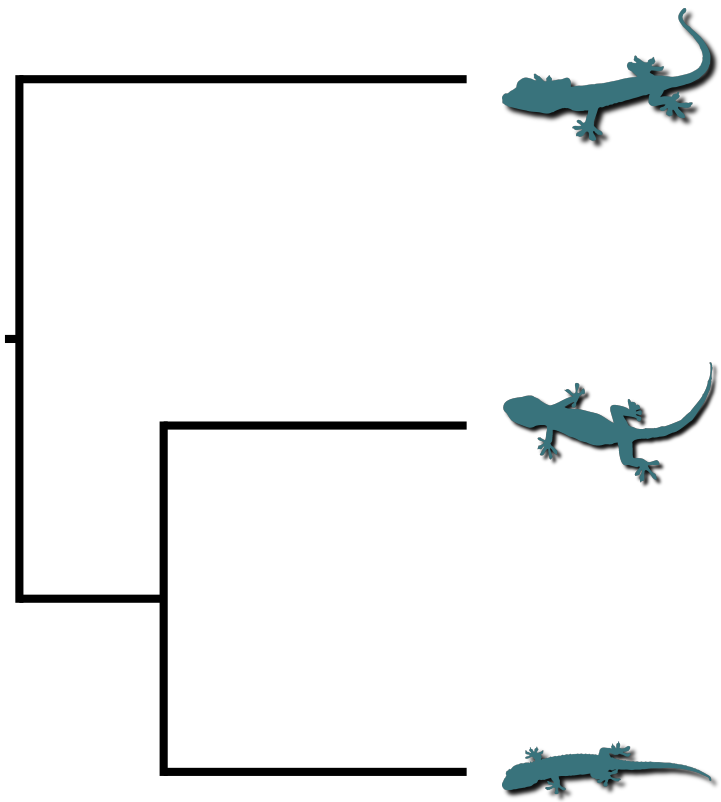
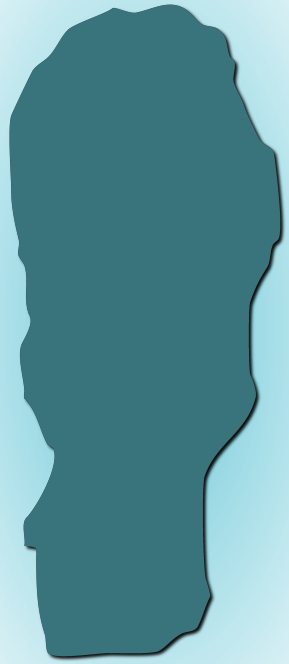
SSB 2018 Comparative Phylogeography Workshop

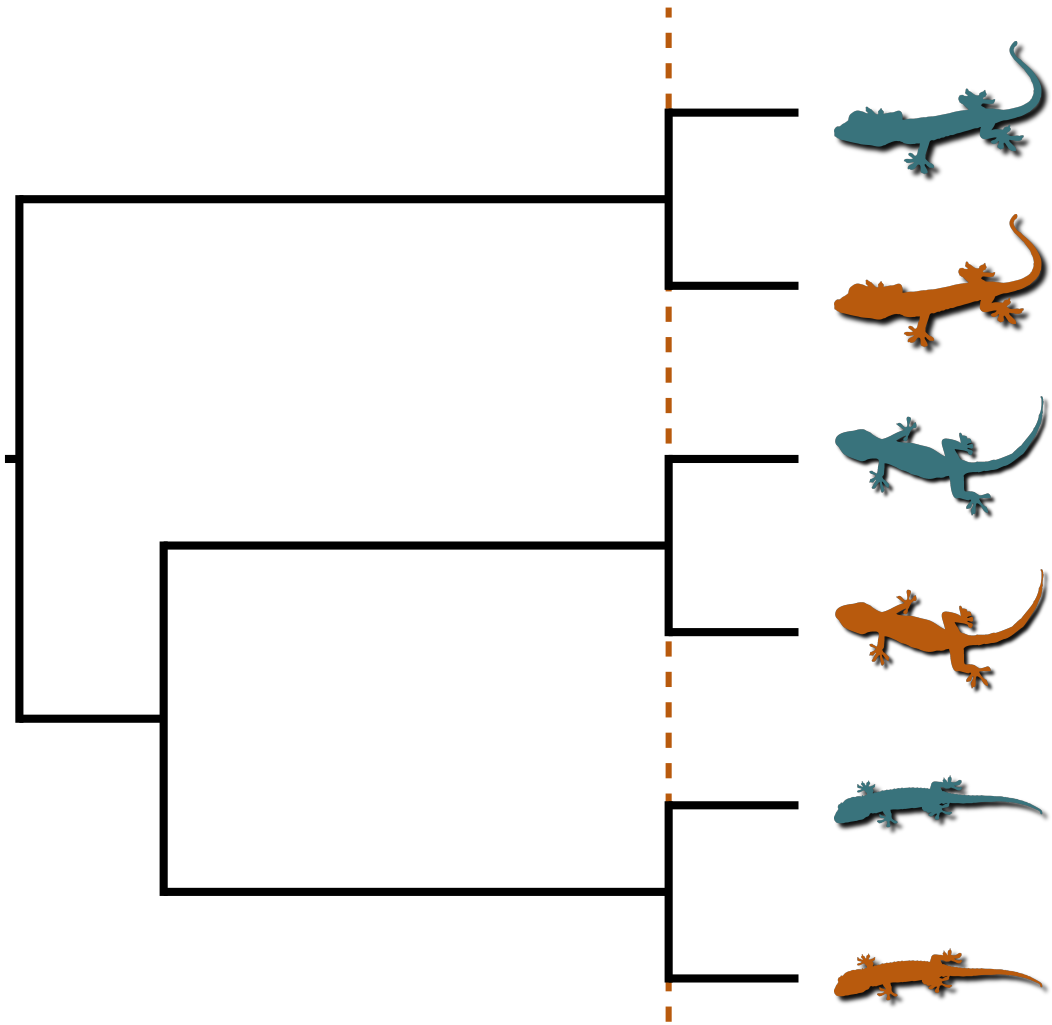
Jamie R. Oaks

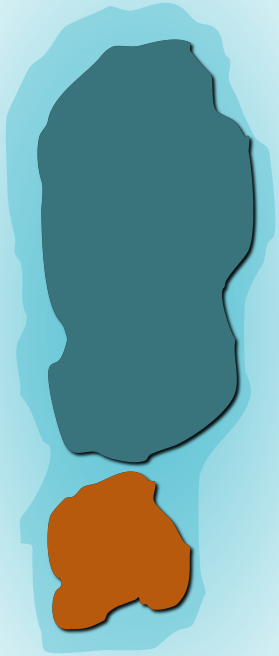
Department of Biological Sciences &
Museum of Natural History, Auburn University

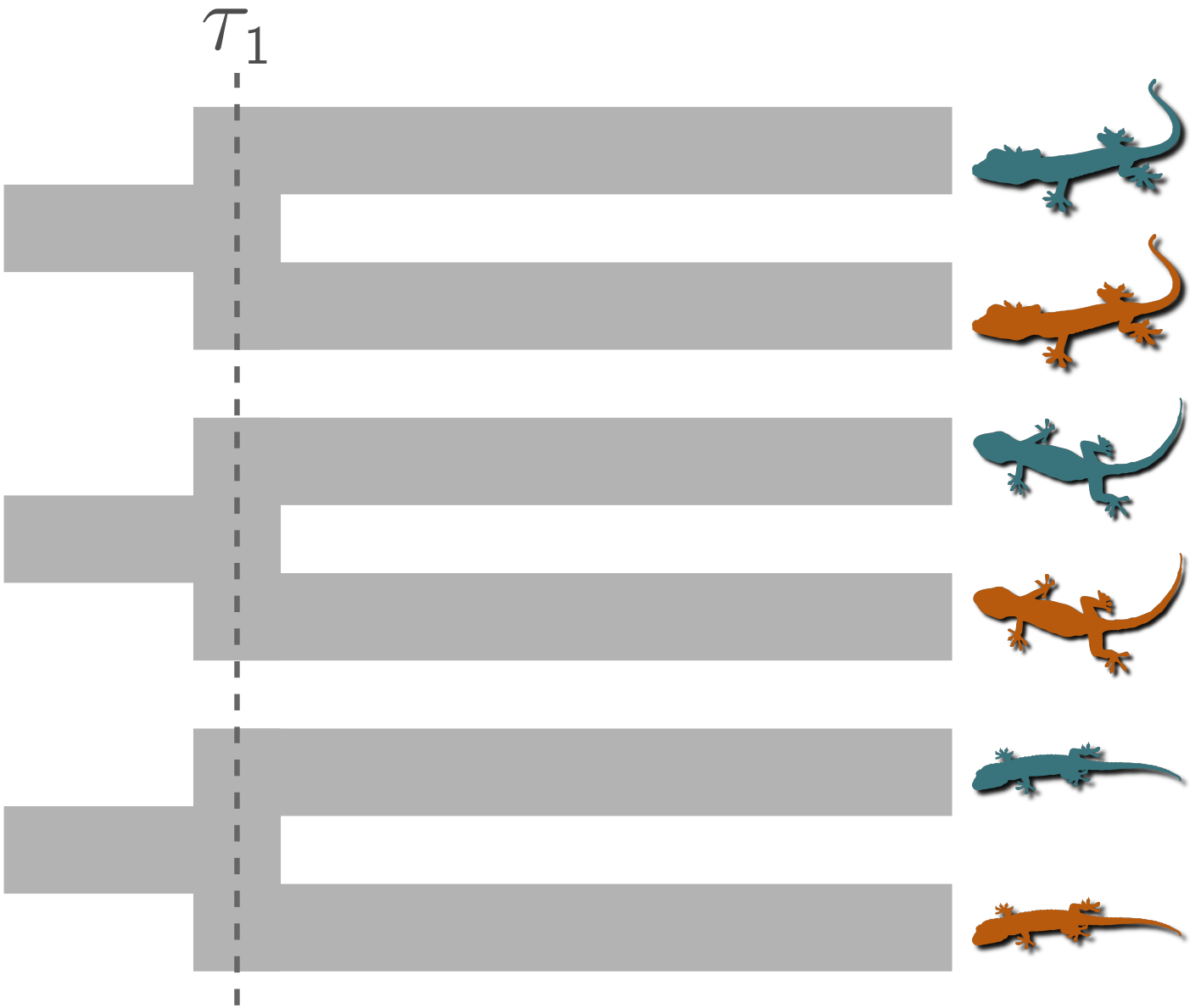
June 4, 2018

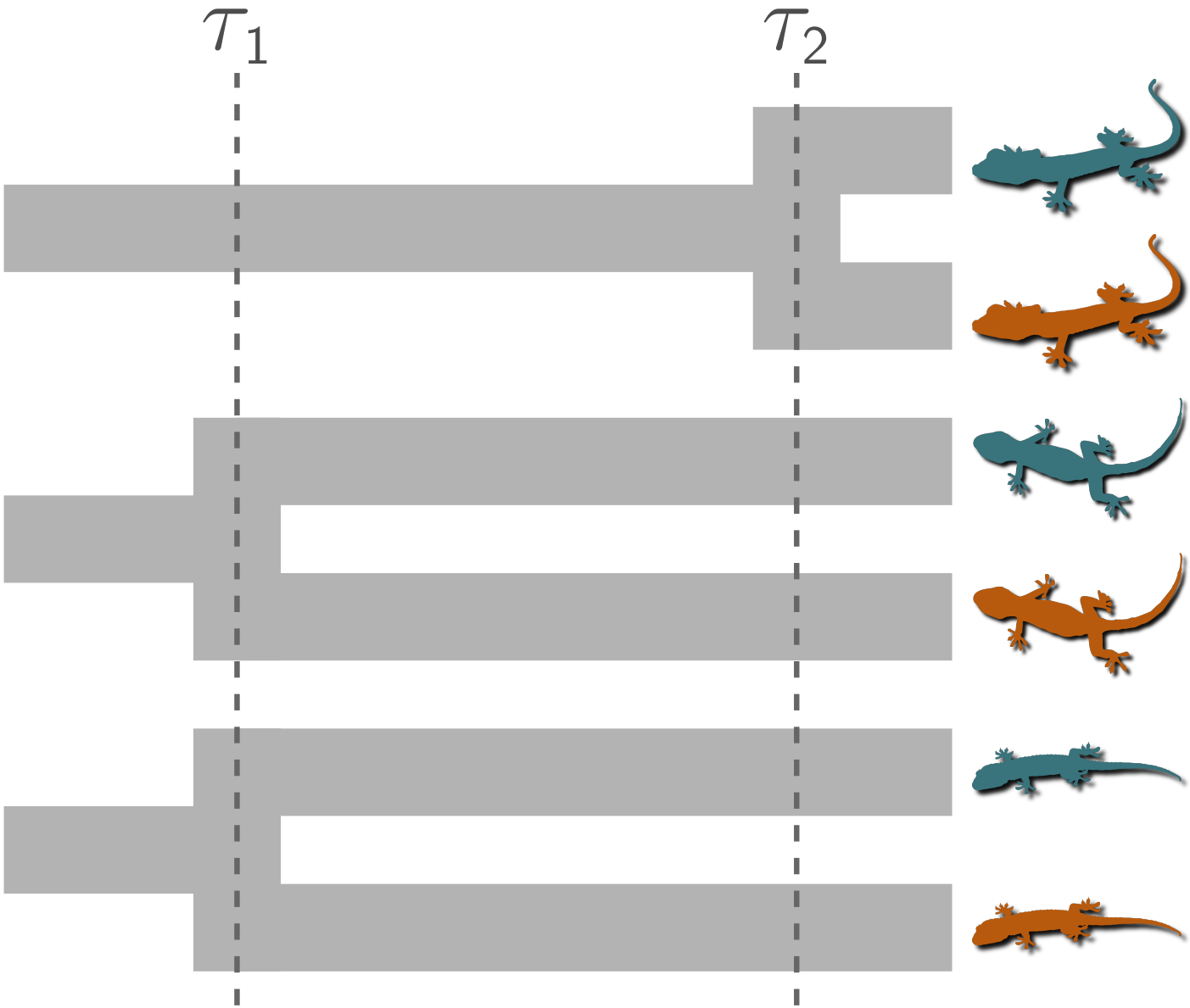


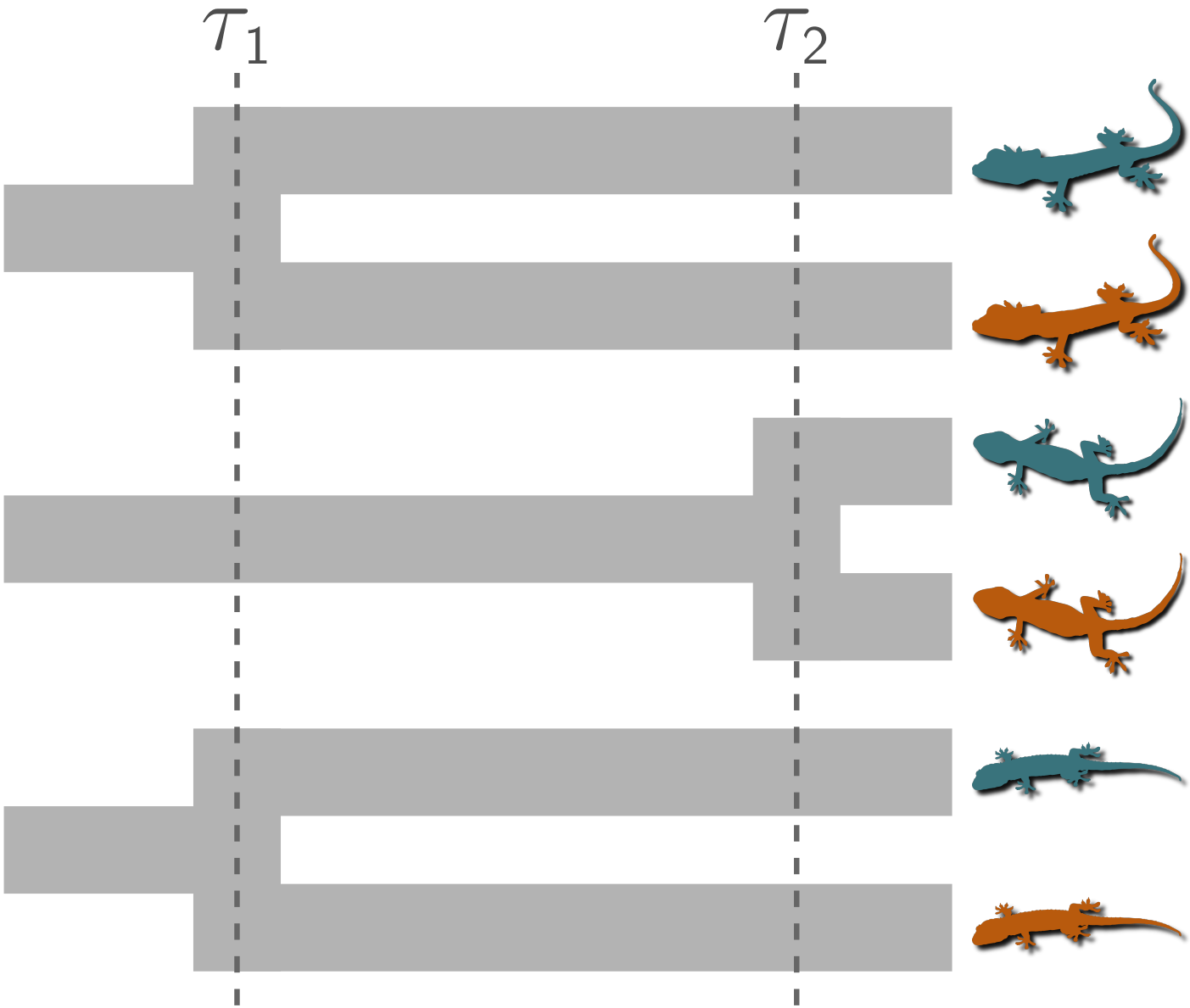


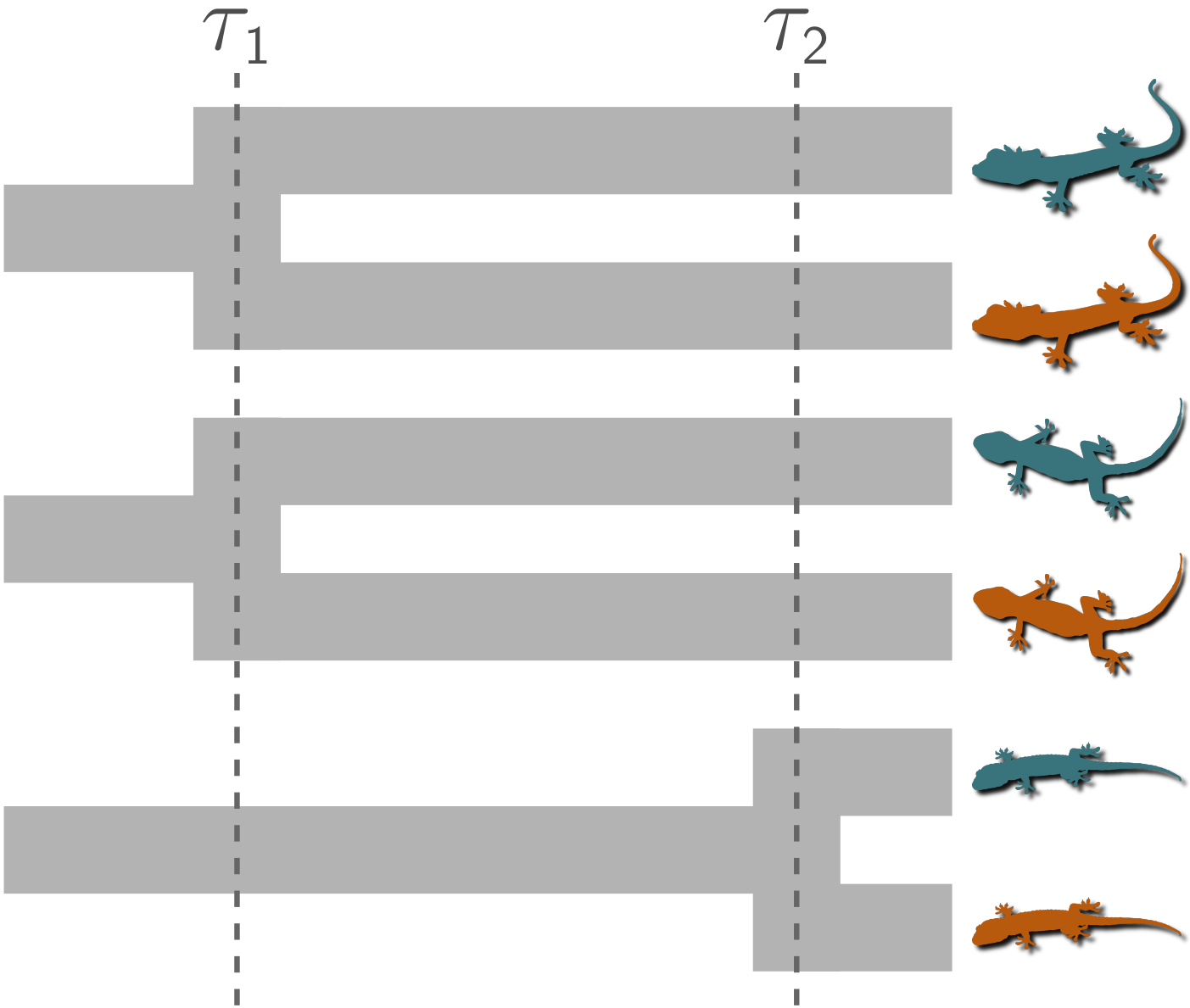


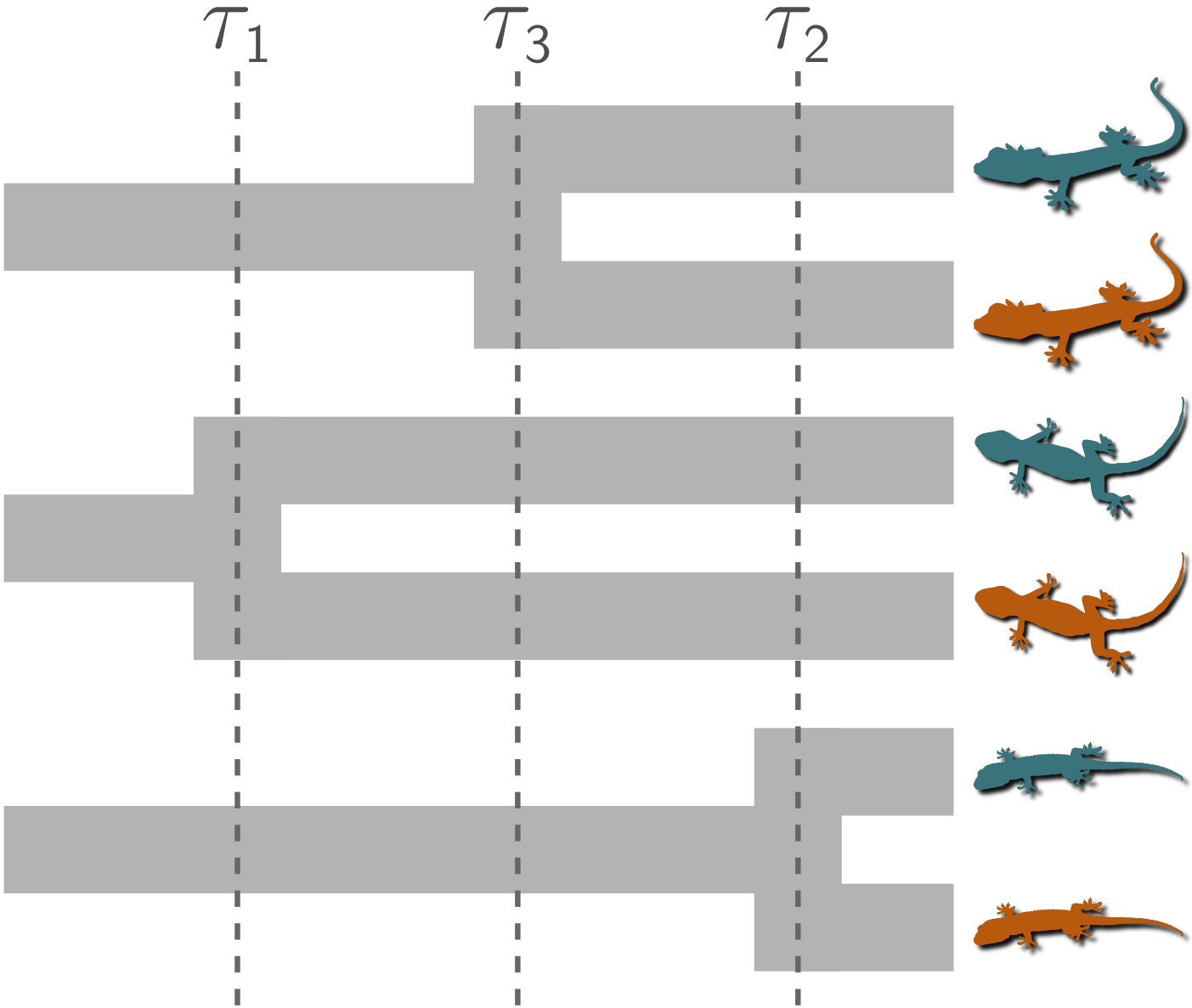



τ_1

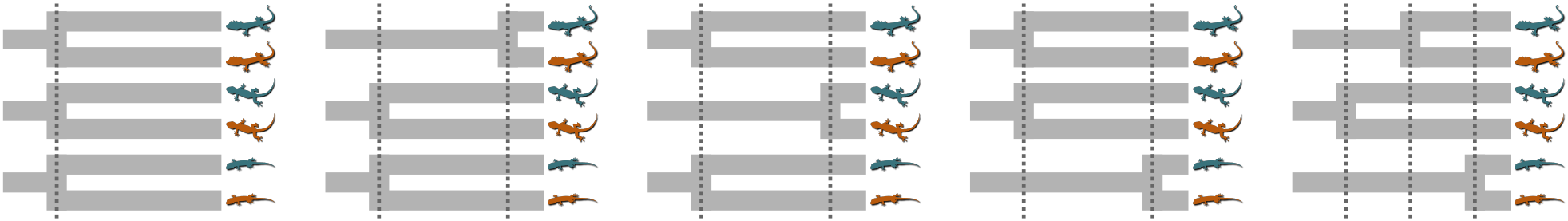

τ_1
τ_2

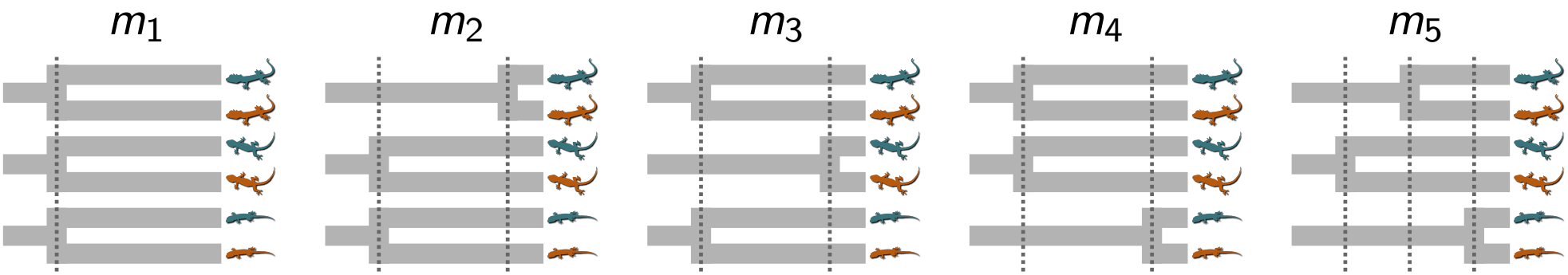

τ_1
τ_2


τ_1
τ_2

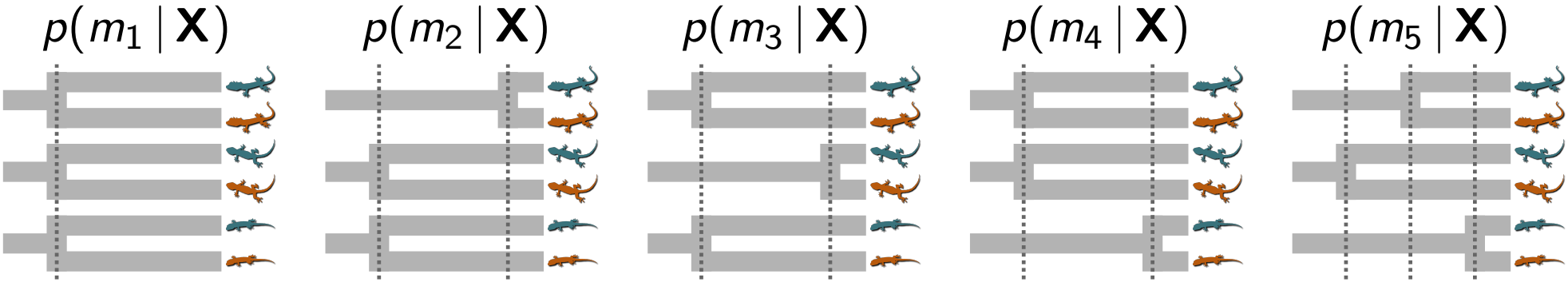

τ_1
τ_3
τ_2

m_1 m_2 m_3 m_4 m_5



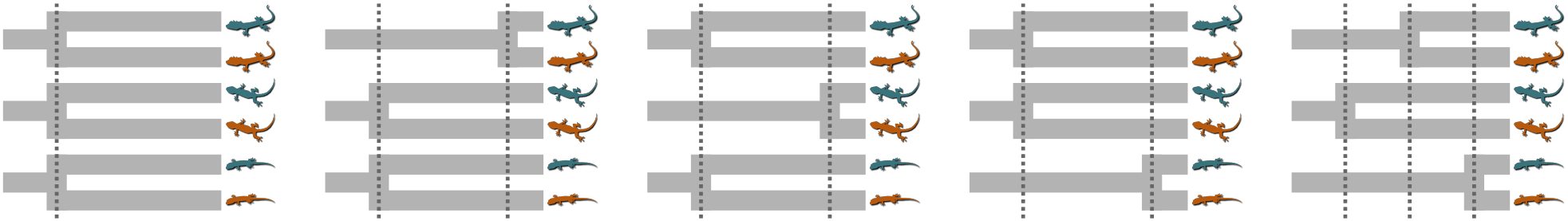


We want to infer the model and divergence times given genetic data



We want to infer the model and divergence times given genetic data

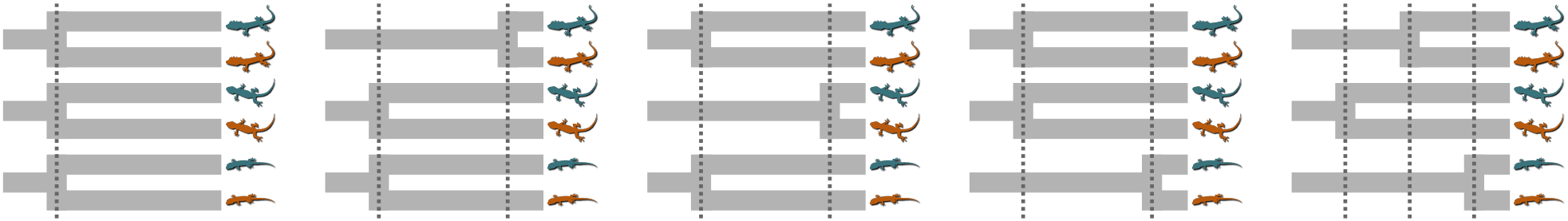
$p(m_1 \mid \mathbf{X})$ $p(m_2 \mid \mathbf{X})$ $p(m_3 \mid \mathbf{X})$ $p(m_4 \mid \mathbf{X})$ $p(m_5 \mid \mathbf{X})$



We want to infer the model and divergence times given genetic data

$$p(m_i \mid \mathbf{X}) \propto p(\mathbf{X} \mid m_i)p(m_i)$$

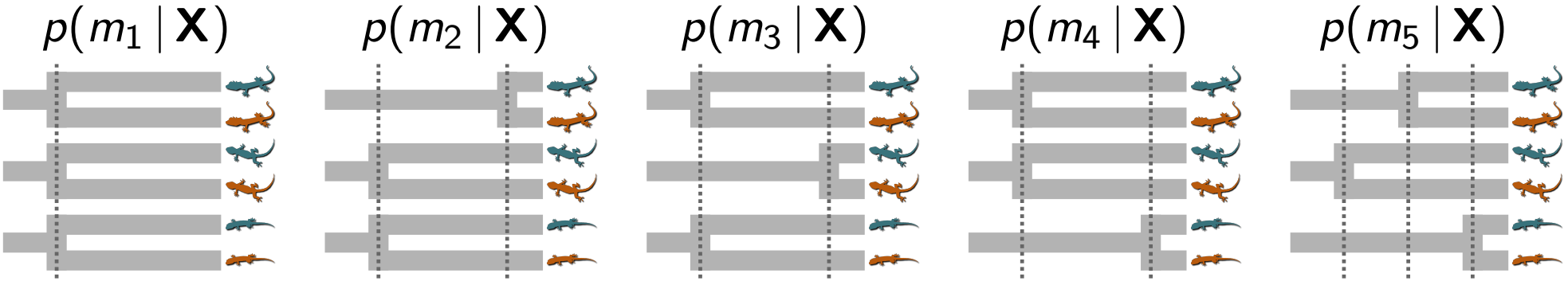
$p(m_1 \mid \mathbf{X})$ $p(m_2 \mid \mathbf{X})$ $p(m_3 \mid \mathbf{X})$ $p(m_4 \mid \mathbf{X})$ $p(m_5 \mid \mathbf{X})$



We want to infer the model and divergence times given genetic data

$$p(m_i \mid \mathbf{X}) \propto p(\mathbf{X} \mid m_i)p(m_i)$$

$$p(\mathbf{X} \mid m_i) = \int_\theta p(\mathbf{X} \mid \theta, m_i)p(\theta \mid m_i)d\theta$$



We want to infer the model and divergence times given genetic data

$$p(m_i \mid \mathbf{X}) \propto p(\mathbf{X} \mid m_i)p(m_i)$$

$$p(\mathbf{X} \mid m_i) = \int_{\theta} p(\mathbf{X} \mid \theta, m_i)p(\theta \mid m_i)d\theta$$

- Divergence times
- Gene trees
- Substitution parameters
- Demographic parameters

- Let's assume we are interested in the probability of a coin we have not seen landing heads-side up when it is flipped (θ)

- Let's assume we are interested in the probability of a coin we have not seen landing heads-side up when it is flipped (θ)
- Before flipping, we decide to compare two models that vary in our prior assumptions about the probability of the coin landing heads up

- Let's assume we are interested in the probability of a coin we have not seen landing heads-side up when it is flipped (θ)
- Before flipping, we decide to compare two models that vary in our prior assumptions about the probability of the coin landing heads up
- We assume:
  1. The coin is probably fair

     M_1: $\theta \sim \mathrm{Beta}(5.0, 5.0)$

  2. the coin is weighted to land tails side up most of time

     M_2: $\theta \sim \mathrm{Beta}(1.0, 5.0)$

- We do 100 flips and 50 land heads up

- We do 100 flips and 50 land heads up
- Now, we can calculate the posterior distribution for the probability of landing heads up under both our models

- We do 100 flips and 50 land heads up
- Now, we can calculate the posterior distribution for the probability of landing heads up under both our models

$$p(\theta \mid D, M_i) = \frac{p(D \mid \theta, M_i)p(\theta \mid M_i)}{p(D \mid M_i)}$$

- We do 100 flips and 50 land heads up
- Now, we can calculate the posterior distribution for the probability of landing heads up under both our models

$$p(\theta \mid D, M_i) = \frac{p(D \mid \theta, M_i)p(\theta \mid M_i)}{p(D \mid M_i)}$$

- We see the posterior distribution of θ is very robust to our prior assumptions
- https://kerrycobb.github.io/beta-binomial-web-demo/

- However, we want to compare the ability of the models to explain the data

- However, we want to compare the ability of the models to explain the data
- We need to average (integrate) the likelihood density function over all possible values of θ, weighting by the prior

- However, we want to compare the ability of the models to explain the data
- We need to average (integrate) the likelihood density function over all possible values of θ, weighting by the prior

$$p(D \mid M_1) = \int_{\theta} p(D \mid \theta, M_1) p(\theta \mid M_1) \mathrm{d}\theta$$

- However, we want to compare the ability of the models to explain the data
- We need to average (integrate) the likelihood density function over all possible values of θ, weighting by the prior

$$p(D \mid M_1) = \int_{\theta} p(D \mid \theta, M_1) p(\theta \mid M_1) \mathrm{d}\theta$$

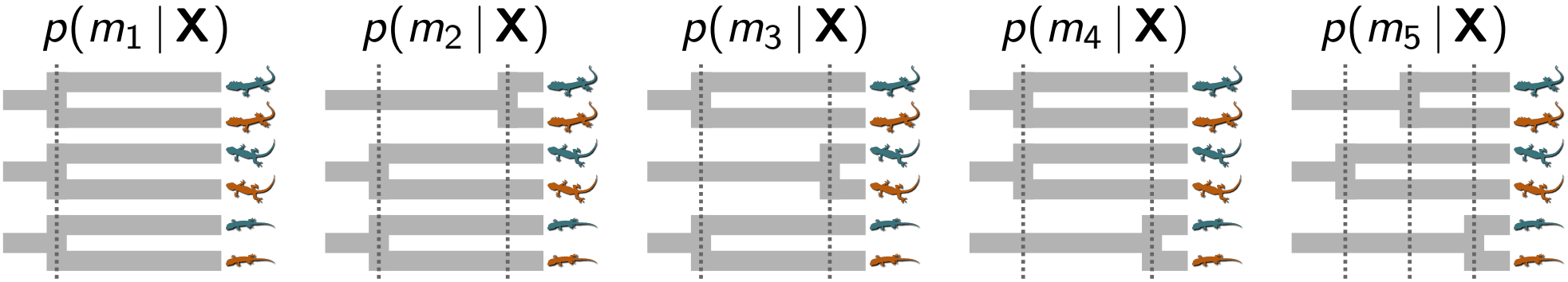
- The marginal likelihoods of our model is sensitive to the prior no matter how much we flip the coin.

- Why do we care about the marginal likelihood?

- Why do we care about the marginal likelihood?
- It's *the evidence* that updates our prior to give us the posterior probability of the model

- Why do we care about the marginal likelihood?
- It's *the evidence* that updates our prior to give us the posterior probability of the model

$$p(M_1 \mid D) = \frac{p(D \mid M_1)p(M_1)}{p(D \mid M_1)p(M_1) + p(D \mid M_2)p(M_2)}$$



We want to infer the model and divergence times given genetic data

$$p(m_i \mid \mathbf{X}) \propto p(\mathbf{X} \mid m_i)p(m_i)$$

$$p(\mathbf{X} \mid m_i) = \int_\theta p(\mathbf{X} \mid \theta, m_i)p(\theta \mid m_i)d\theta$$

- Divergence times
- Gene trees
- Substitution parameters
- Demographic parameters

Ecoevolity: Estimating evolutionary coevality

[1] R. M. Neal (2000). *Journal of Computational and Graphical Statistics* 9: 249–265

[2] D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932

Ecoevolity: Estimating evolutionary coevality

- CTMC model of characters evolving along genealogies
- Coalescent model of genealogies branching within populations
- Dirichlet-process prior across divergence models
- Gibbs sampling[1] to numerically sample models
- Analytically integrate over genealogies[2]

[1] R. M. Neal (2000). *Journal of Computational and Graphical Statistics* 9: 249–265

[2] D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932

Ecoevolity: Estimating evolutionary coevality

- CTMC model of characters evolving along genealogies
- Coalescent model of genealogies branching within populations
- Dirichlet-process prior across divergence models
- Gibbs sampling[1] to numerically sample models
- Analytically integrate over genealogies[2]

- *Goal: Fast, full-likelihood Bayesian method to infer patterns of co-diversification from genome-scale data*

[1] R. M. Neal (2000). *Journal of Computational and Graphical Statistics* 9: 249–265

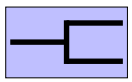
[2] D. Bryant et al. (2012). *Molecular Biology and Evolution* 29: 1917–1932

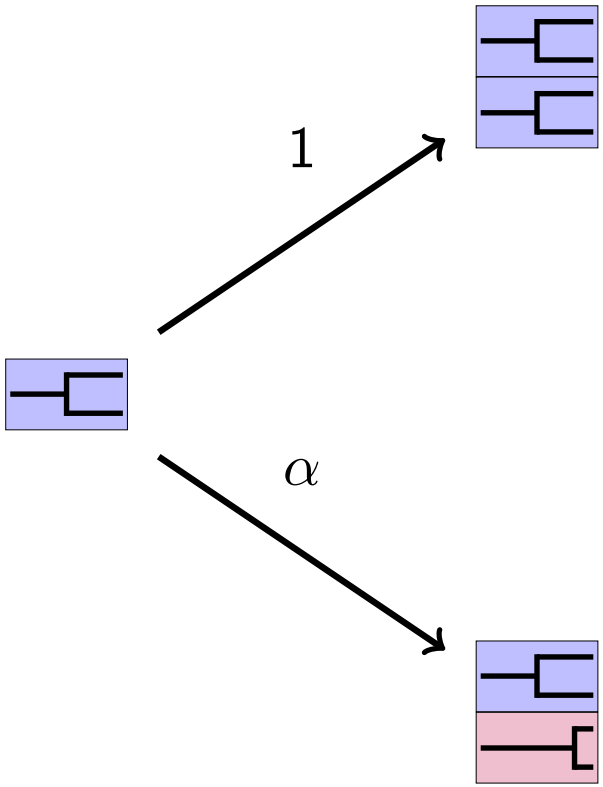
Sampling divergence models

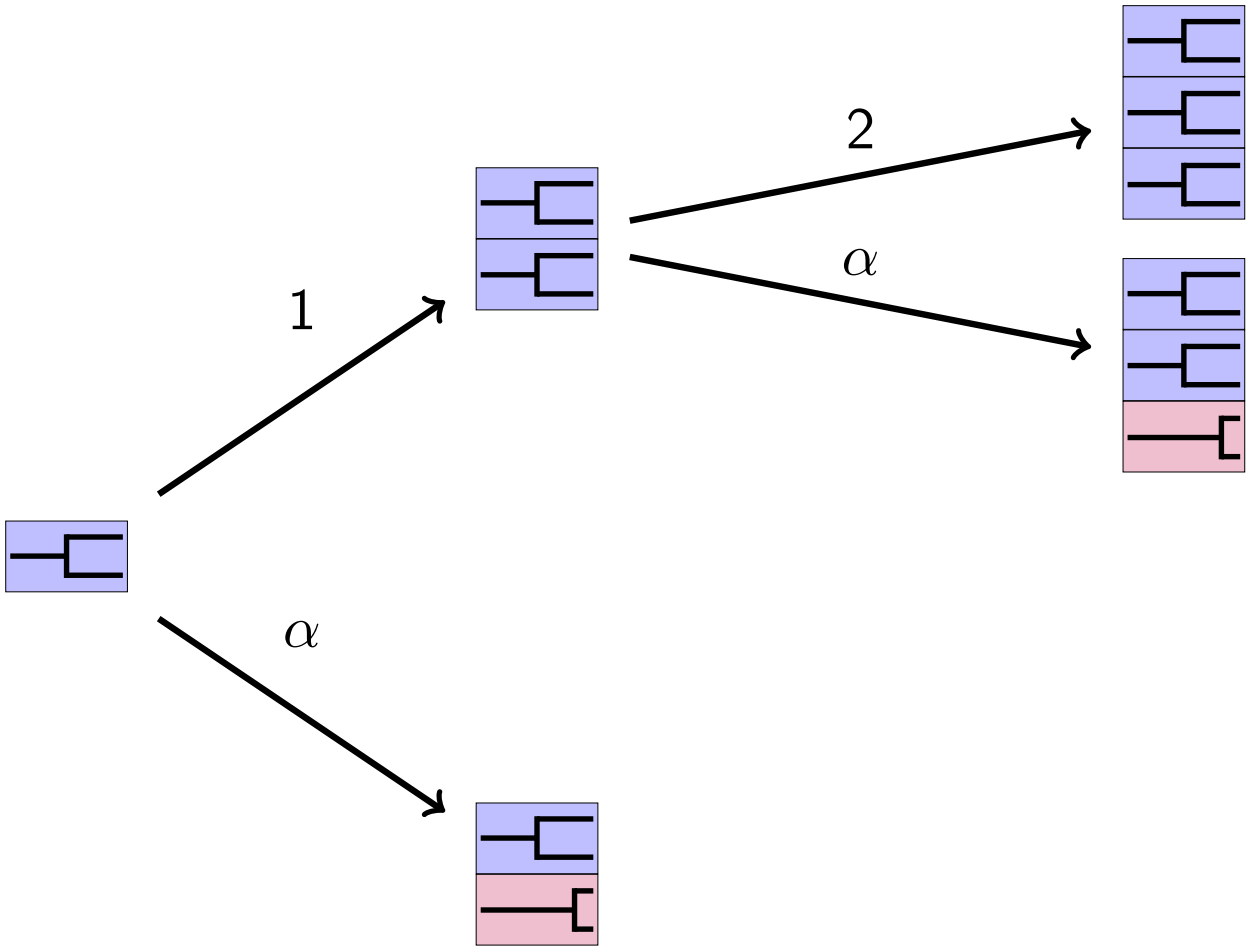
Sampling divergence models

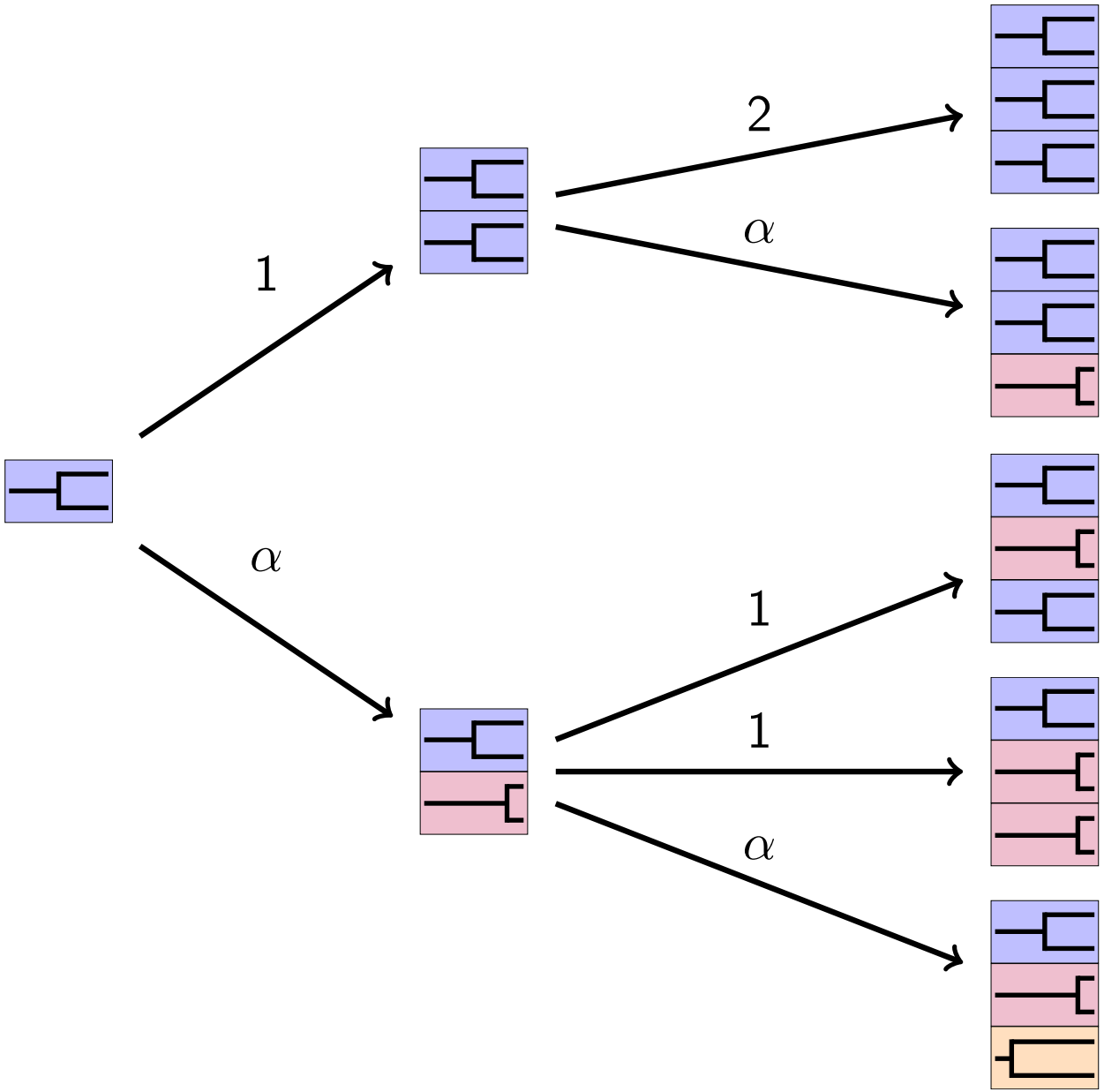
- The divergence models are ways of assigning our taxa to events
- A Dirichlet process prior (DPP) model is a convenient and flexible solution
  - Common Bayesian approach to assigning variables to an unknown number of categories
  - Controlled by "concentration" parameter: α

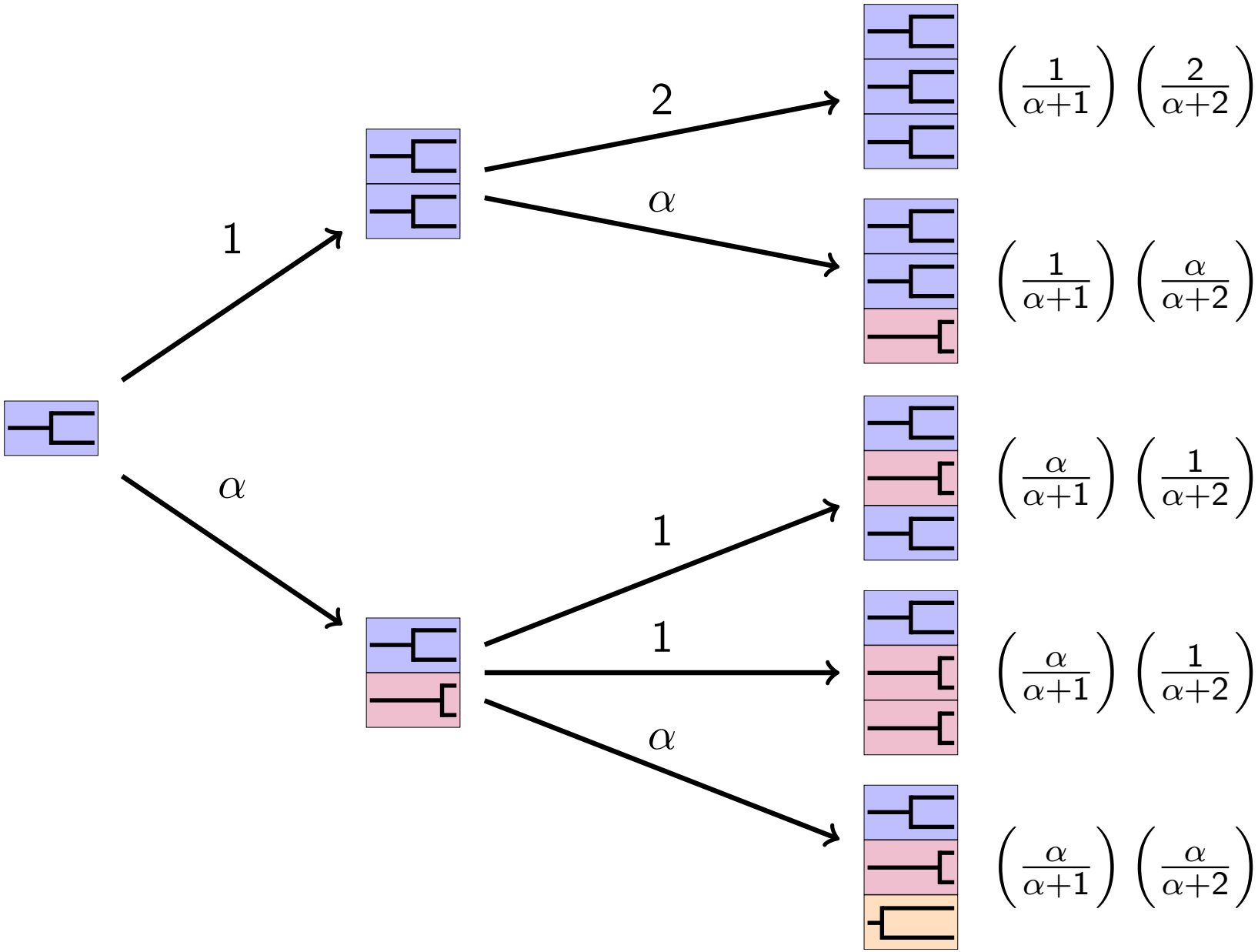
Peter Dirichlet



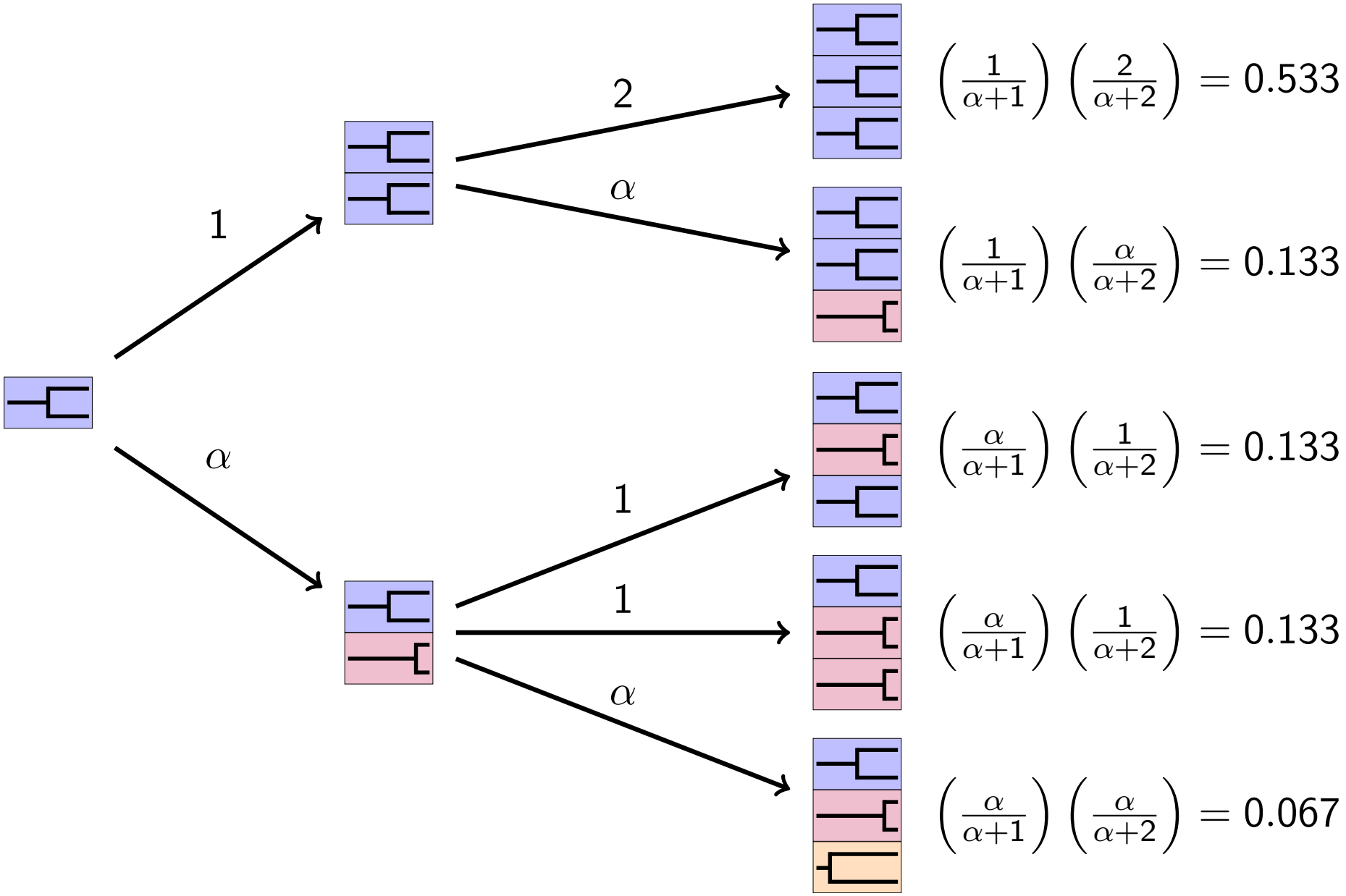

1
α


1
α
2
α

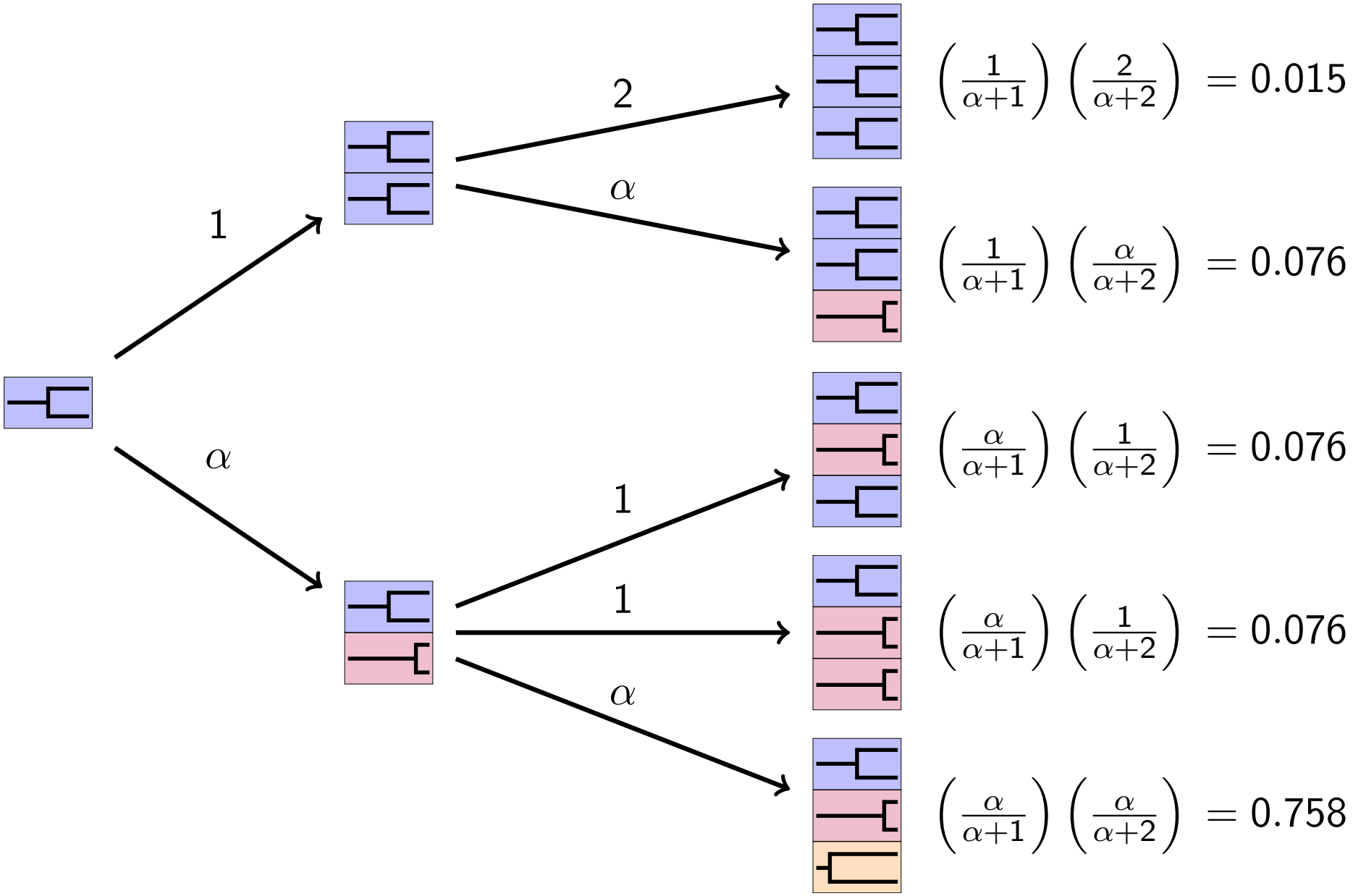

1
α
2
α
1
1
α


1
α
2
α
1
1
α
$\left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right)$
$\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)$

$\alpha = 0.5$

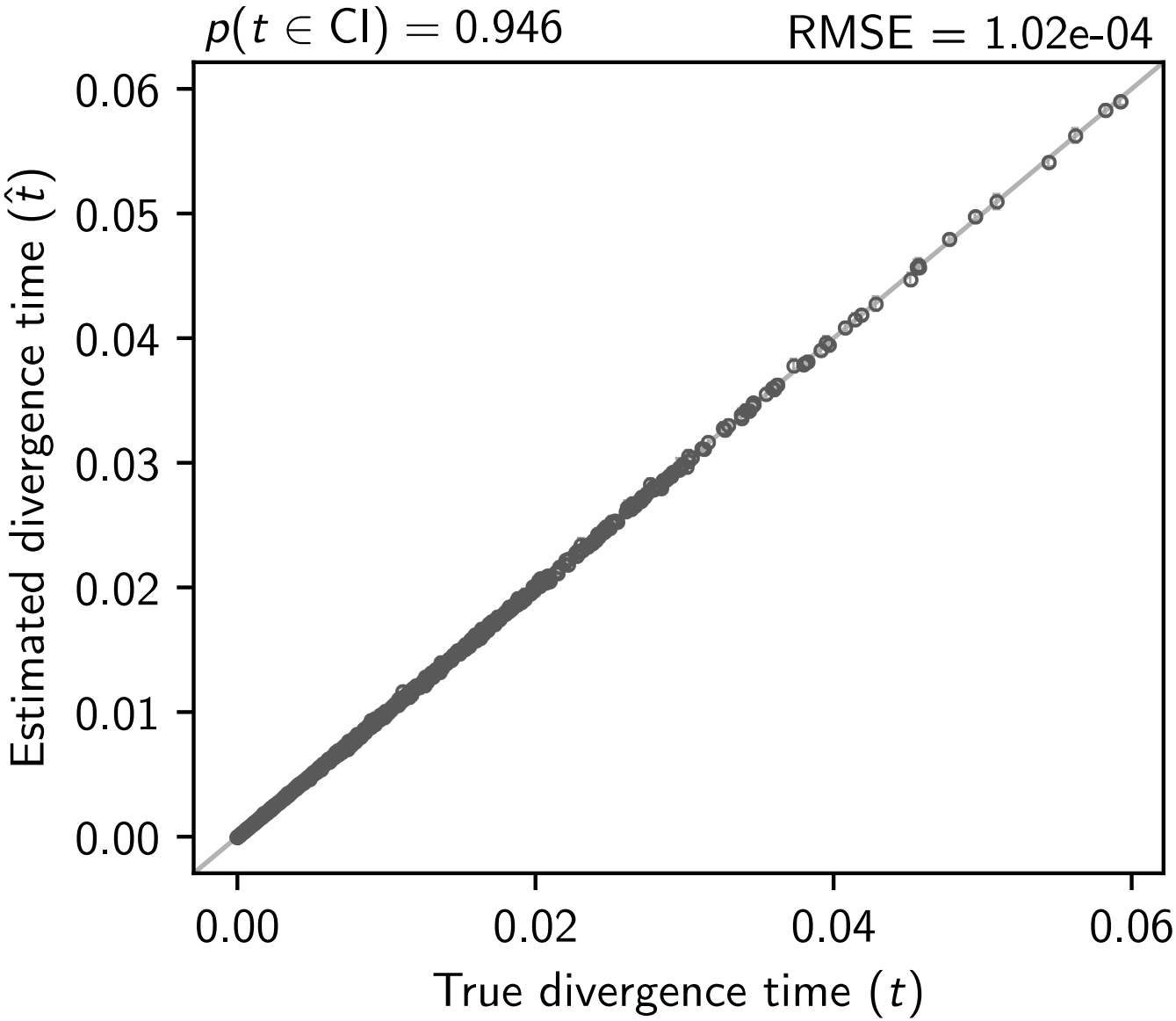

1
α
2
α
1
1
α
$\left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right) = 0.533$
$\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right) = 0.133$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right) = 0.133$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right) = 0.133$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right) = 0.067$

$\alpha = 10.0$

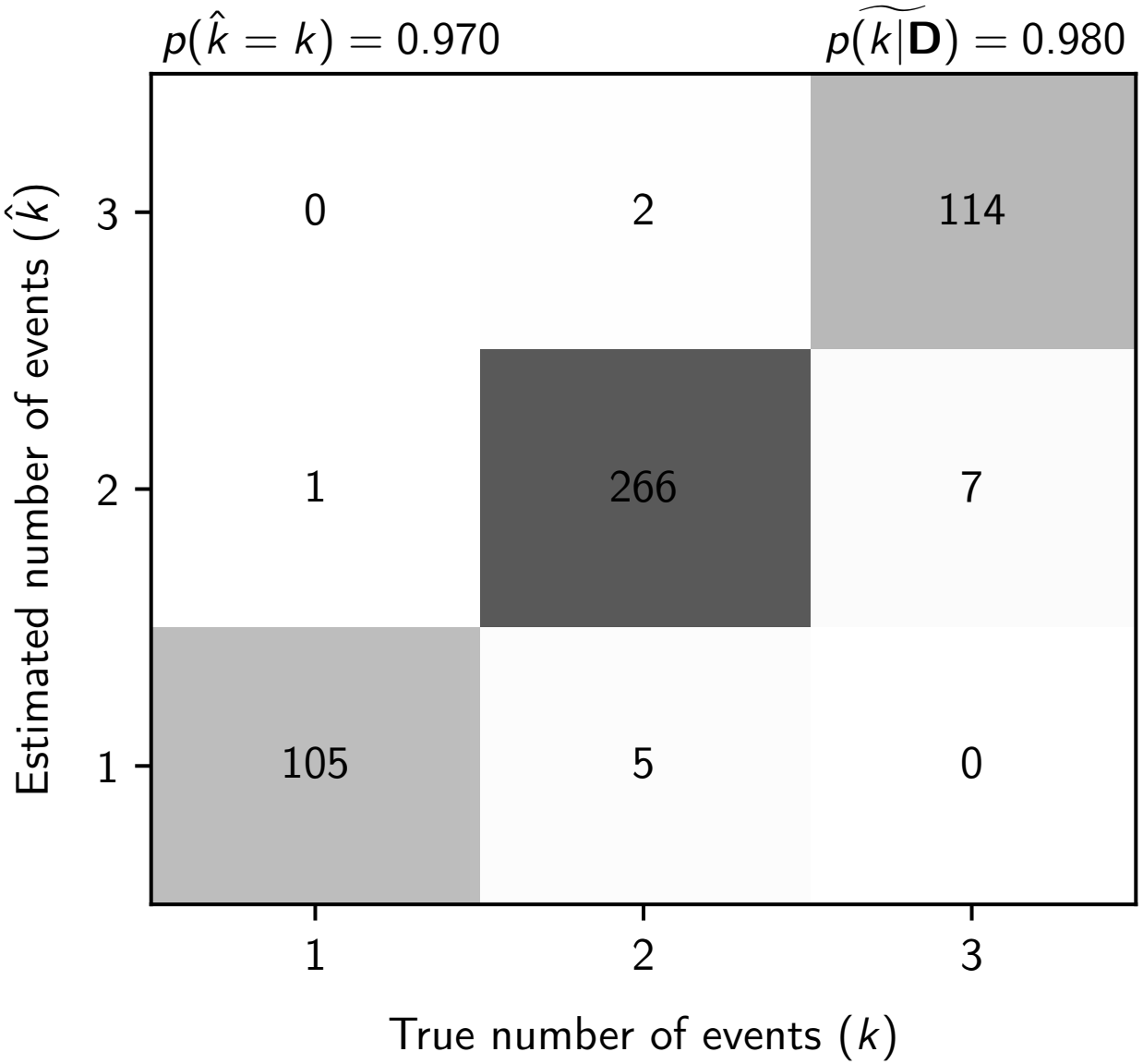

1
α
2
α
1
1
α
$\left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right) = 0.015$
$\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right) = 0.076$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right) = 0.076$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right) = 0.076$
$\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right) = 0.758$

Does it work?

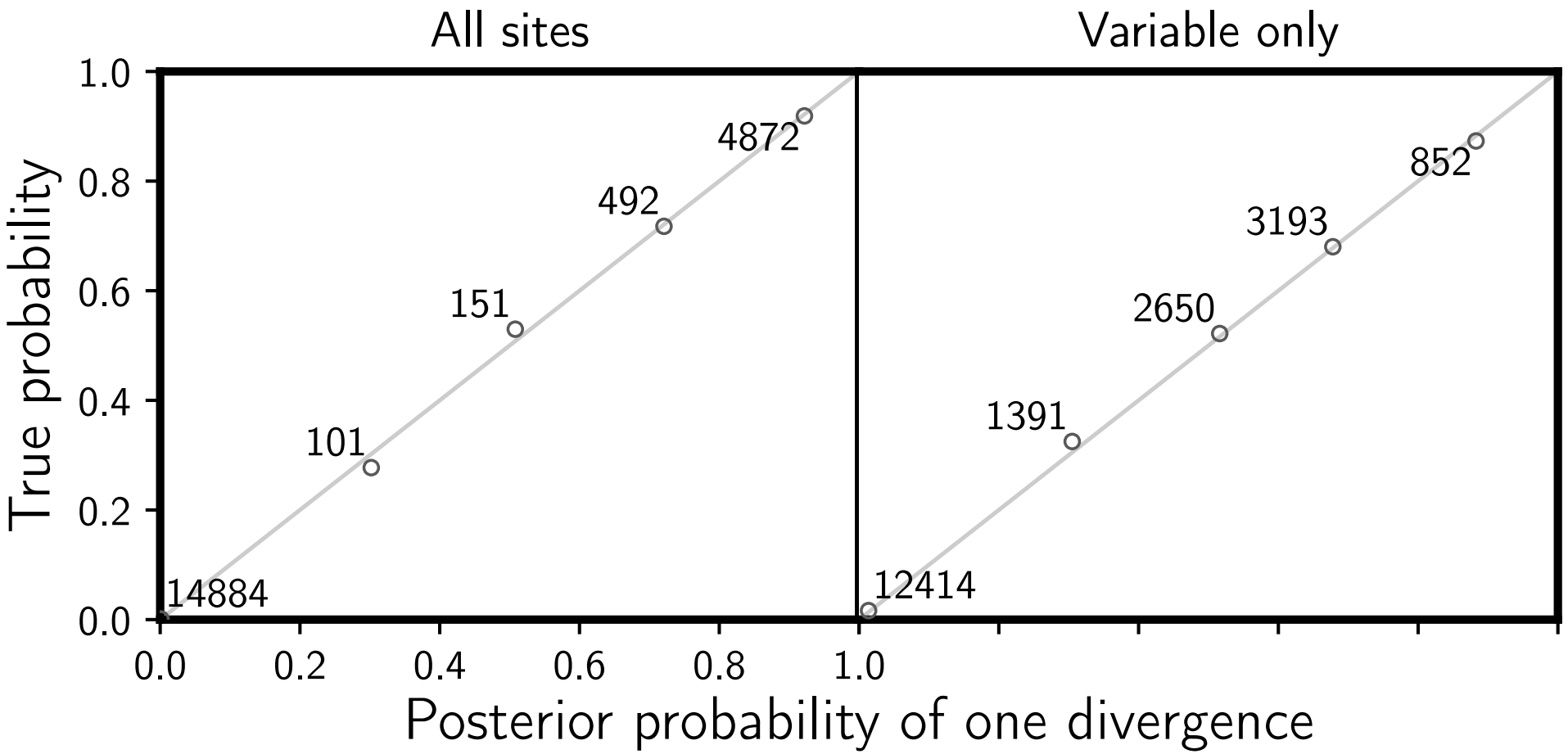
Ecoevolity: Simulation-based assessment

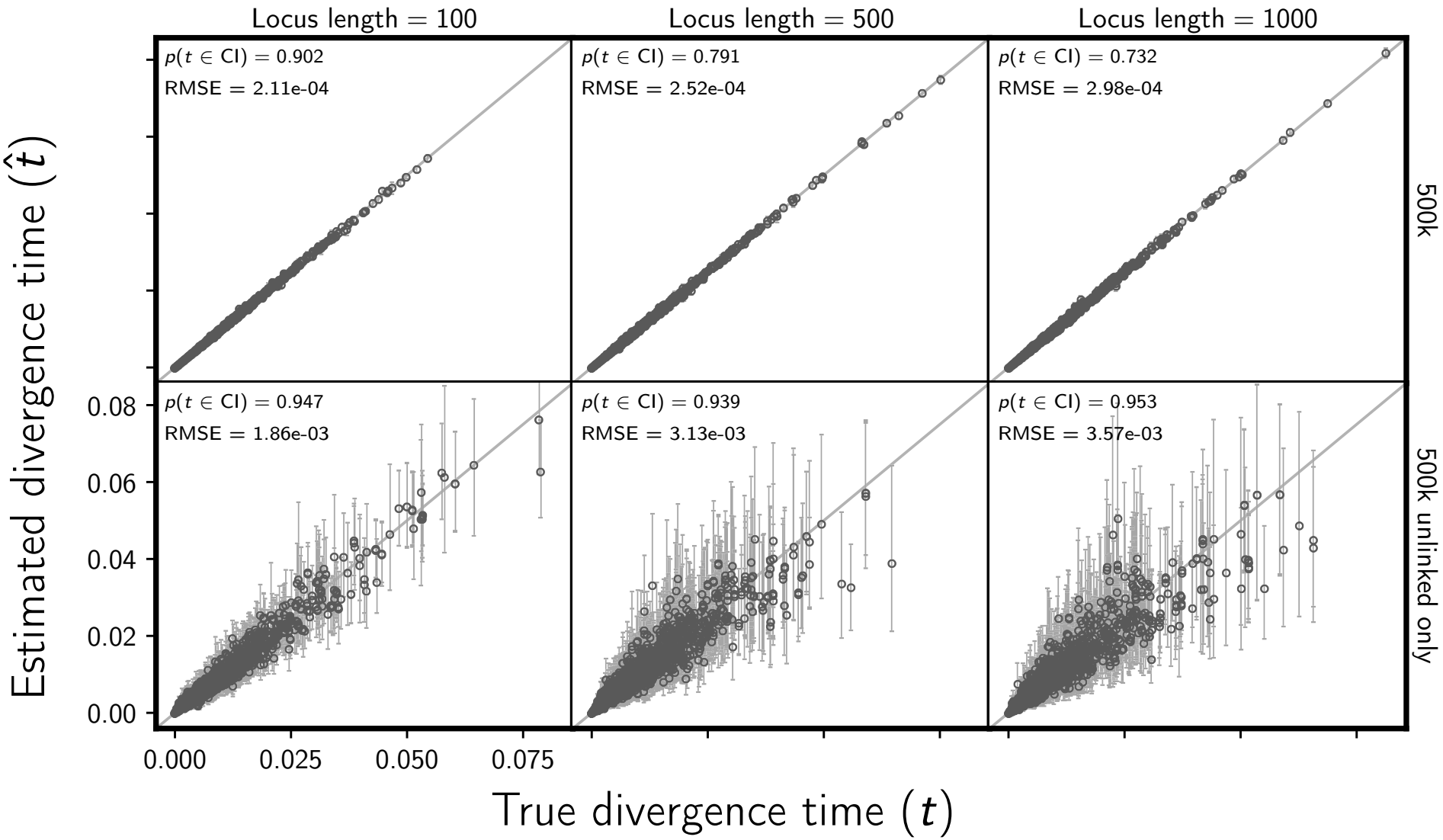


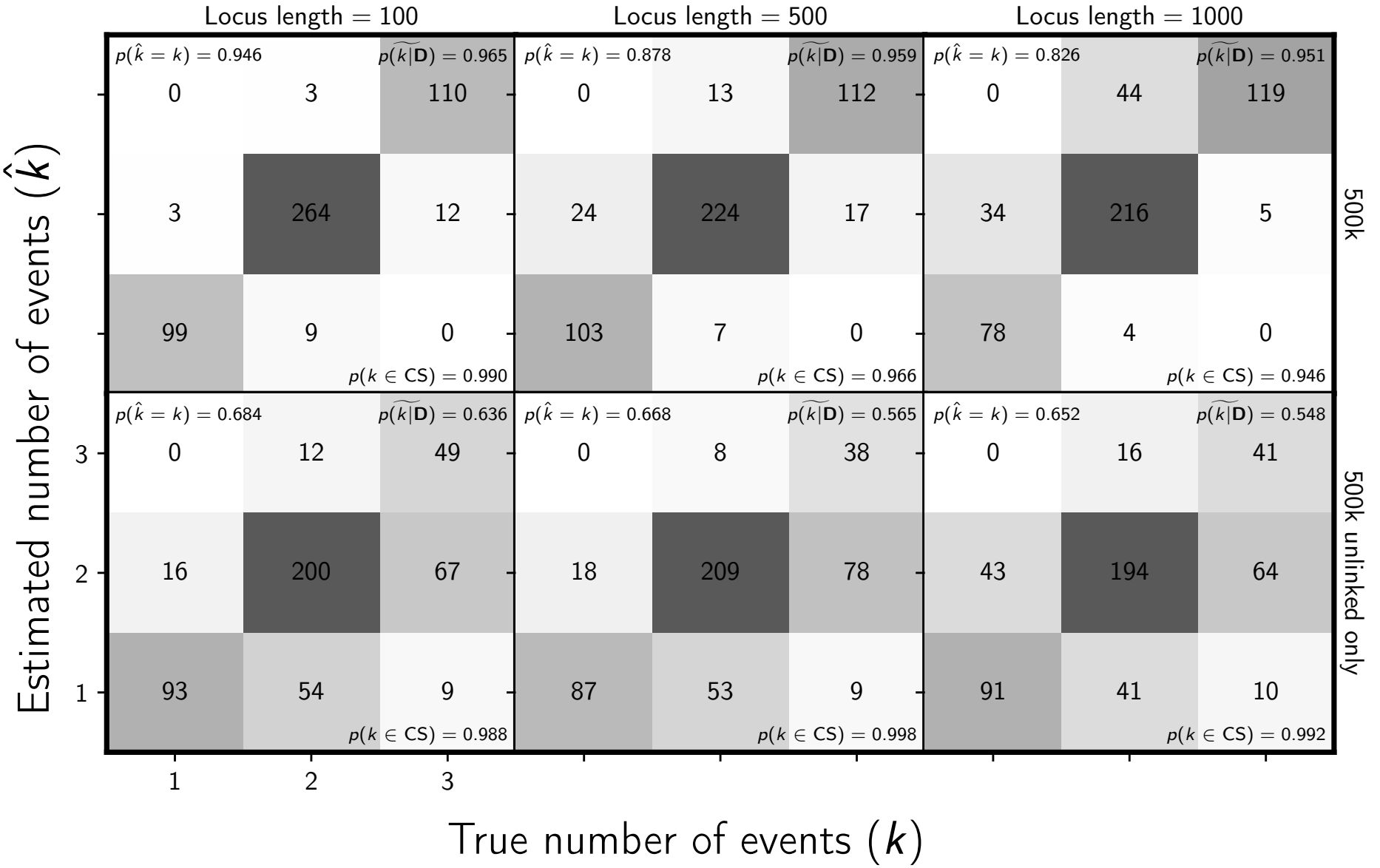
Ecoevolity: Simulation-based assessment



Ecoevolity: Simulation-based assessment




Locus length = 100
Locus length = 500
Locus length = 1000
p(t ∈ CI) = 0.902
RMSE = 2.11e-04
p(t ∈ CI) = 0.791
RMSE = 2.52e-04
p(t ∈ CI) = 0.732
RMSE = 2.98e-04
p(t ∈ CI) = 0.947
RMSE = 1.86e-03
p(t ∈ CI) = 0.939
RMSE = 3.13e-03
p(t ∈ CI) = 0.953
RMSE = 3.57e-03
500k
500k unlinked only
Estimated divergence time (t̂)
True divergence time (t)
0.00
0.02
0.04
0.06
0.08
0.000
0.025
0.050
0.075


Locus length = 100
Locus length = 500
Locus length = 1000
500k
500k unlinked only
$p(\hat{k} = k) = 0.946$
$p(\widetilde{k|\mathbf{D}}) = 0.965$
0 3 110
3 264 12
99 9 0
$p(k \in \text{CS}) = 0.990$
$p(\hat{k} = k) = 0.878$
$p(\widetilde{k|\mathbf{D}}) = 0.959$
0 13 112
24 224 17
103 7 0
$p(k \in \text{CS}) = 0.966$
$p(\hat{k} = k) = 0.826$
$p(\widetilde{k|\mathbf{D}}) = 0.951$
0 44 119
34 216 5
78 4 0
$p(k \in \text{CS}) = 0.946$
$p(\hat{k} = k) = 0.684$
$p(\widetilde{k|\mathbf{D}}) = 0.636$
0 12 49
16 200 67
93 54 9
$p(k \in \text{CS}) = 0.988$
$p(\hat{k} = k) = 0.668$
$p(\widetilde{k|\mathbf{D}}) = 0.565$
0 8 38
18 209 78
87 53 9
$p(k \in \text{CS}) = 0.998$
$p(\hat{k} = k) = 0.652$
$p(\widetilde{k|\mathbf{D}}) = 0.548$
0 16 41
43 194 64
91 41 10
$p(k \in \text{CS}) = 0.992$
Estimated number of events ($\hat{k}$)
3
2
1
1 2 3
True number of events (k)

Caveats

- We have to make strong prior assumptions about the relative rates of mutation among our taxa.
- The model assumes no migration after divergence