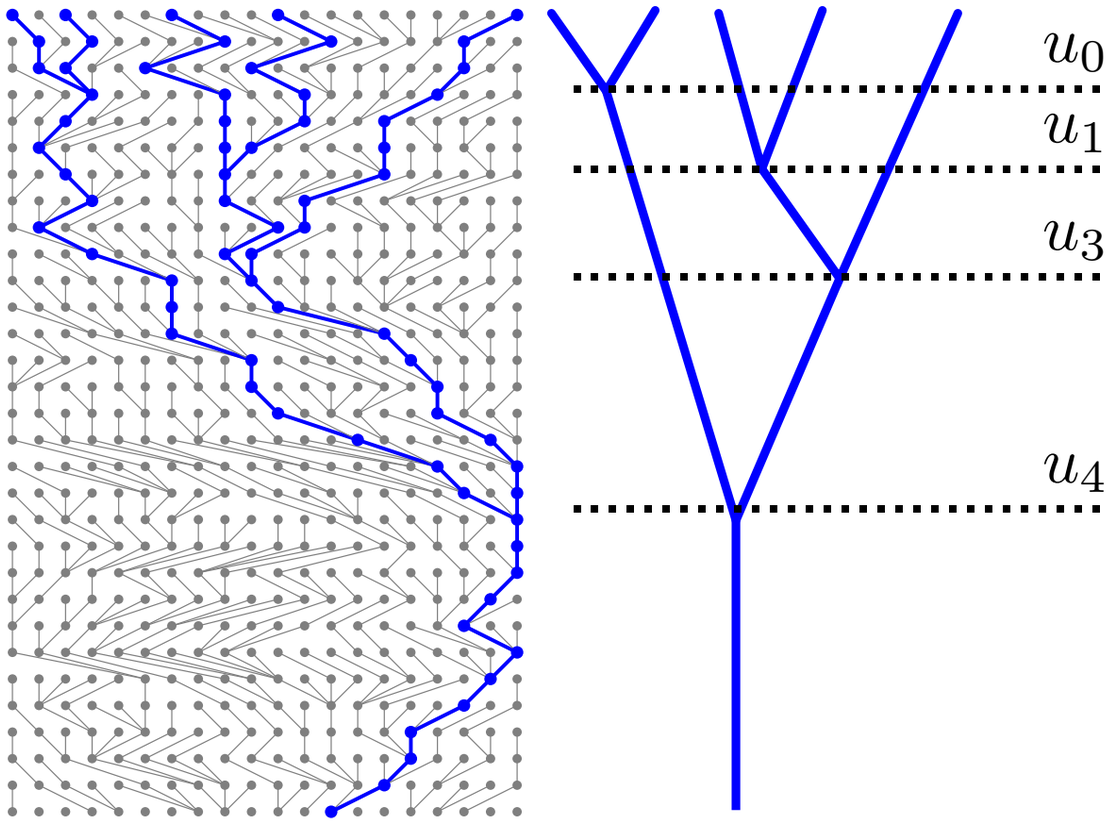
Extension of the basic coalescent

Peter Beerli
Florida State University
#MolEvol2017 MBL Woods Hole

Kingman's coalescent

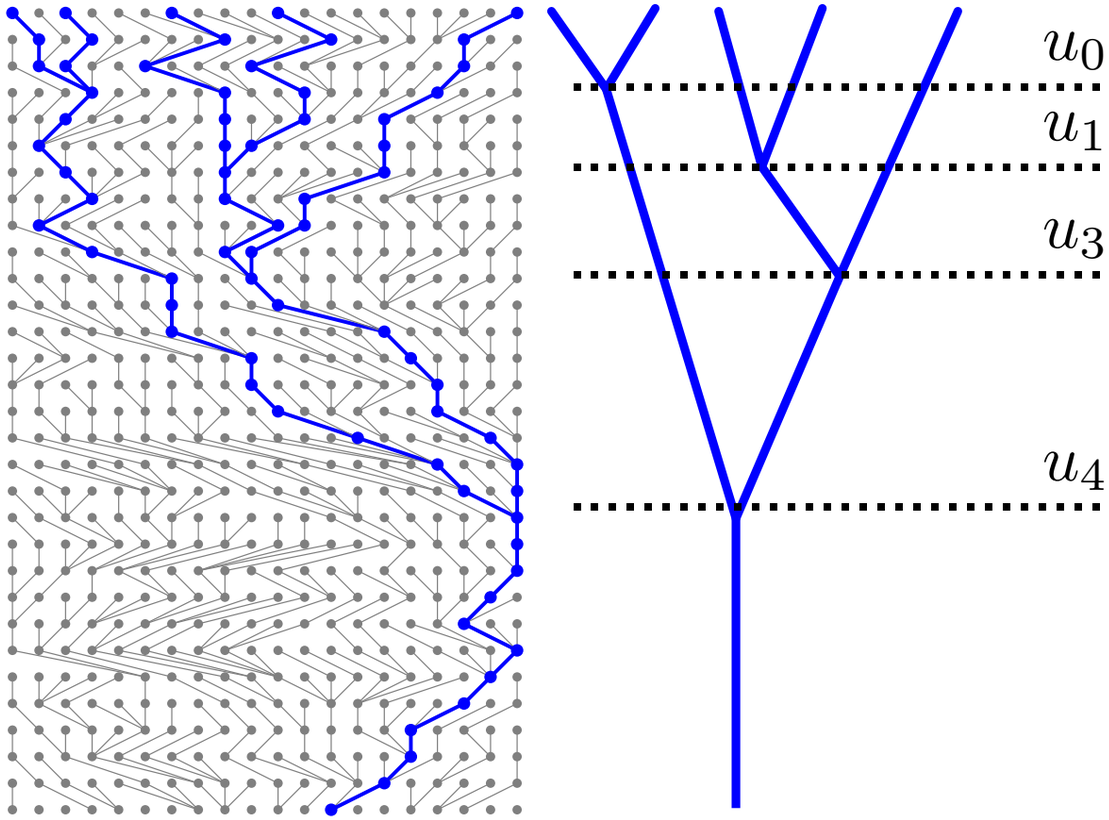


$$\mathrm{P}(G|\Theta) = \prod_{j=0}^{T} e^{-u_j \frac{k_j(k_j-1)}{\Theta}} \frac{2}{\Theta}$$

$$\Theta = 4N_e\mu$$

- calculate the probability that we wait the time interval u until a coalescent
- calculate the probability of the particular coalescent event
- multiply these probabilities for all time intervals

Kingman's coalescent



$$\mathrm{P}(G|N) = \prod_{j=0}^{T}$$

= Waiting time for coalescent event

= Probability of coalescent event

- calculate the probability that we wait the time interval u until a coalescent
- calculate the probability of the particular coalescent event
- multiply these probabilities for all time intervals

Extensions of the basic coalescence

Extensions of the basic coalescence

Extensions of the basic coalescence

Extensions of the basic coalescence

Extensions of the basic coalescence

- Population growth (two parameters), fluctuations, bottlenecks
- Migration among populations (potentially thousands, parameters)
- Population splitting (many parameters)
- Recombination (parameters)
- Shortcut methods
- Genomics and the coalescence

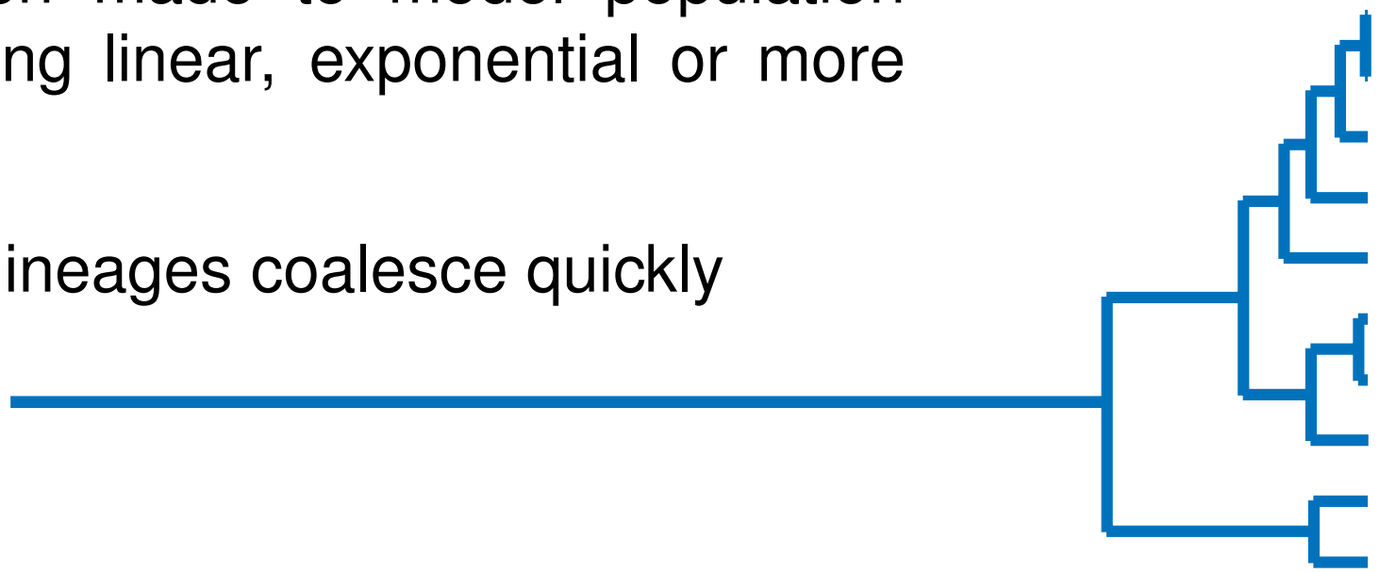
Extensions of the basic coalescent

Populations are rarely completely stable through time, and attempts have been made to model population growth or shrinkage using linear, exponential or more general approaches.

Extensions of the basic coalescent

Populations are rarely completely stable through time, and attempts have been made to model population growth or shrinkage using linear, exponential or more general approaches.

- In a small population lineages coalesce quickly

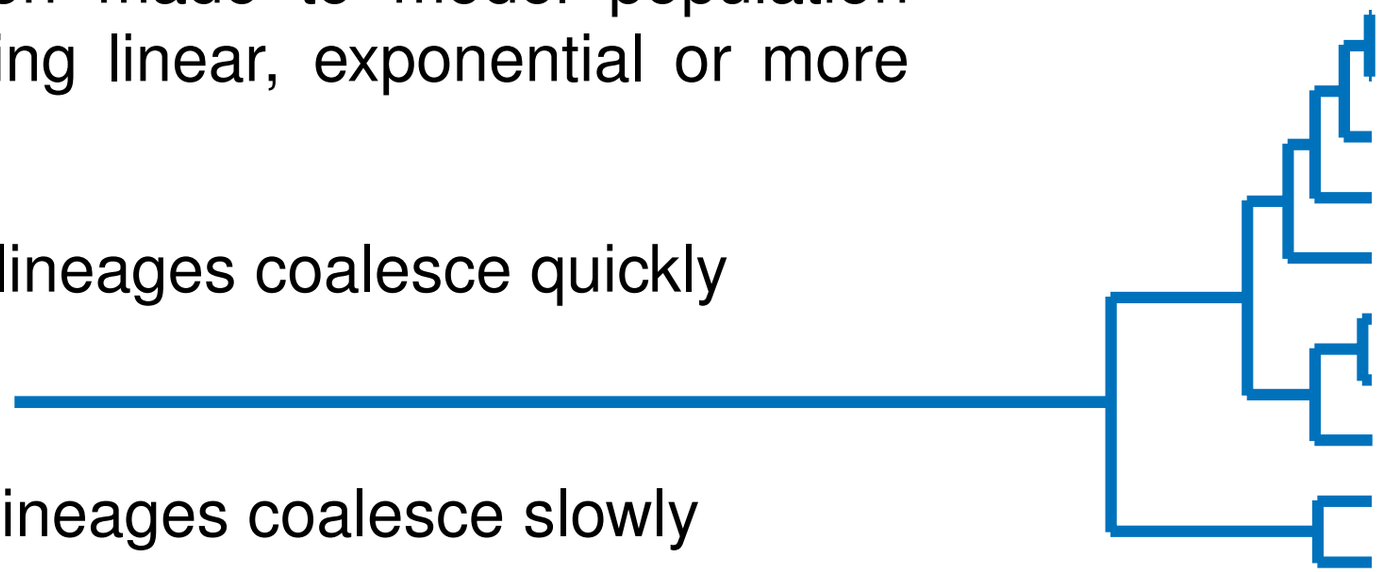


This leaves a signature in the data. We can exploit this and estimate the population growth rate g jointly with the current population size Θ.

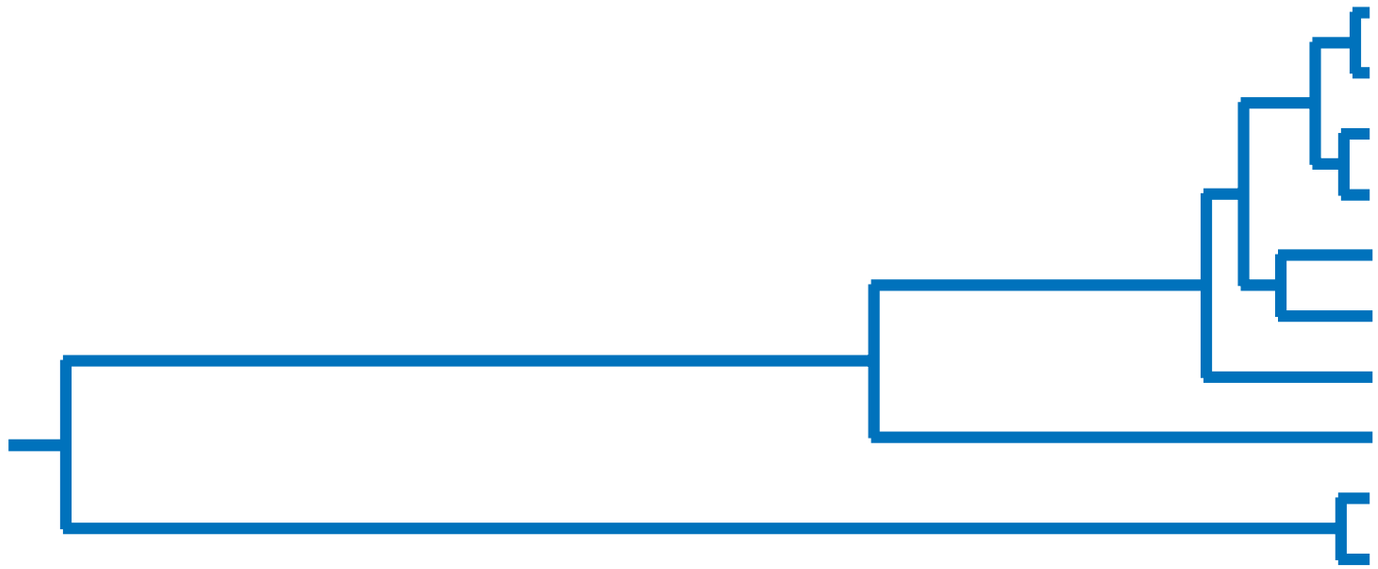
Extensions of the basic coalescent

Populations are rarely completely stable through time, and attempts have been made to model population growth or shrinkage using linear, exponential or more general approaches.

- In a small population lineages coalesce quickly



- In a large population lineages coalesce slowly



This leaves a signature in the data. We can exploit this and estimate the population growth rate g jointly with the current population size Θ.

Extensions of the basic coalescent

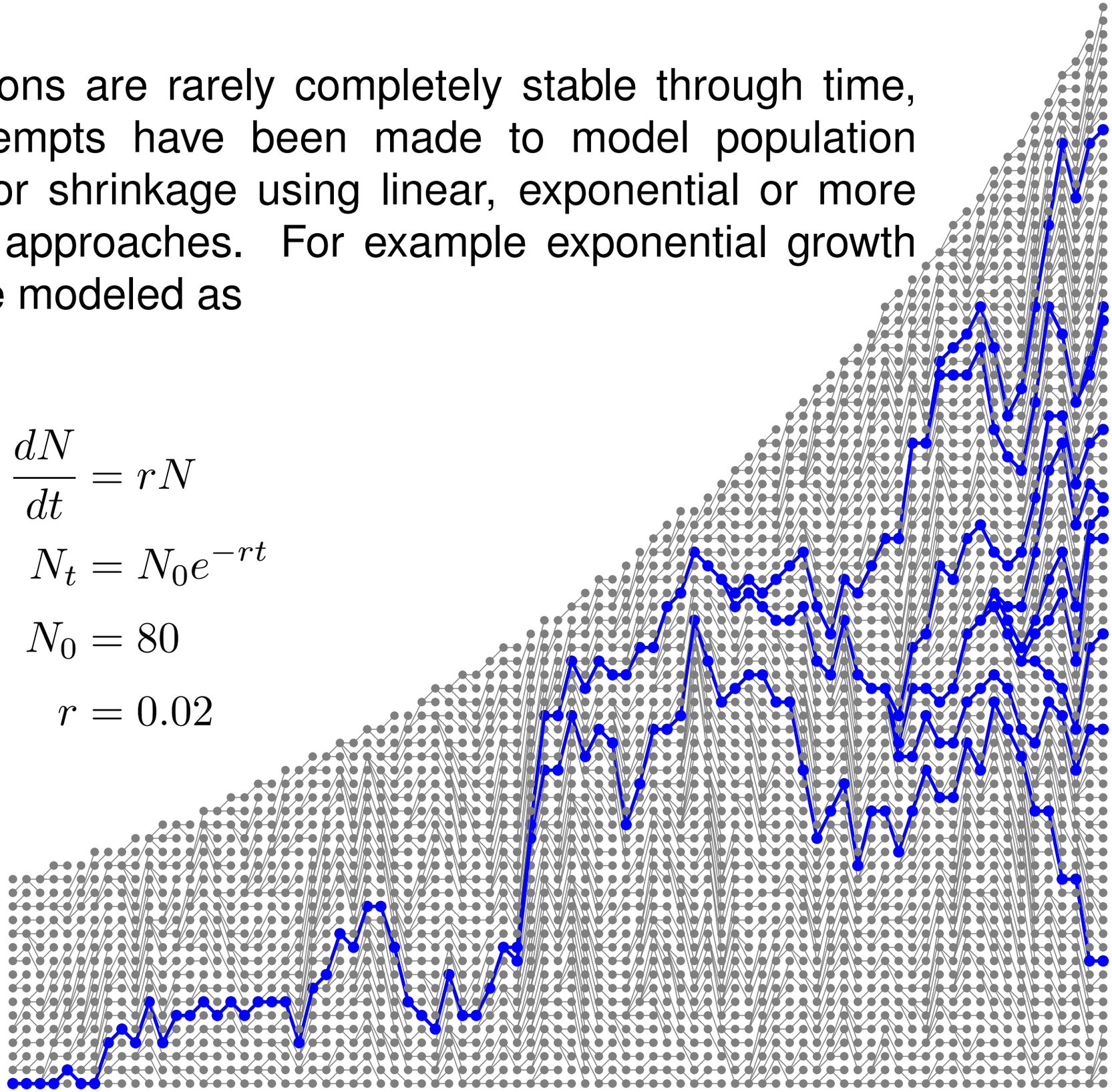
Populations are rarely completely stable through time, and attempts have been made to model population growth or shrinkage using linear, exponential or more general approaches. For example exponential growth could be modeled as

$$\frac{dN}{dt} = rN$$

$$N_t = N_0 e^{-rt}$$

$$N_0 = 80$$

$$r = 0.02$$



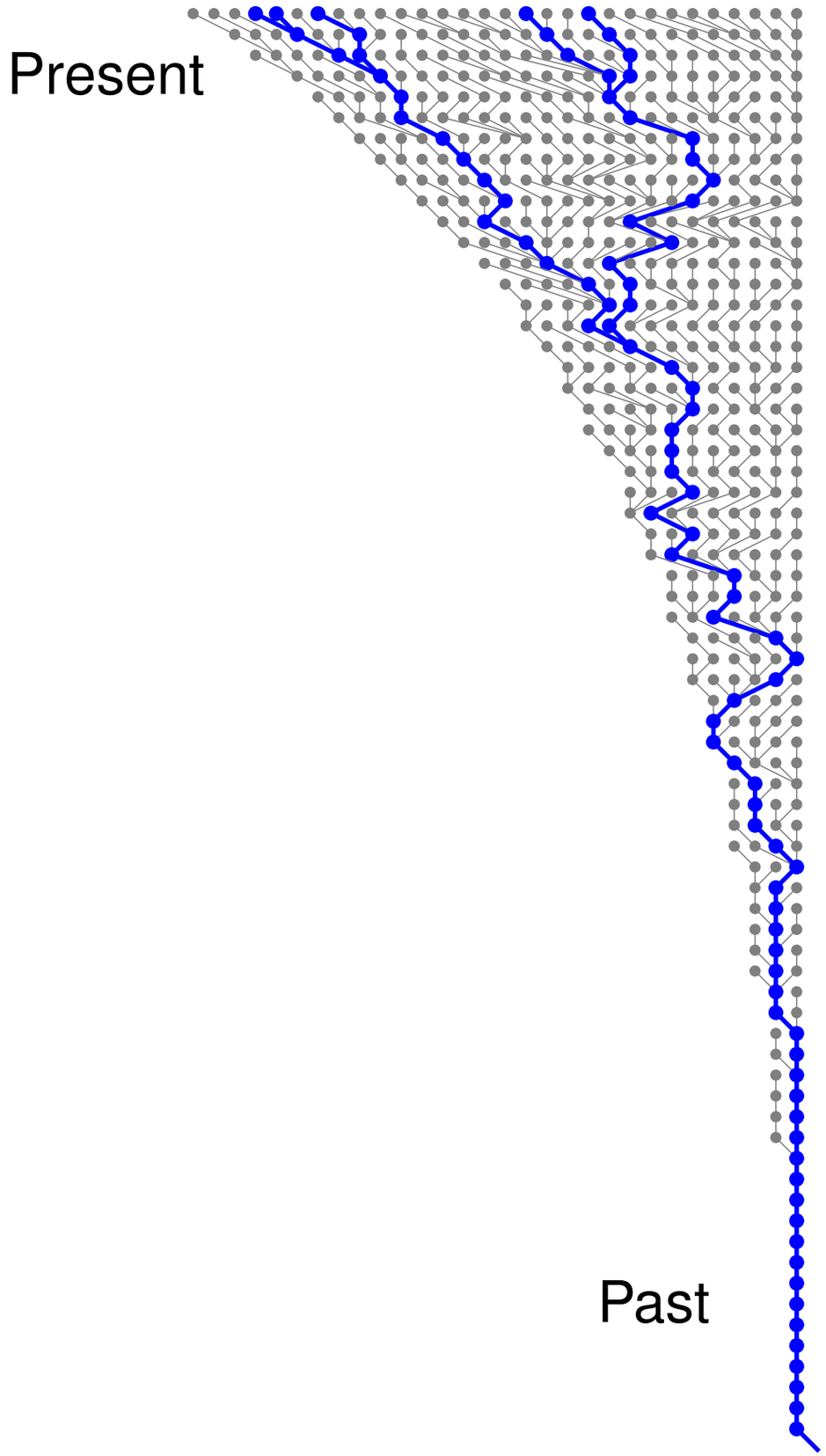
Extensions of the basic coalescent

For constant population size we found

$$\mathrm{p}(\mathrm{G}|\Theta) = \prod_{\mathrm{j}} \mathrm{e}^{-\mathrm{u_j}\frac{\mathrm{k}(\mathrm{k}-1)}{\Theta}} \frac{2}{\Theta}$$

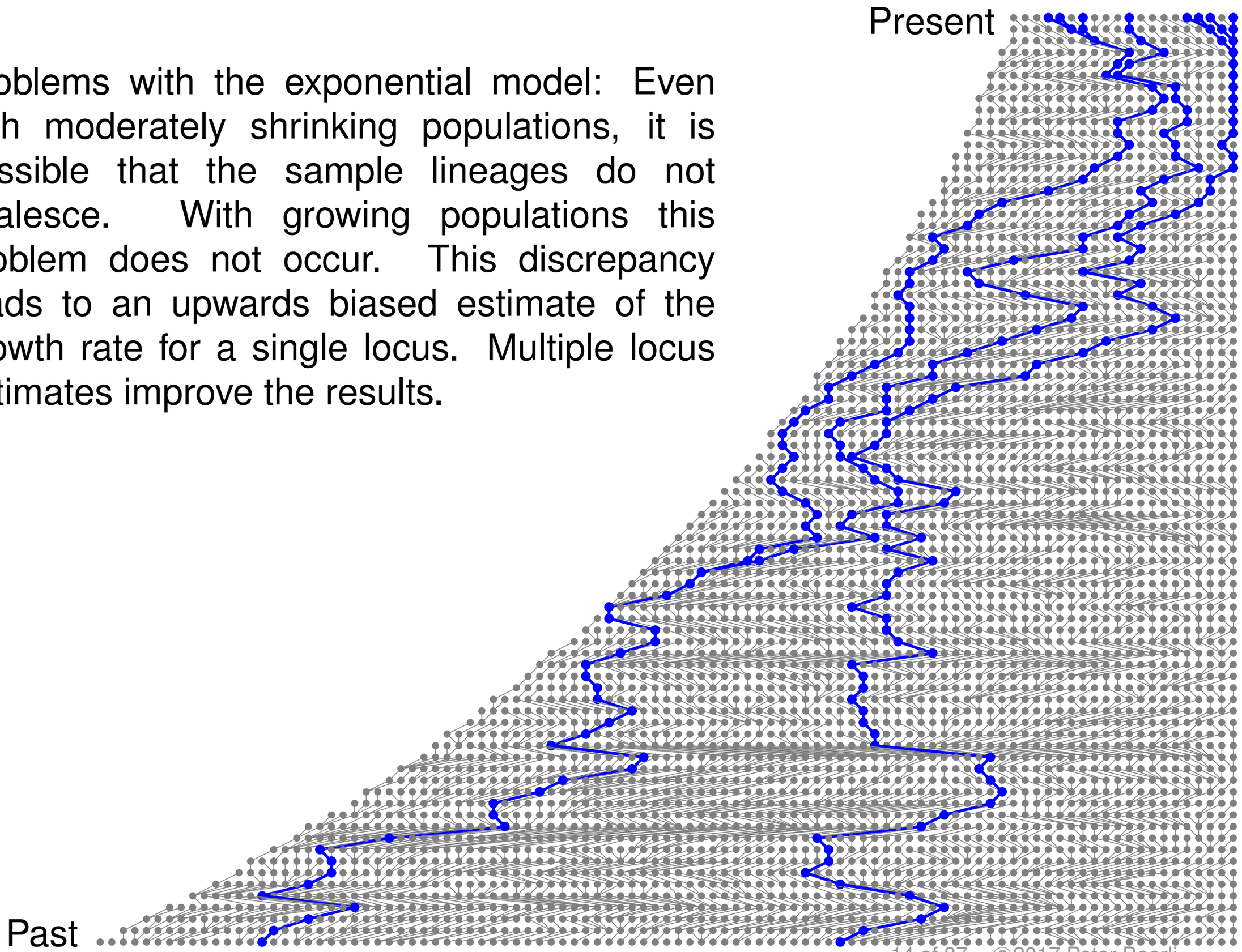
Relaxing the constant size to exponential growth and using $g = r/\mu$ leads to

$$\mathrm{p}(\mathrm{G}|\Theta_0, \mathrm{g}) = \prod_{\mathrm{j}} \mathrm{e}^{-(\mathrm{t_j}-\mathrm{t_{j-1}})\frac{\mathrm{k}(\mathrm{k}-1)}{\Theta_0 \mathrm{e}^{-\mathrm{gt}}}} \frac{2}{\Theta_0 \mathrm{e}^{-\mathrm{gt}}}$$



Extensions of the basic coalescent

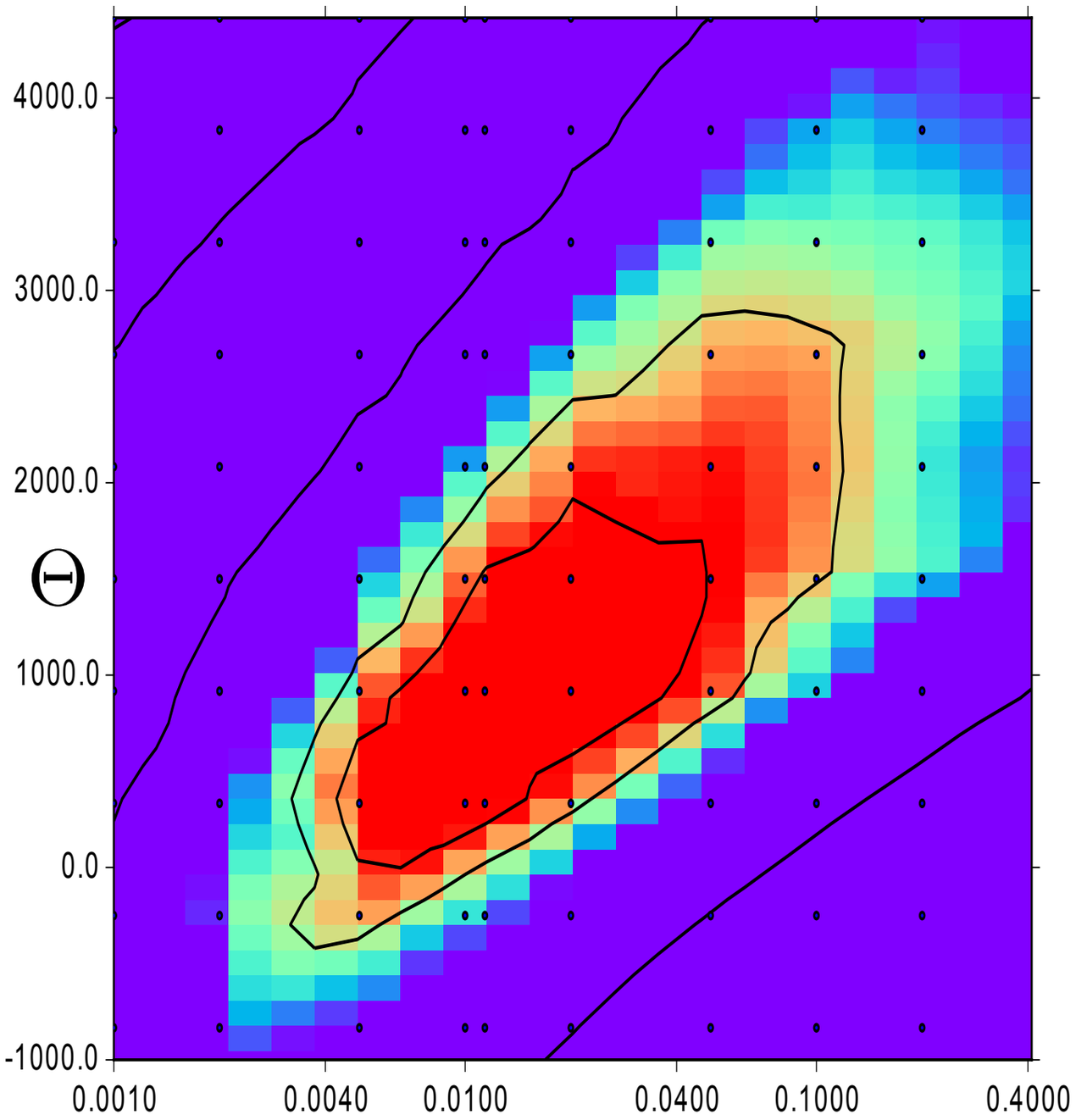
Problems with the exponential model: Even with moderately shrinking populations, it is possible that the sample lineages do not coalesce. With growing populations this problem does not occur. This discrepancy leads to an upwards biased estimate of the growth rate for a single locus. Multiple locus estimates improve the results.

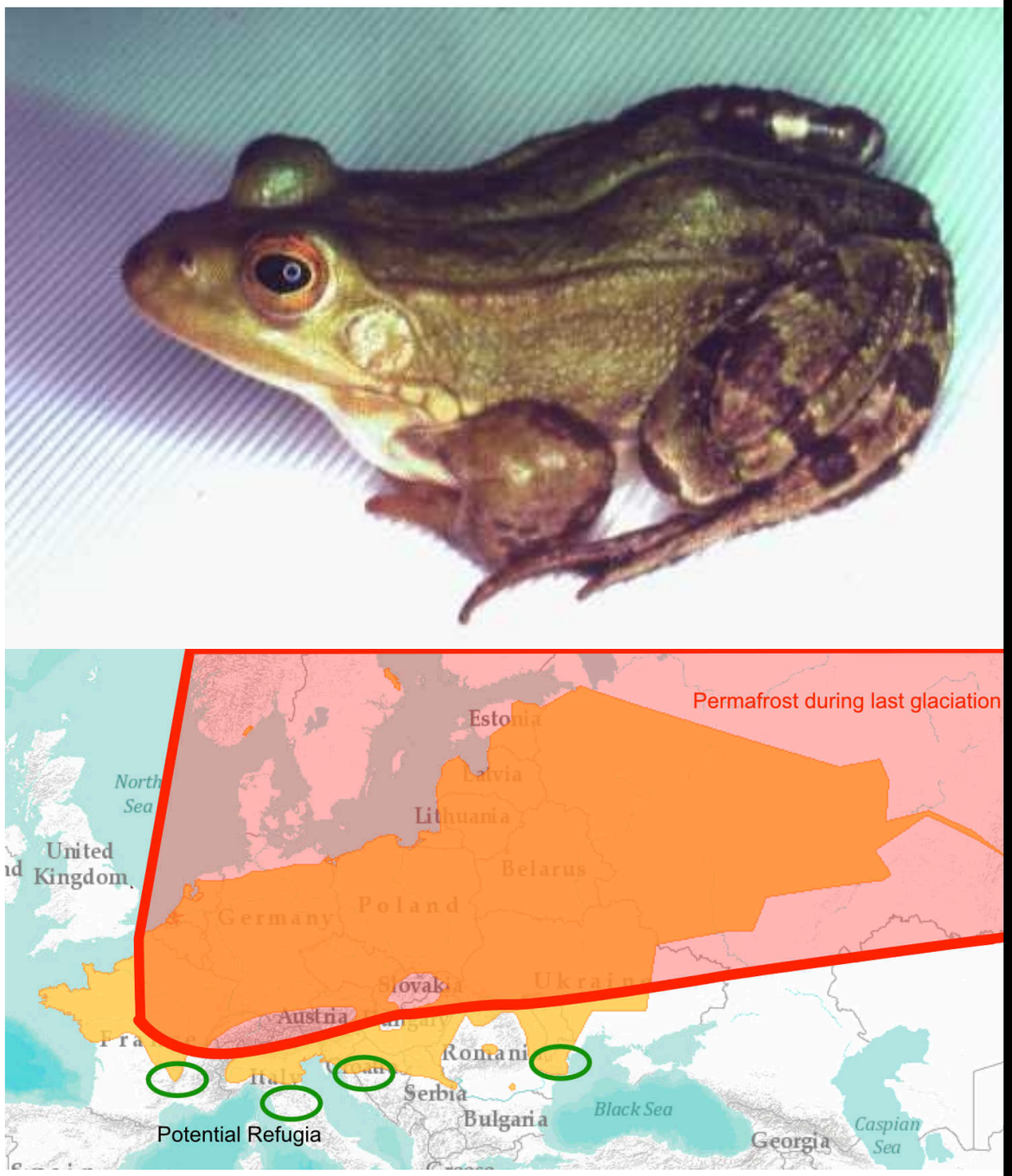


Expansion of *Pelophylax lessonae* in Europe

Grow-A-Frog

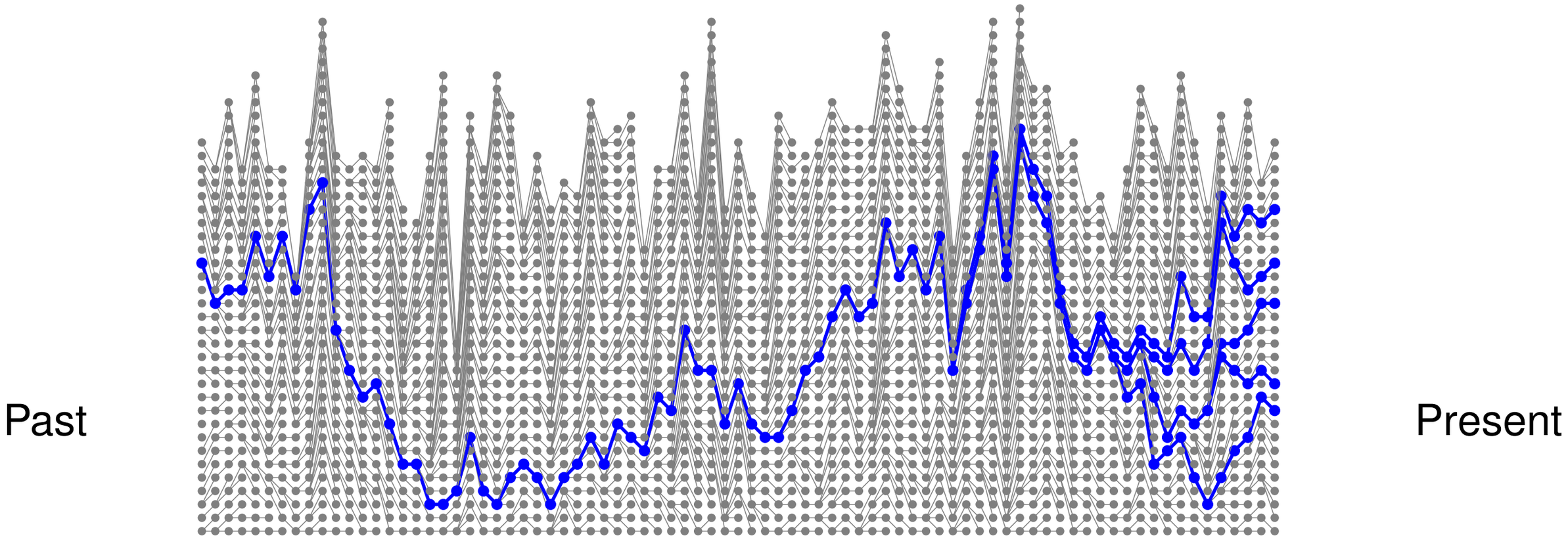
Expansion of *Pelophylax lessonae* in Europe



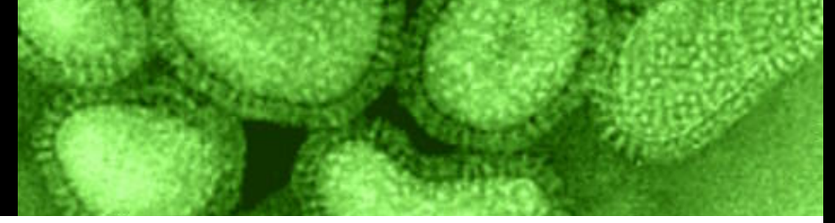


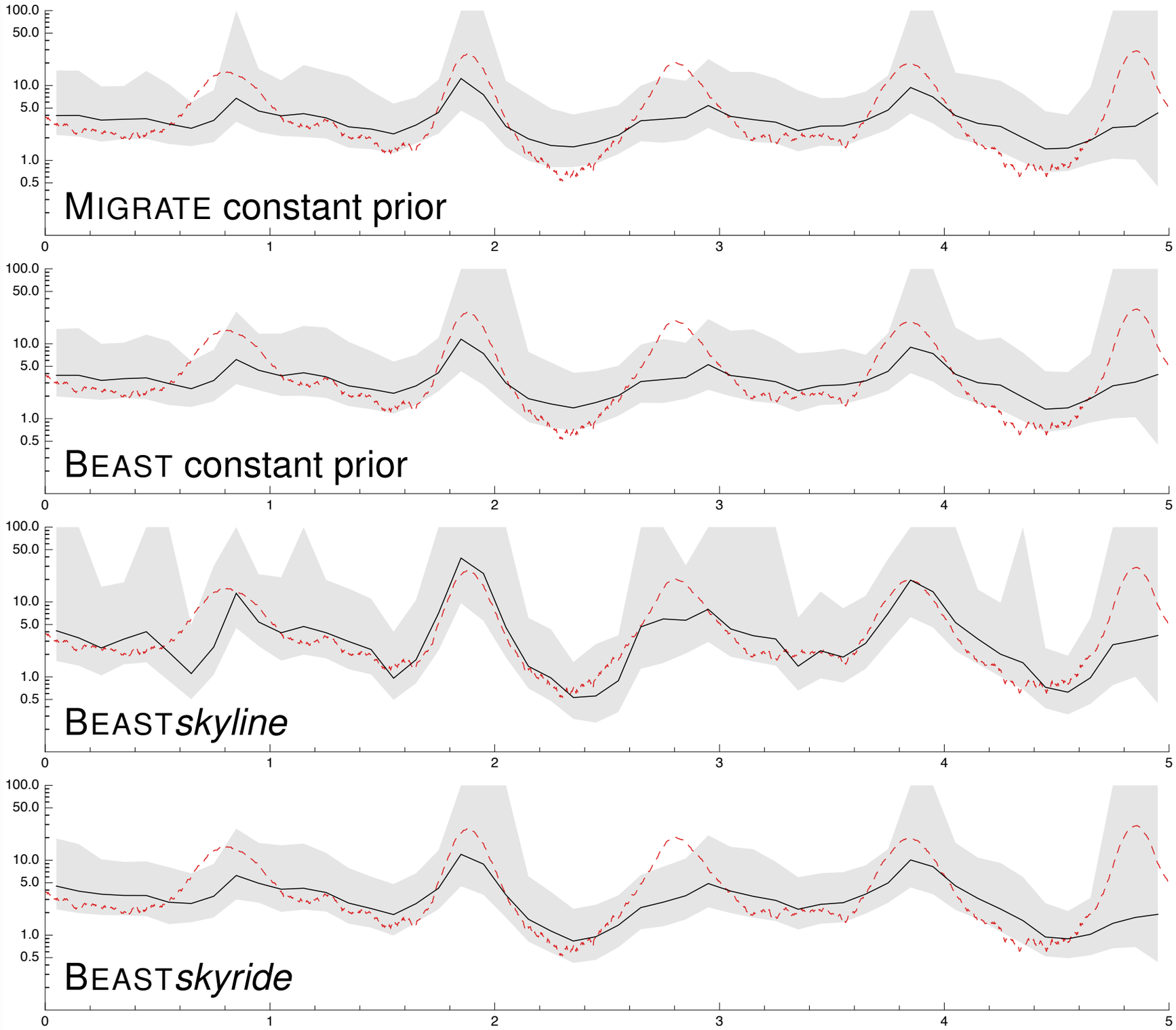
Extensions of the basic coalescent

Random fluctuations of the population size are most often ignored. BEAST (and to some extent MIGRATE) can handle such scenarios. BEAST is using a full parametric approach (skyride, skyline) whereas MIGRATE uses a non-parametric approach for its skyline plots that has the tendency to smooth the fluctuations too much, compared to BEAST.



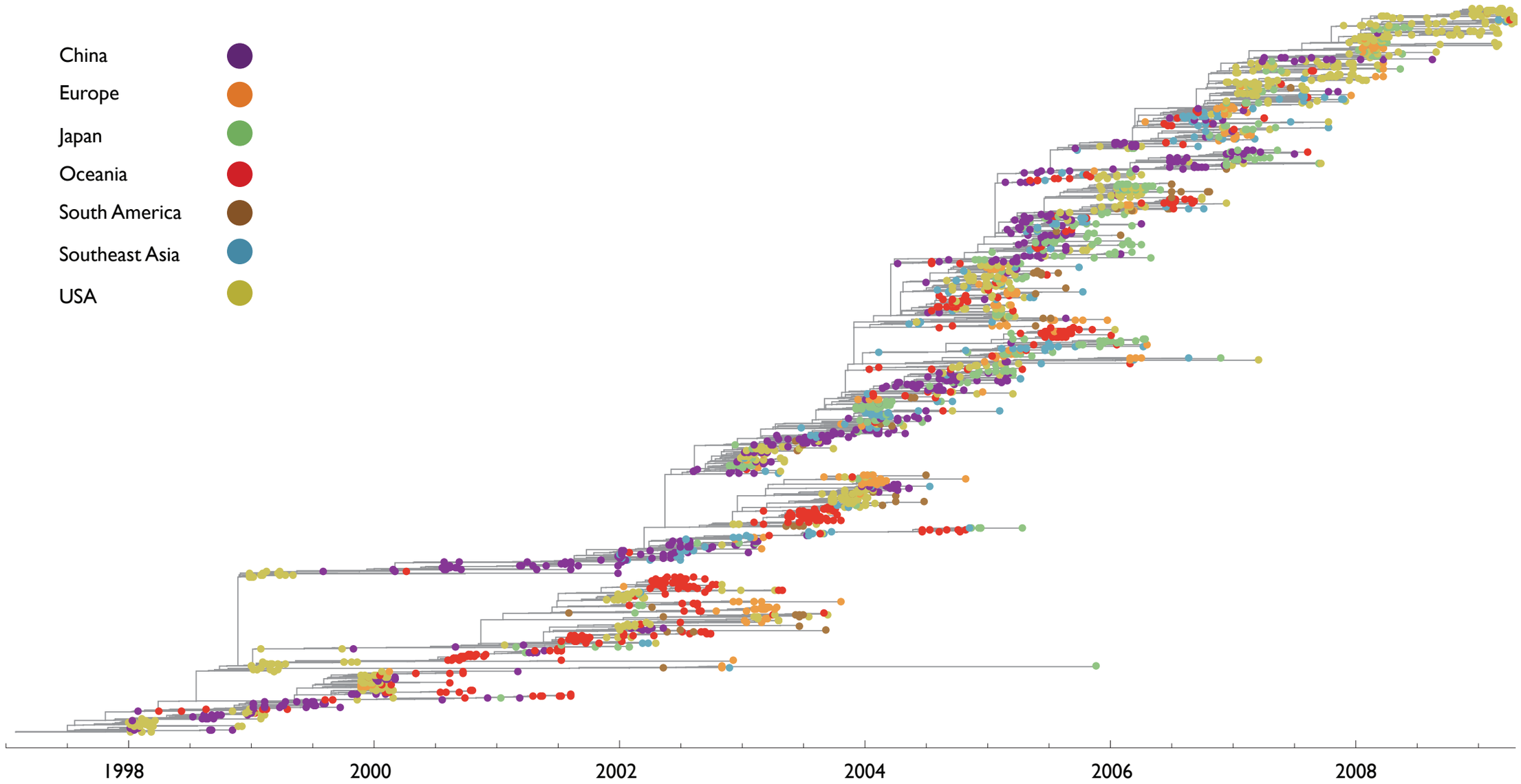
Extensions of the basic coalescent





Comparison of the skyline plots of simulated influenza dynamics analyzed by MIGRATE and BEAST. The x-axis is the time in years and the y-axis is effective population size. The data are sequences from 250 individuals sampled at regular intervals over 5 years. The dashed curve is the actual population size deduced from the true genealogy; black lines are the mean results of MIGRATE or BEAST; gray area is the 95% credibility interval. BEAST *skyline* matches the actual population size better than all other methods. Simulation and graphs courtesy of Trevor Bedford.

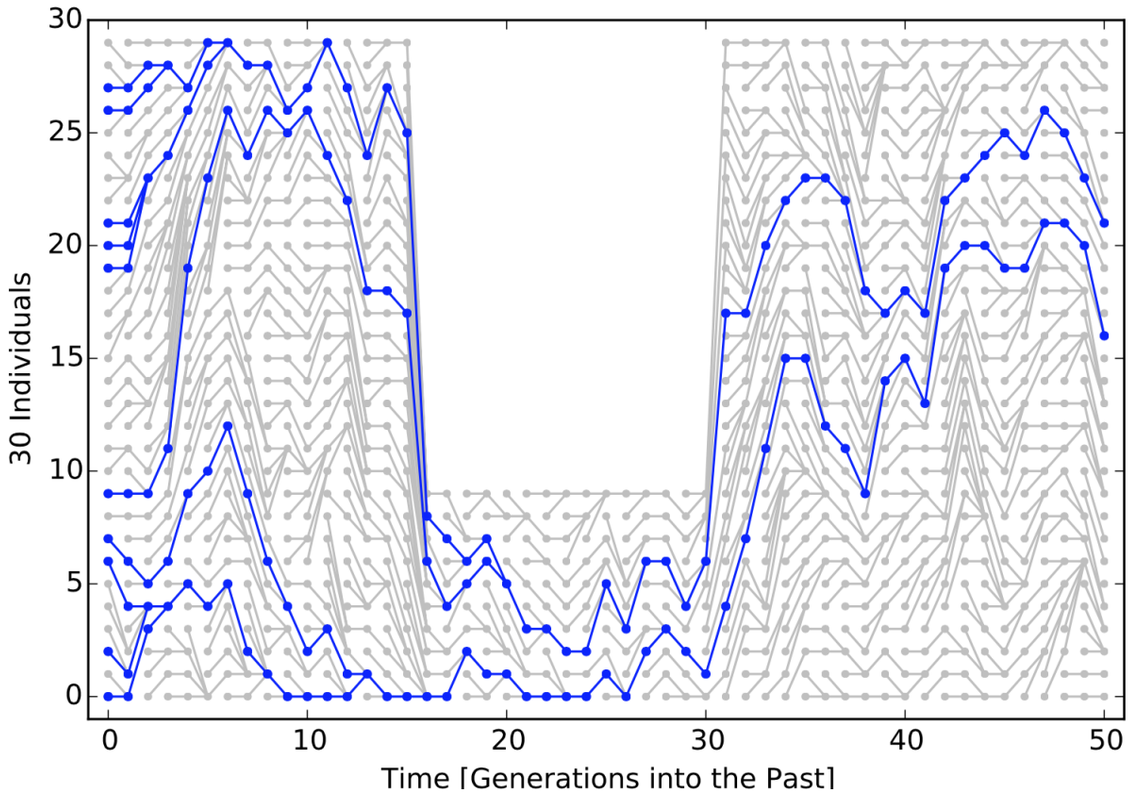
Influenza H3N2 word-wide distribution through time

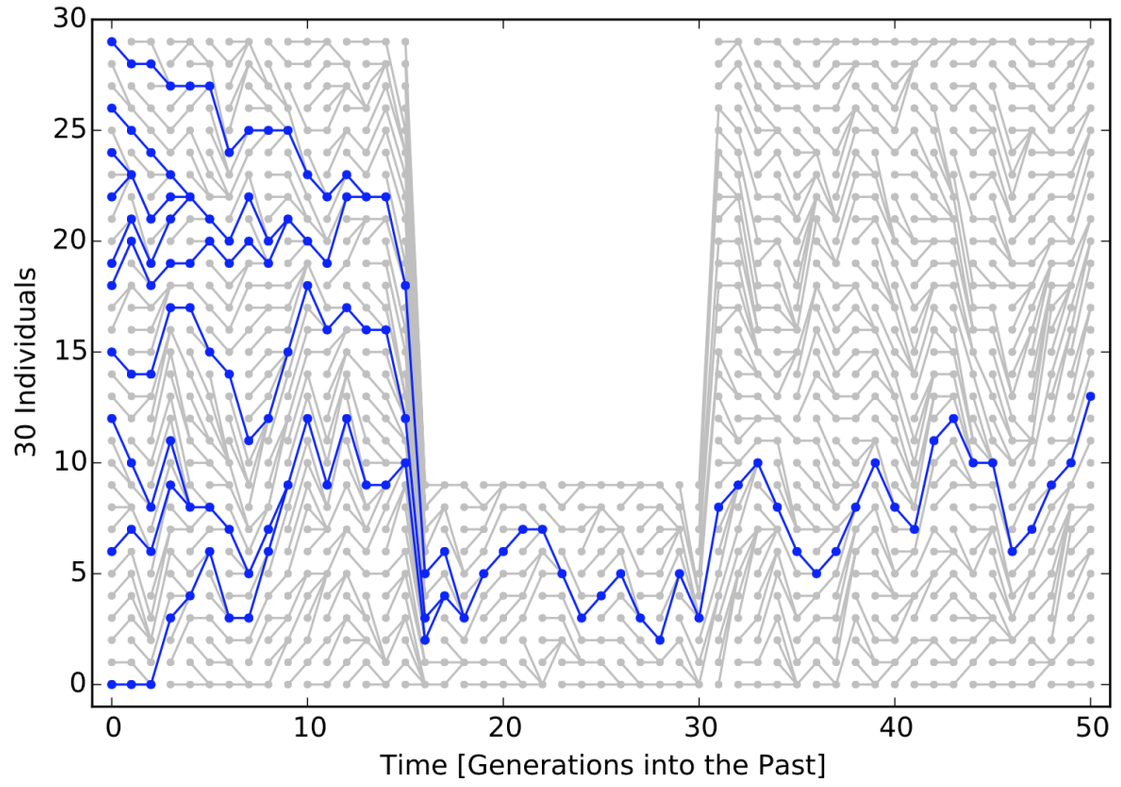


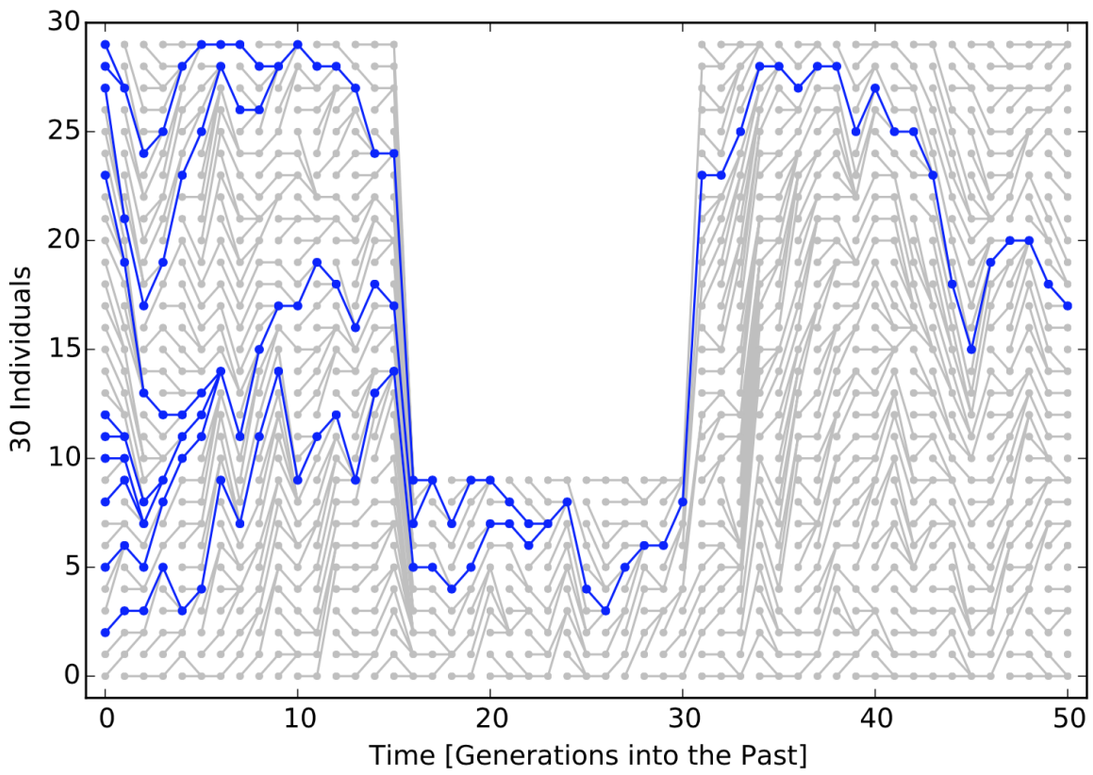
Bedford et al. 2010 Global migrational dynamics underlie evolution and persistence of human influenza A (H3N2). – ©2017 Peter Beerli

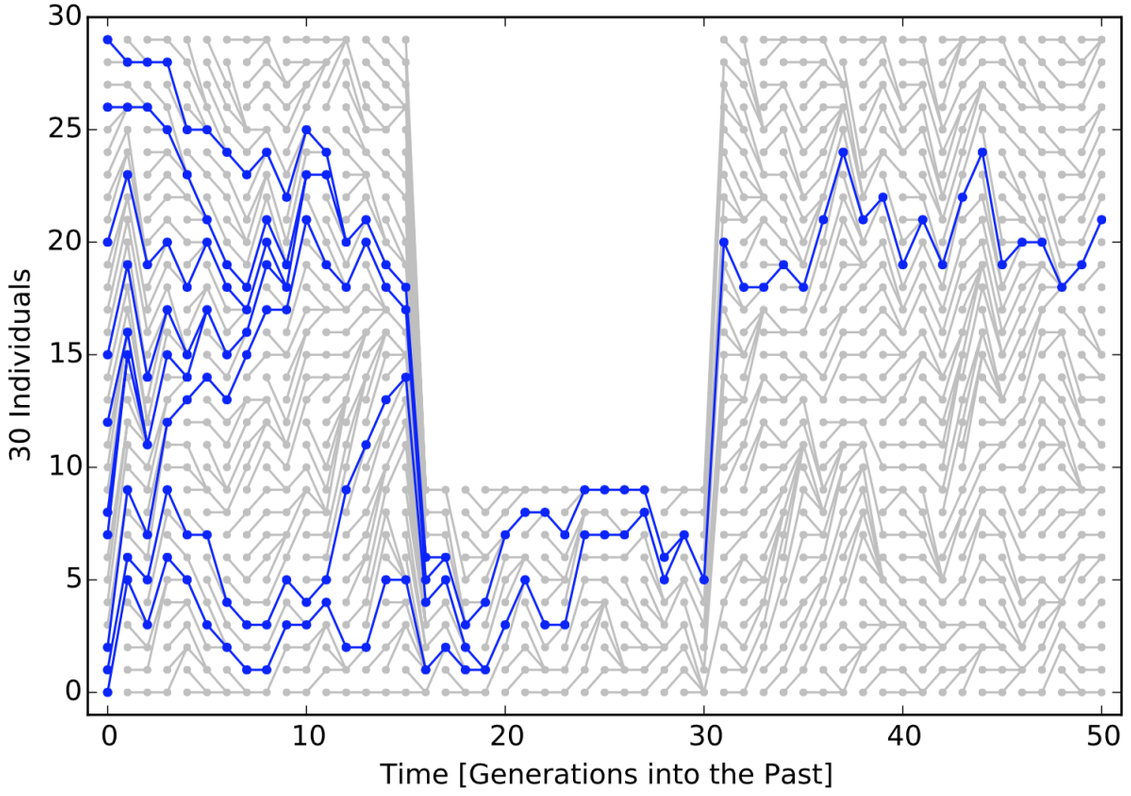
Extensions of the basic coalescent

Bottlenecks



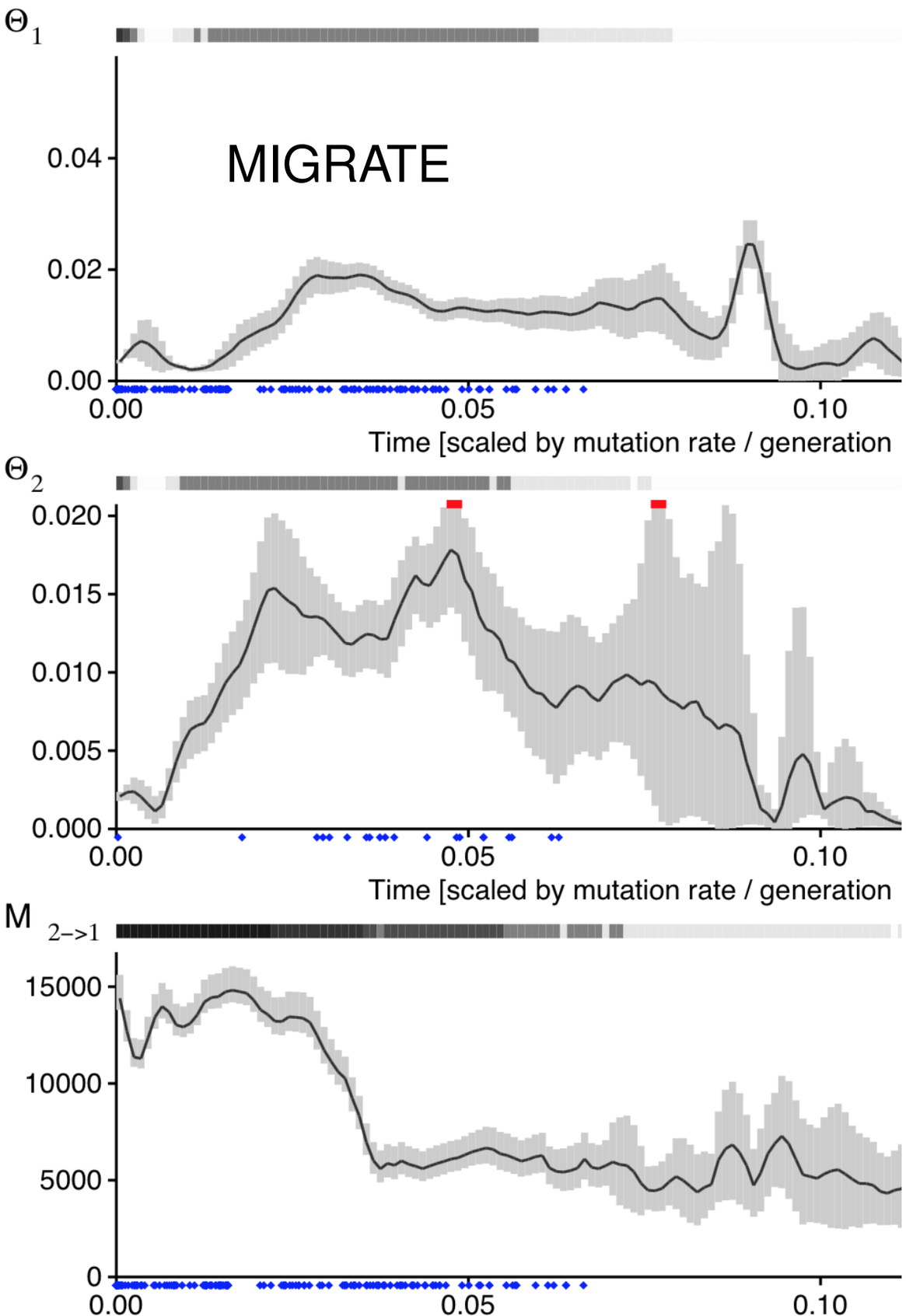


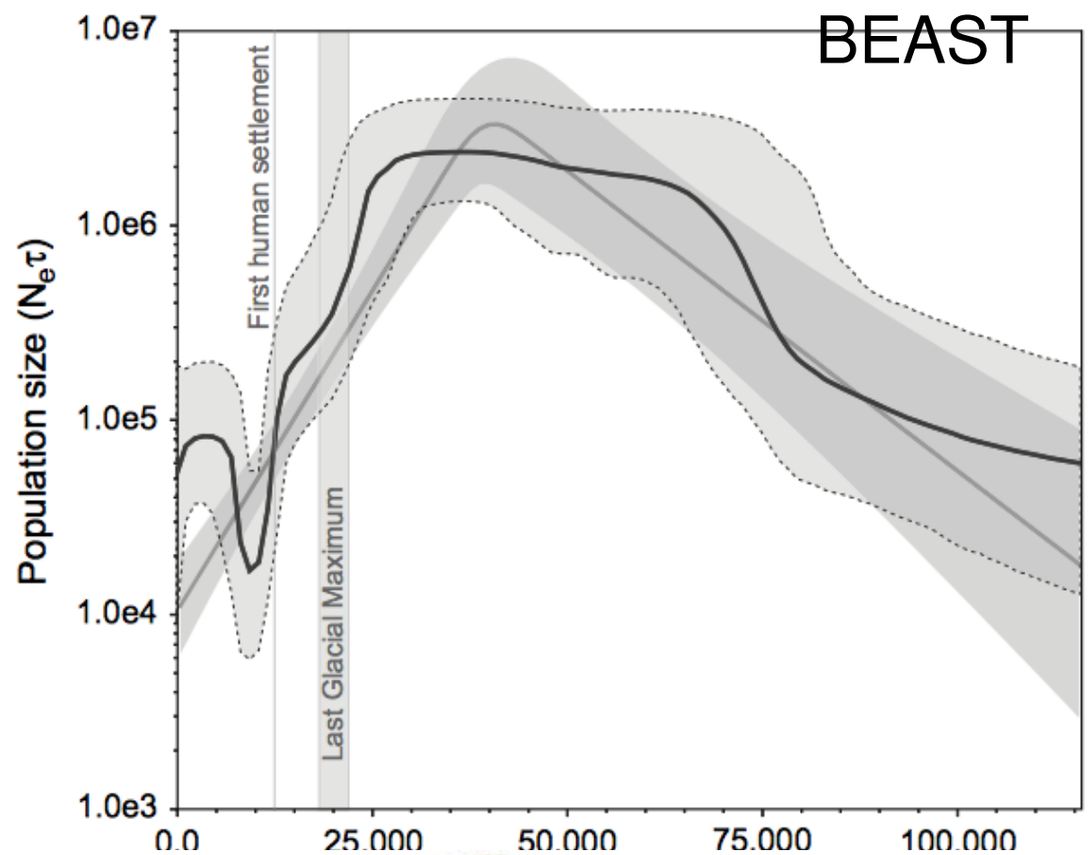




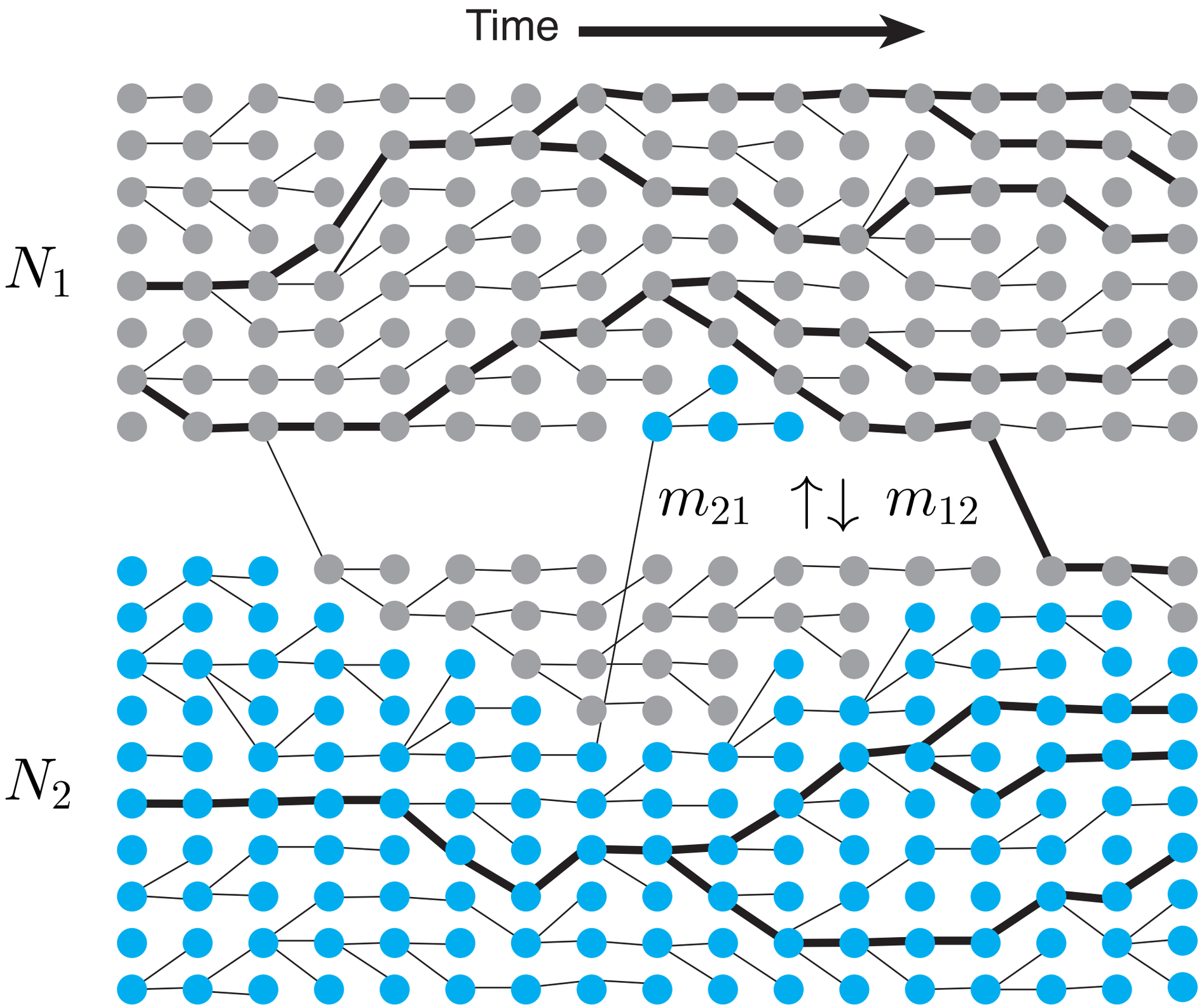
Extensions of the basic coalescent

Skyline plots

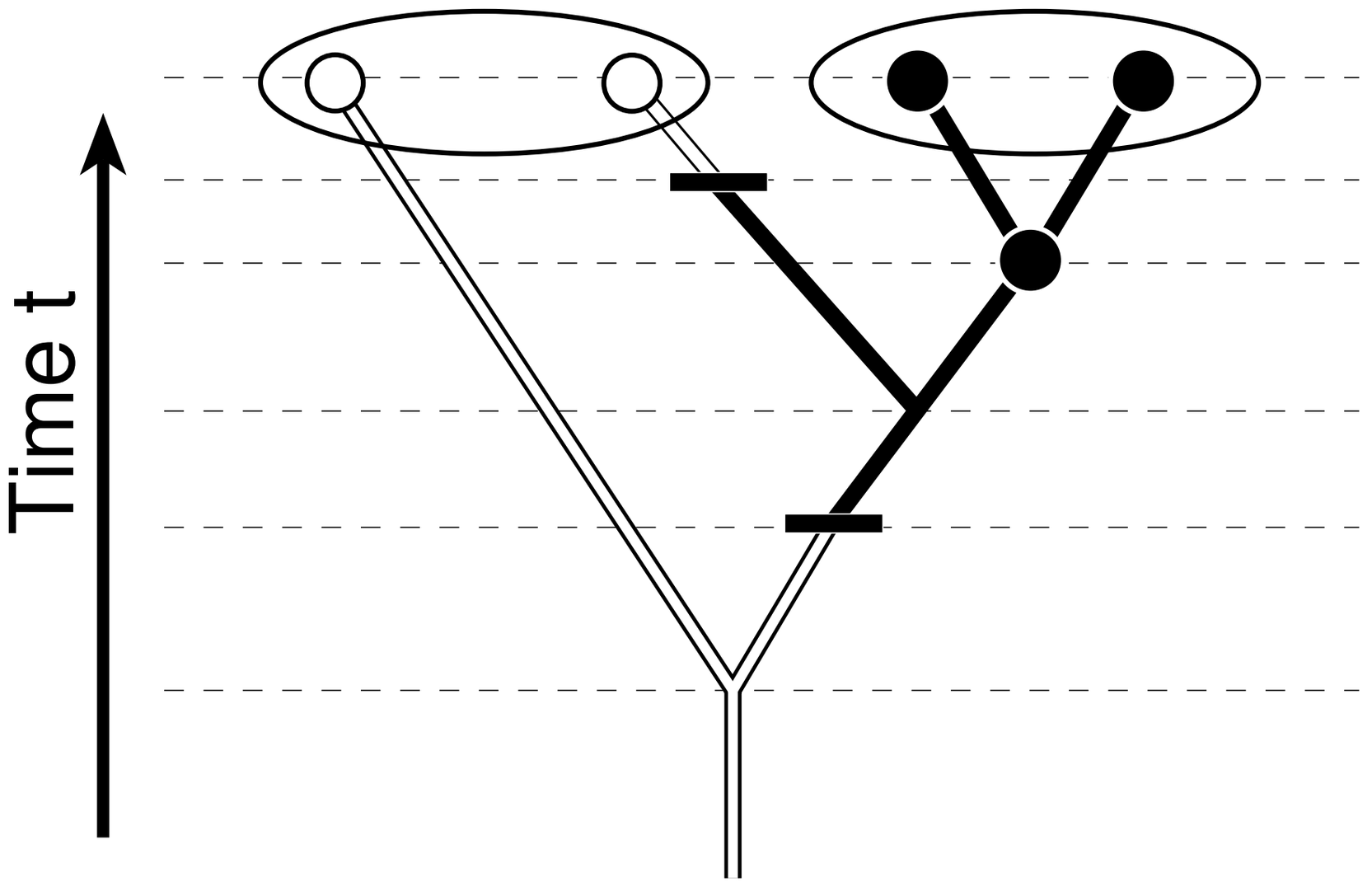




Extensions of the basic coalescent



Extensions of the basic coalescent



Δt	k_1	k_2
u_1	2	2
u_2	1	3
u_3	1	2
u_4	1	3
u_5	2	0

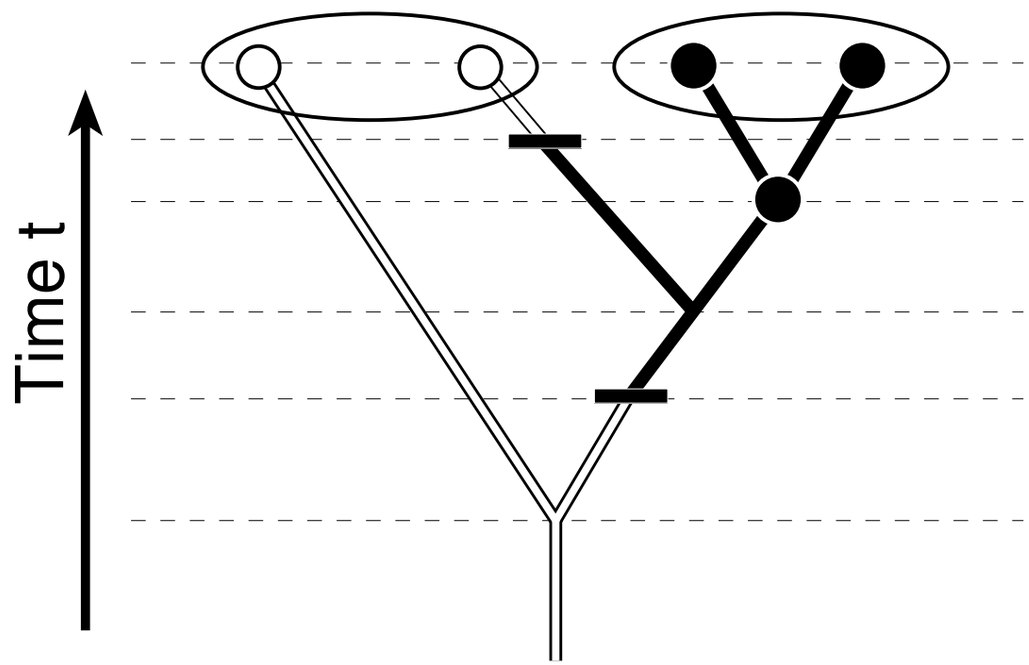
Extensions of the basic coalescent

The single population coalescence rate is

$$\frac{k(k-1)}{4N}.$$

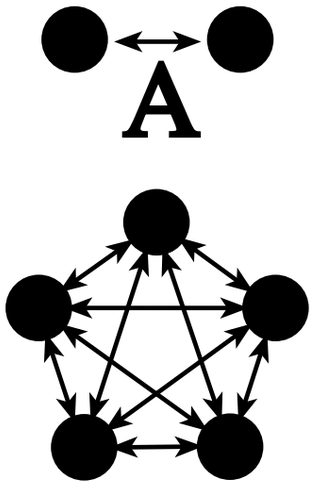
Changes for two populations to

$$\frac{k_1(k_1-1)}{\Theta_1} + \frac{k_2(k_2-1)}{\Theta_2} + k_1 M_{2,1} + k_2 M_{1,2}$$



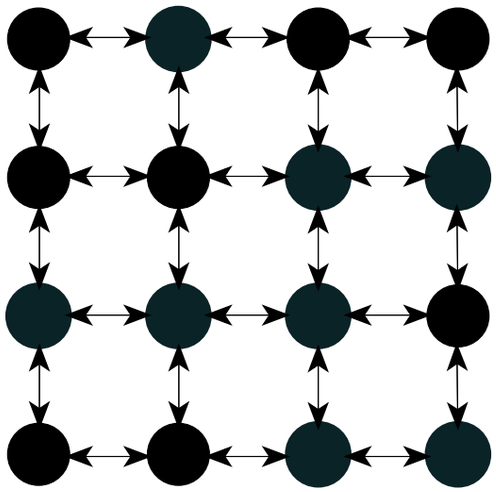
Δt	k_1	k_2
u_1	2	2
u_2	1	3
u_3	1	2
u_4	1	3
u_5	2	0

Structured populations

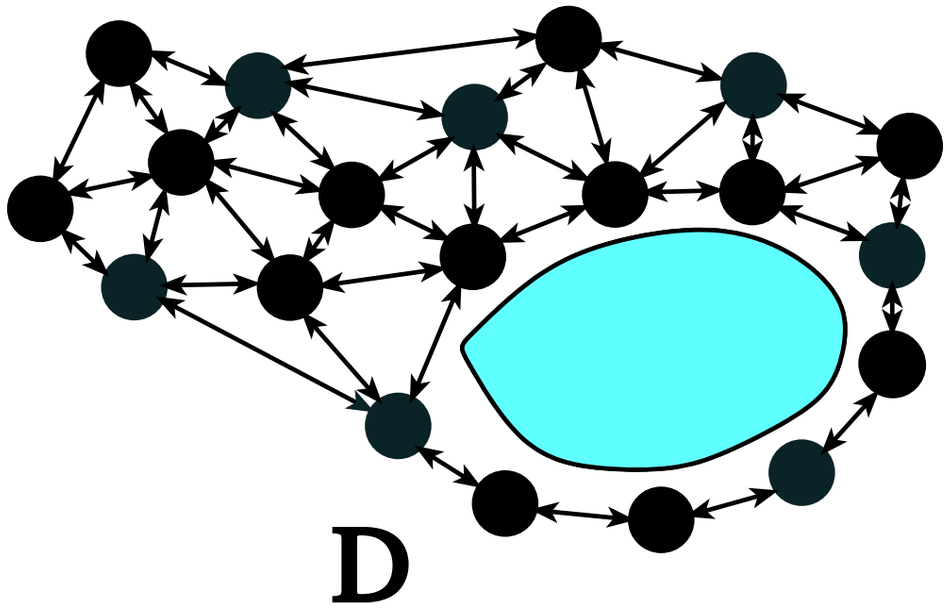


A

B

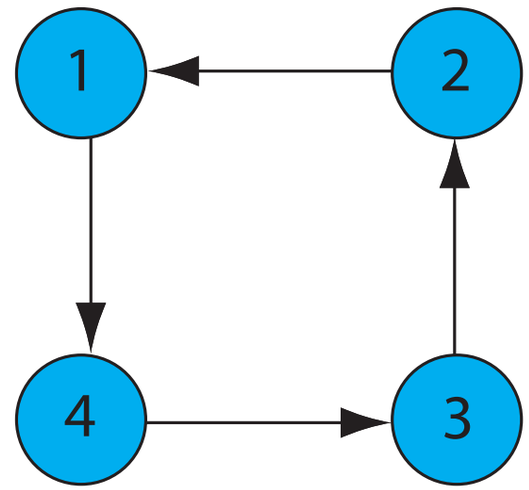


C

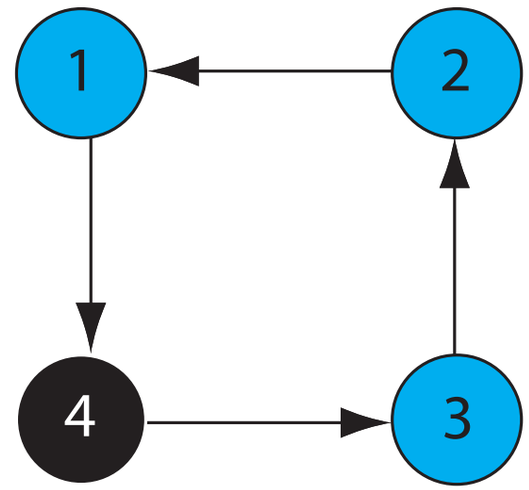


D

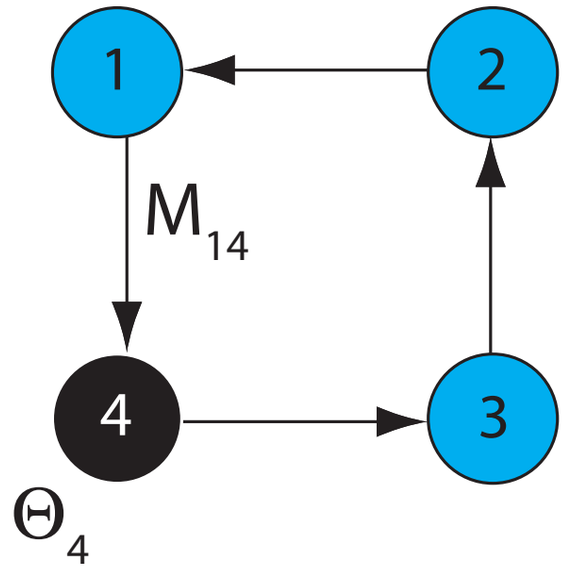
Bayesian inference



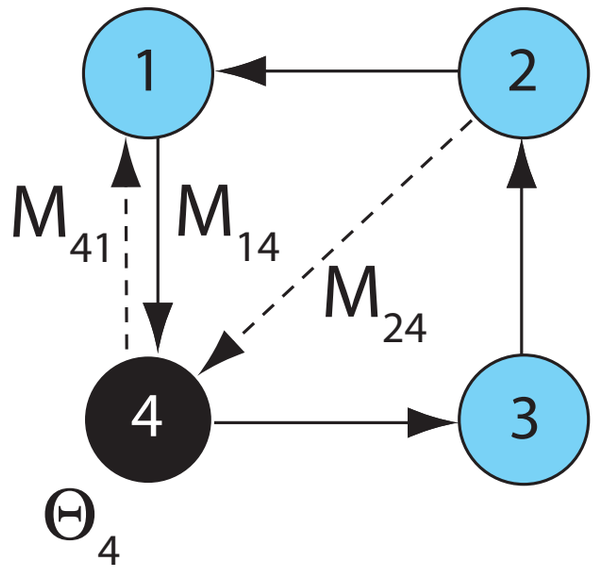
Bayesian inference



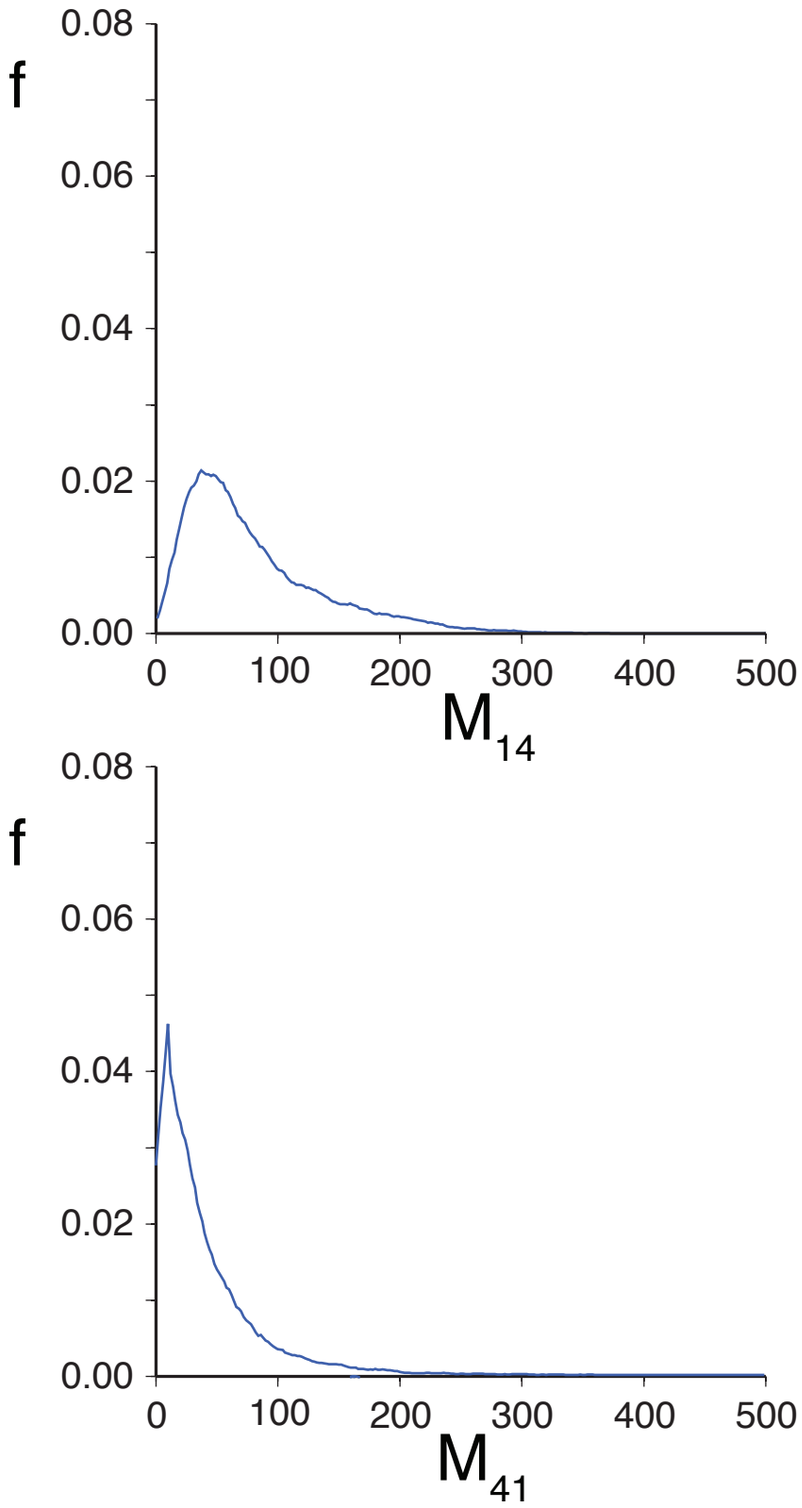
Bayesian inference

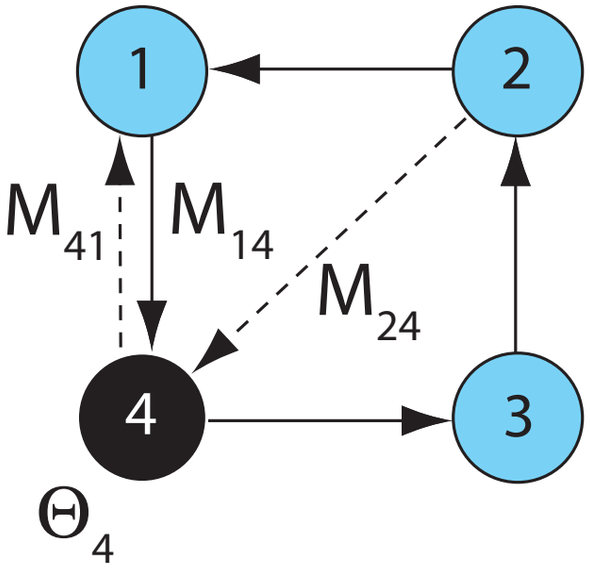


Bayesian inference

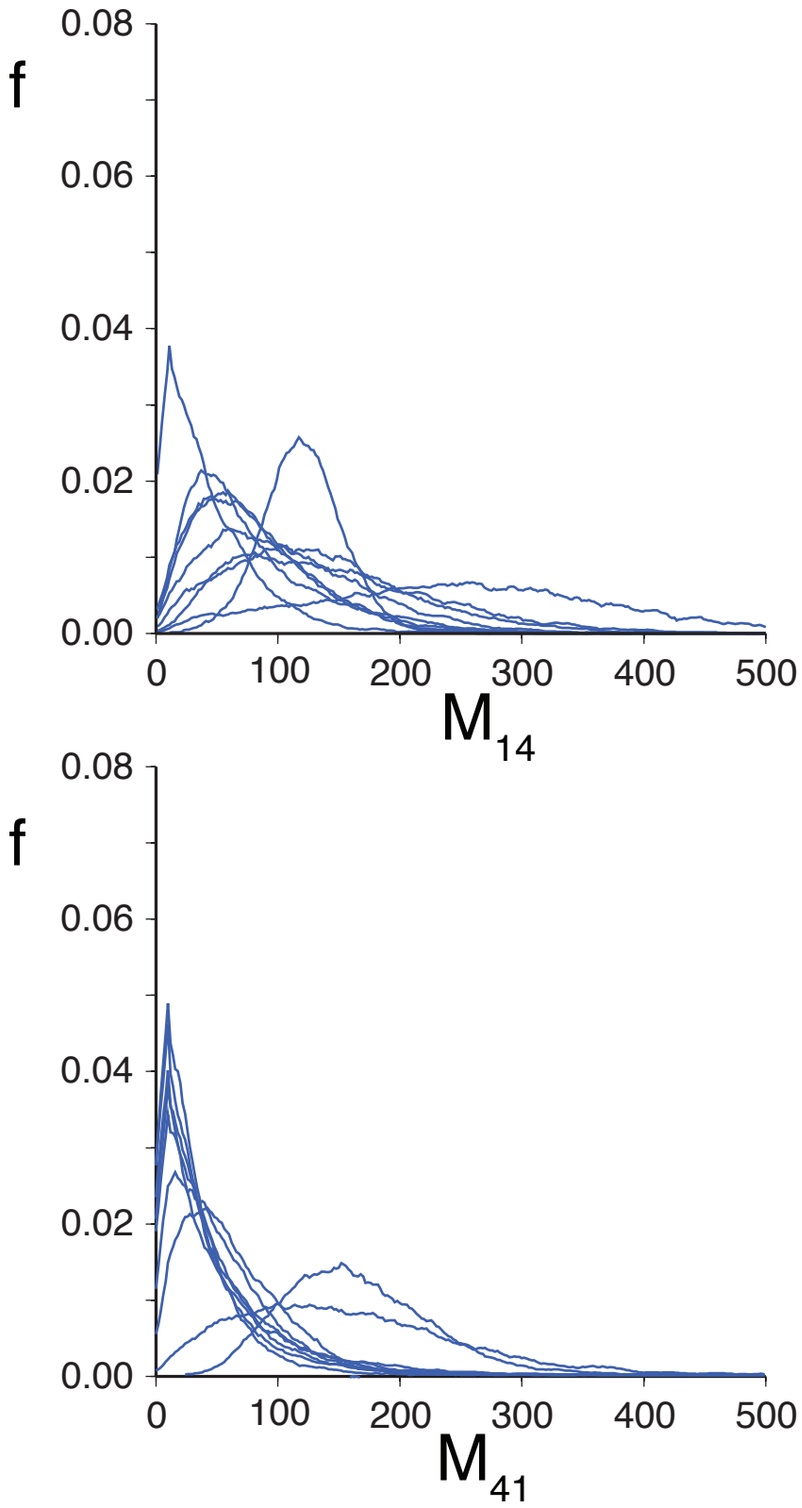


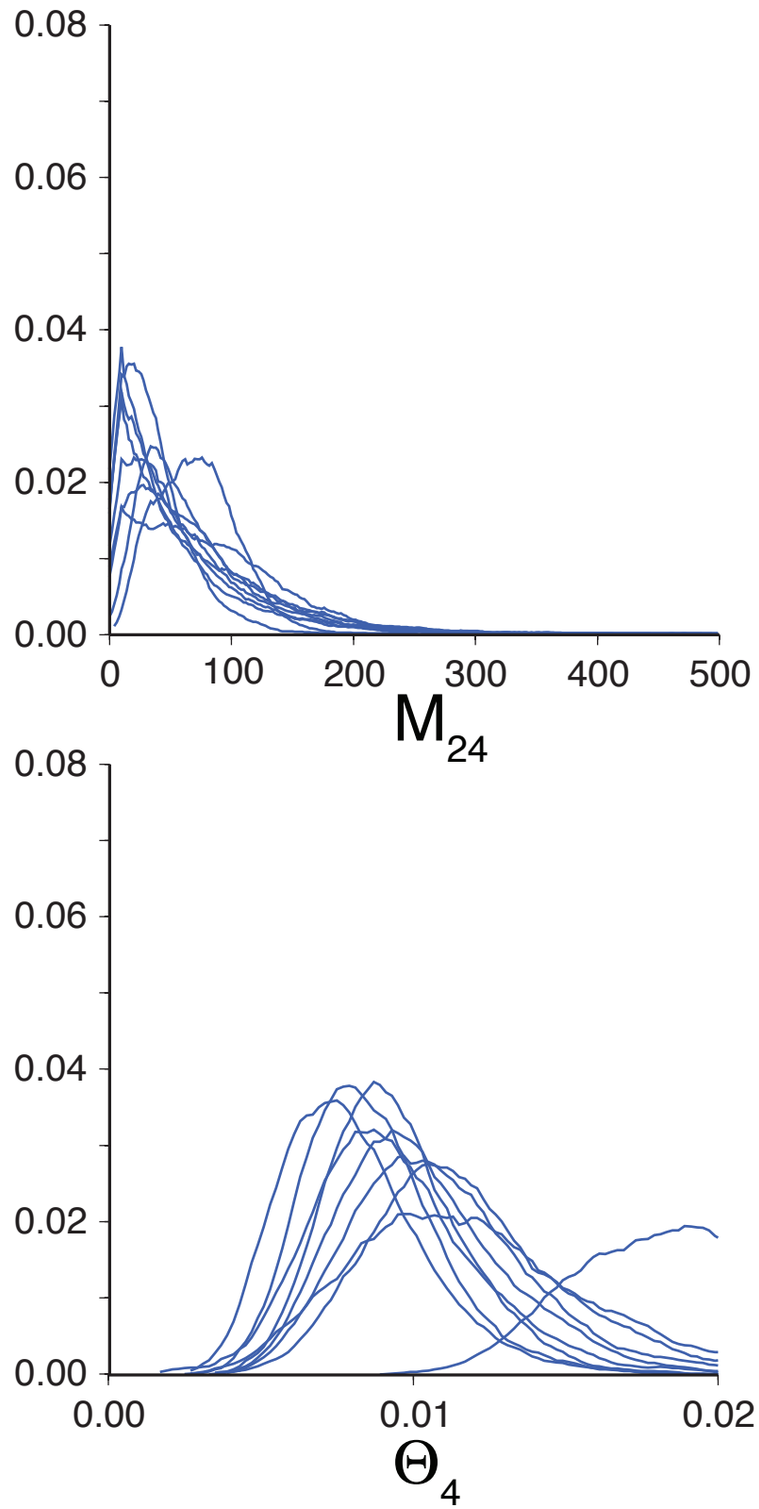
Bayesian inference

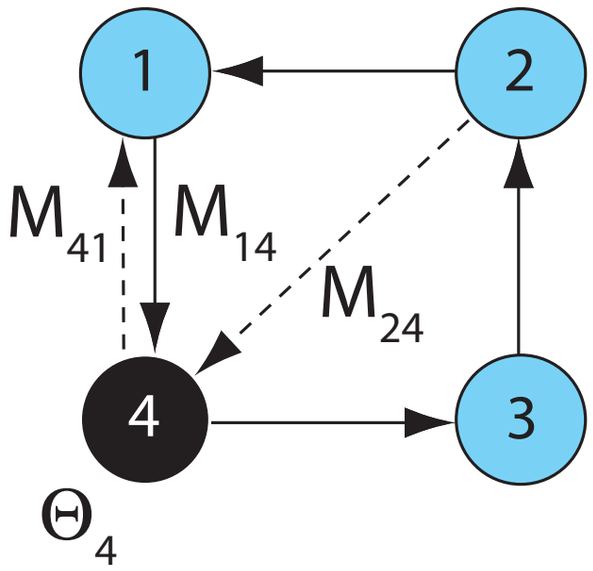




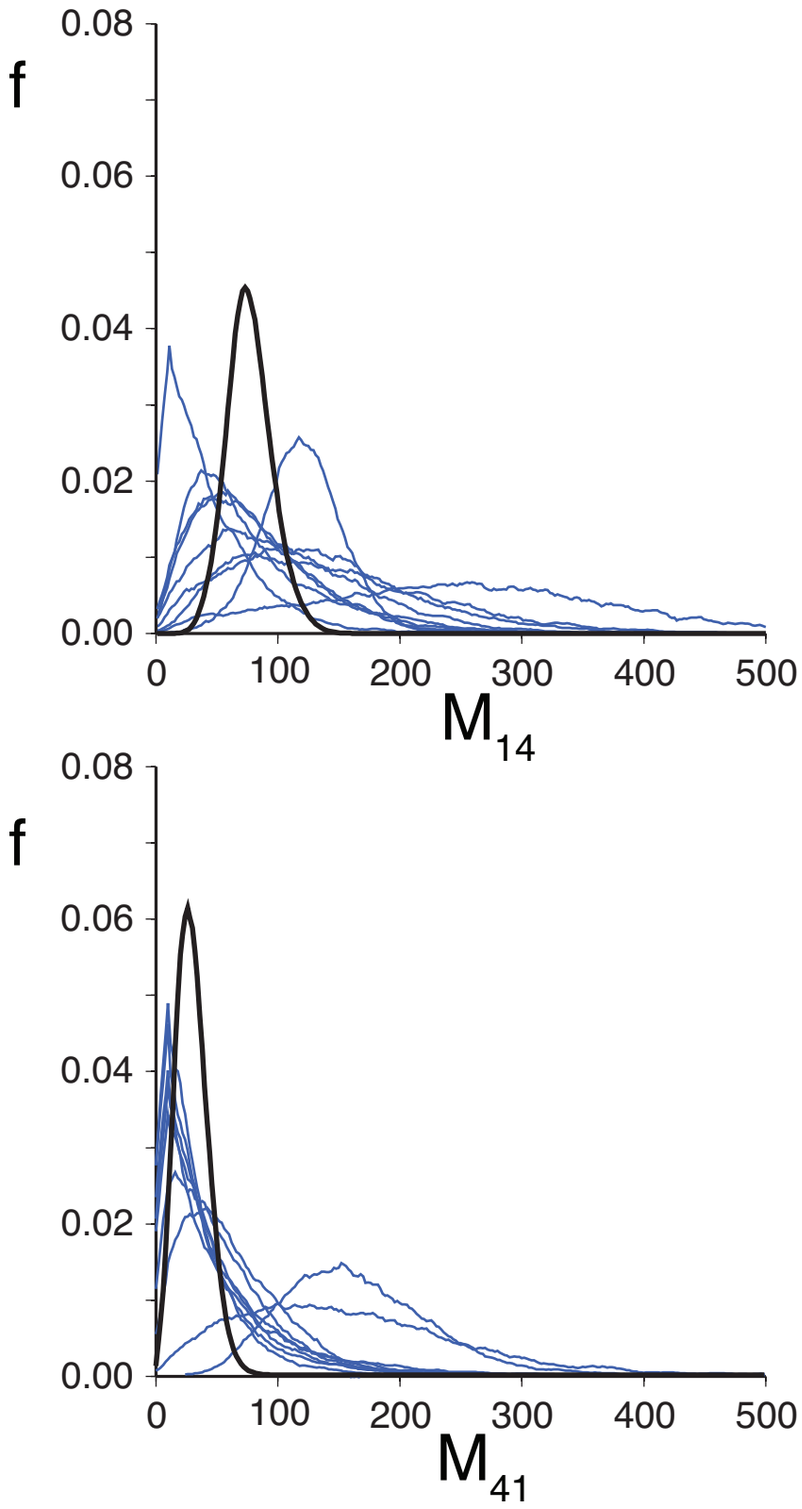
Bayesian inference

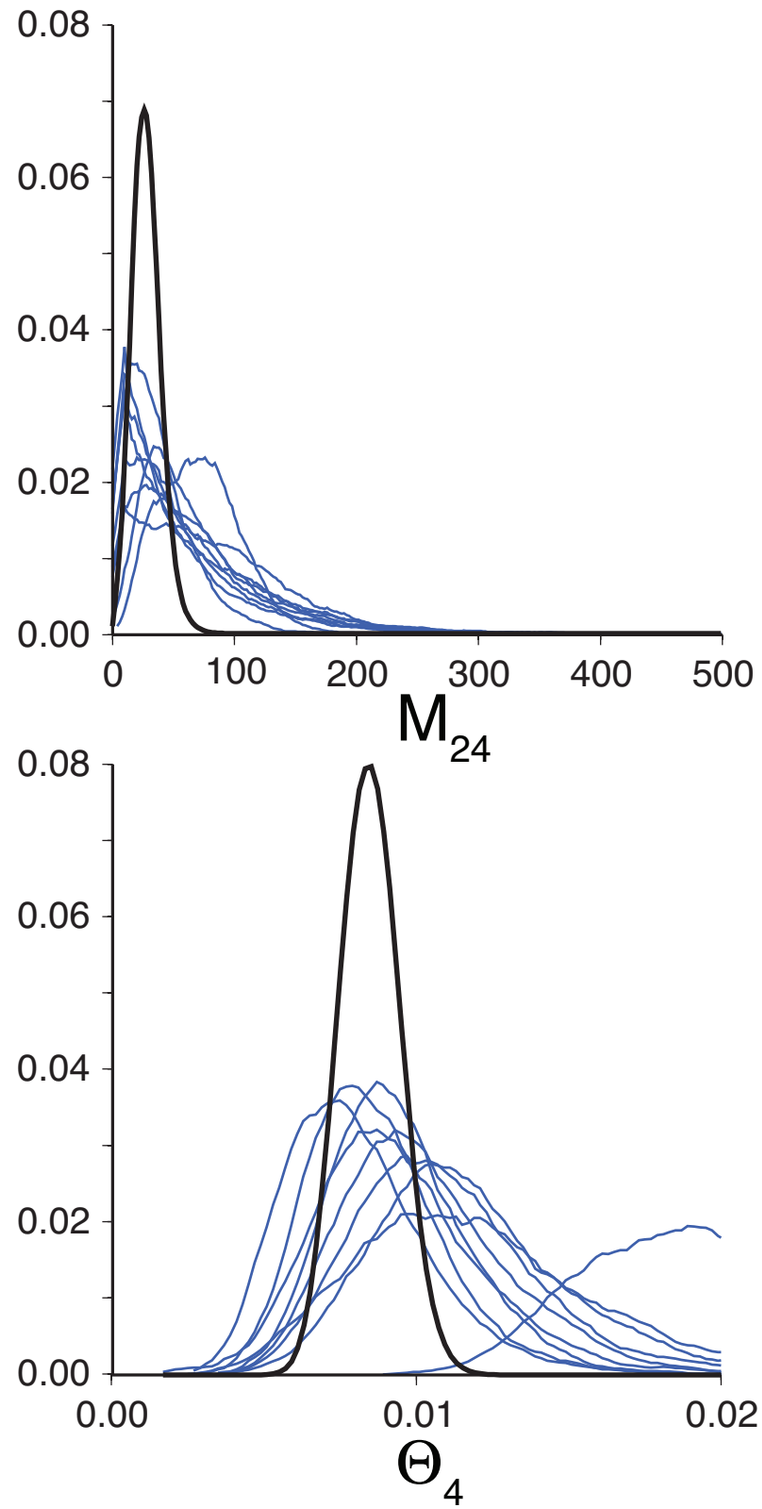


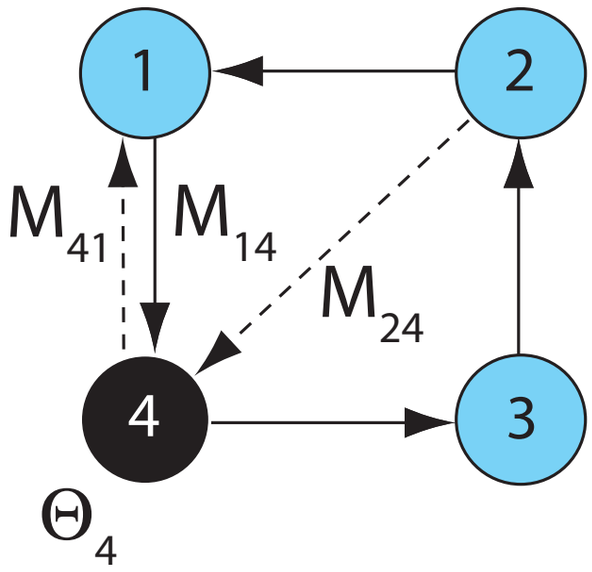




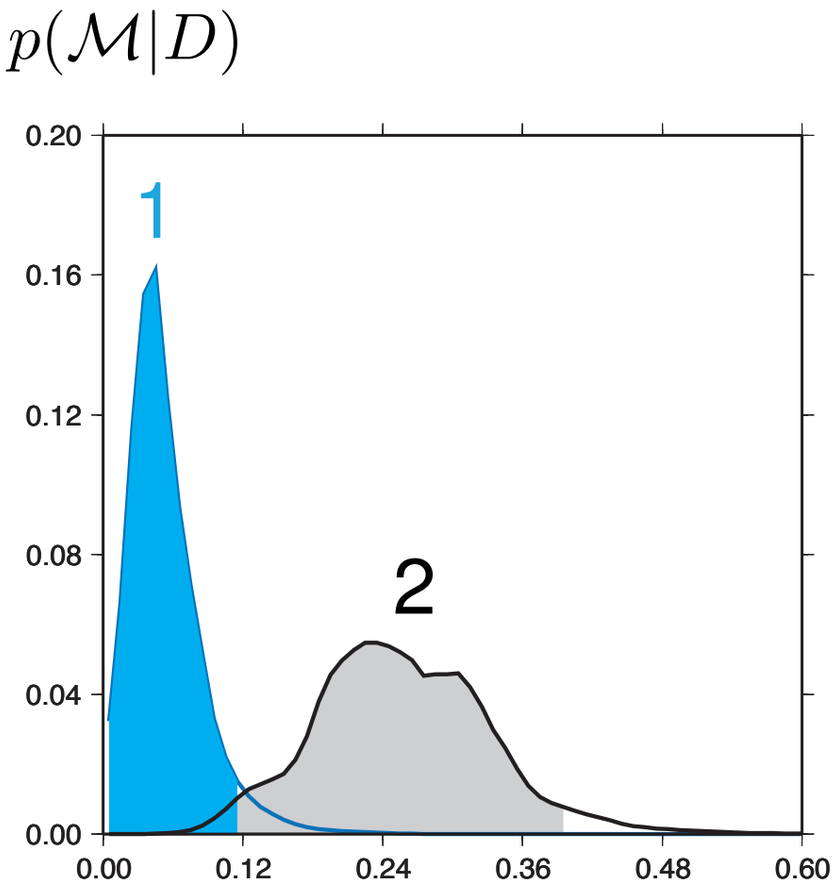
Bayesian inference

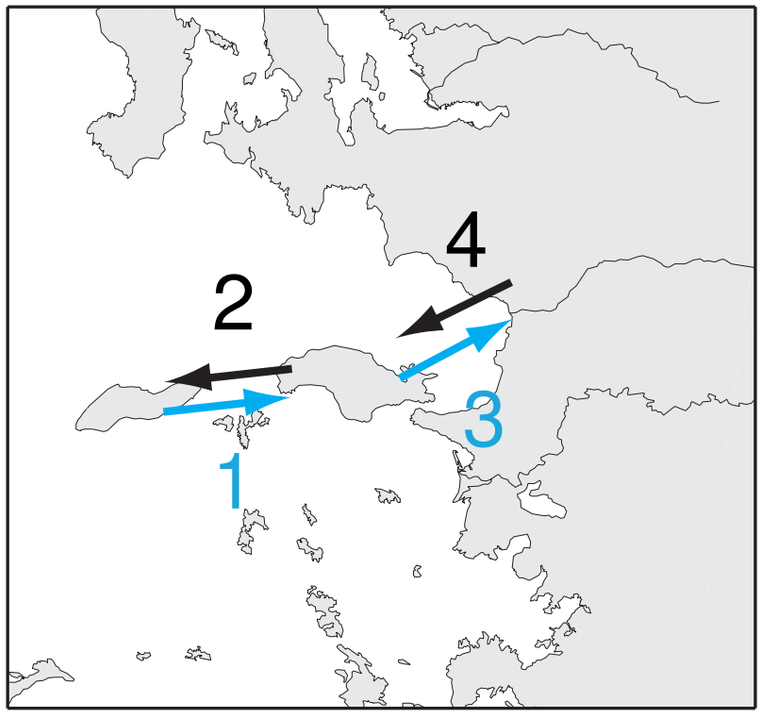


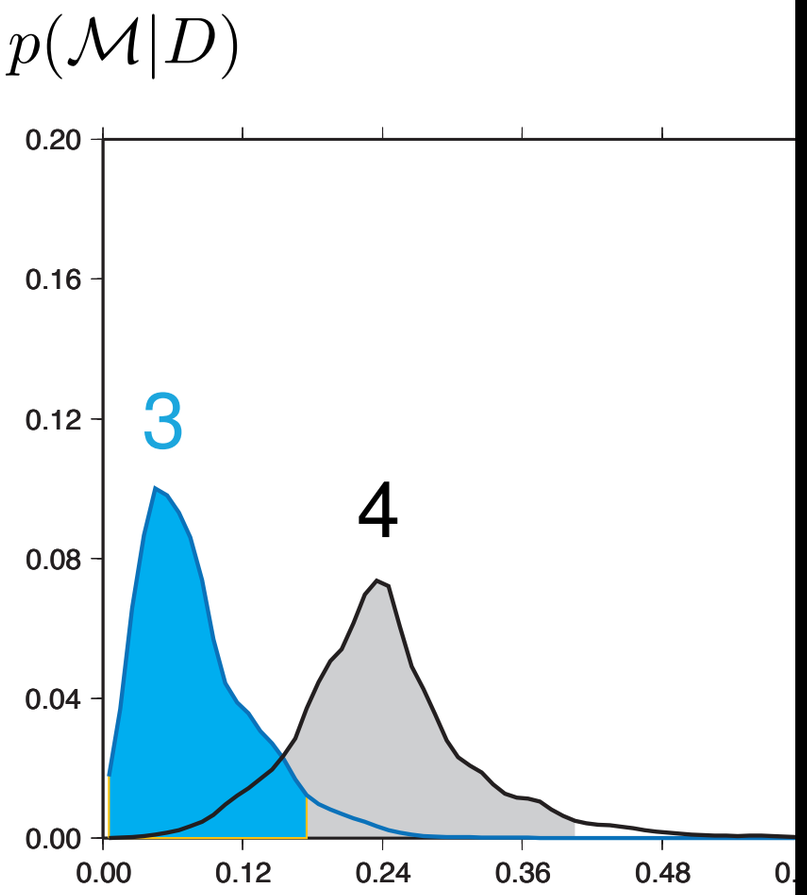




Obvious migration pattern

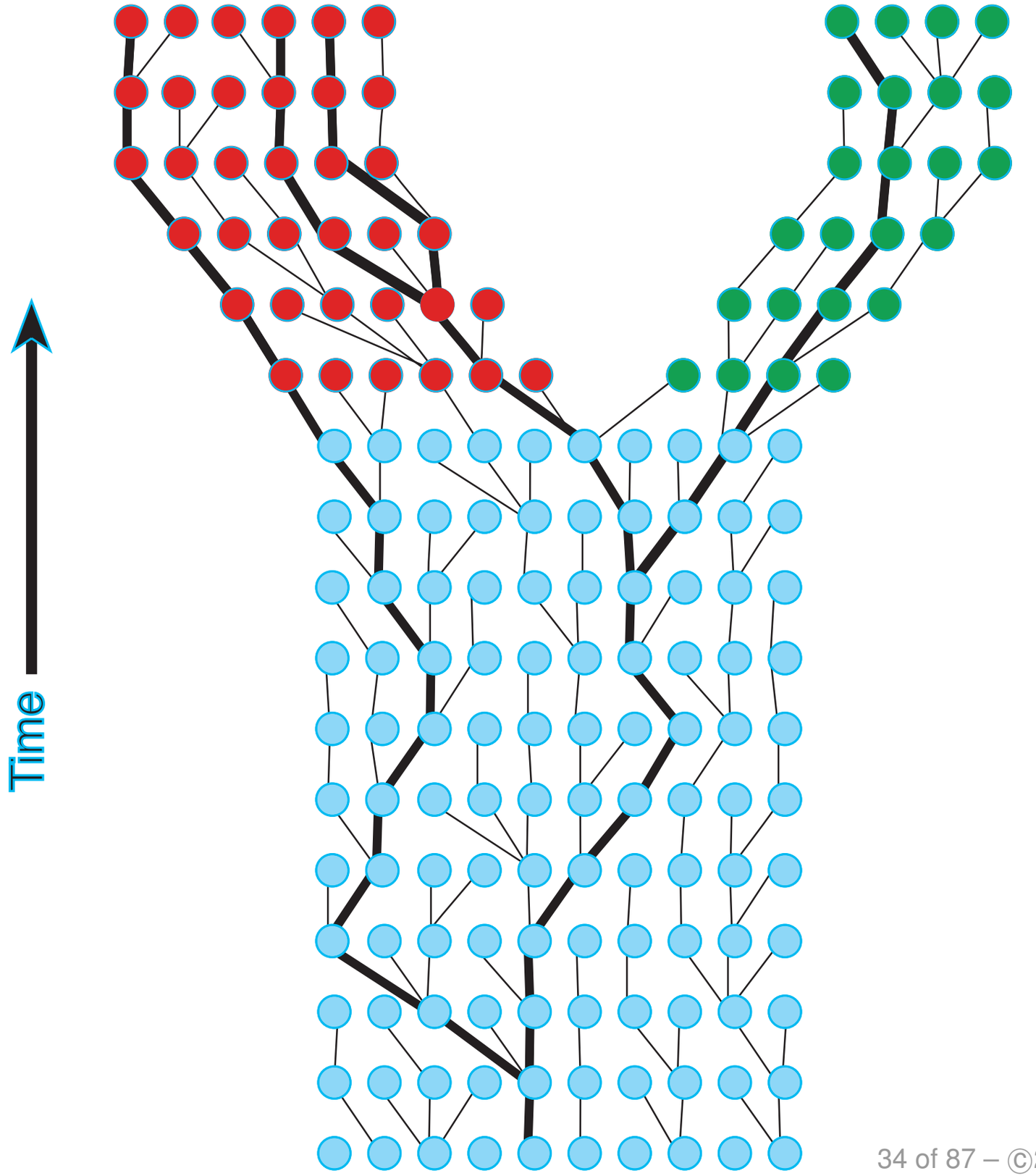




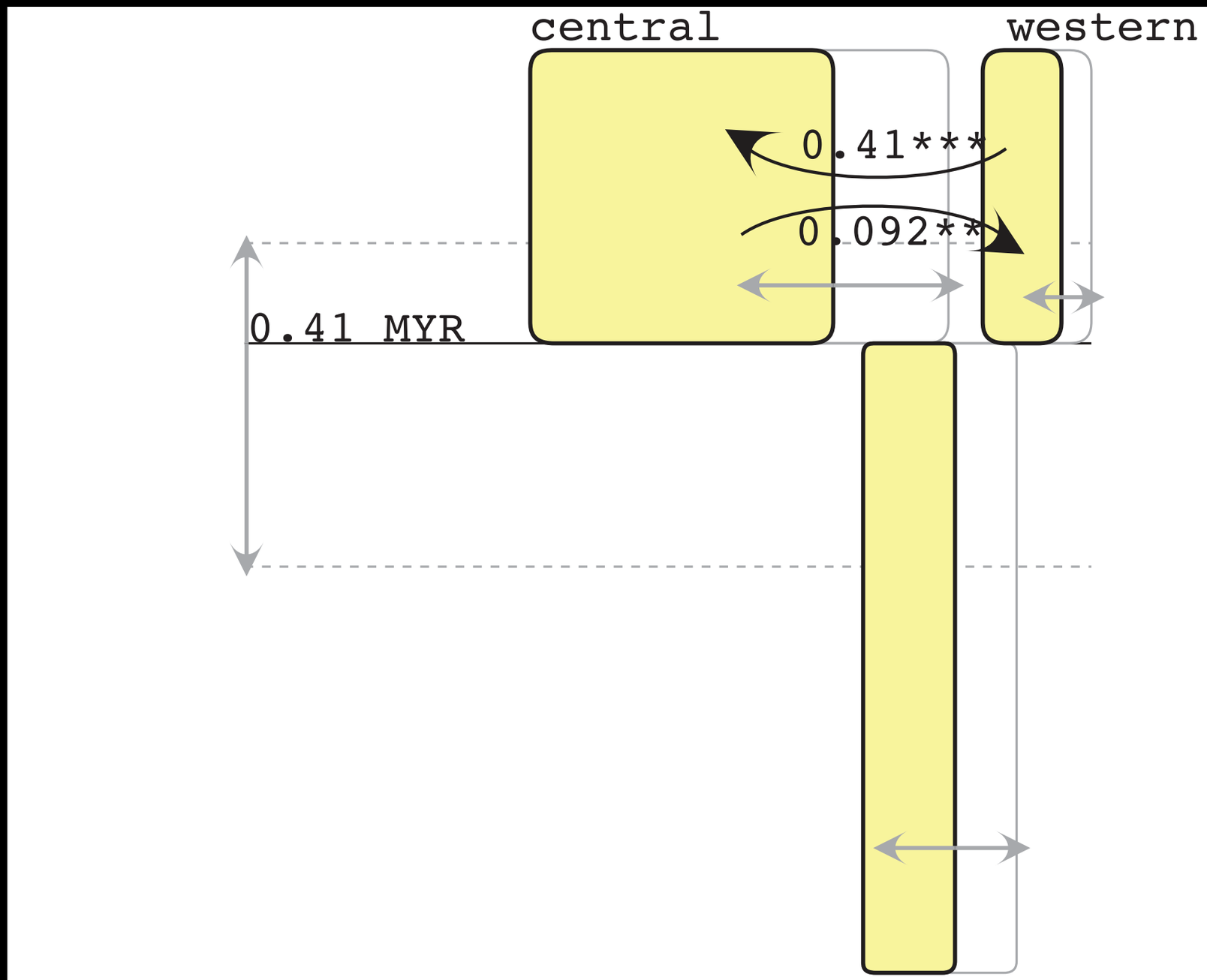


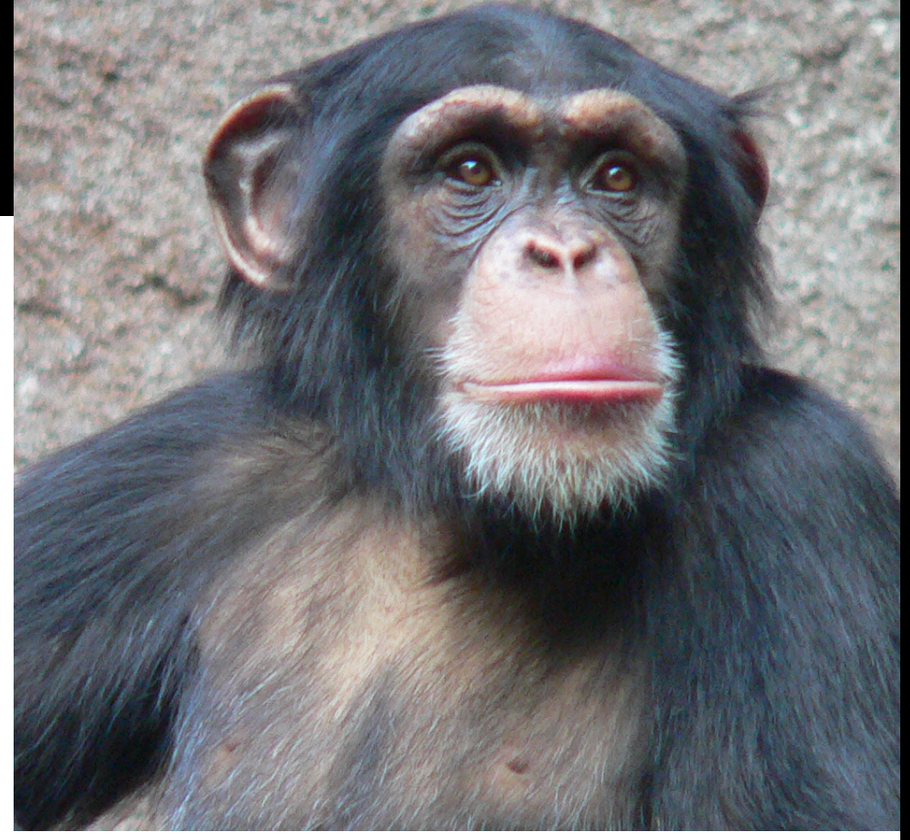
Extensions of the basic coalescent

Population splitting



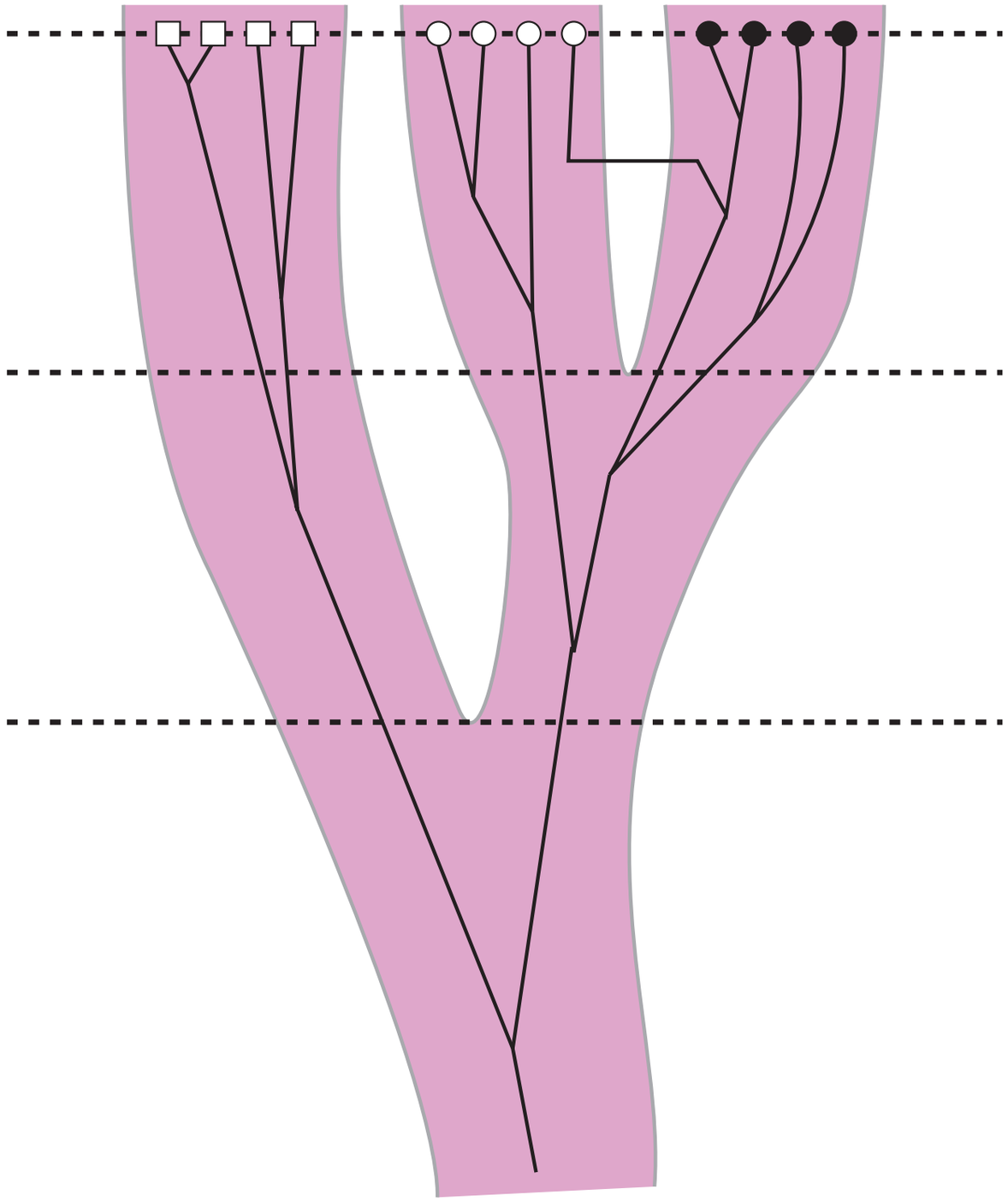
Population splitting





IM: isolation with migration; co-estimation of divergence parameters, population sizes and migration rates. Not all datasets can separate migration from divergence, and multiple loci are helpful.

Population splitting



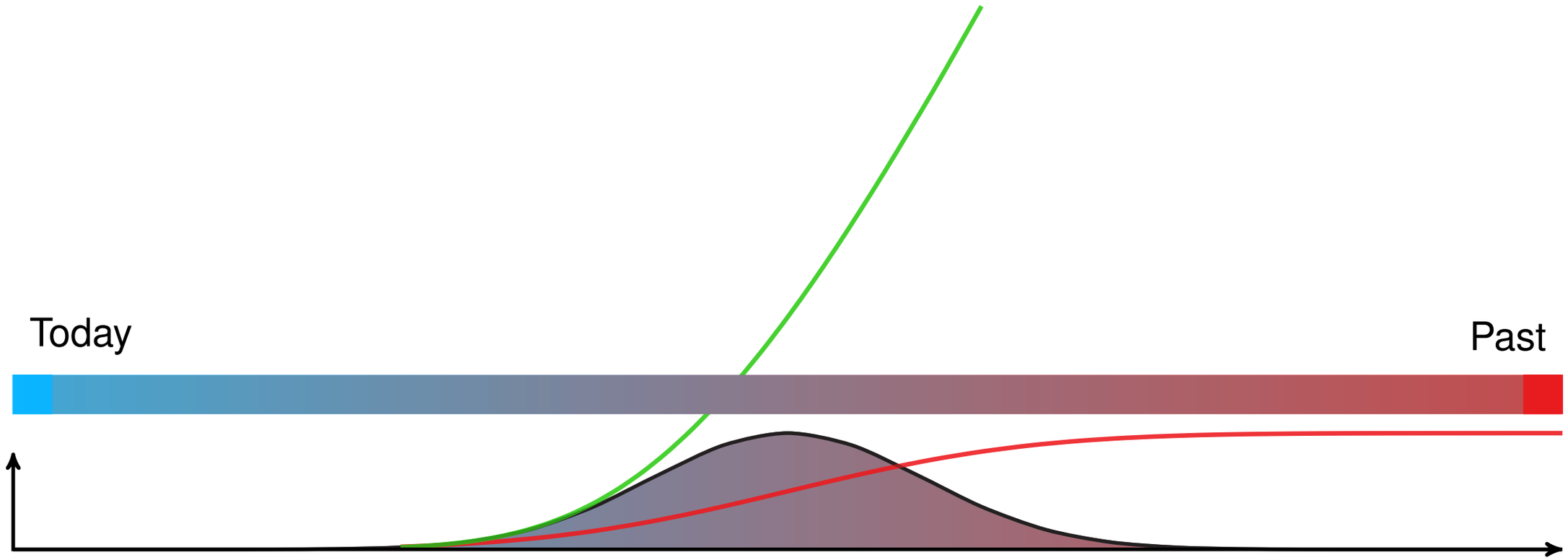
Population splitting

if we consider only a single individual that is today in population **A**. We also know that its ancestor was a member of population **B** then it will be only a matter of time to change the population label, but when?

(Palczewski, Ashki, and Beerli [in prep.] An alternative population fission model to the isolation with migration model.)

Population splitting

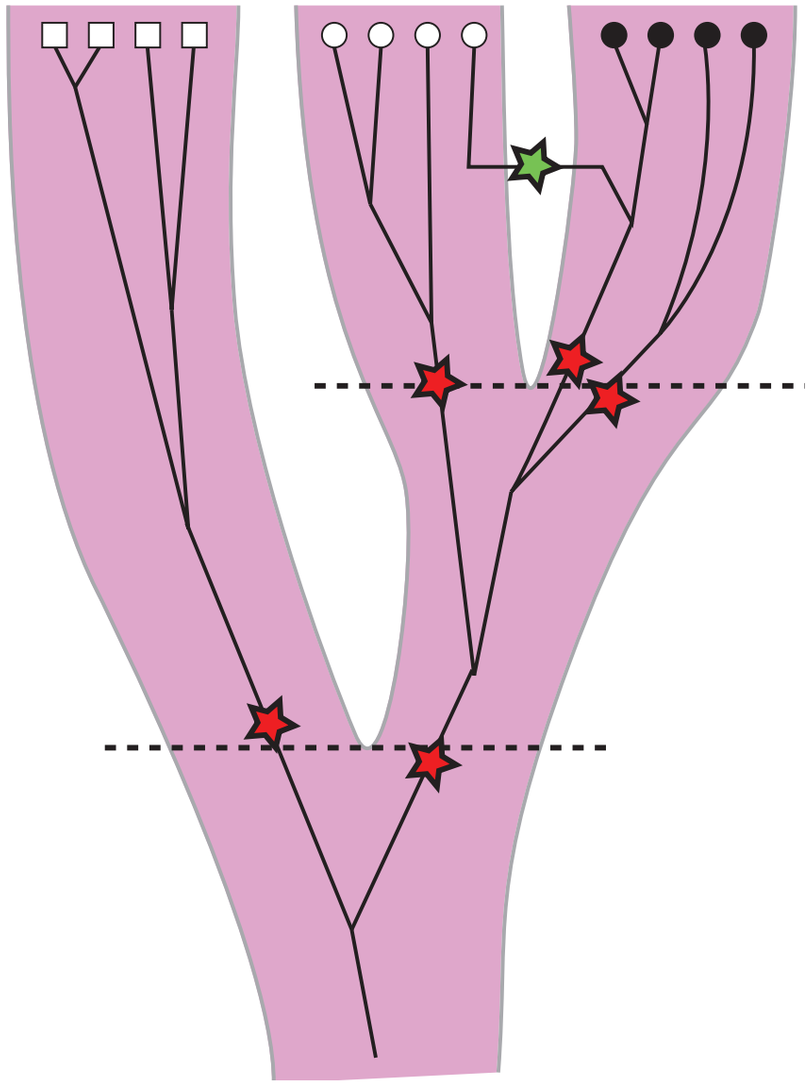
Looking backwards in time we could think about the risk of **A** turning into **B** which becomes larger and larger the further back in time the lineage goes. In the coalescence framework we are well accustomed to that thinking: we use the risk of a coalescent or the risk of a migration event. This risk can be expressed using the hazard function (or failure rate). Here we use the hazard function of the Normal distribution.



(Palczewski, Ashki, and Beerli [in prep.] An alternative population fission model to the isolation with migration model.)

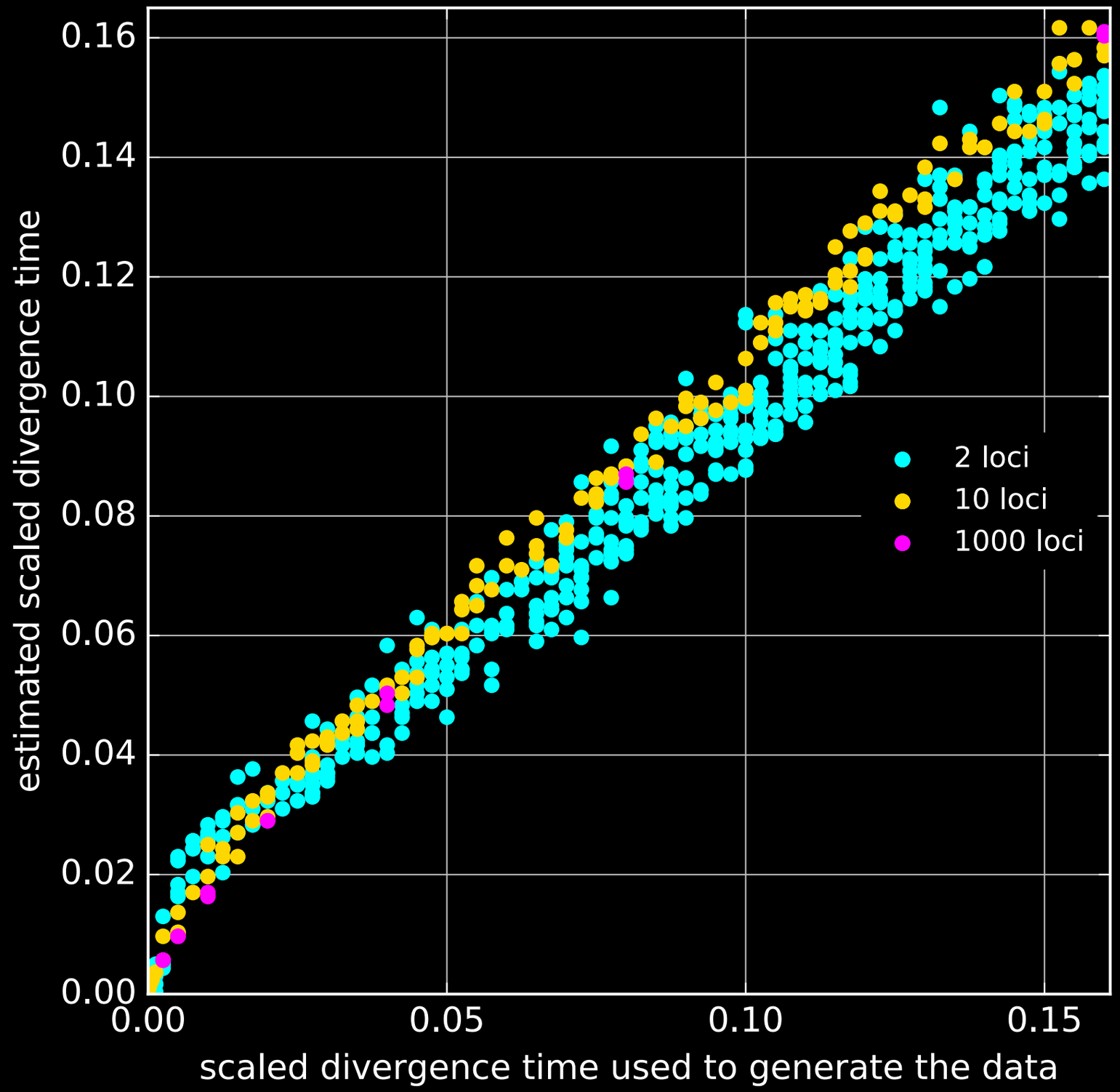
Population splitting

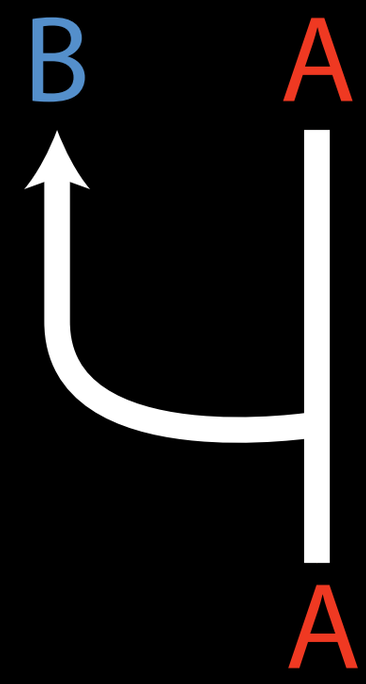
One lineage is easy, but what about the genealogy? Each lineage is at risk of being in the ancestral population, thus we need to consider coalescences, migration events, and population label changing events. This results in genealogies that are realizations of migration and population splitting events.



(Palczewski, Ashki, and Beerli [in prep.] An alternative population fission model to the isolation with migration model.)

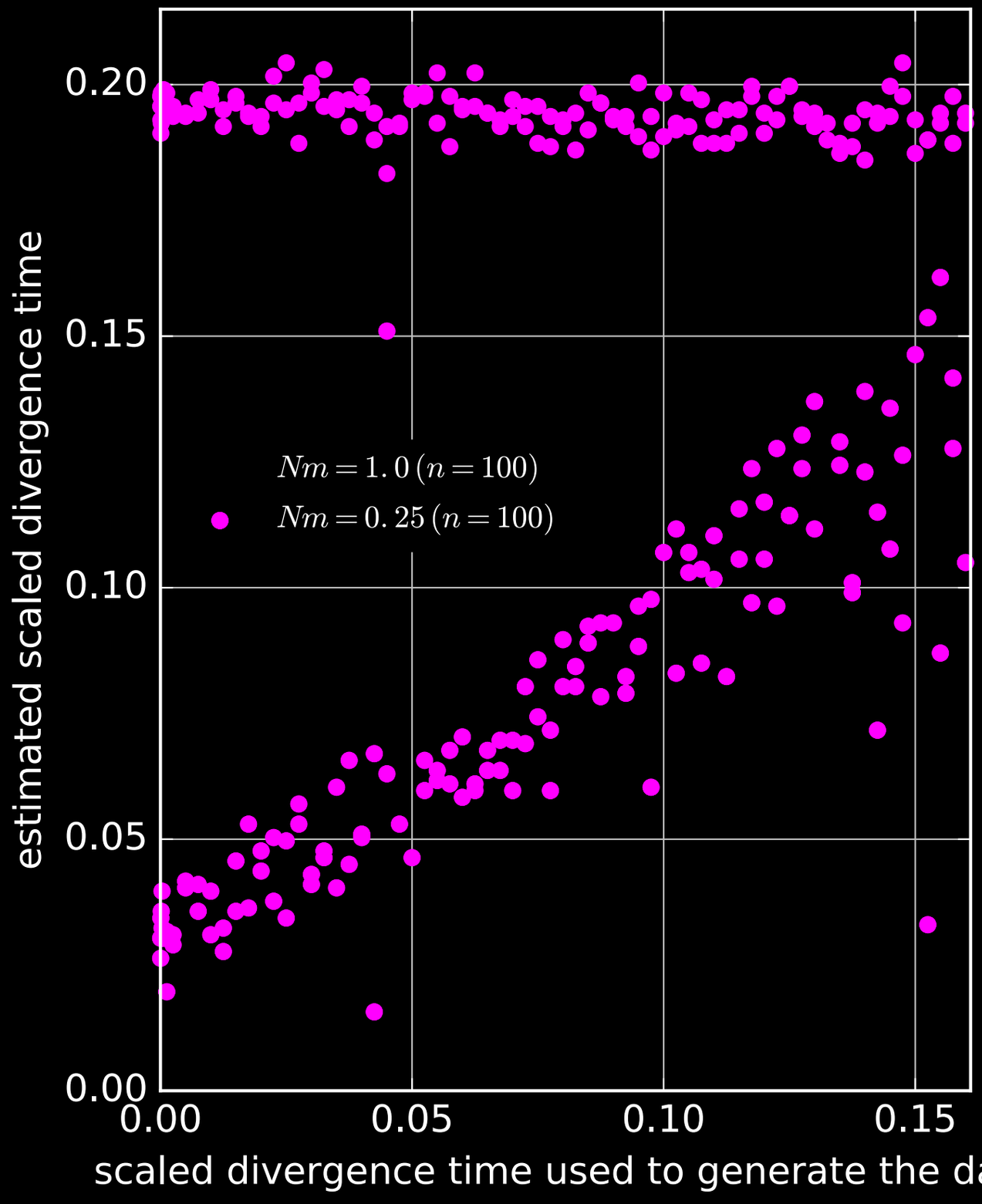
Estimated versus simulated divergence times

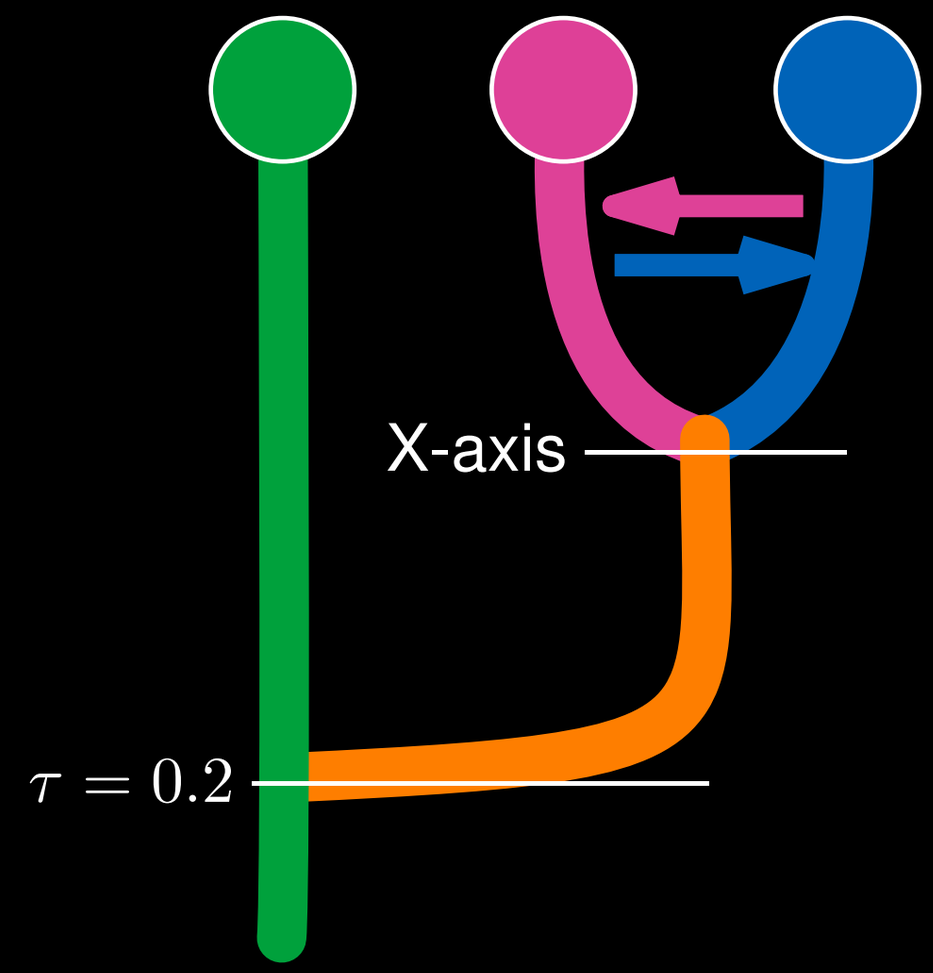




(Beerli, Ashki, and Palczewski [in prep.] Population divergence estimation using individual lineage label switching.)

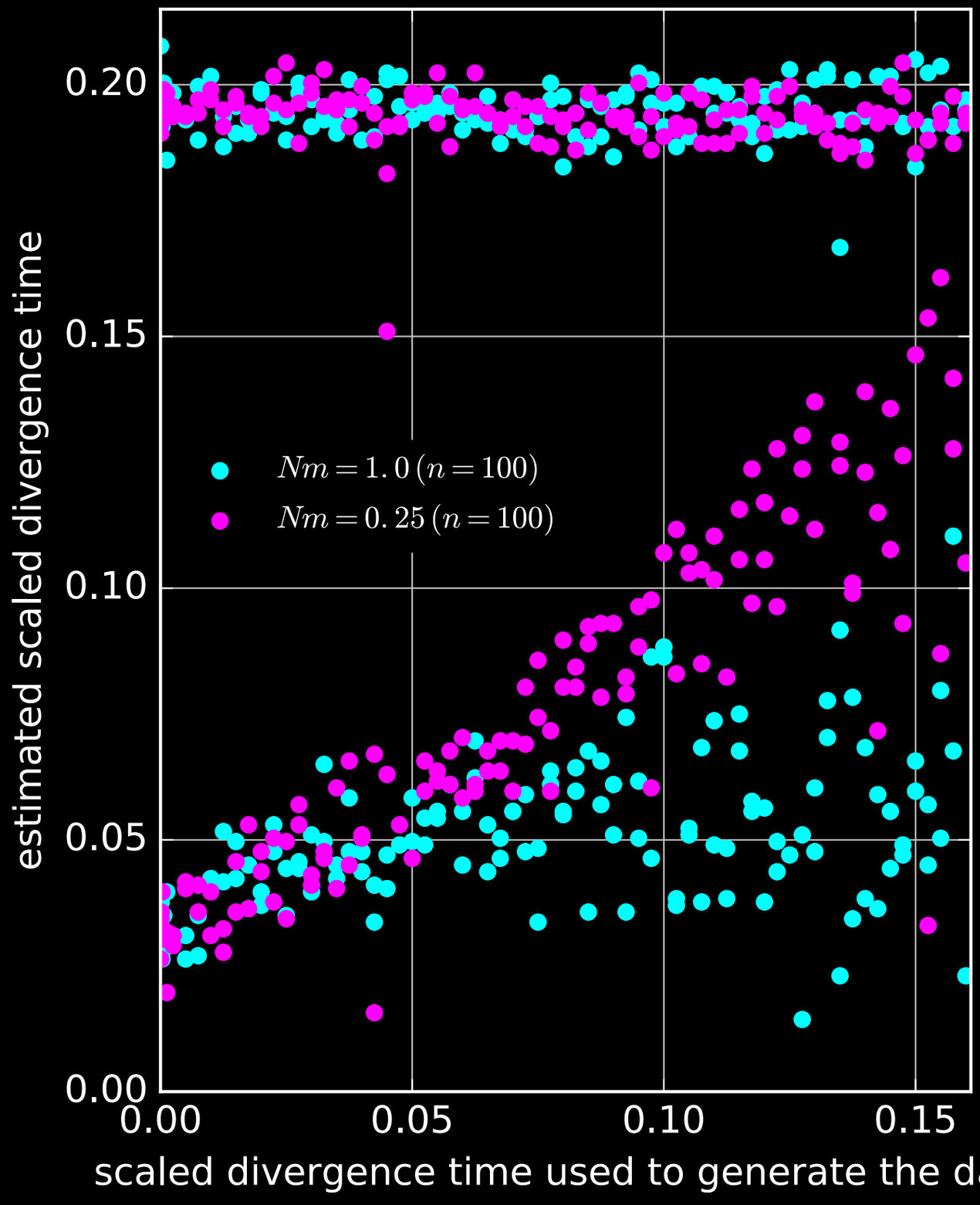
Estimated versus simulated divergence times

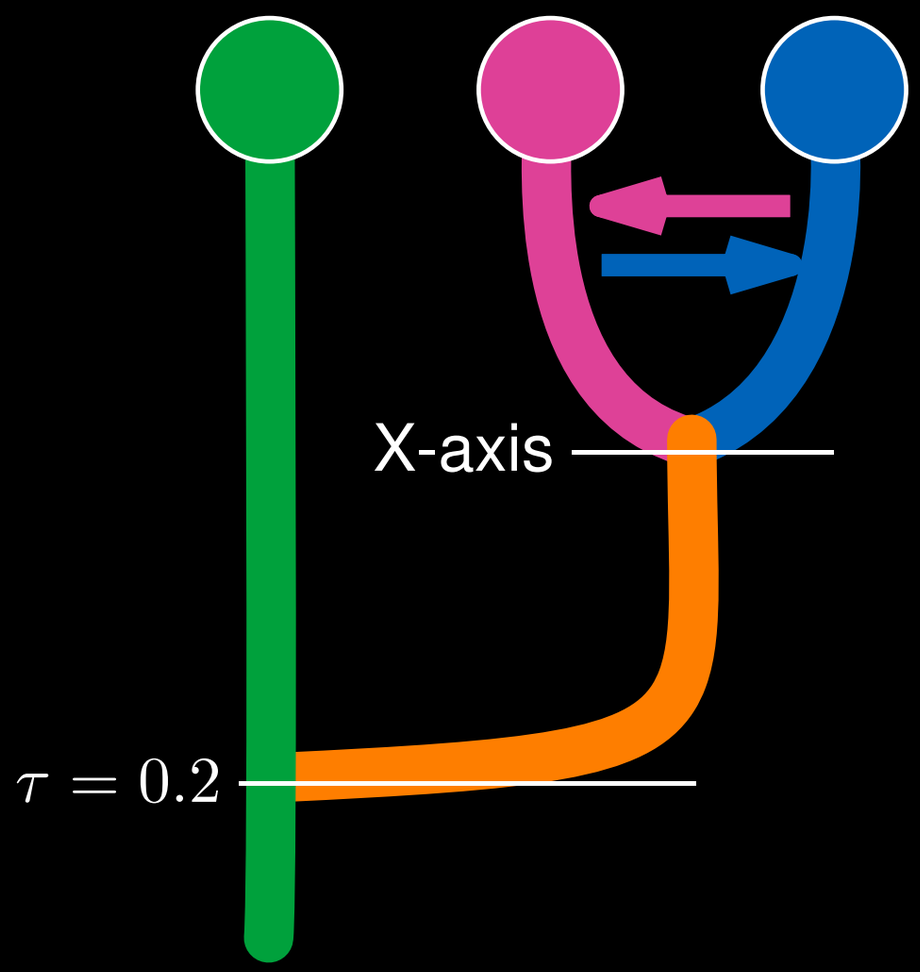




(Beerli, Ashki, and Palczewski [in prep.] Population divergence estimation using individual lineage label switching.)

Estimated versus simulated divergence times





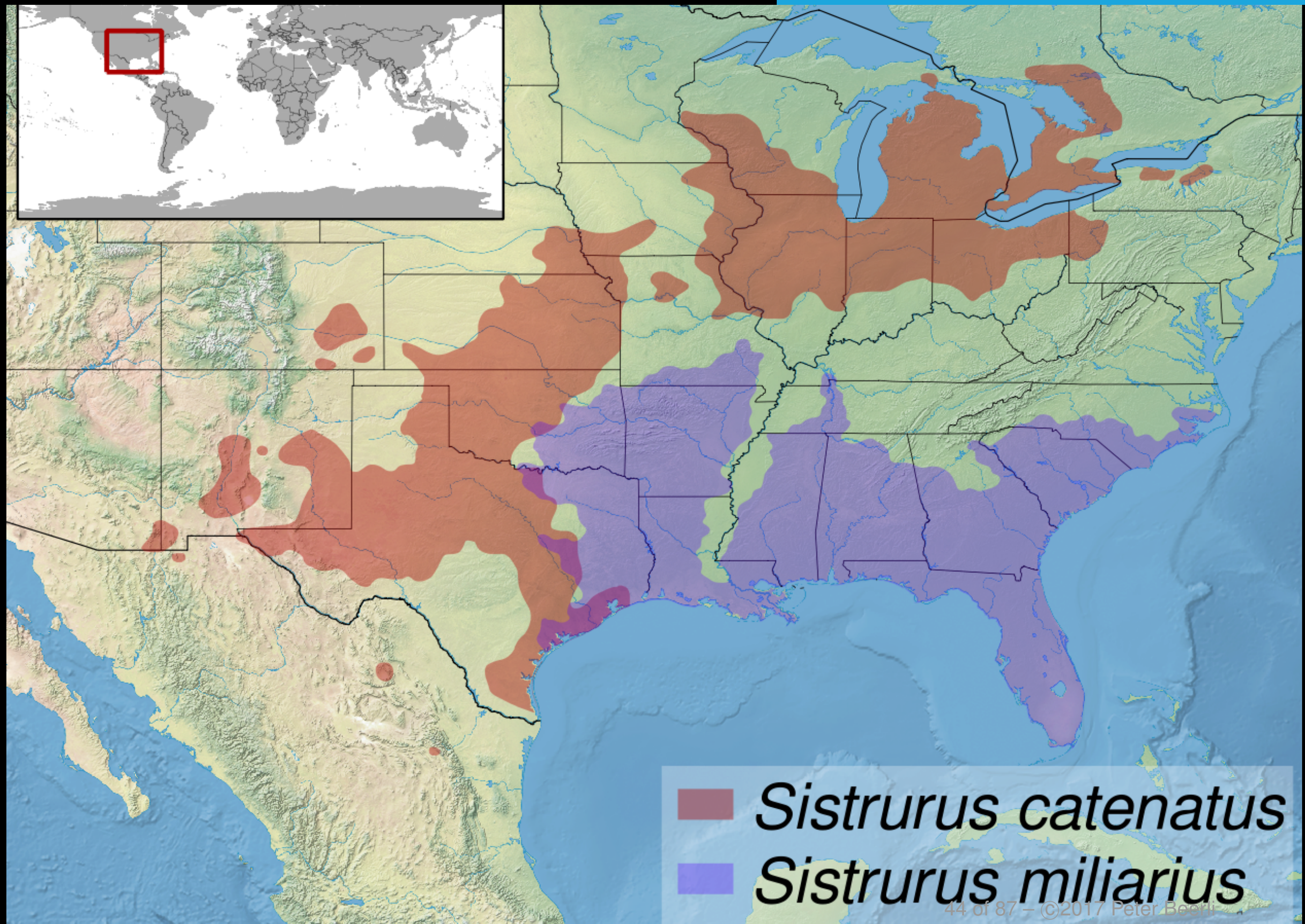
(Beerli, Ashki, and Palczewski [in prep.] Population divergence estimation using individual lineage label switching.)

Population splitting

Lisle Gibbs, Ohio (Kubatko et al. 2011)

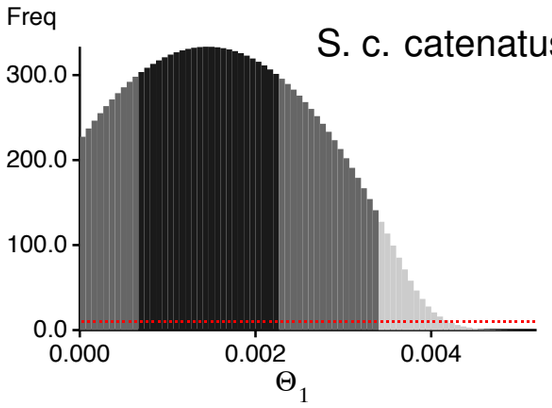
Population splitting

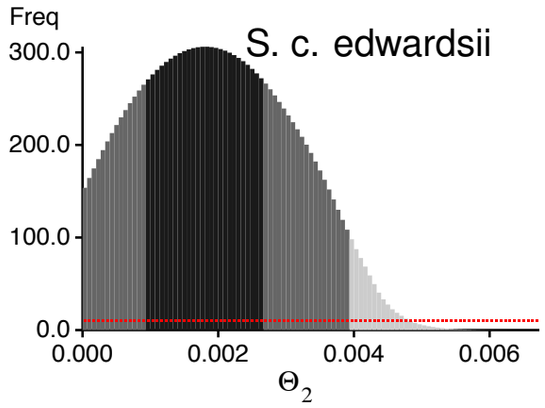
Wikipedia: Sistrurus

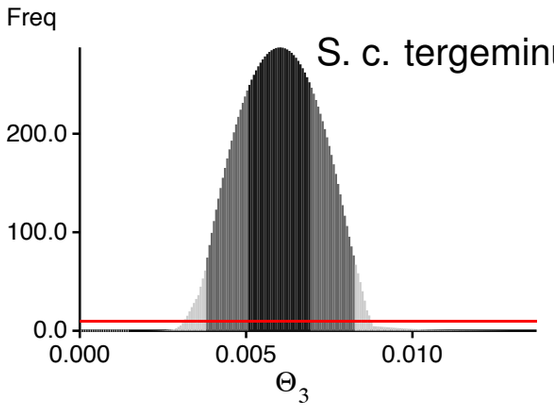


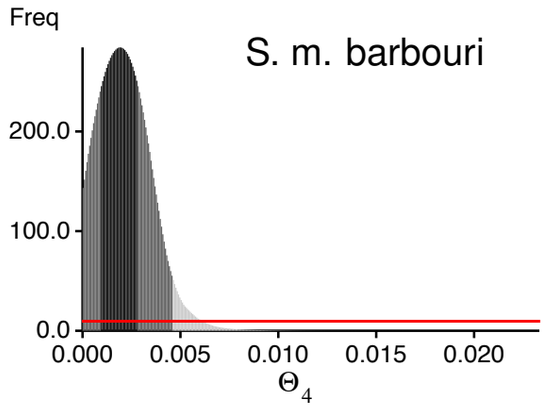
Population splitting

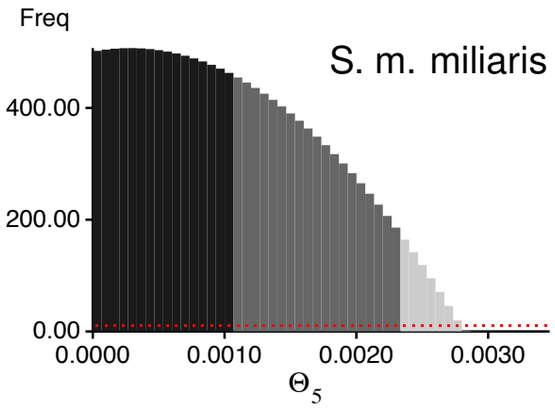
Pygmy rattle snakes

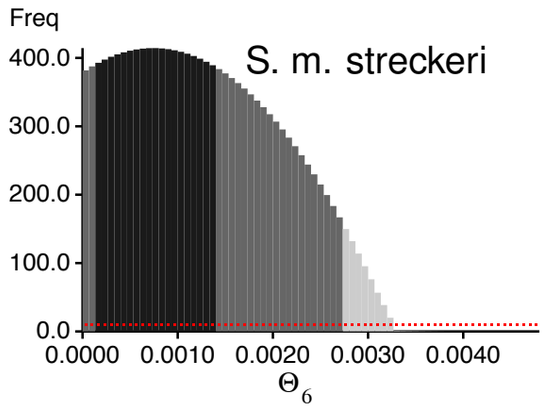


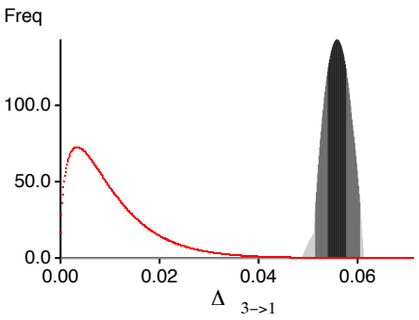


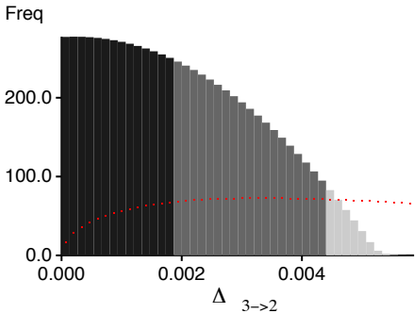


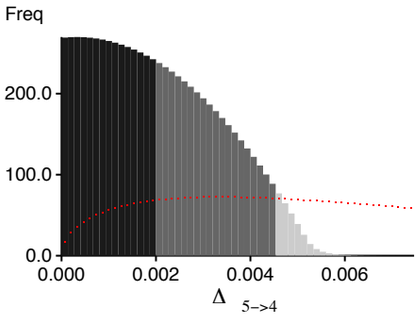


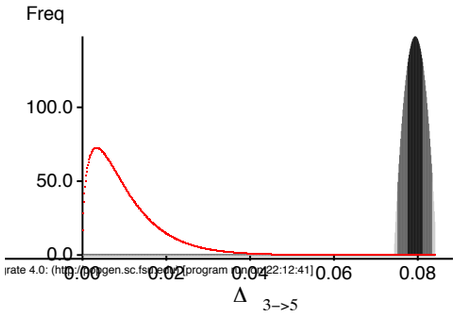








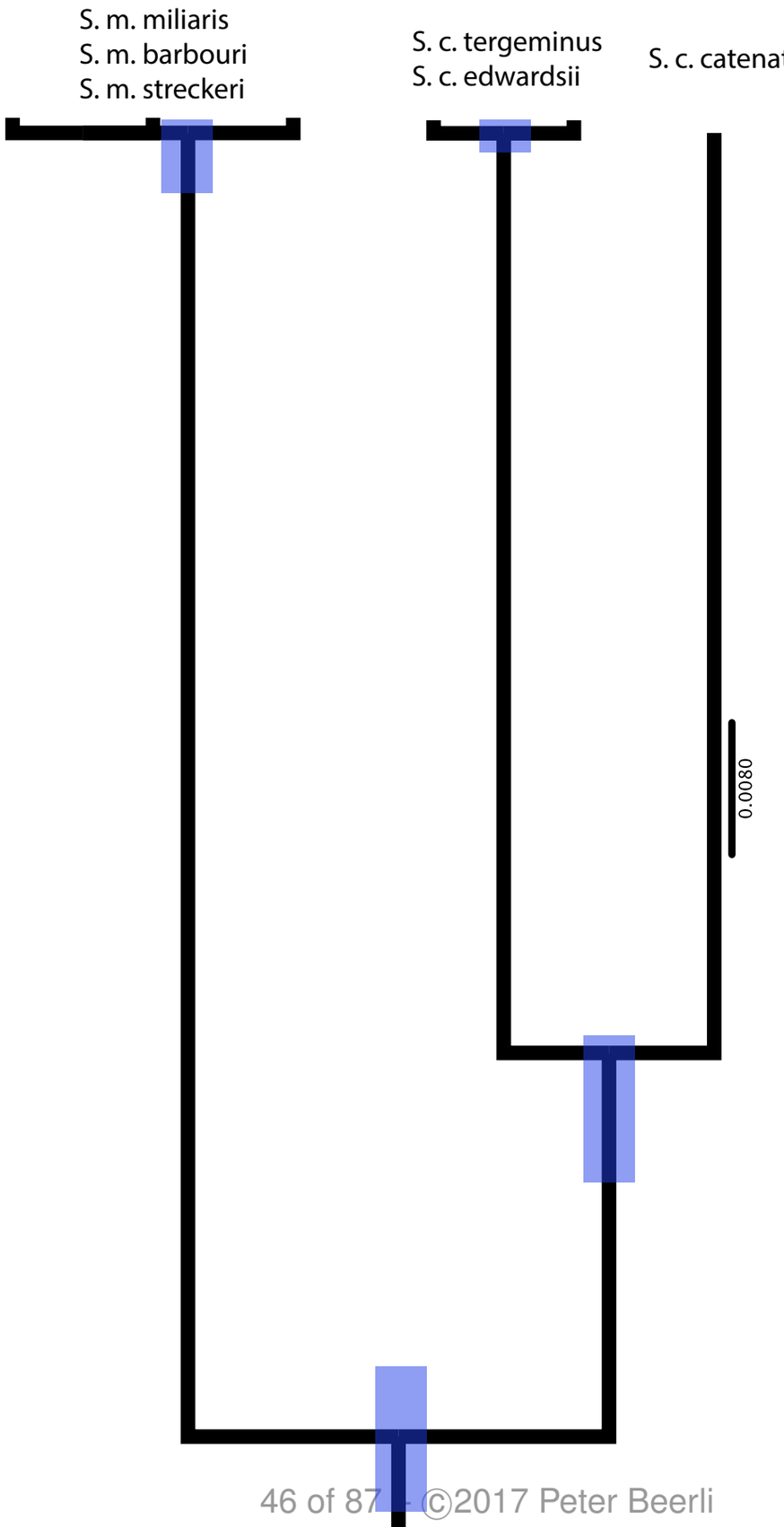




Estimation of splitting dates of 6 subspecies of pygmy rattle snakes using MIGRATE (data from Kubatko et al. 2011)

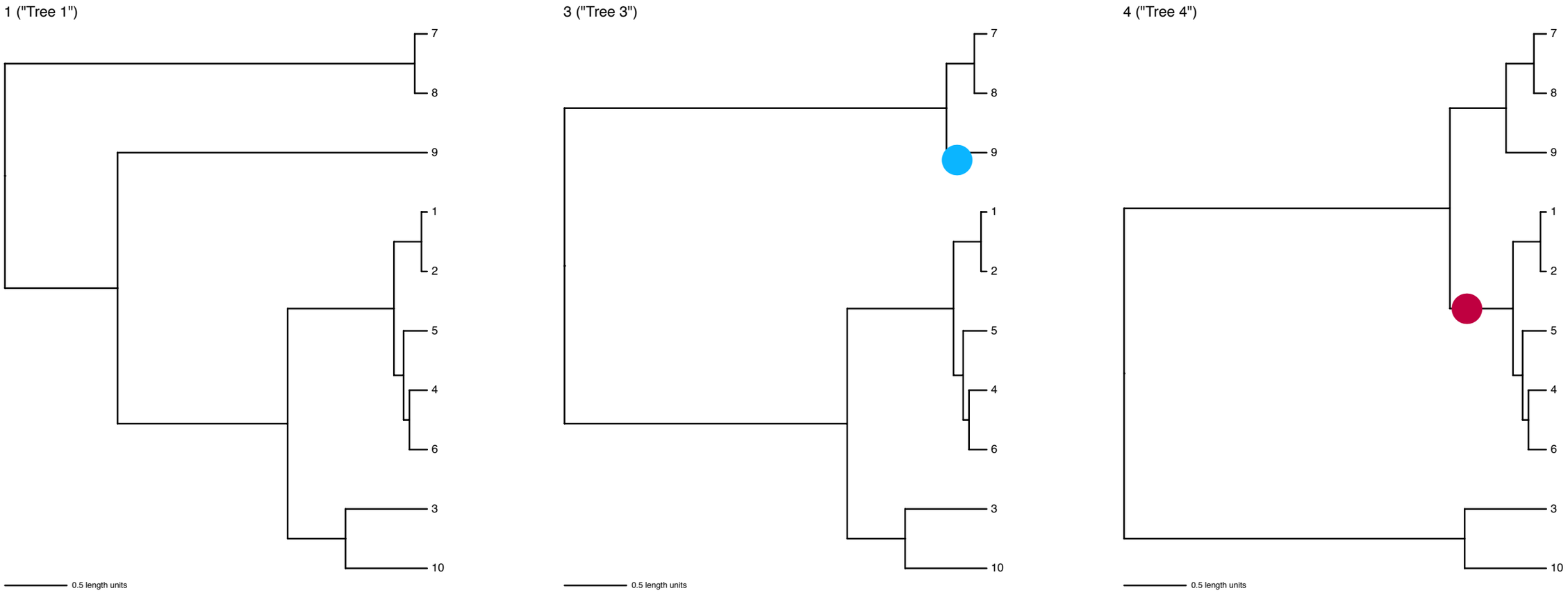
Population splitting

Pygmy rattle snakes



Estimation of splitting dates of 6 subspecies of pygmy rattle snakes using MIGRATE (data from Kubatko et al. 2011)

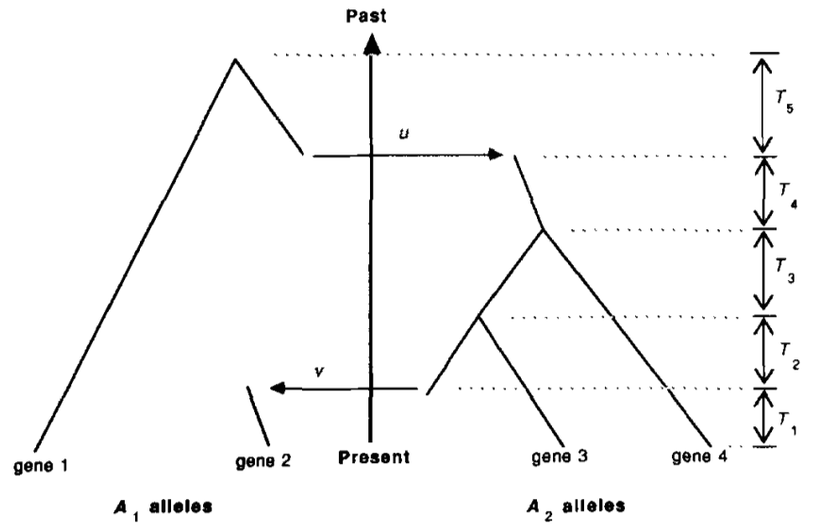
Coalescent and Recombination



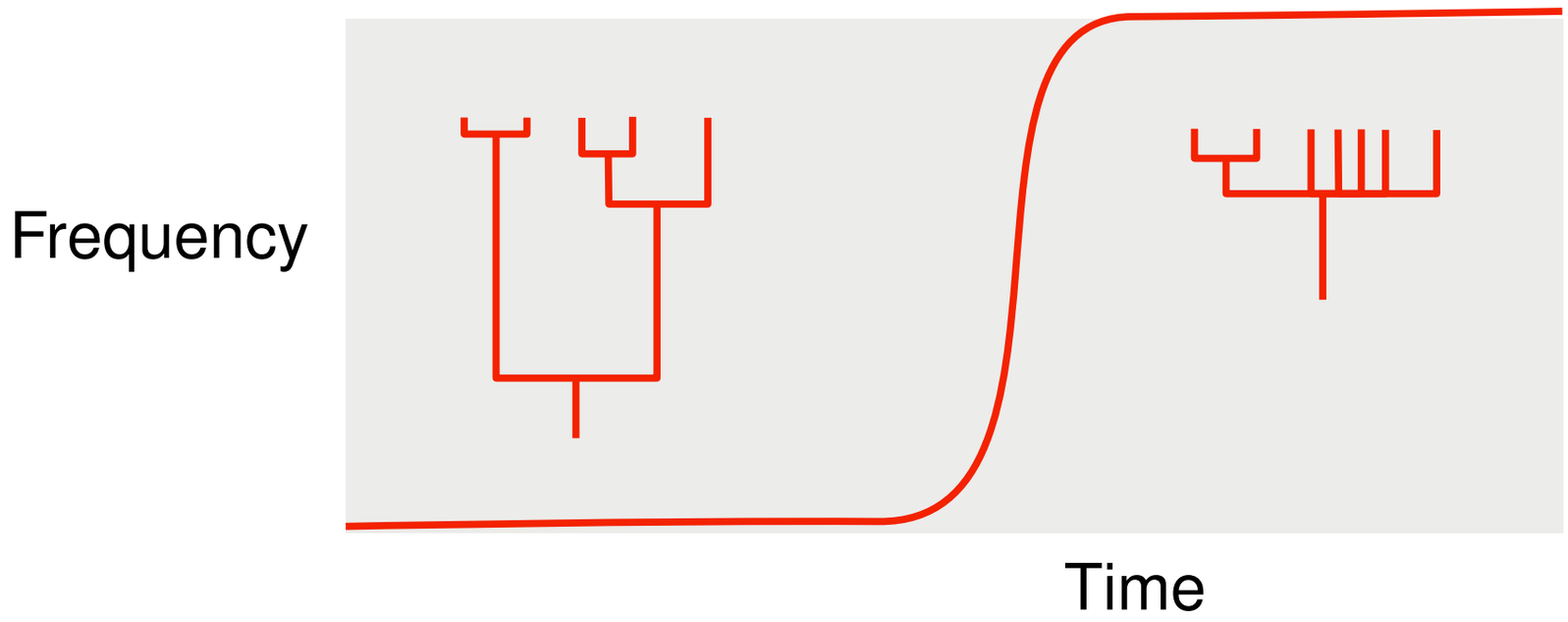
Programs that analyze recombination: LAMARC (Kuhner et al. 2006). [see also last section of lecture]

Coalescent and Selection

balancing selection: We can treat the the observed selection classes as 'populations' and the migration rate will become a measure of selection pressure. (Darden, Kaplan, and Hudson 1988)

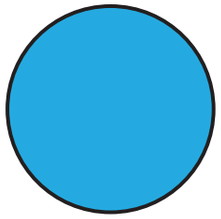


positive selection:

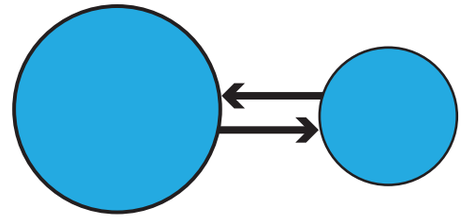


So many models – so little time

Structured vs non-structured populations



A single population allows free interbreeding of all individuals, mutation accumulate approximately by $N \times \mu$ where N is the population size, and μ is the mutation rate per generation. Highly variable populations persist longer and can resist catastrophes better.



A structured population restricts interbreeding to the subpopulations. Variability in a subpopulation is gained about $N_{\text{subpop}} \times (m + \mu)$ where m is the immigration rate per generation. With very high immigration rates the structured population behaves like a single population. If N_{subpop} is small the risk of extinction is high, but such systems are often more resistant to extinction by a parasite/virus/bacteria because the transmission of these is slowed down compared to a single population.

Location versus Population

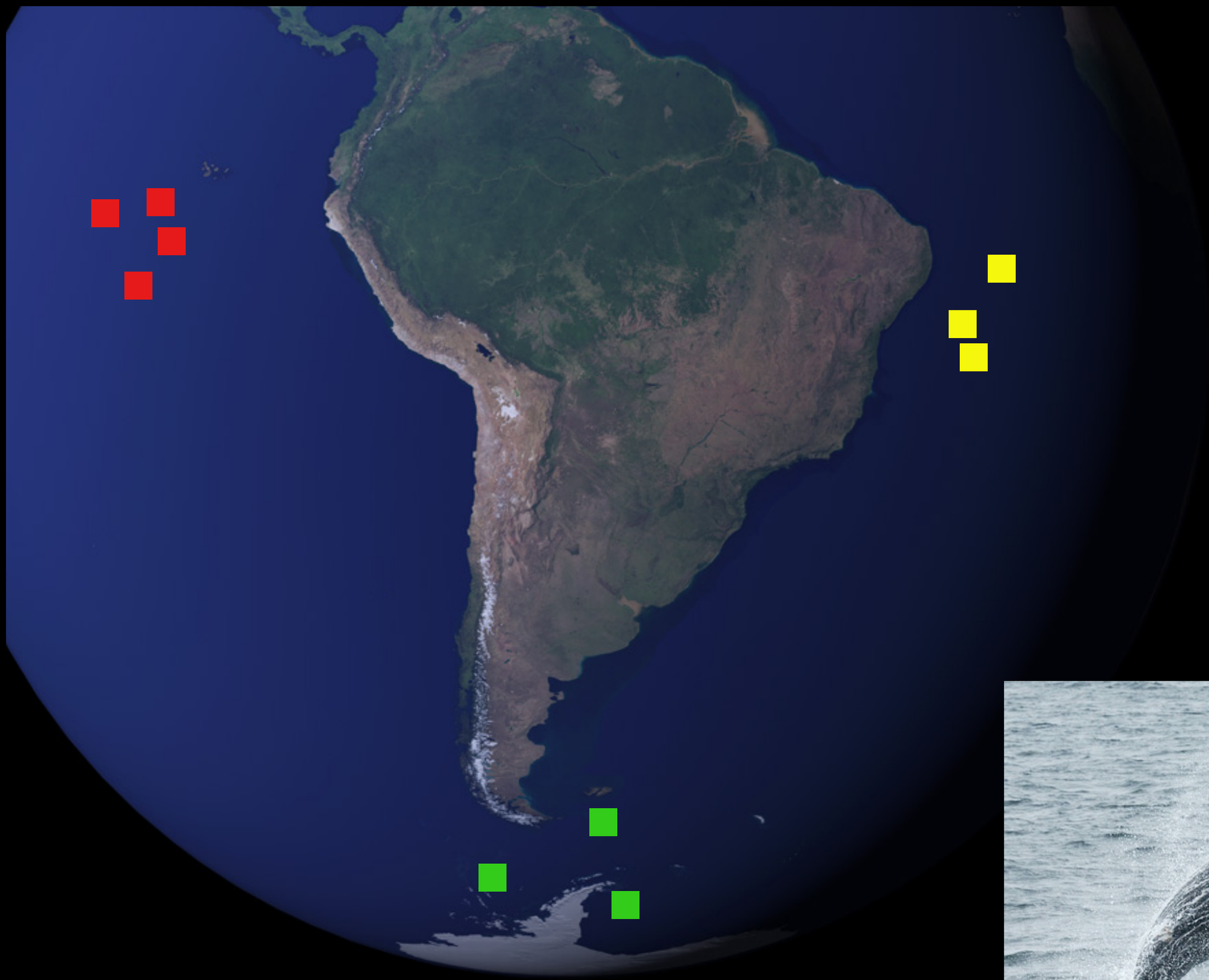
Location versus Population

Location ≈ Population

Location versus Population

Picture: Brandt – ©2017 Peter Beerli

Location $\stackrel{?}{=}$ Population



Picture: Bilanol – ©2017 Peter Beard

Model comparison

With a criterium such as likelihood we can compare nested models. In phylogenetics, we commonly use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different from each other, or which mutation model provides the best answers among the tested models.

Kass and Raftery (1995) popularized the Bayes Factor as a Bayesian alternative to the LRT.

Betting and Odds Ratios

Knew that we ventured on such dangerous seas
That if we wrought out life 'twas ten to one
William Shakespeare (Henry IV).

circa. 1594, by Michelangelo Merisi da Caravaggi

Bayesian Odds Ratios

Using Bayes' theorem:

$$\mathrm{p}(\mathrm{M}_1|\mathrm{X}) = \frac{\mathrm{p}(\mathrm{M}_1)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}$$

we can express support of one model over another as a ratio:

$$\frac{\mathrm{p}(\mathrm{M}_1|\mathrm{X})}{\mathrm{p}(\mathrm{M}_2|\mathrm{X})} = \frac{\frac{\mathrm{p}(\mathrm{M}_1)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}}{\frac{\mathrm{p}(\mathrm{M}_1)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}}$$

Bayesian Odds Ratios

Using Bayes' theorem:

$$p(M_1|X) = \frac{p(M_1)p(X|M_1)}{p(X)}$$

we can express support of one model over another as a ratio:

$$\frac{p(M_1|X)}{p(M_2|X)} = \frac{\frac{p(M_1)p(X|M_1)}{p(X)}}{\frac{p(M_1)p(X|M_1)}{p(X)}}$$

$$\text{Posterior Odds } \frac{p(M_1|X)}{p(M_2|X)} = \text{Prior Odds } \frac{p(M_1)}{p(M_2)} \times \text{Bayes Factor } \frac{p(X|M_1)}{p(X|M_2)}$$

Bayesian Odds Ratios

$$\frac{p(M_1|X)}{p(M_2|X)} = \frac{\frac{p(M_1)p(X|M_1)}{p(X)}}{\frac{p(M_1)p(X|M_1)}{p(X)}}$$

$$\text{Posterior Odds } \frac{p(M_1|X)}{p(M_2|X)} = \text{Prior Odds } \frac{p(M_1)}{p(M_2)} \times \text{Bayes Factor } \frac{p(X|M_1)}{p(X|M_2)}$$

Bayes factor

We can use the posterior odds ratio or equivalently the Bayes factors for model comparison:

$$\mathrm{BF} = \frac{\mathrm{p(X|M_1)}}{\mathrm{p(X|M_2)}} \qquad \mathrm{LBF} = 2\ln \mathrm{BF} = 2\ln\left(\frac{\mathrm{p(X|M_1)}}{\mathrm{p(X|M_2)}}\right)$$

The magnitude of BF gives us evidence against hypothesis M_2

$$\mathrm{LBF} = 2\ln \mathrm{BF} = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

Marginal likelihood approximation

So why are we not all running BF analyses instead of the AIC, BIC, DIC, FIC, GIC, LRT, ...?

Marginal likelihood

So why are we not all running BF analyses instead of the AIC, BIC, DIC, FIC, GIC, LRT, ...?

Typically, it is rather difficult to calculate the marginal likelihoods with good accuracy, because most often we only approximate the posterior distribution using Markov chain Monte Carlo (MCMC).
In MCMC we need to know only differences and therefore we typically do not need to calculate the denominator to calculate the Posterior distribution $\mathrm{p}(\Theta|\mathrm{X})$:

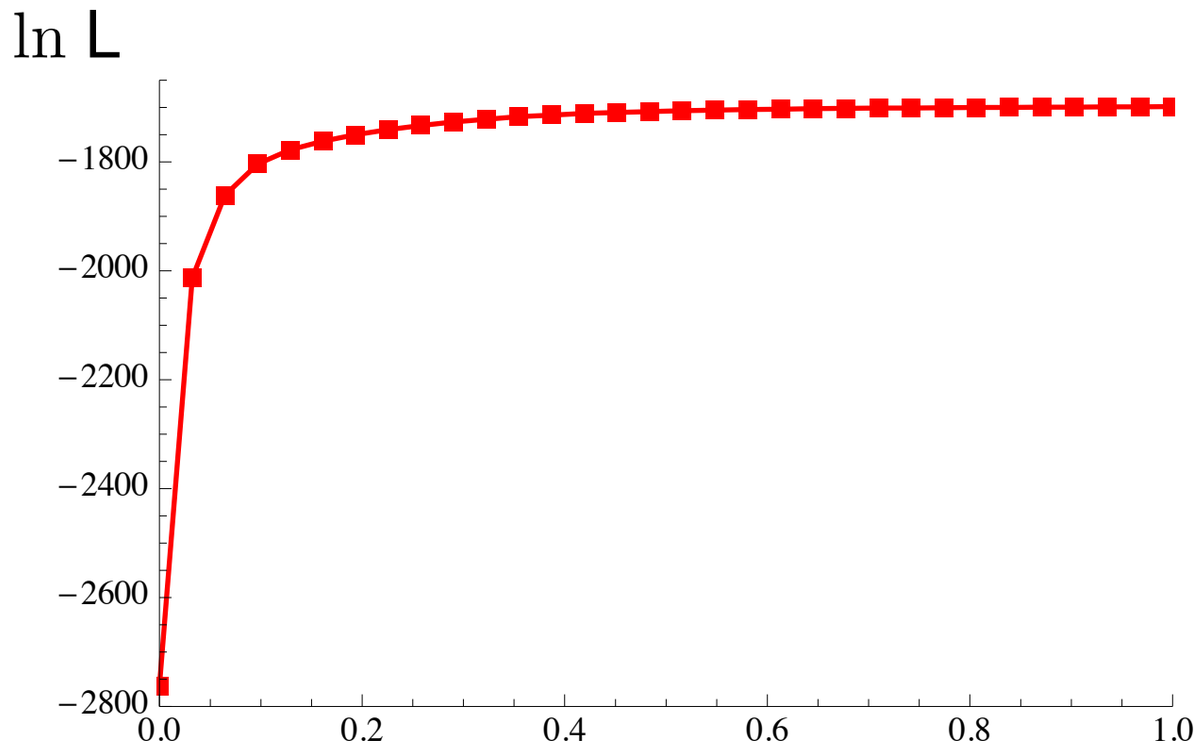
$$\mathrm{p}(\Theta|\mathrm{X},\mathrm{M}) = \frac{\mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)}{\mathrm{p}(\mathrm{X}|\mathrm{M})} = \frac{\mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)}{\int_\Theta \mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)\mathrm{d}\Theta}$$

where $\mathrm{p}(\mathrm{X}|\mathrm{M})$ is the marginal likelihood.

Thermodynamic integration

Thermodynamic integration (Path sampling) [Gelman and Meng 1997, Lartillot et al. 2006]: method is tedious to compute because several MCMC chains are needed. Results are accurate and reproducible with small variance when MCMC runs were run long enough.

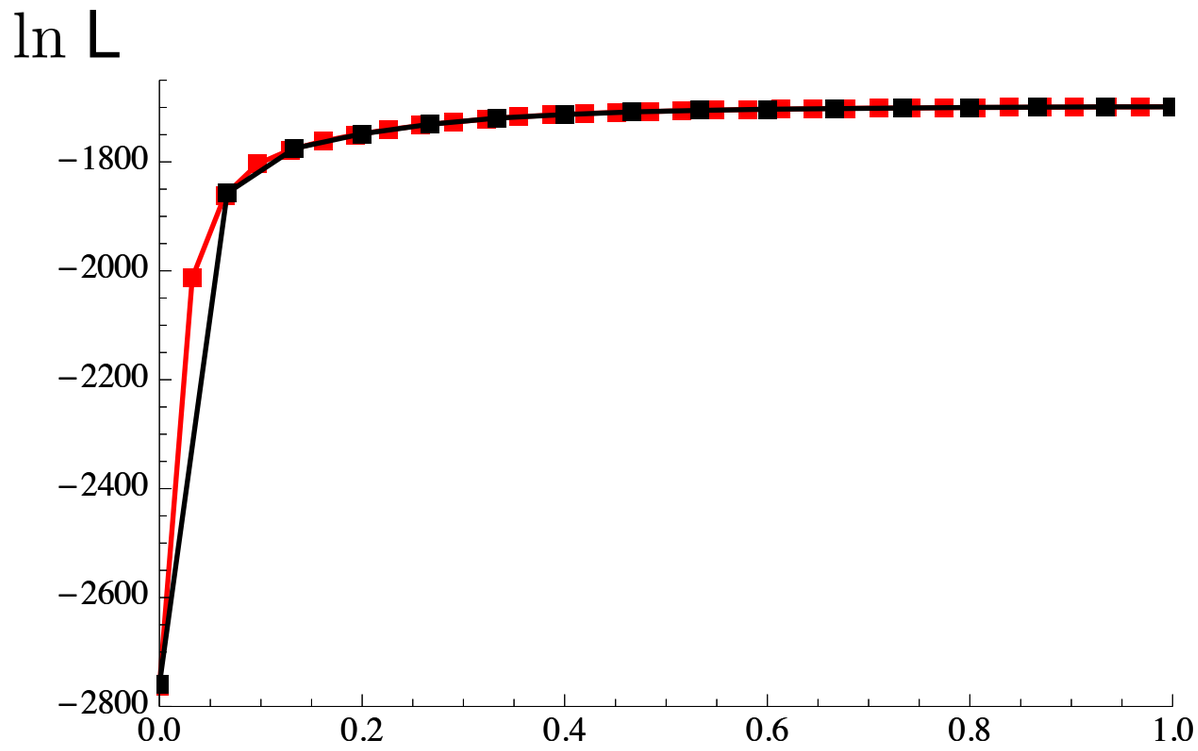
$$\ln p(X|M_i) = \int_0^1 \mathbb{E}(\ln p_t(X|M_i))dt$$



Thermodynamic integration

Thermodynamic integration (Path sampling) [Gelman and Meng 1997, Lartillot et al. 2006]: method is tedious to compute because several MCMC chains are needed. Results are accurate and reproducible with small variance when MCMC runs were run long enough.

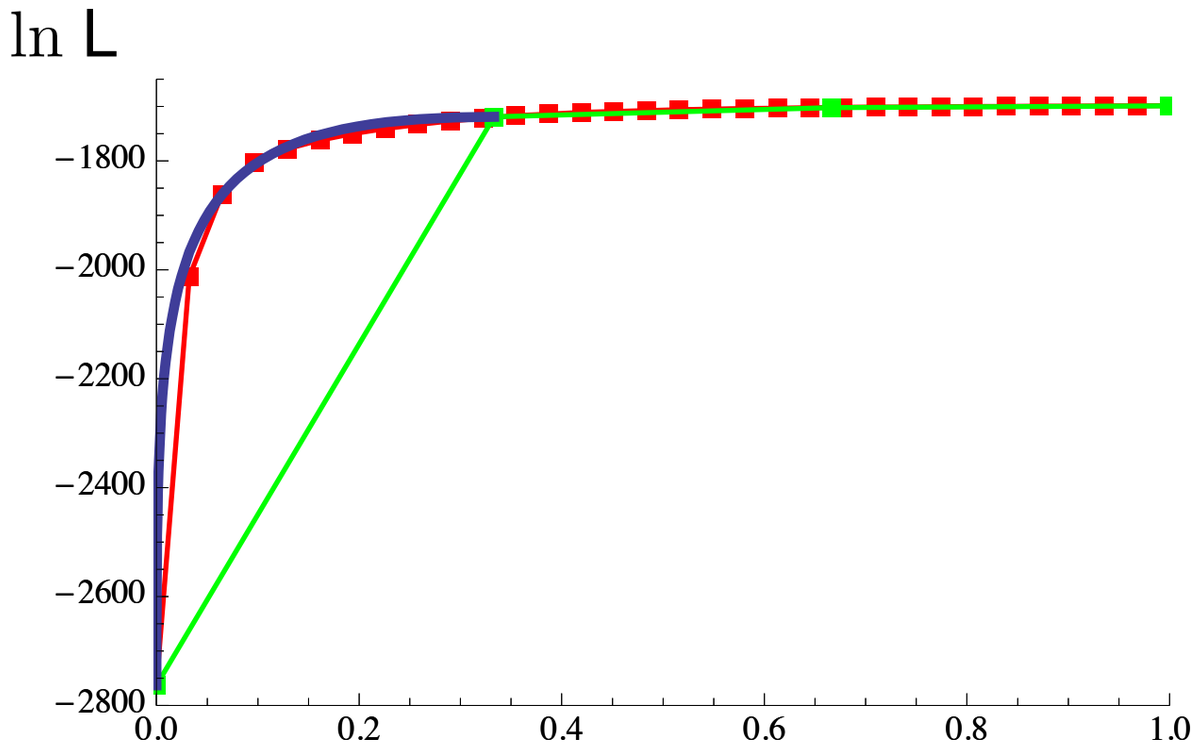
$$\ln p(X|M_i) = \int_0^1 \mathbb{E}(\ln p_t(X|M_i))dt$$



Thermodynamic integration

Thermodynamic integration (Path sampling) [Gelman and Meng 1997, Lartillot et al. 2006]: method is tedious to compute because several MCMC chains are needed. Results are accurate and reproducible with small variance when MCMC runs were run long enough.

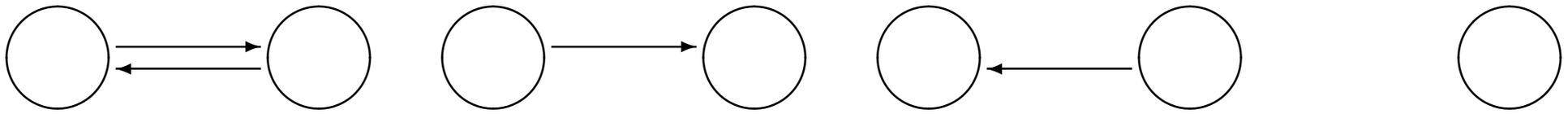
$$\ln p(X|M_i) = \int_0^1 \mathbb{E}(\ln p_t(X|M_i))dt$$



A simple example

We want to establish a direction of geneflow between 2 populations.

We generate 4 hypotheses



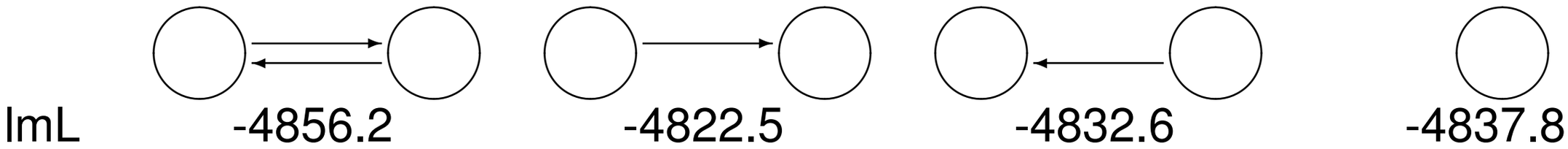
We collect data from individuals in the two populations

Analyze the data in MIGRATE

A simple example

Recipe: starting with the finished dish

of the 4 hypotheses:

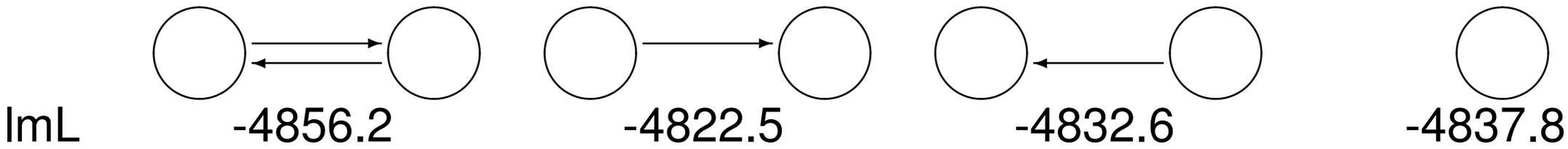


Data was simulated using the second model (2) from the left.

A simple example

Recipe: starting with the finished dish

of the 4 hypotheses:



The best model (highest lmL) is the model second from left (model 2).

We can calculate the log Bayes factor for two leftmost models as

$$LBF_{12} = 2(lmL_1 - lmL_2) = 2(-4856.2 - -4822.5) = -67.4$$

The value suggests that we should strongly prefer model 2 over model 1.

Data was simulated using the second model from the left (model 2).

A simple example

Recipe:

1. Pick the hypothesis with largest number of parameters

2. Set priors and run parameters (use heated chains) so that you are comfortable with the result (converged, etc)

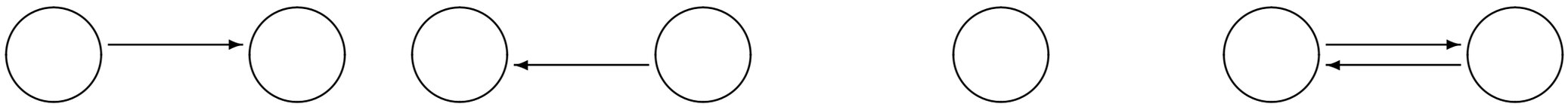
3. Record the log marginal likelihood from the output.

4. Pick next hypothesis, adjust migration model, and run and record the log marginal likelihood.

5. Repeat (4) until all log marginal likelihoods are calculated

6. Compare the log marginal likelihoods, for example order the hypothesis accordingly, or calculate the model probability

A simple example



lmL	-4822.5	-4832.6	-4837.8	-4856.2
P(model)	0.99	0.01	0.0	0.0

Model probability (Burnham and Anderson 2002) calculation:

$$\mathrm{P}(M_i) = \frac{\exp(lmL_i)}{\sum_j \exp(lmL_j)} = \frac{mL_i}{\sum_j mL_j}$$

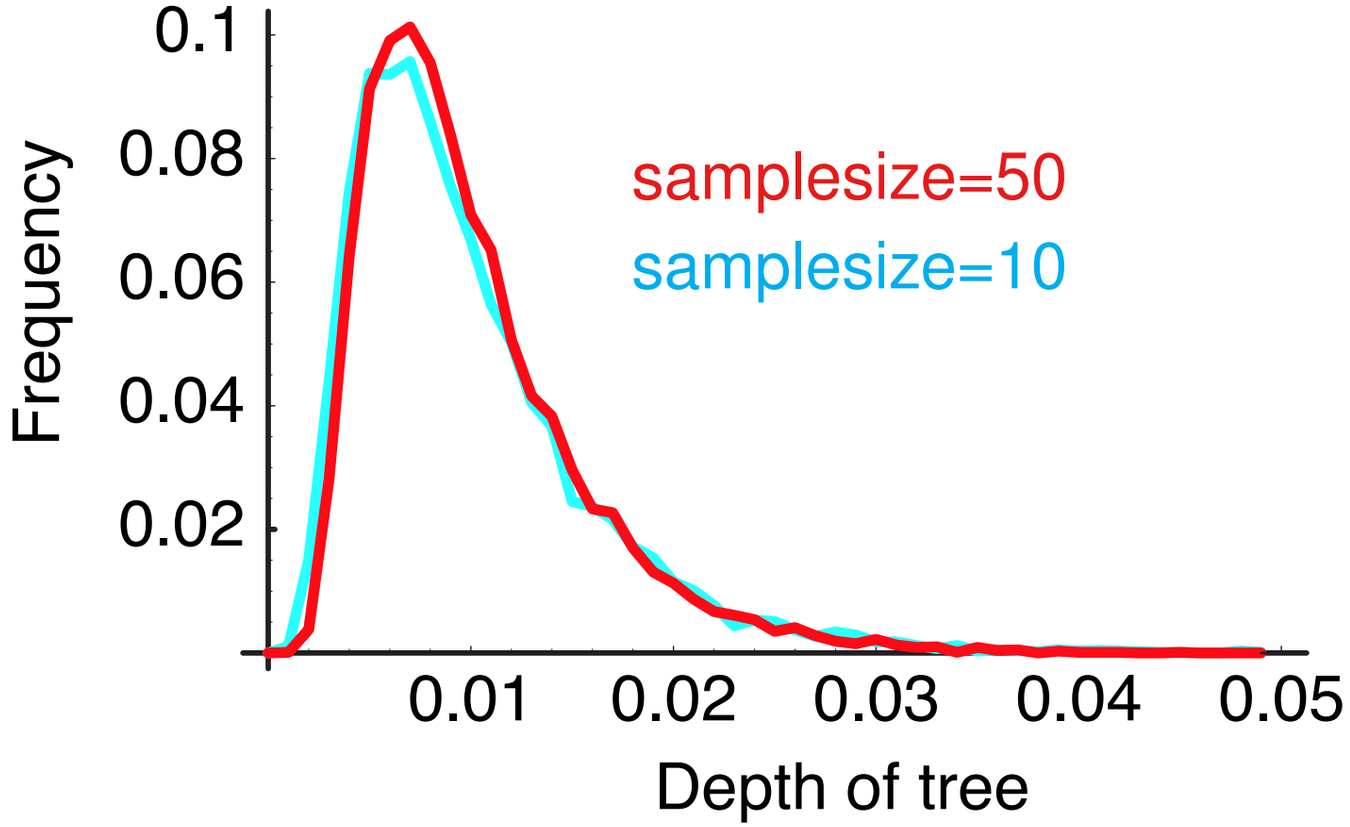
Robustness of the coalescence

Violating assumptions

- Required samples
- Recombination
- Selection

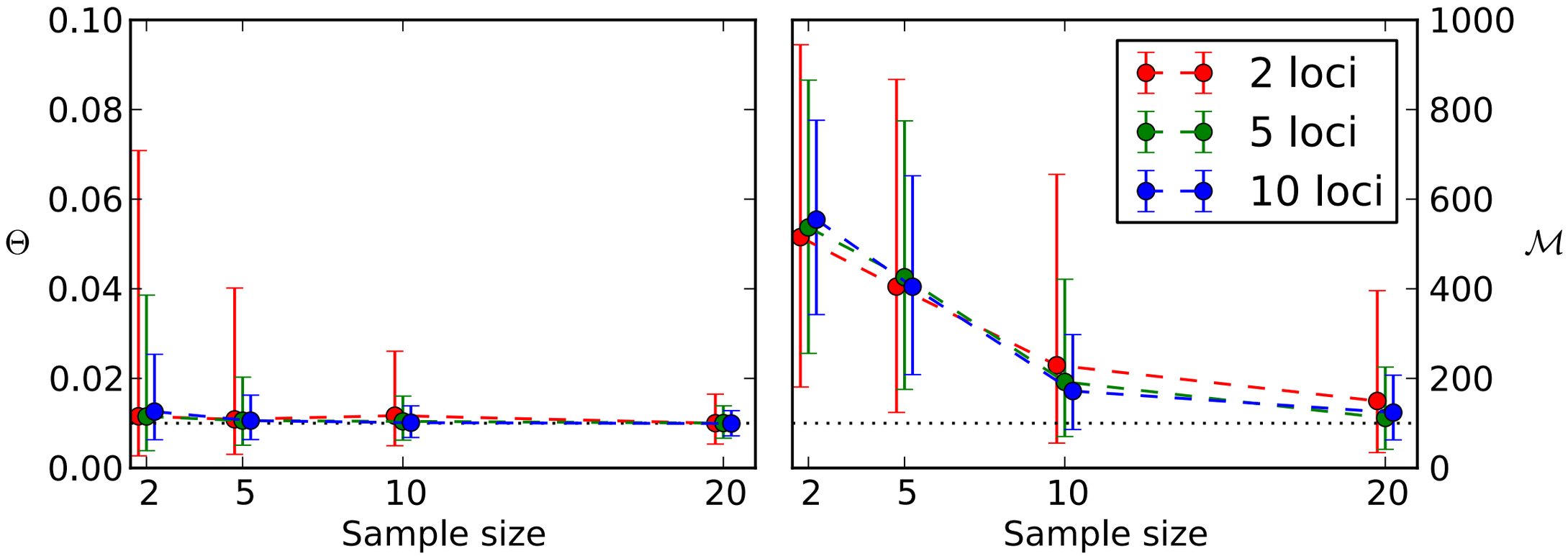
Required samples is small

- The time to the most recent common ancestor is robust to different sample sizes.
- Simulated sequence data from a single population have shown that after 8 individuals you should better add another locus than more individuals.



Felsenstein (2005)
Pluzhnikov and Donnelly (1996)

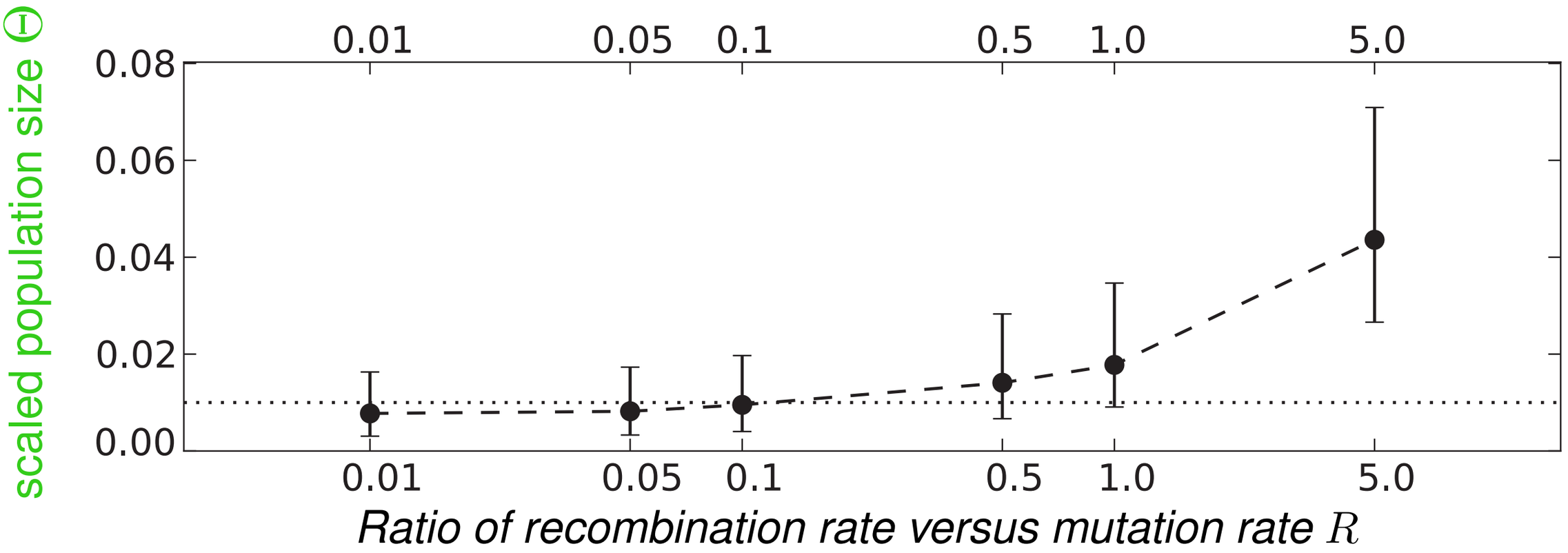
Required number of samples is small



Medium variability DNA dataset: Mutation-scaled population size Θ and mutation-scaled migration rate M versus sample size for 2, 5, and 10 loci. The true $\Theta_T = 0.01$ is marked with the dotted gray line; $M = 100$

Ignoring recombination

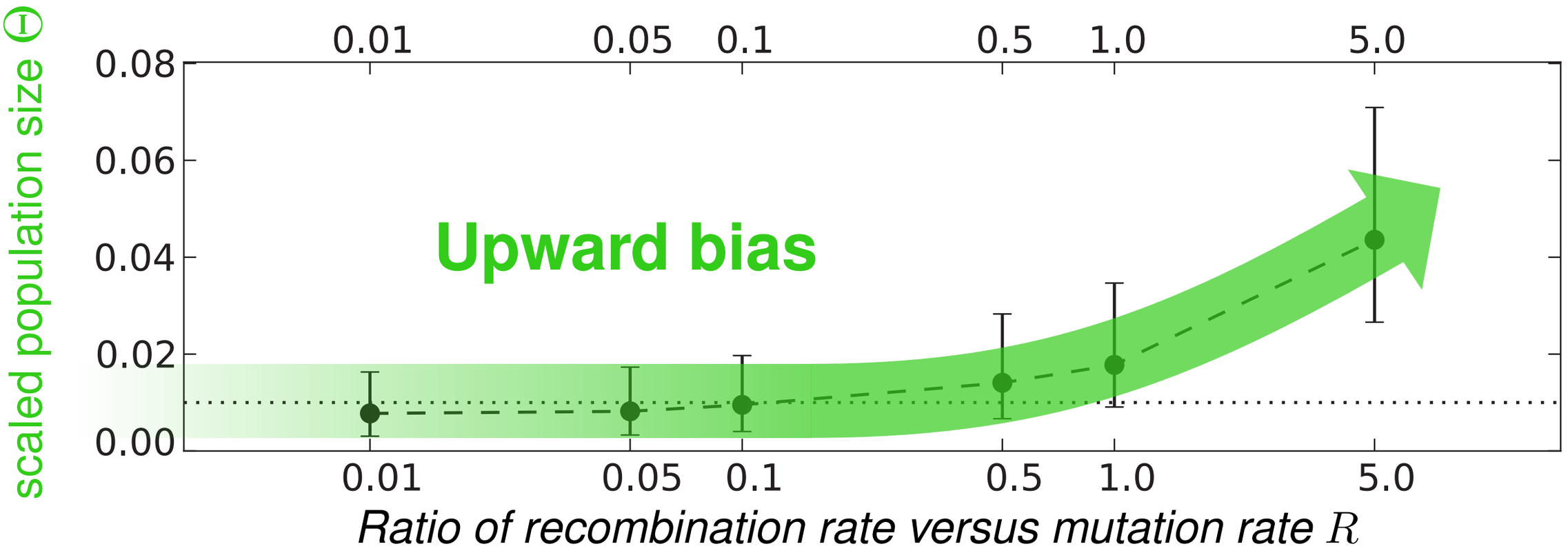
~500 simulated datasets



Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Ignoring recombination

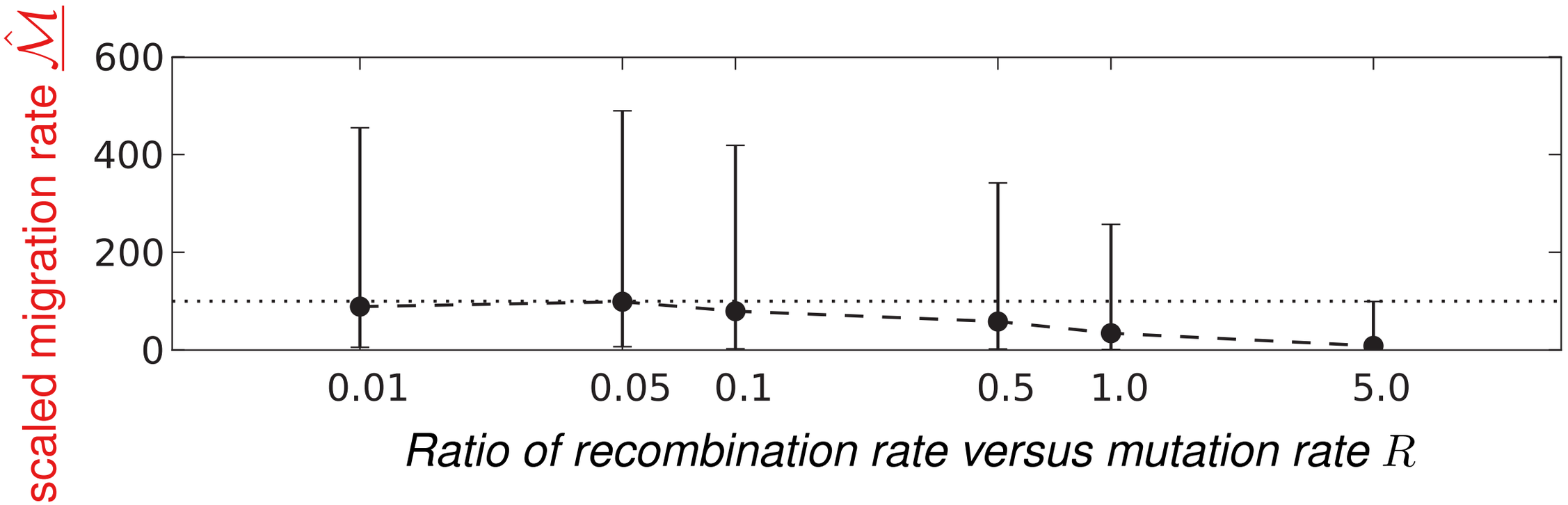
~500 simulated datasets



Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Ignoring recombination

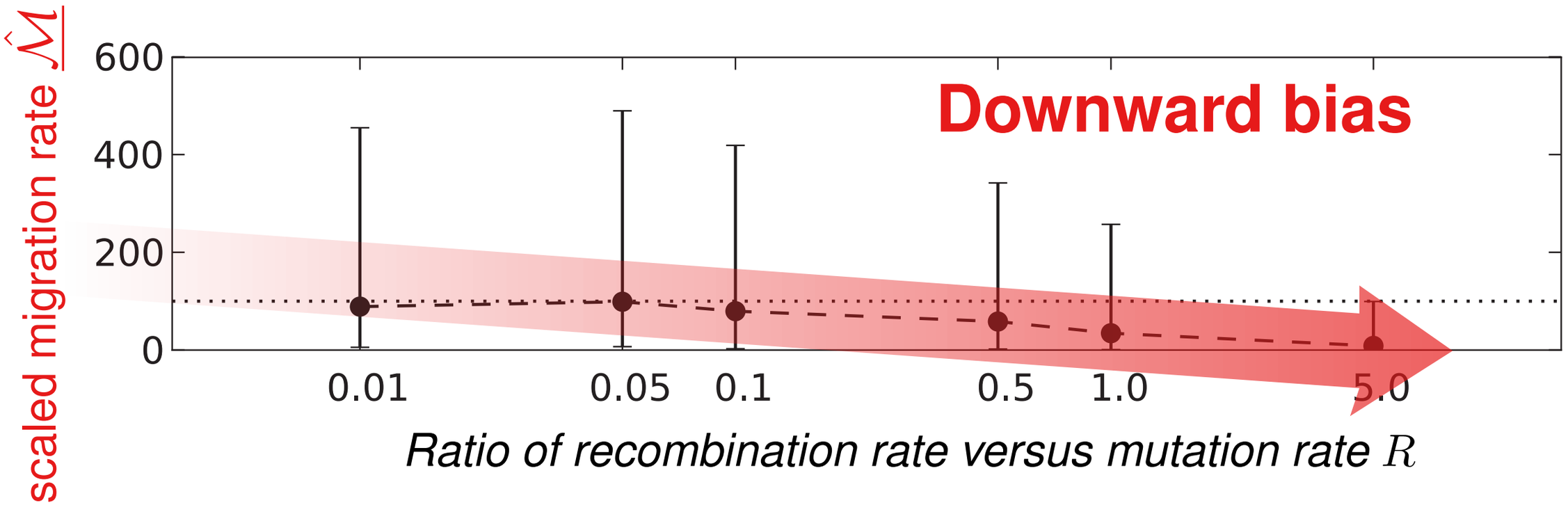
~500 simulated datasets



Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Ignoring recombination

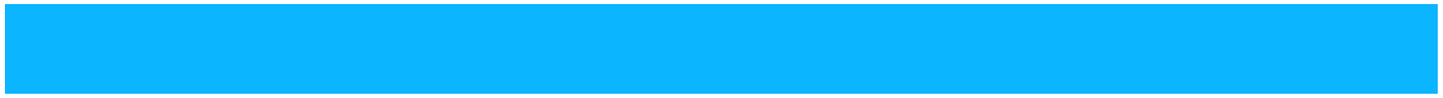
~500 simulated datasets



Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Breaking up long sequences

Calculate the log marginal likelihoods $\ln mL$ of models of interest and compare them. This is familiar to phylogeneticists who use mutation model partitions, but here they are analyzed independently.

H_0: 1 locus  $\ln mL = -1938$

H_1: 2 loci  $\ln mL = -1878$

H_2: 3 loci  $\ln mL = -1934$

Breaking up long sequences

Calculate the log marginal likelihoods $\ln mL$ of models of interest and compare them. This is familiar to phylogeneticists who use mutation model partitions, but here they are analyzed independently.

H_0 : $\ln mL = -1938$

H_{1a} : $\ln mL = -1878$

H_{1b} : $\ln mL = -1918$

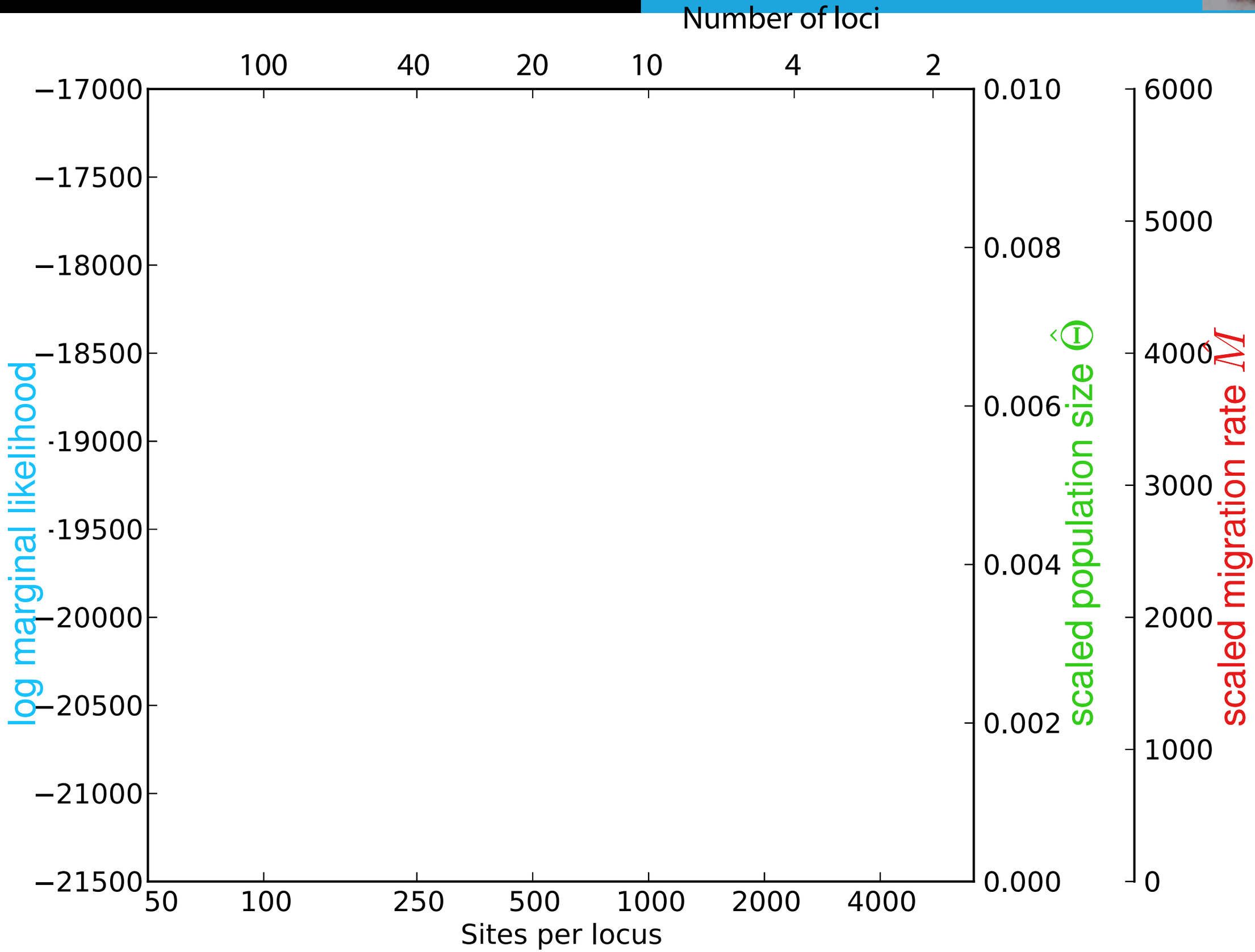
H_2 : $\ln mL = -1934$

Sorting the log marginal likelihoods: $H_{1a} > H_{1b} > H_2 > H_0$

Chopping a real dataset

D. melanogaster Chr2L

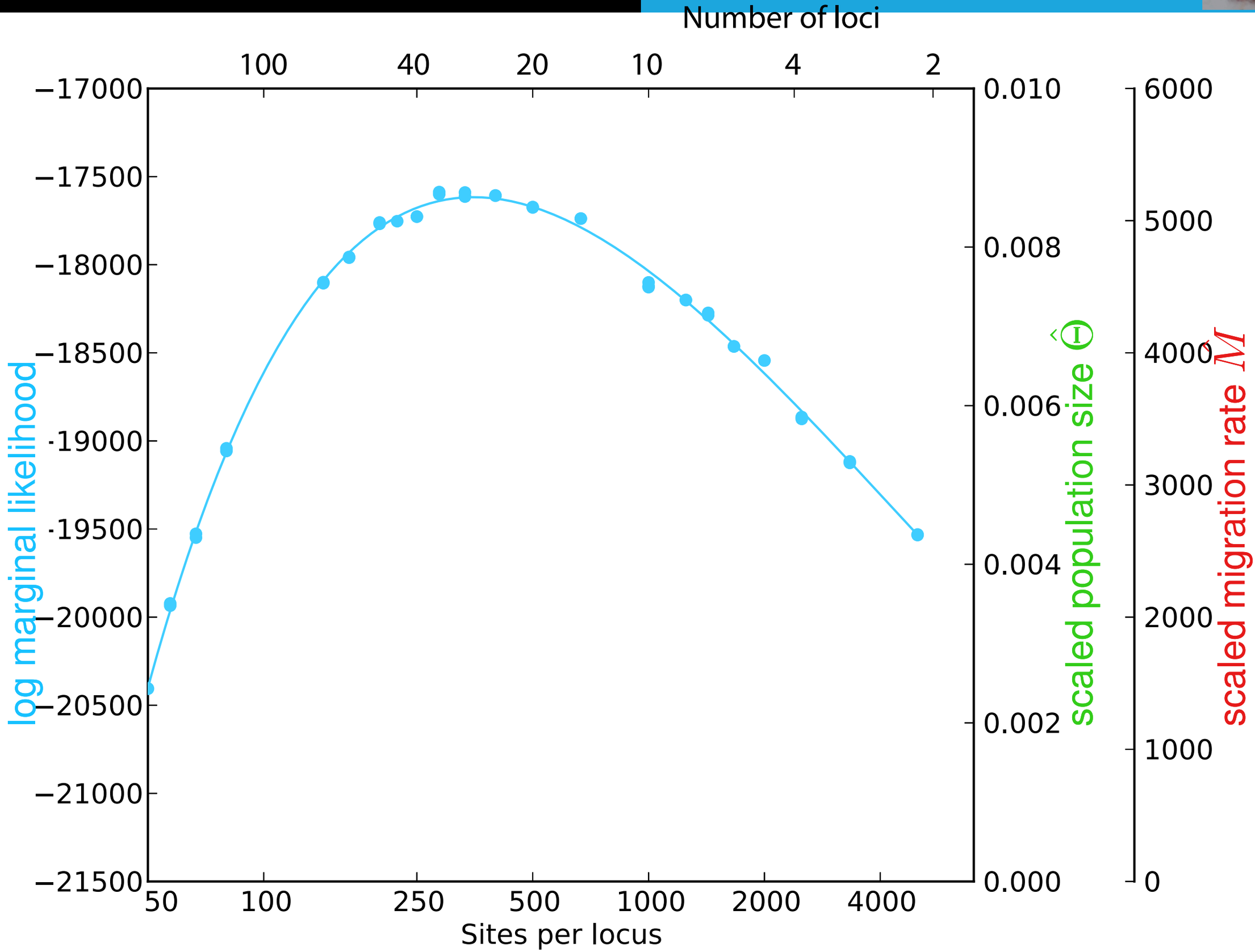
position: $5 \times 10^6 + 10,000bp$



Chopping a real dataset

D. melanogaster Chr2L

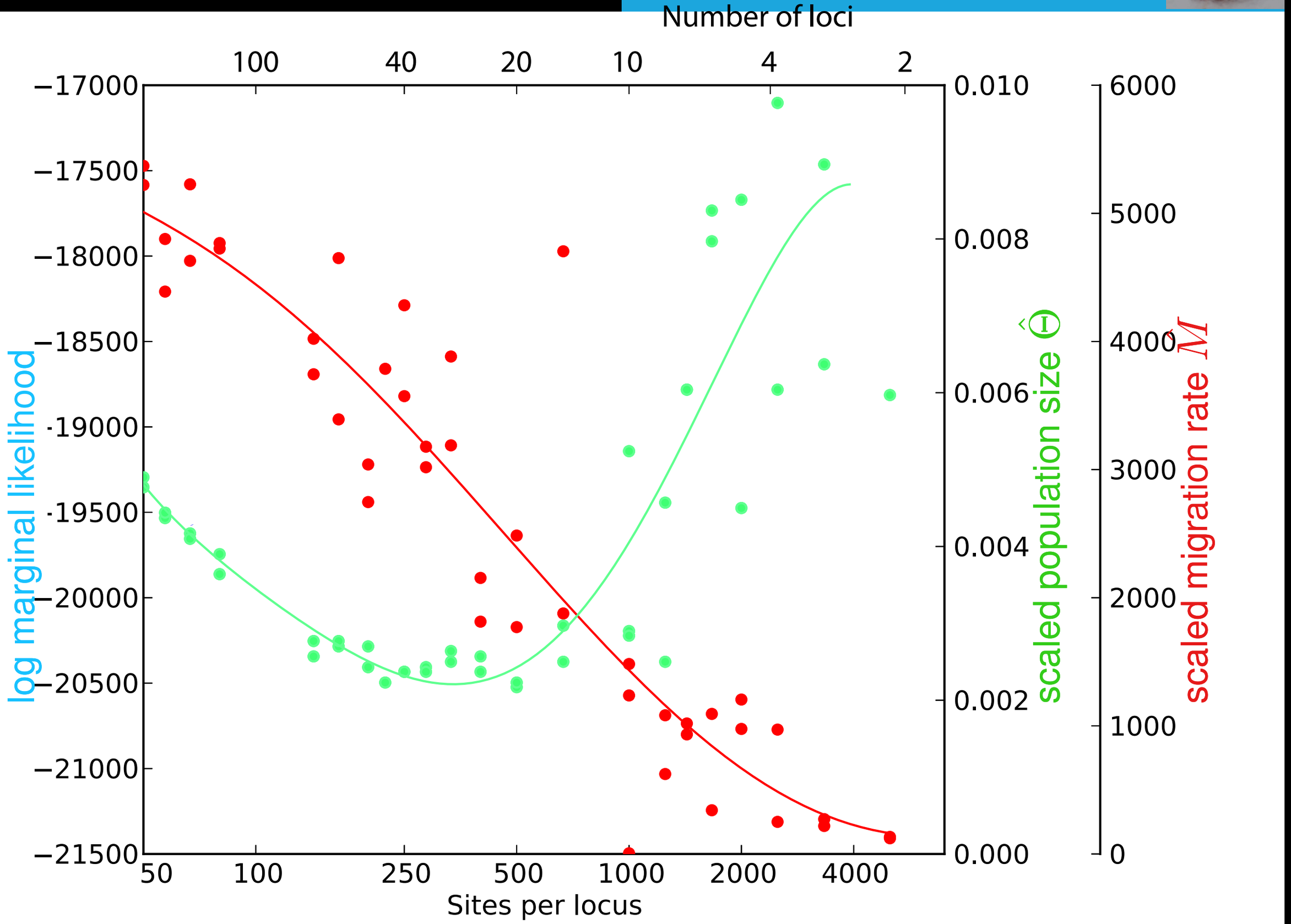
position: $5 \times 10^6 + 10,000bp$



Chopping a real dataset

D. melanogaster Chr2L

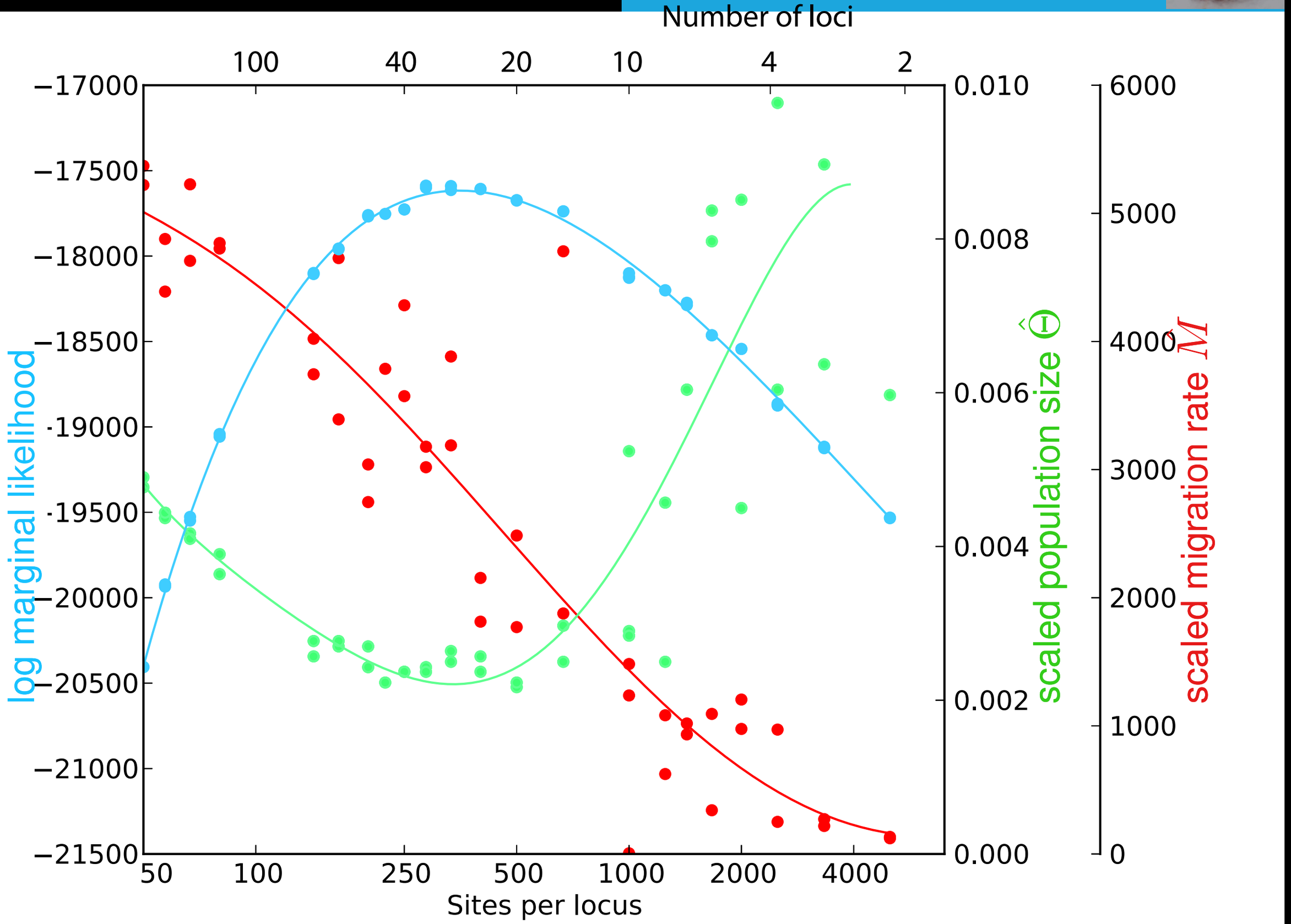
position: $5 \times 10^6 + 10,000bp$



Chopping a real dataset

D. melanogaster Chr2L

position: $5 \times 10^6 + 10,000bp$



Ignored selection

The standard coalescent assumes neutral mutations and also exchangeable number of offspring, loci under selection will violate both tenets. In the allele frequency spectrum literature recently there is a strong push on looking at signals of selection, which seems still very difficult in 'traditional' coalescence approaches.

- A new mutation that has a positive effect will replace some of the variability present in the population. All linked sites will suffer a drop in effective population size.

- A new mutation that has a negative effect and will be most likely removed , also resulting in a reduction of variability (and population size)

This is used in genome-wide selection scans, but influence of population growth, population structure on such estimates are not well studied.

Outlook

- We will have a lab tonight where you will differentiate between 8 simple population models that include ”speciation” (or population splitting) with and without migration using a data set of complete genomes of Zika viruses.

- (On the http://popgen.sc.fsu.edu website, check out “Bayes factors” and “Parallel migrate”, there is also a Google support group to look up answers, ask questions and receive answers [mostly by me])