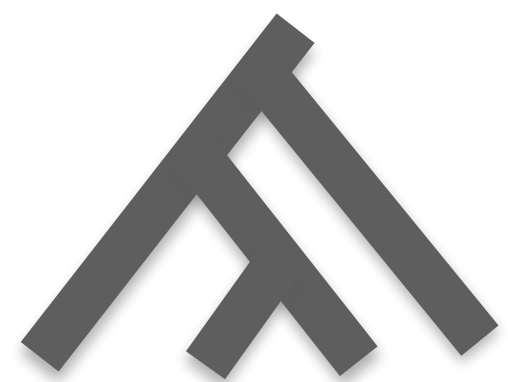
limited to
small num.
of taxa



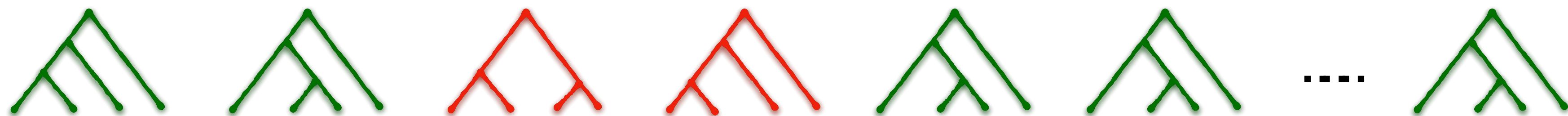
usually from
MSAs

Chromosome

Our problem: For a species tree and gene trees **ordered by the positions**;
detect segment(s) with deviations & violations!



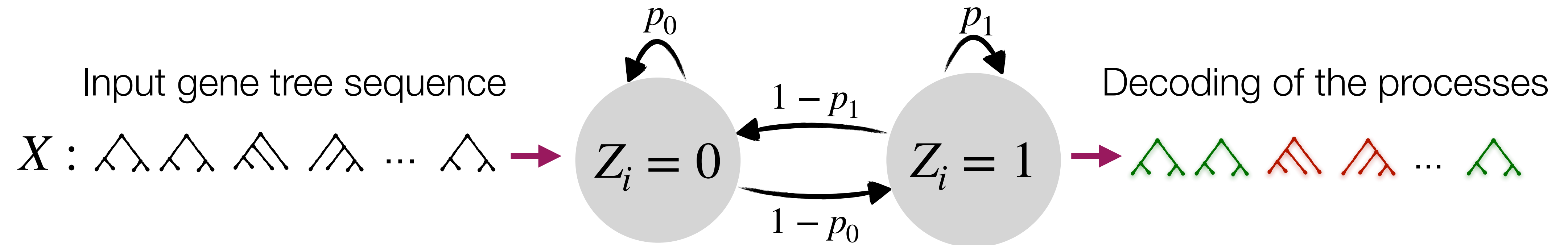
Species Tree



Gene Trees

Phlag: a hidden Markov model

A **two-state HMM** if we reasonably assume Z is to be Markovian:

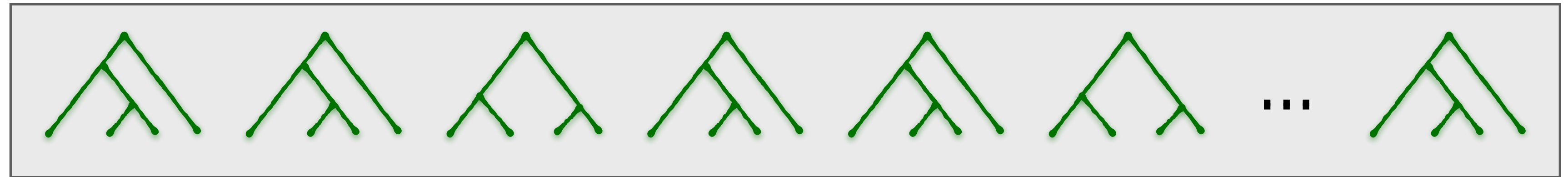


We are *ideally* interested in $P(Z | X, \theta_0, \theta_1)$, thus $P(X_i | \theta_{Z_i})$, and Z as Viterbi path

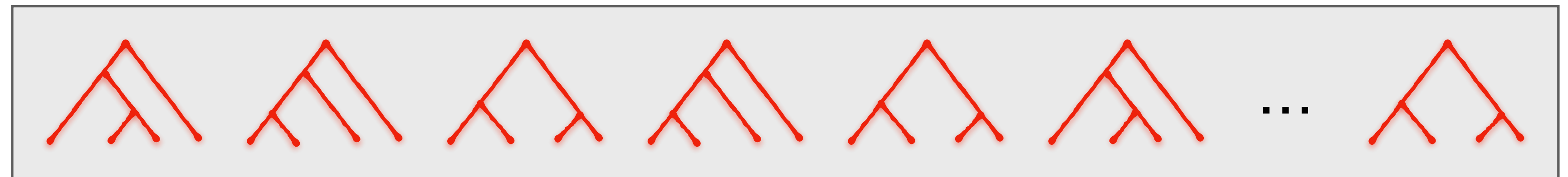
$P(X_i | \theta_1)$ emission probabilities **are not accessible** for the alternative process

Computing $P(X_i | \theta_0)$ under MSC is **possible** [Wu 2012], **but not scalable**

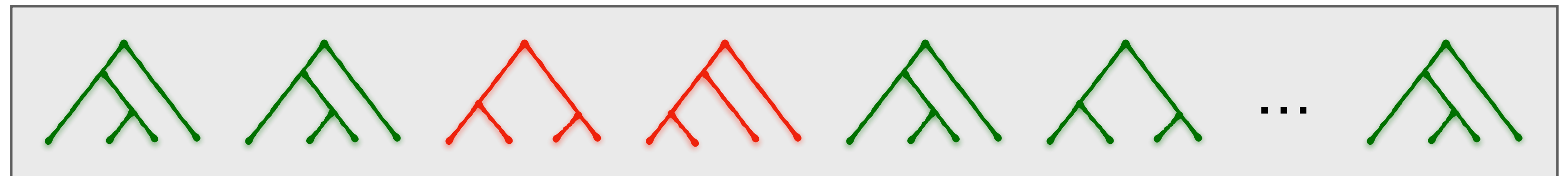
MSC(T):
 G_i^0



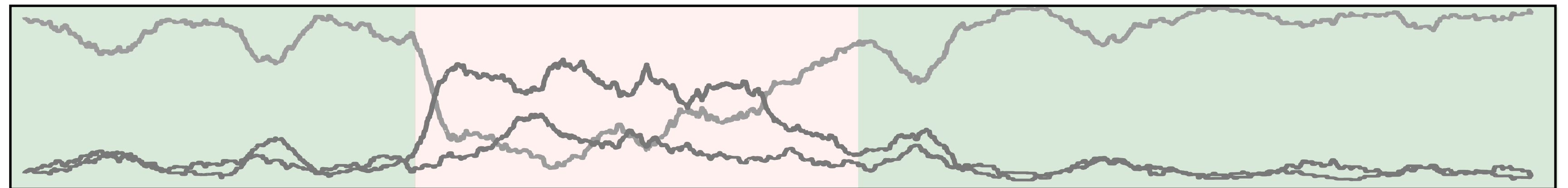
Unkown process:
 G_i^1



Mixture process:
 (X_i, Z_i)



Emissions:
 $F(X_i, b_j) = Y_i^{(j)}$



Position on chromosome

Connecting the emission distribution to MSC

$Y_i | Z_i = 1 \sim \text{Cat}(e_1)$ is unknown

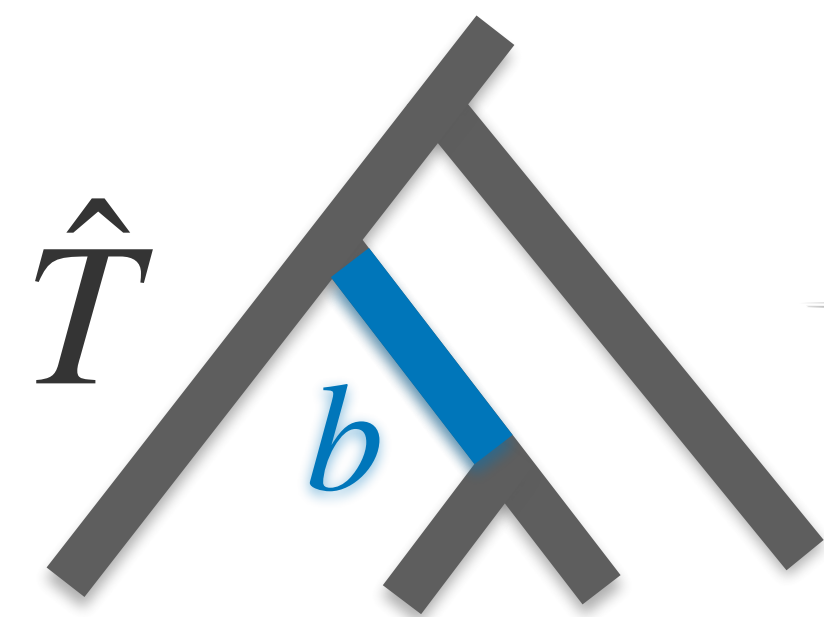
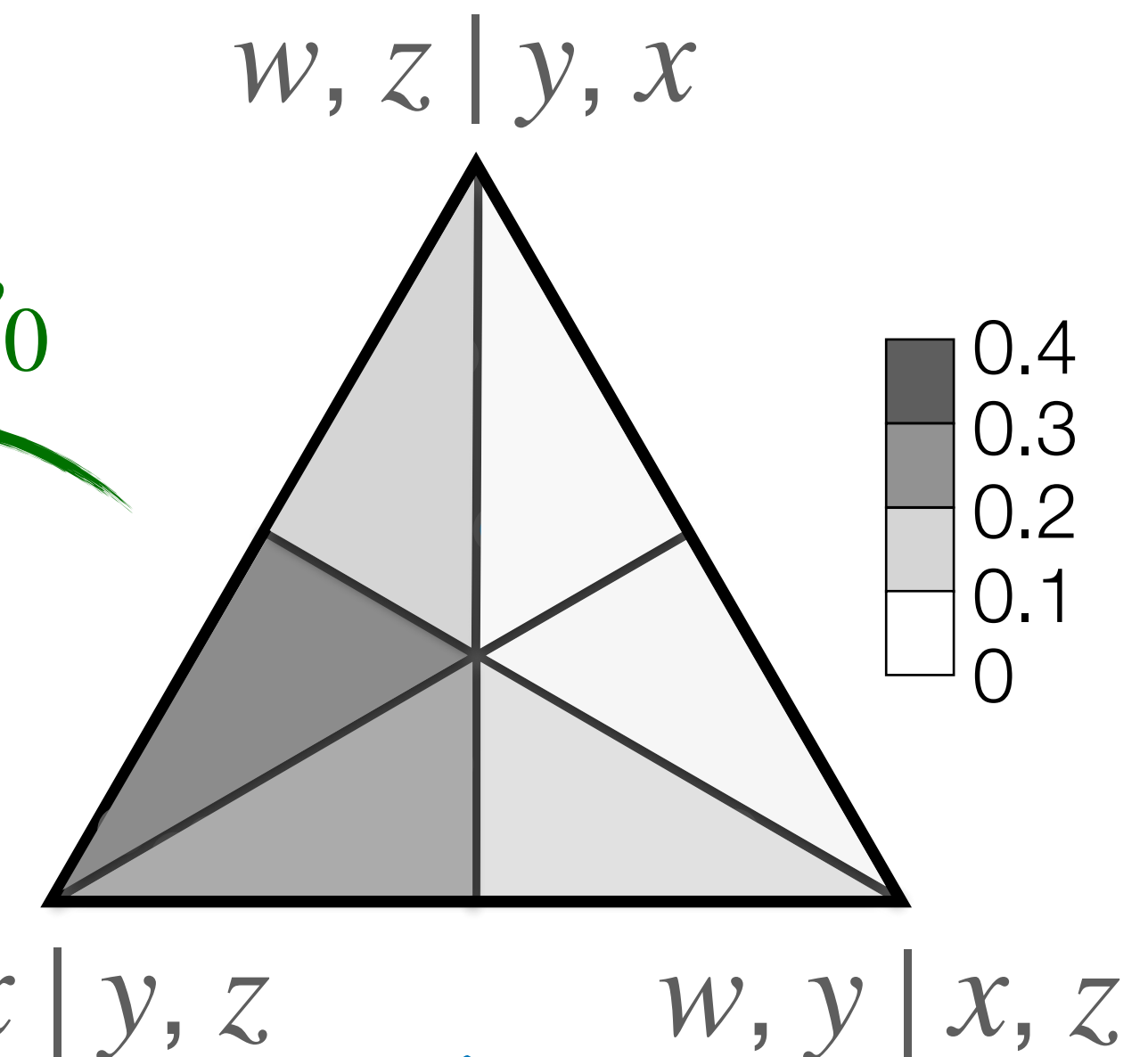
(estimate e_1 from the data)

$Y_i | Z_i = 0 \sim \text{Cat}(e_0)$ follows $\text{MSC}(T)$

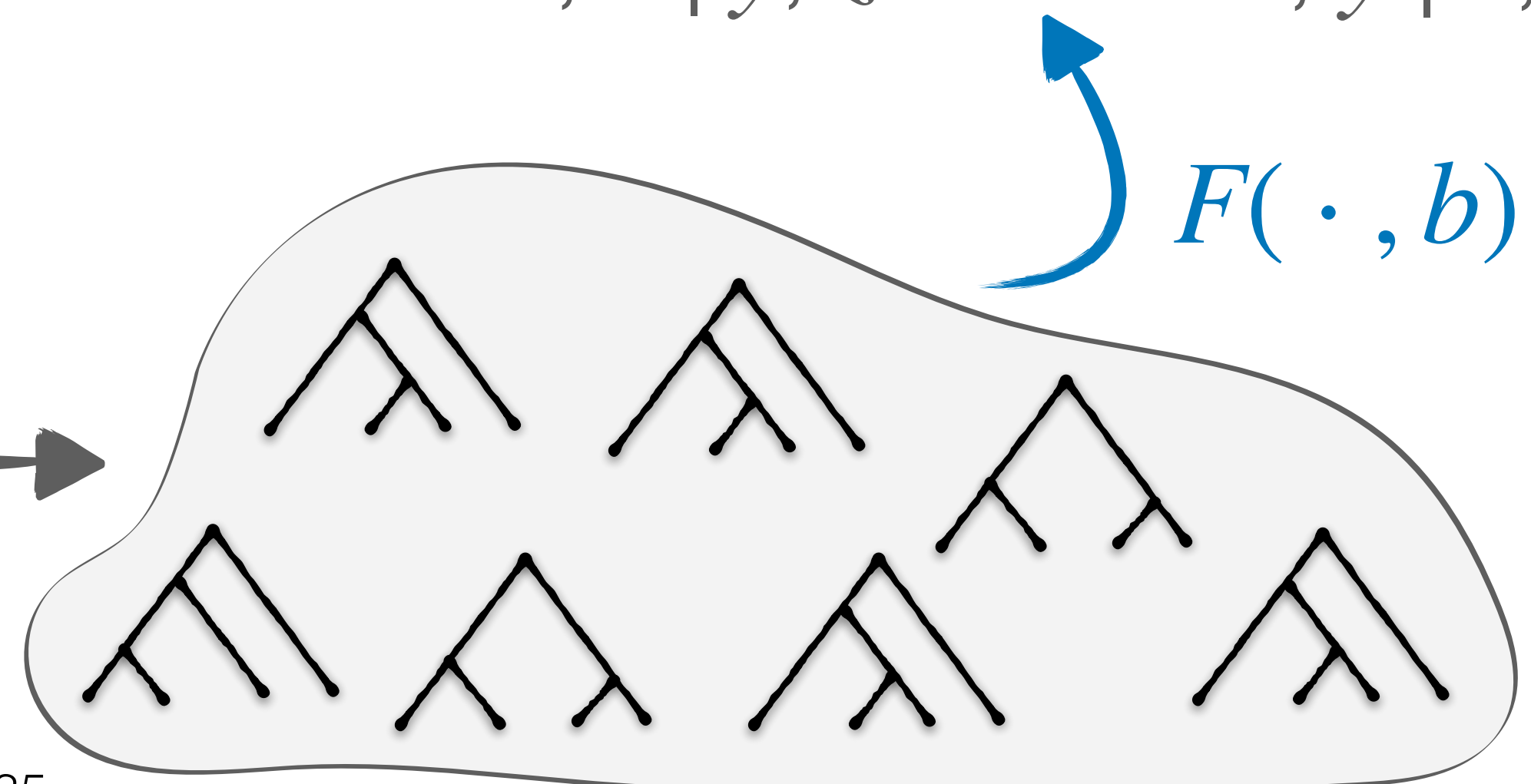
We have an estimate $\hat{T}$

- assuming correct topology
- branch lengths (CU) may differ

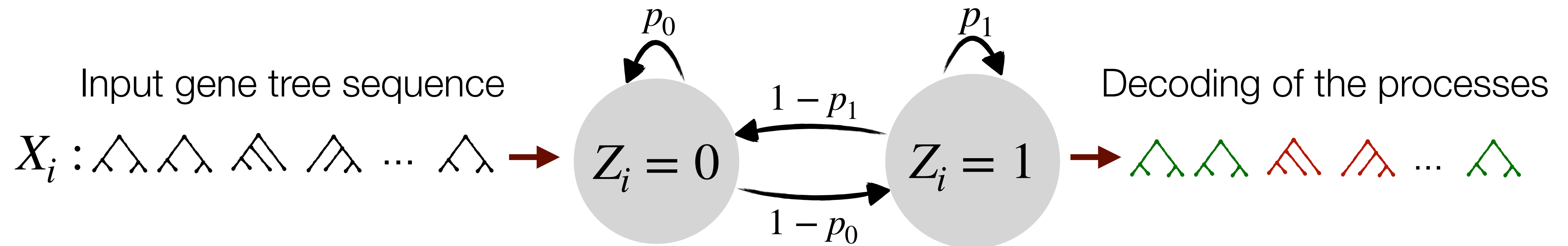
estimate
parameters e_0



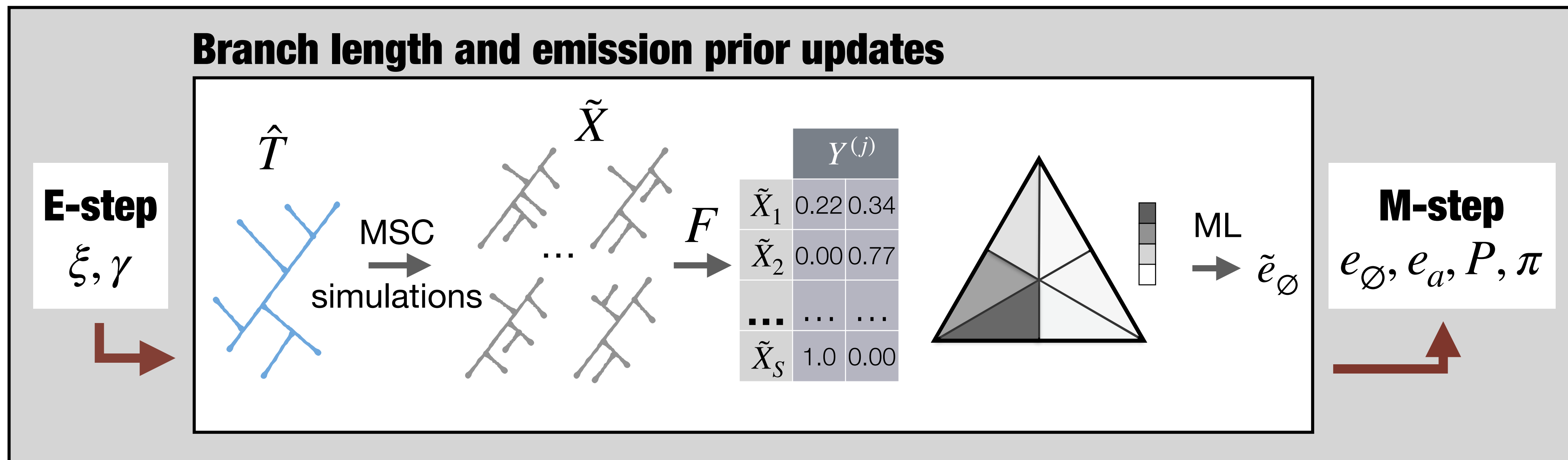
simulate gene trees
under MSC



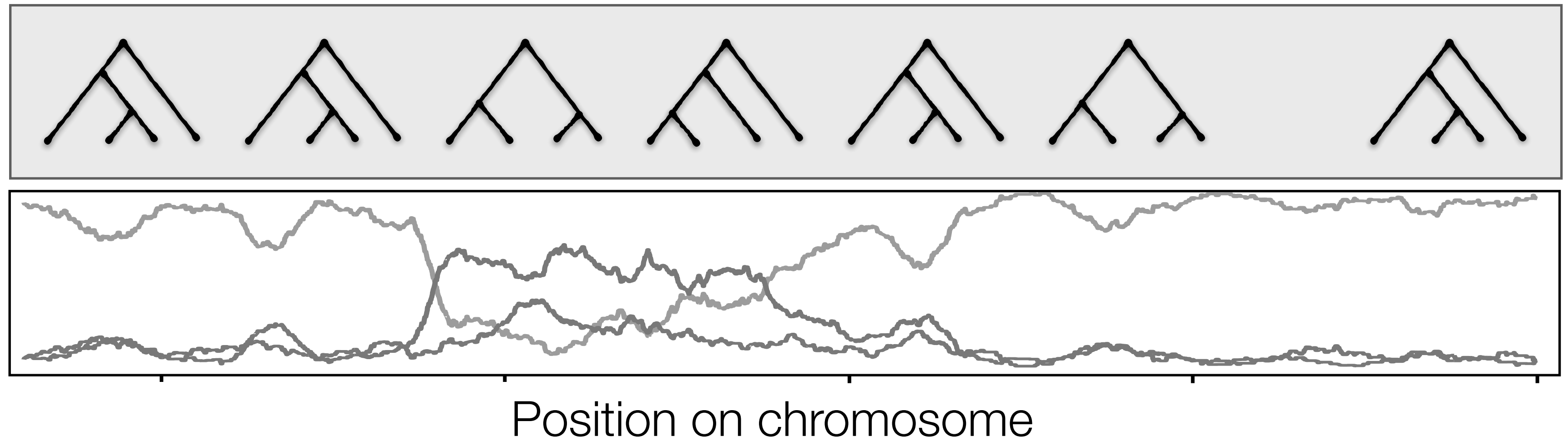
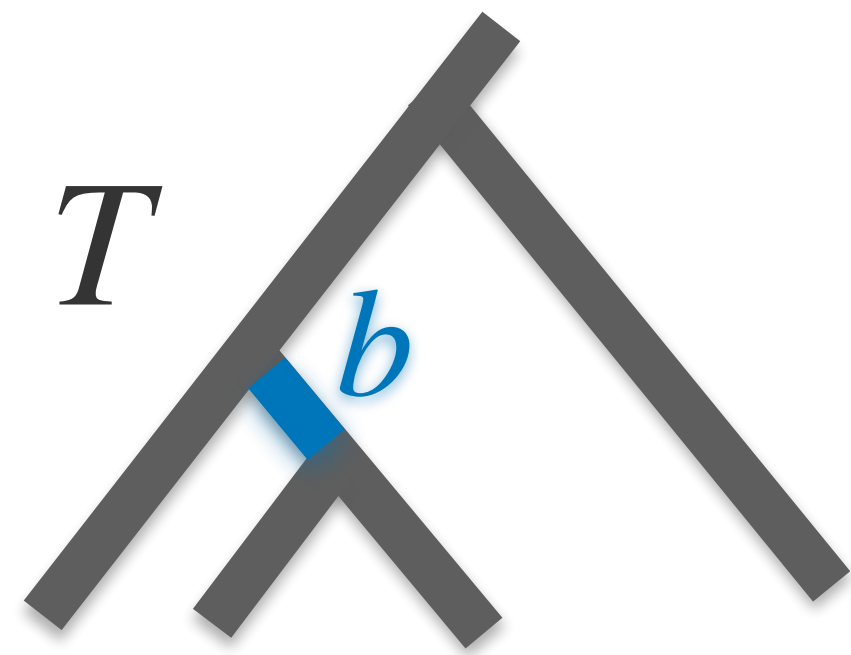
Summary of the algorithm



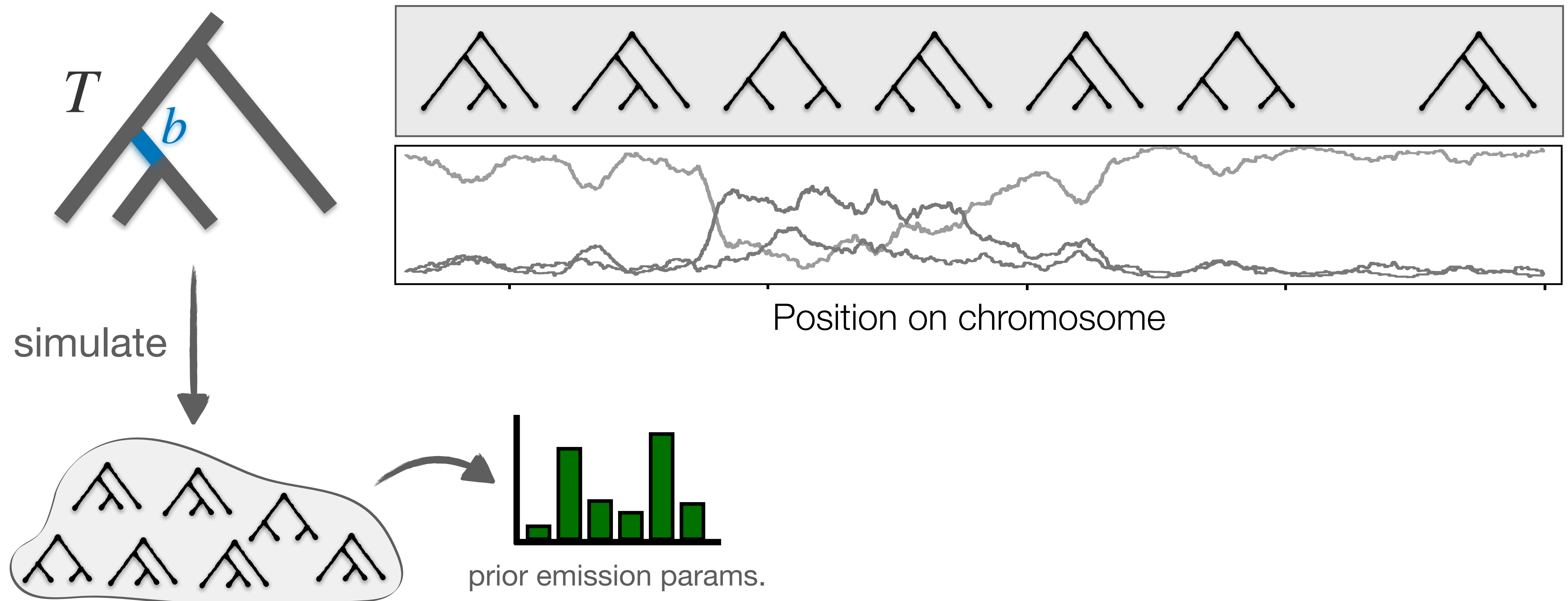
EM algorithm



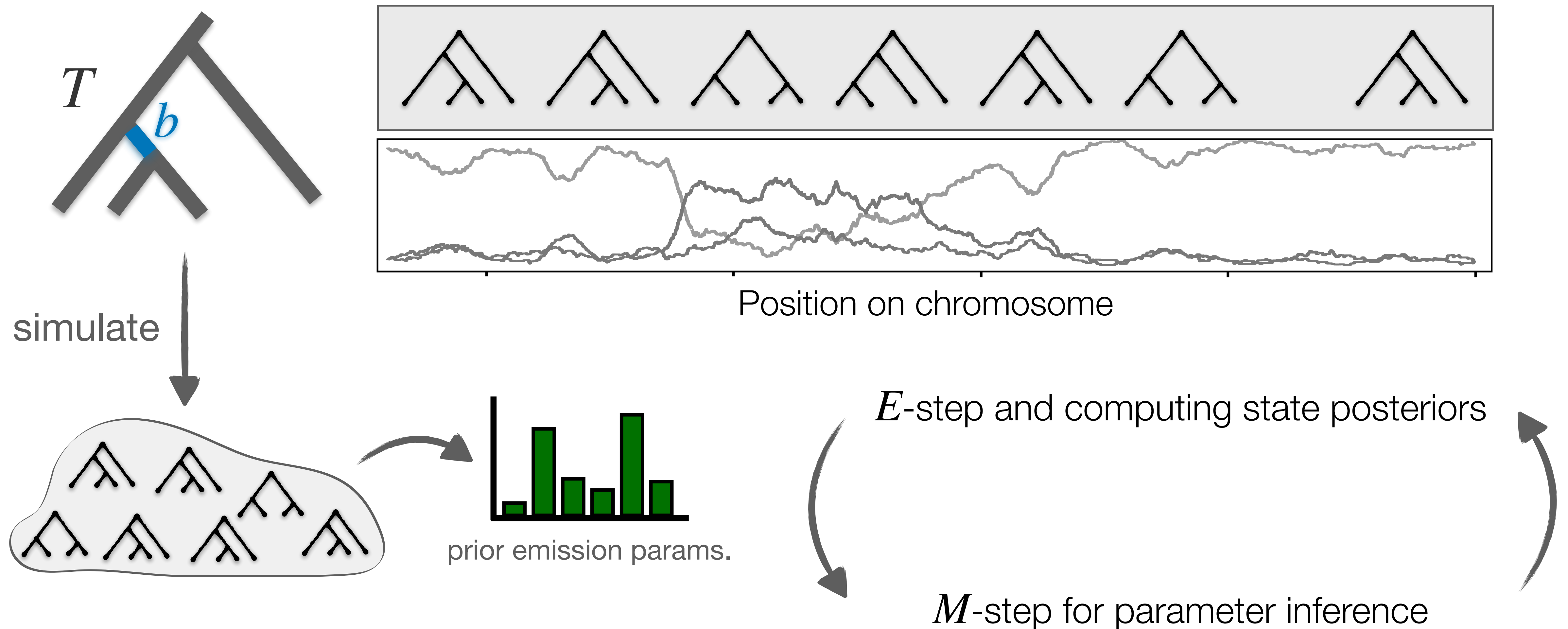
Summary of the Phlag



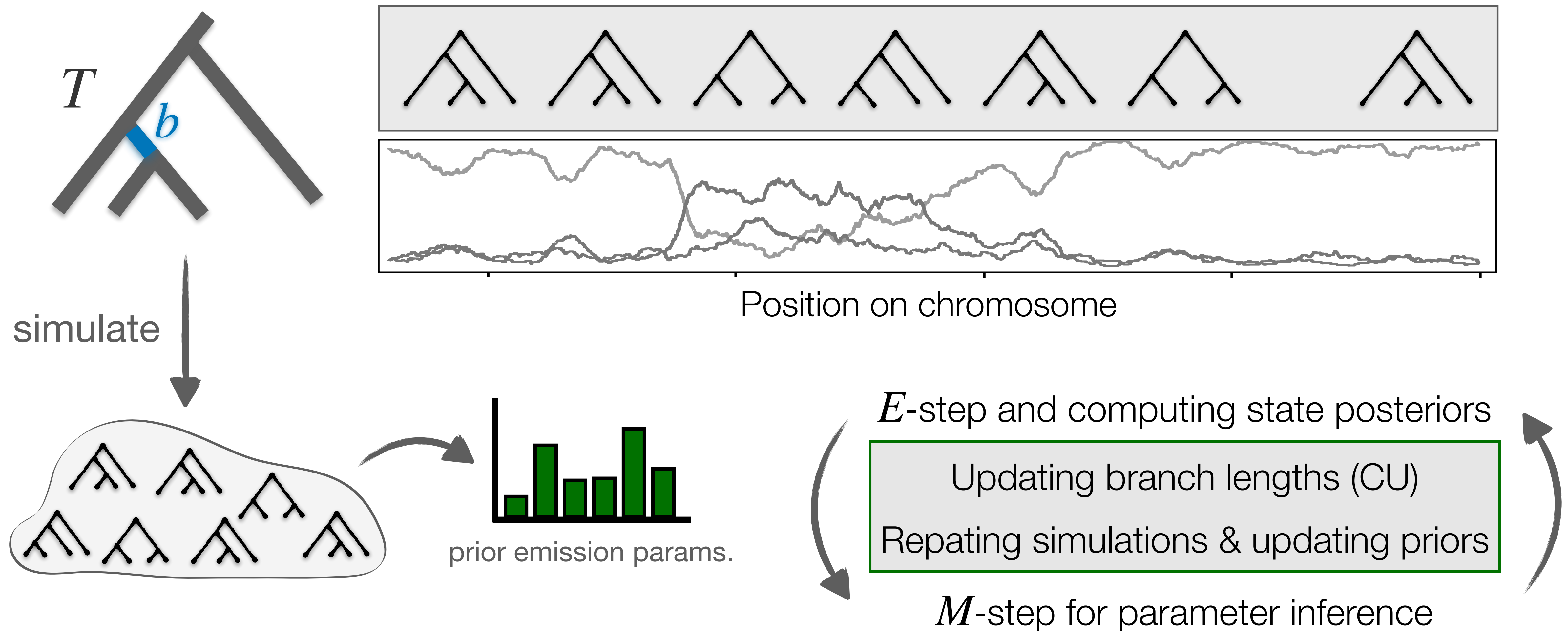
Summary of the Phlag



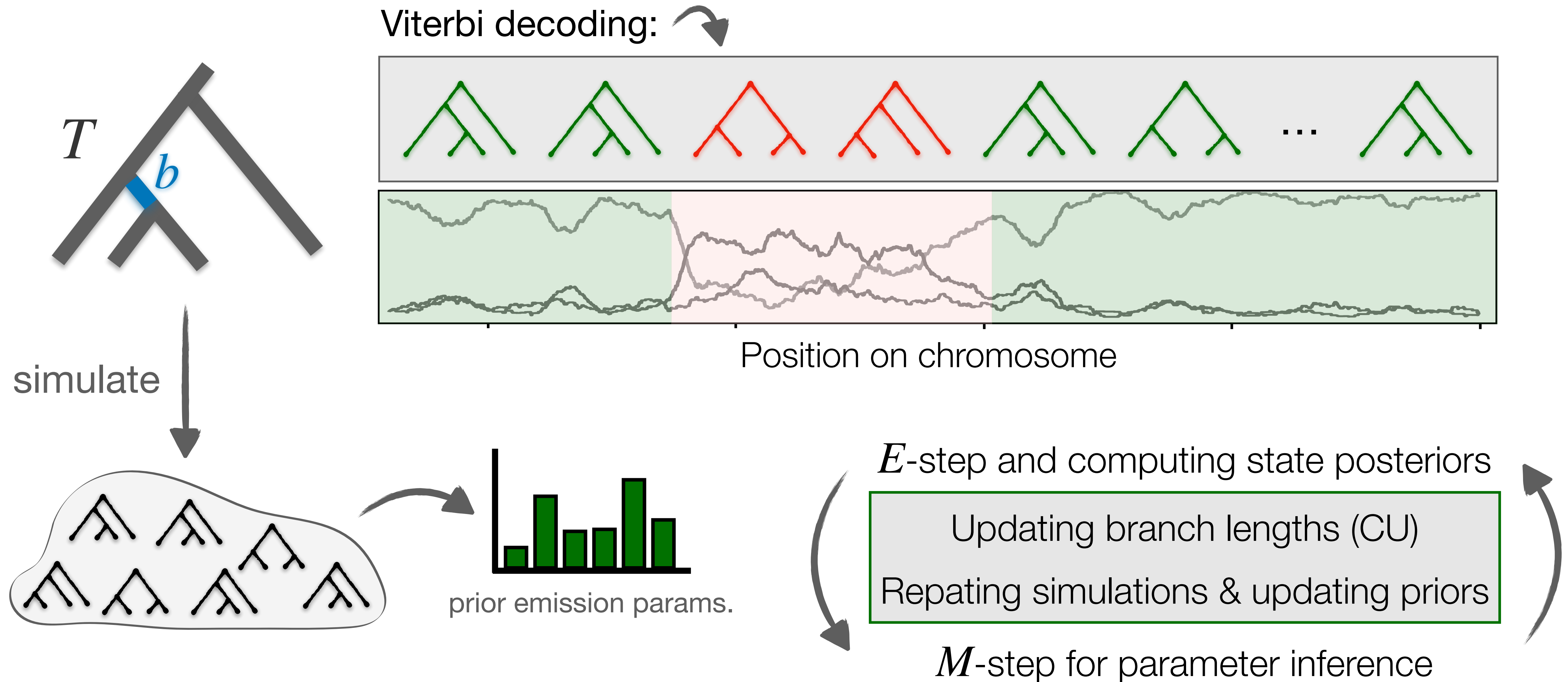
Summary of the Phlag



Summary of the Phlag

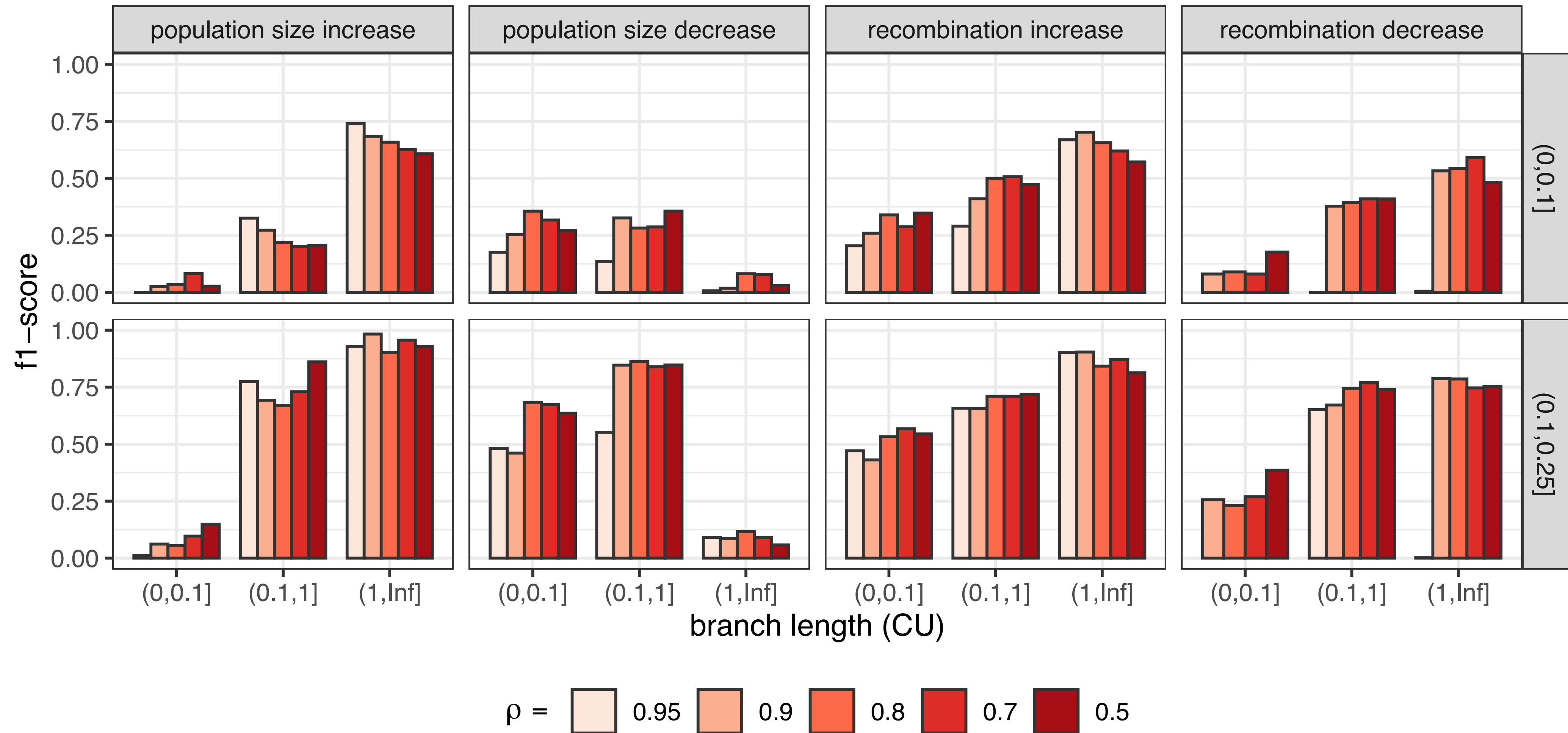


Summary of the Phlag



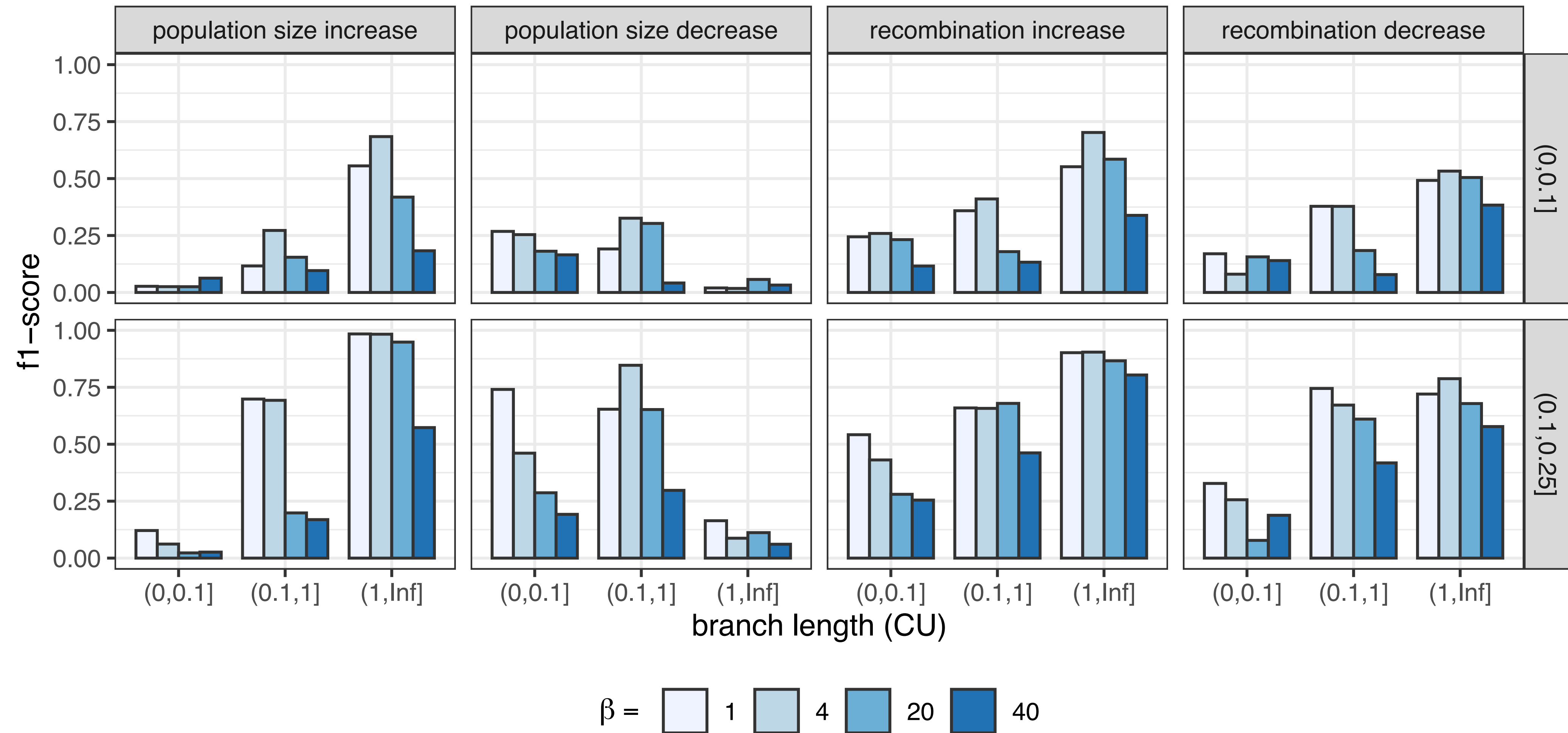
Transition matrix hyperparameters

Changing ρ (default: 0.90), fixed $\beta = 4$



Transition matrix hyperparameters

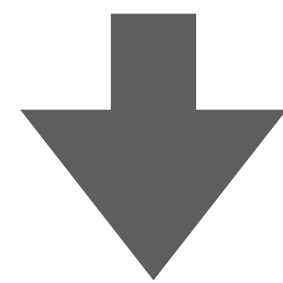
Changing β (default: 4), fixed $\rho = 0.9$



Transition matrix hyperparameters

$$P = \begin{pmatrix} p_0 & 1 - p_0 \\ 1 - p_1 & p_1 \end{pmatrix} \rightarrow \pi^* = (\rho, 1 - \rho) \quad \rho = \frac{1 - p_1}{2 - p_0 - p_1}$$

$$N\rho(1 - p_0) = \beta \text{ yields } p_0 = 1 - \beta/(N\rho)$$



$$p_0 = \frac{N\rho - \beta}{N\rho} \text{ and } p_1 = \frac{\beta - N + N\rho}{N(\rho - 1)}$$

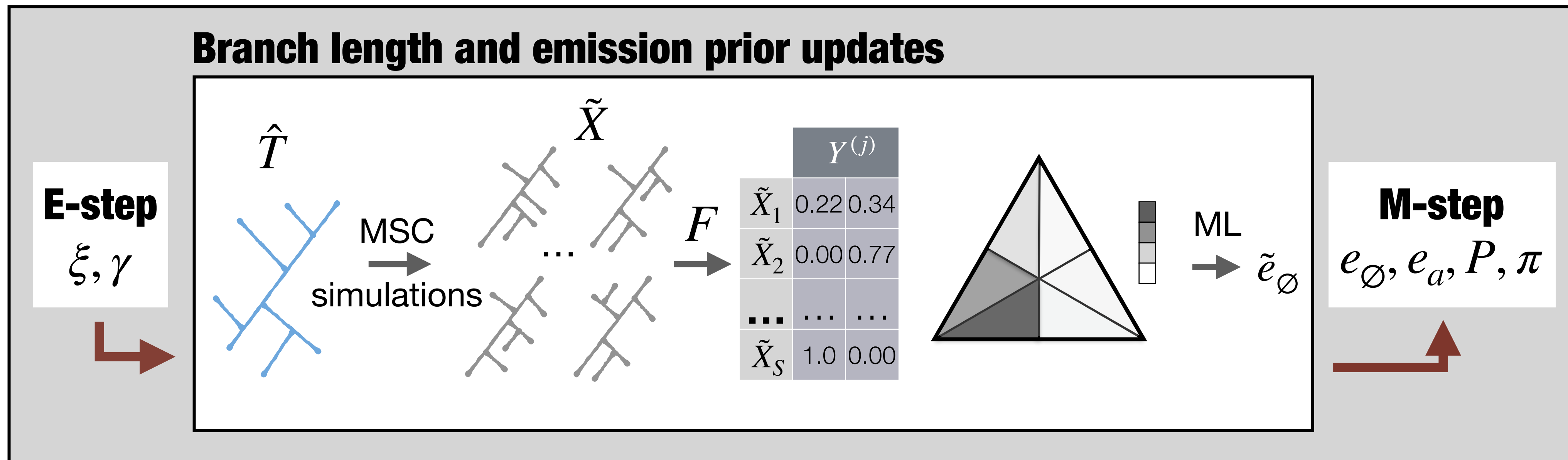
$$p_0 \sim \text{Beta}(\alpha_0, \beta), p_1 \sim \text{Beta}(\alpha_1, \beta)$$

$$\mathbb{E}[p_0] = \frac{\alpha_0}{\alpha_0 + \beta} \text{ and } \mathbb{E}[p_1] = \frac{\alpha_1}{\alpha_1 + \beta}$$

$$\alpha_0 = N\rho - \beta \text{ and } \alpha_1 = N(1 - \rho) - \beta$$

More on EM and empirical Bayes approach

EM algorithm



More on EM and empirical Bayes approach

Algorithm 1 Inference of parameters and decoding. INITIALIZE initializes parameters. Each EM iteration requires $O(mN)$. ESTIMATEBRANCHCU updates CU branch lengths using Equation (4) in $O(nN)$. Each update of $\tilde{e}_\emptyset$ is $O(mnS)$.

```
1: procedure PHLAG( $X, \mathcal{B}', \hat{T}; \rho, \beta, \lambda, L$ )
2:    $\gamma \in [0, 1]^N \leftarrow \mathbf{1}_N$   $\triangleright$  Posterior state prob.
3:    $\hat{T} \leftarrow \text{ESTIMATEBRANCHCU}(X, T, \gamma)$   $\triangleright$  See Equation (4)
4:    $e_\emptyset, e_a, \tilde{e}_\emptyset, P, \pi \leftarrow \text{INITIALIZE}(\hat{T}, \mathcal{B}', \rho, \beta)$ 
5:    $\forall_j : Y_j \leftarrow F(X, \mathcal{B}'_j)$   $\triangleright$  compute statistic (e.g., QQS)
6:   for all  $1 \leq l \leq L$  do
7:      $\gamma, \xi \leftarrow \text{E-STEP}(Y, P, \pi, e_\emptyset, e_a)$ 
8:     if time to update CU lengths then
9:        $\hat{T} \leftarrow \text{ESTIMATEBRANCHCU}(X, T, \gamma)$ 
10:       $\tilde{e}_\emptyset \leftarrow \text{UPDATEEMPIRICALLDISTRIBUTION}(\hat{T})$ 
11:       $e_\emptyset, e_a, P, \pi \leftarrow \text{M-STEP}(Y, \gamma, \xi, \tilde{e}_\emptyset, \rho, \beta, \lambda)$ 
12:       $Z \leftarrow \text{VITERBIDECODING}(Y, e_\emptyset, e_a, P, \pi)$ 
13:      return  $Z, \hat{T}$ 
14: procedure UPDATEEMPIRICALLDISTRIBUTION( $\hat{T}$ )
15:   Simulated  $\tilde{X}$  under MSC parameterized by  $\hat{T}$ 
16:    $\forall_j : \tilde{Y}_j \leftarrow F(\tilde{X}, \mathcal{B}'_j)$ 
17:    $\forall_j : \tilde{e}_\emptyset^{(j)} \leftarrow \text{ML using } \tilde{Y}_j$ 
18:   return  $\tilde{e}_\emptyset$ 
```

Simulation details

Ancestry simulations:

- recombination rate to 5×10^{-9} , generation time at 10 years across the entire tree

Incorporate variation in rates of evolution:

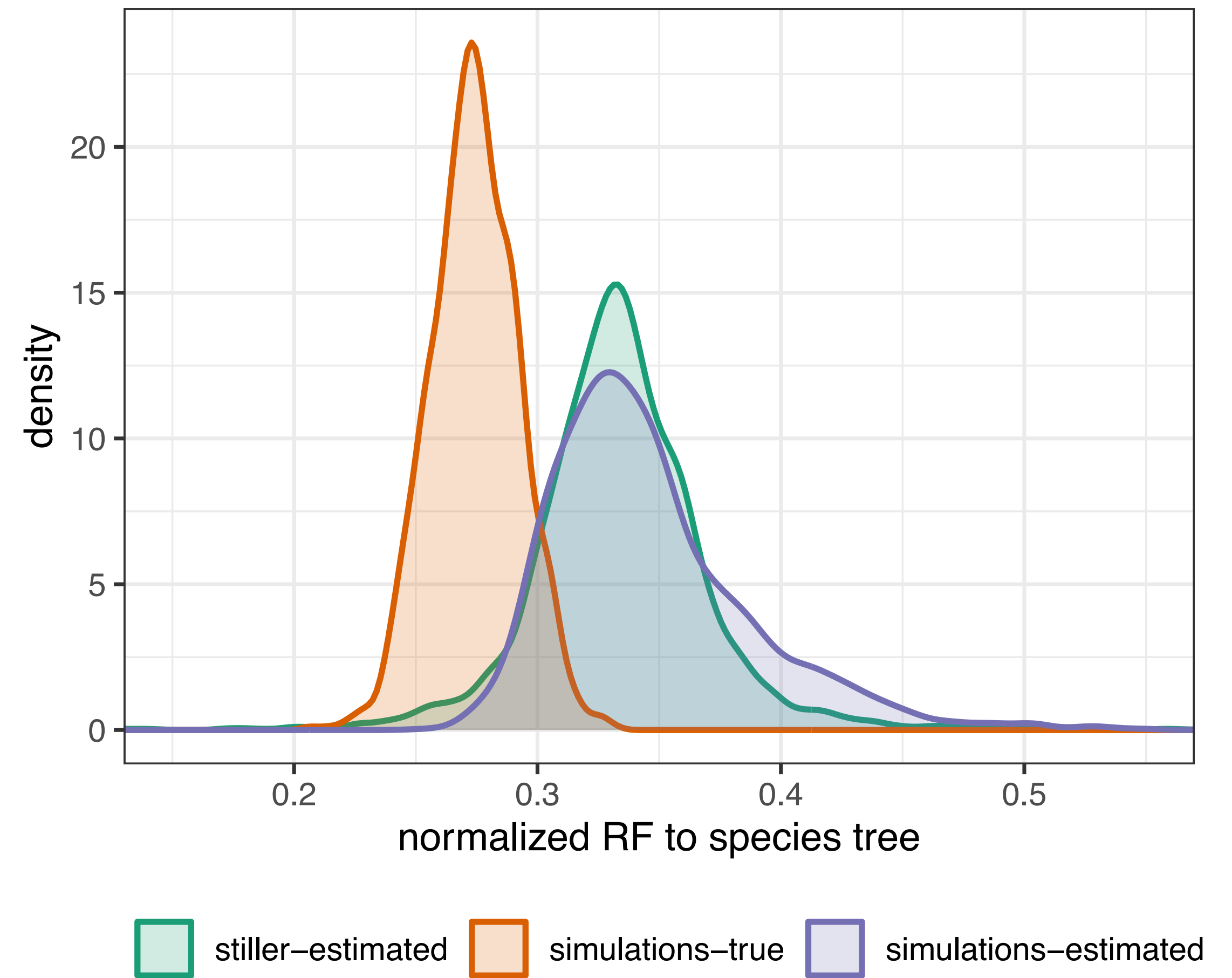
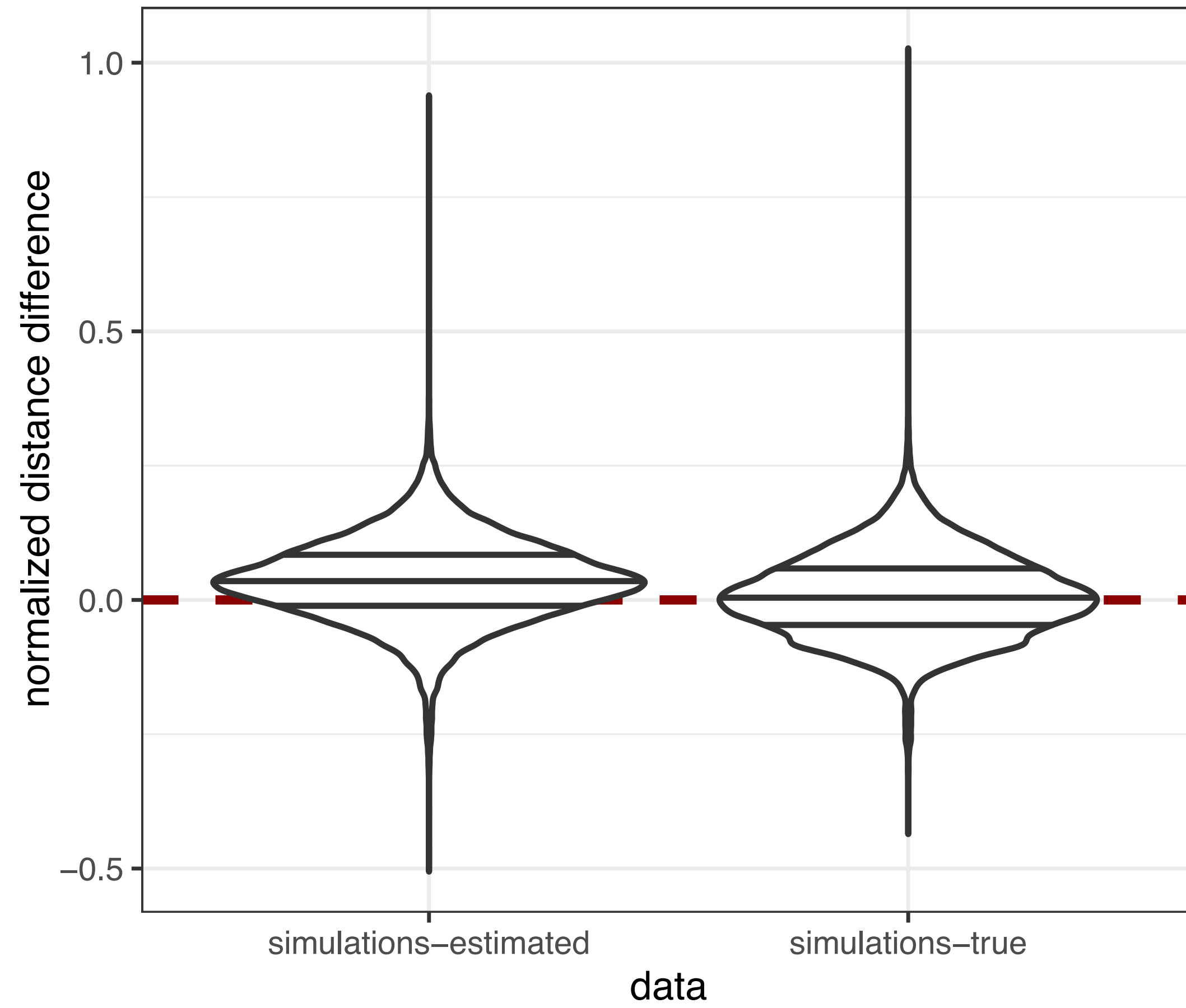
- a genealogy node y sampled at time $t(y)$ within population x is assigned a scaled time

$$t'(y) = t'(x) + r(x)(t(y) - t(x))$$

Alignment simulations:

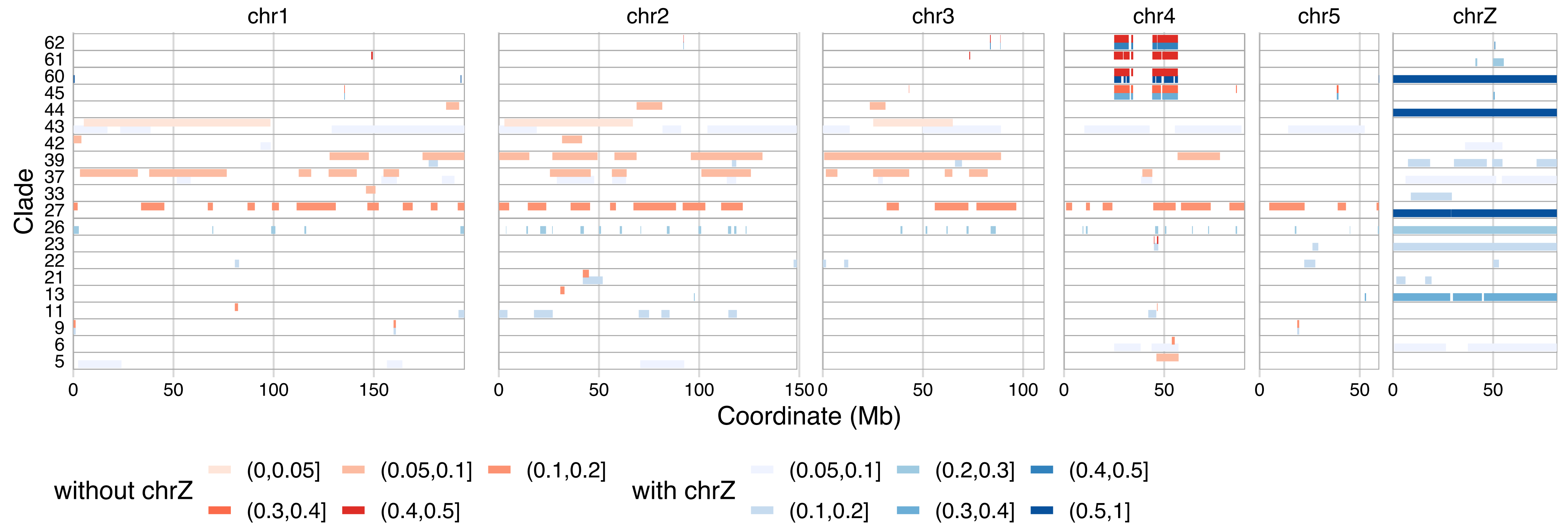
- GTR model with a mean substitution rate of 1.909 per bp per generation
- heterogeneity both across genomic regions and across sites using Gamma models

Discordance between simulated gene trees and T



Bird analysis with and without chrZ

Concatenation & updated prior



Mammalian analysis and QQS values

