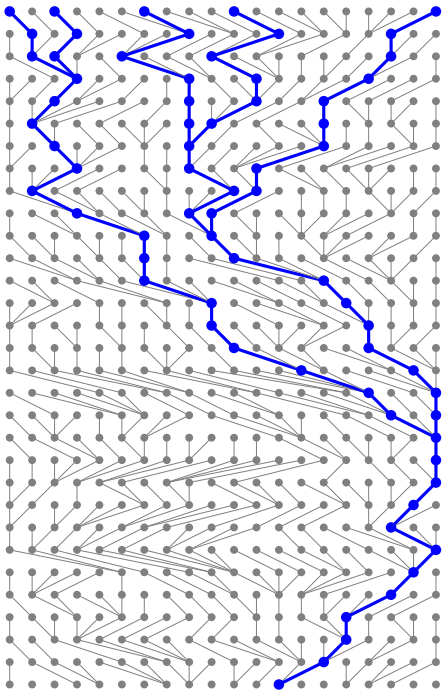
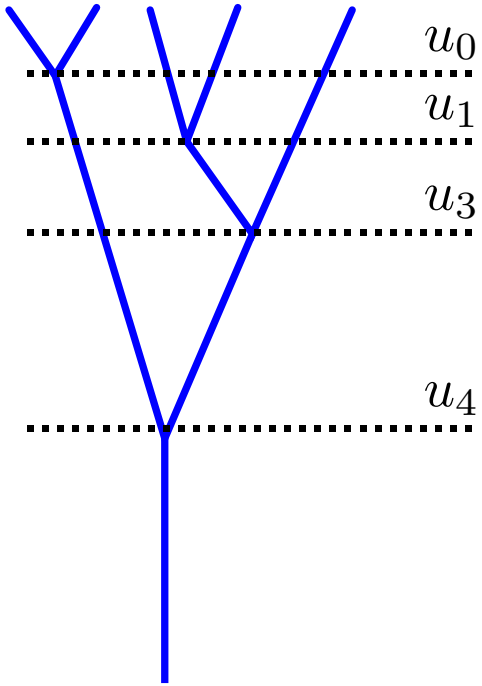
Extension of the basic coalescent

Peter Beerli
Florida State University

#EPONGE2019 Arequipa

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

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

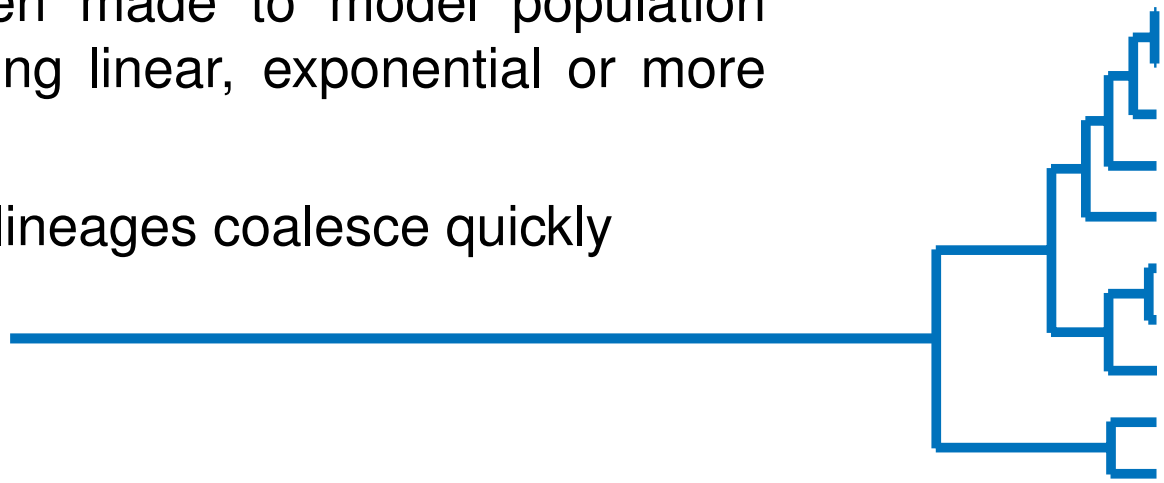
- Population growth (two parameters), fluctuations, bottlenecks
- Migration among populations (potentially thousands, parameters)
- Population splitting (many parameters)
- Recombination (parameters)
- Effect of assumption violation

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.

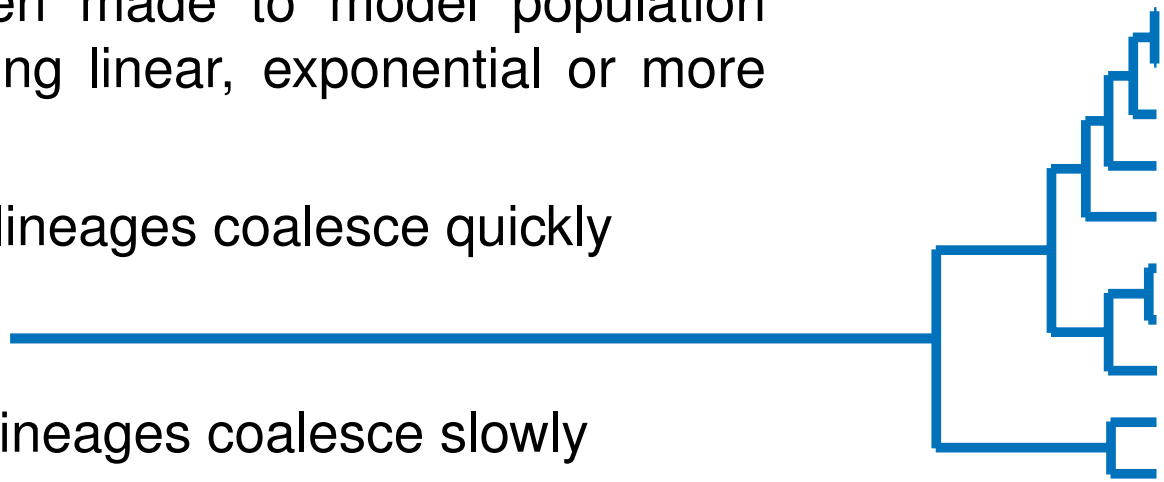
- In a small population lineages coalesce quickly



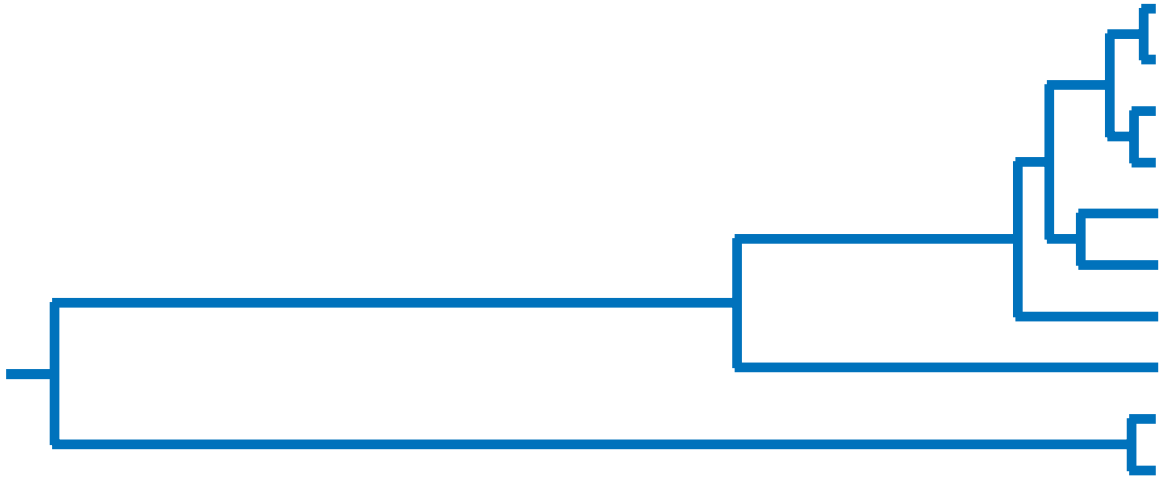
This leaves a signature in the data. We can exploit this and estimate the population growth rate g jointly with the current population size Θ.

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.

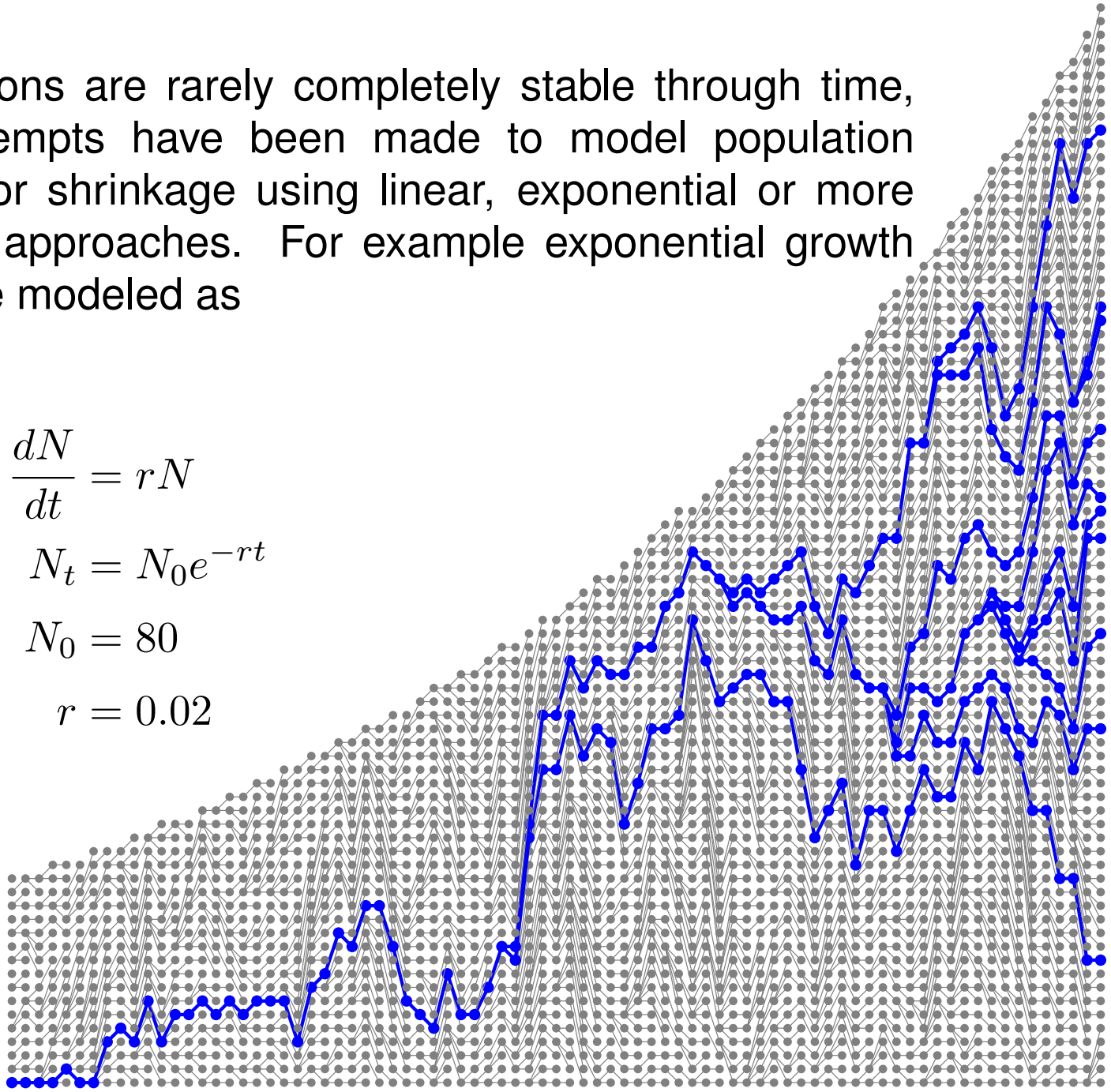
Populations are rarely completely stable through time, and attempts have been made to model population growth or shrinkage using linear, exponential or more general approaches. For example exponential growth could be modeled as

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

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

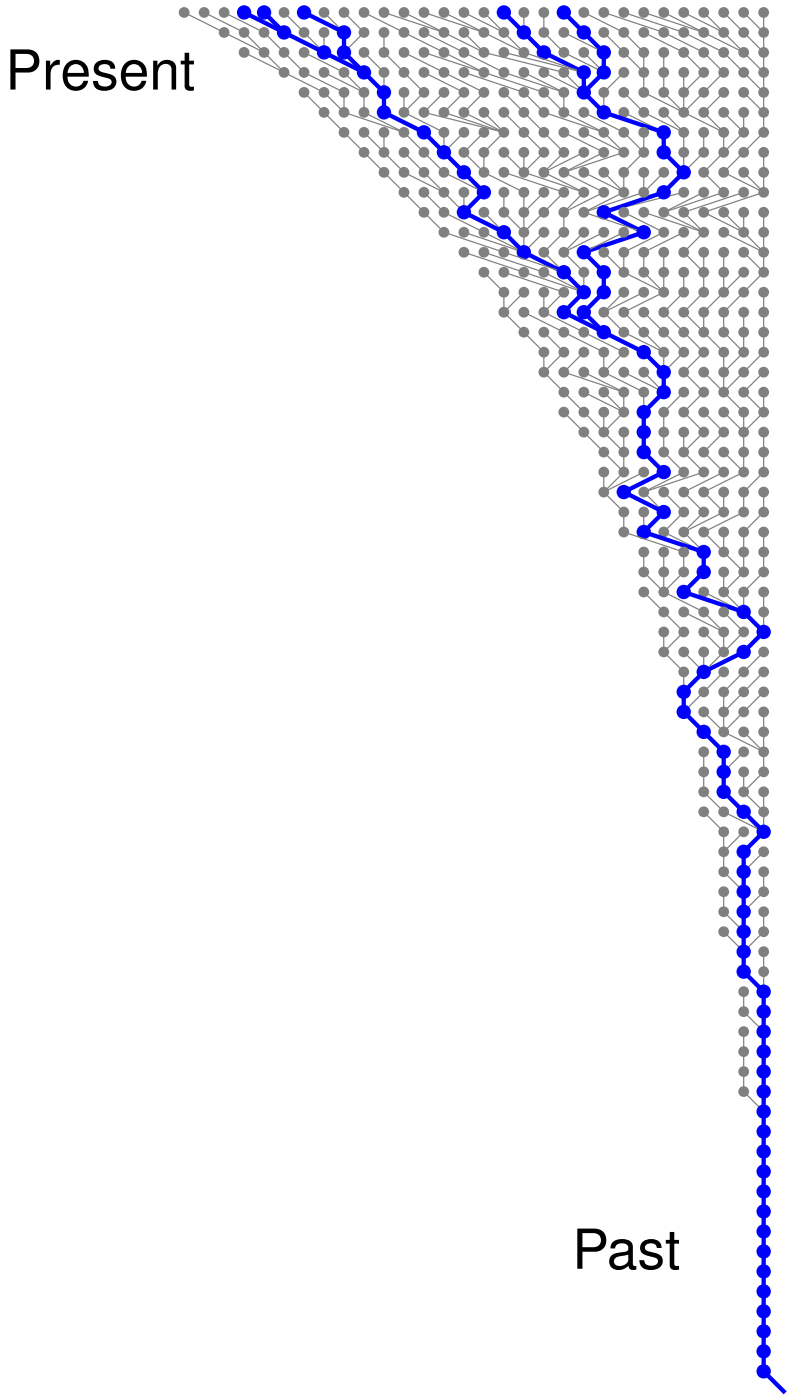
$$N_0 = 80$$

$$r = 0.02$$



Past

Present



For constant population size we found

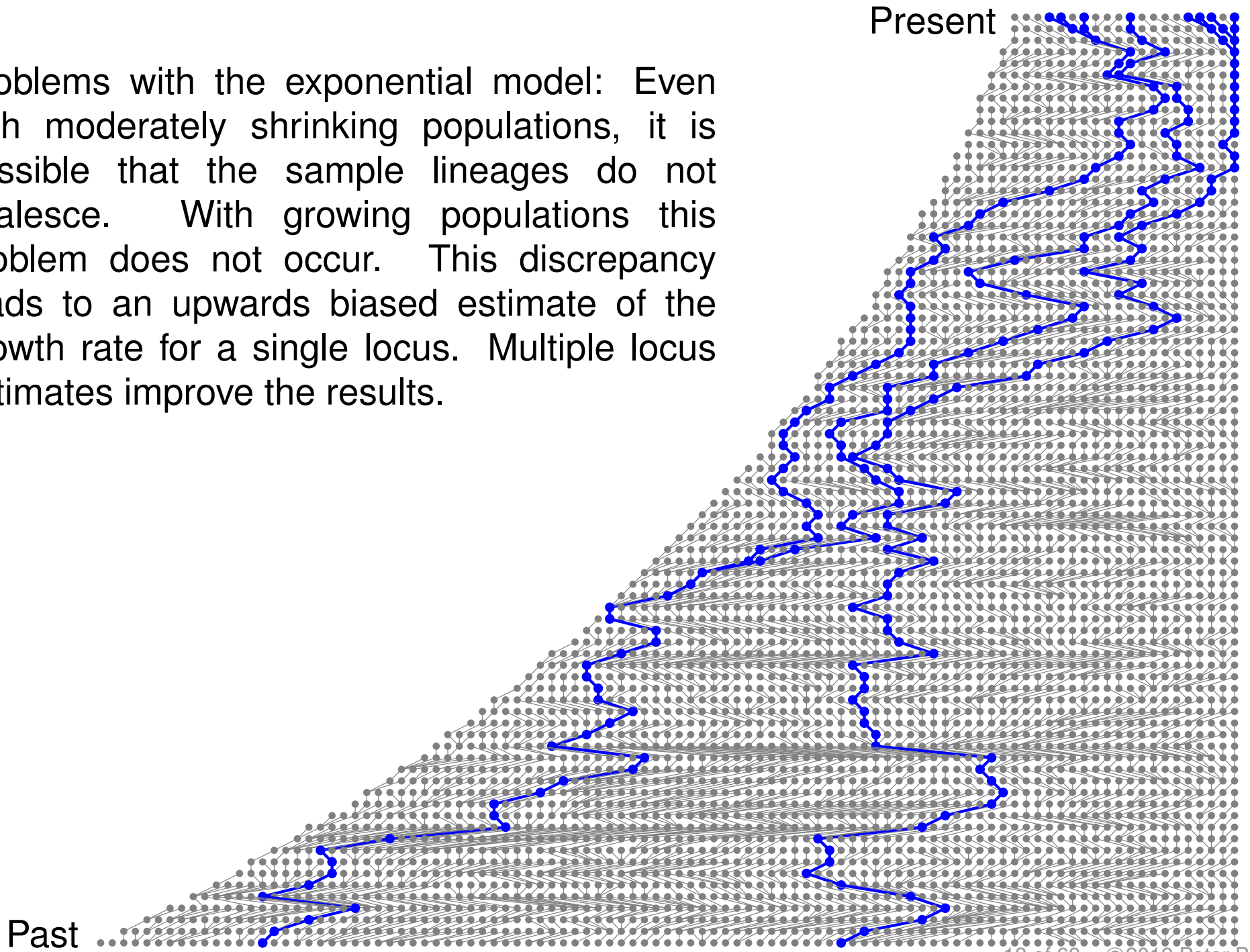
$$\mathrm{p}(\mathrm{G}|\Theta) = \prod_{\mathrm{j}} \mathrm{e}^{-\mathrm{u_j}\frac{\mathrm{k(k-1)}}{\Theta}} \frac{2}{\Theta}$$

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

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

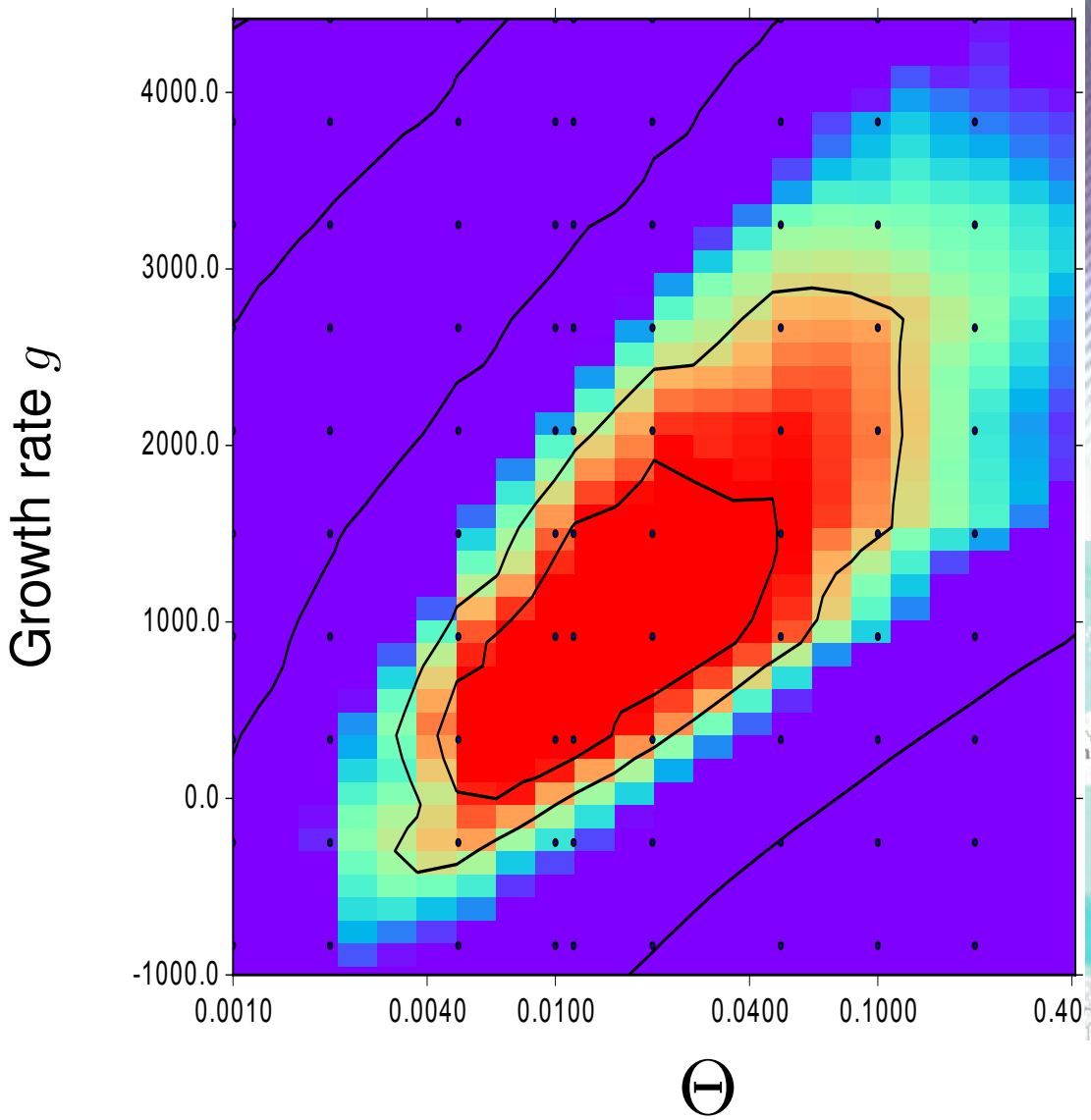
Extensions of the basic coalescent

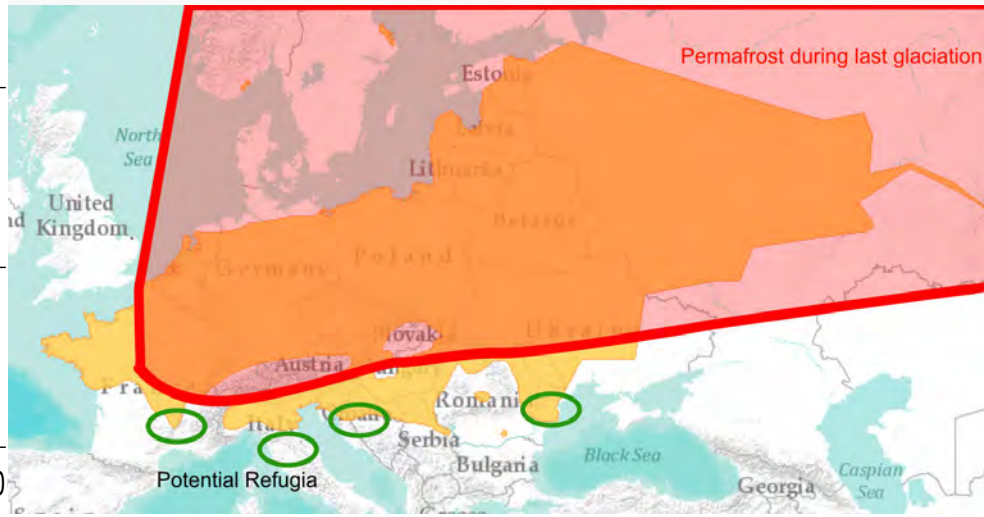
Problems with the exponential model: Even with moderately shrinking populations, it is possible that the sample lineages do not coalesce. With growing populations this problem does not occur. This discrepancy leads to an upwards biased estimate of the growth rate for a single locus. Multiple locus estimates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



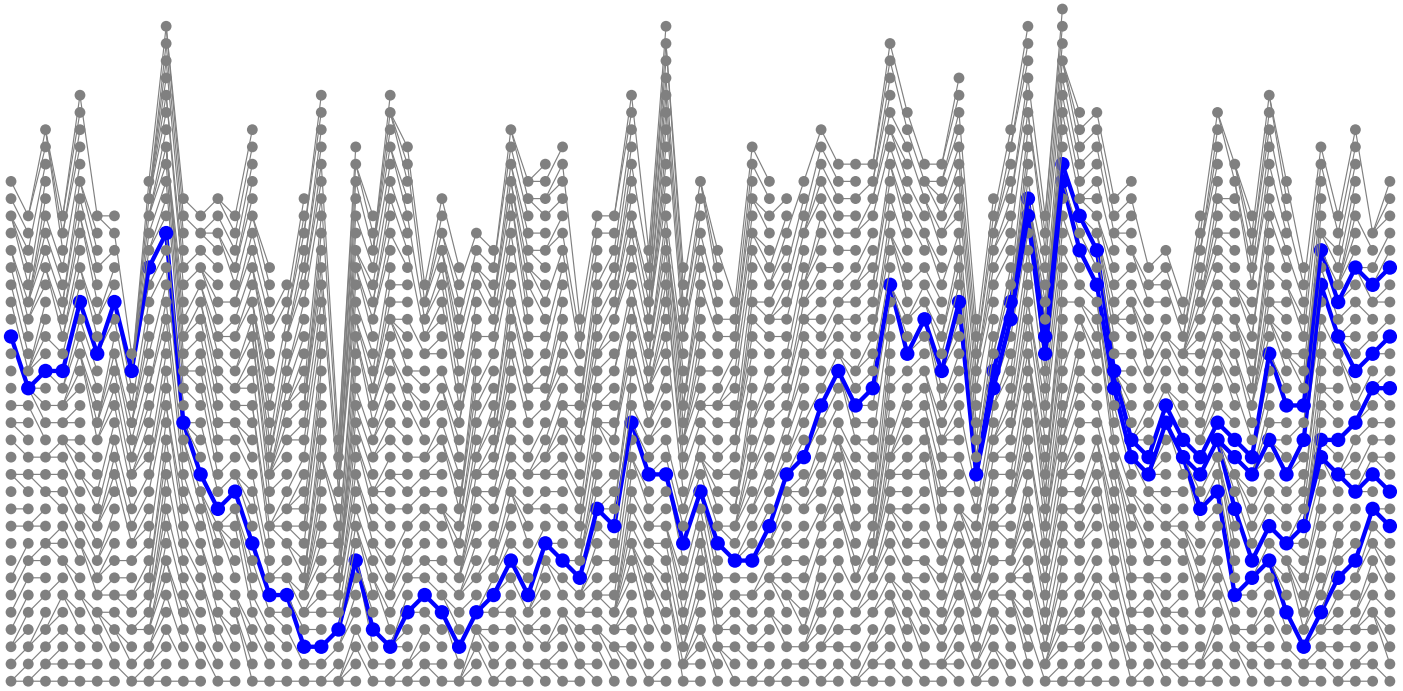
Expansion of *Pelophylax lessonae* in Europe

Expansion of *Pelophylax lessonae* in Europe

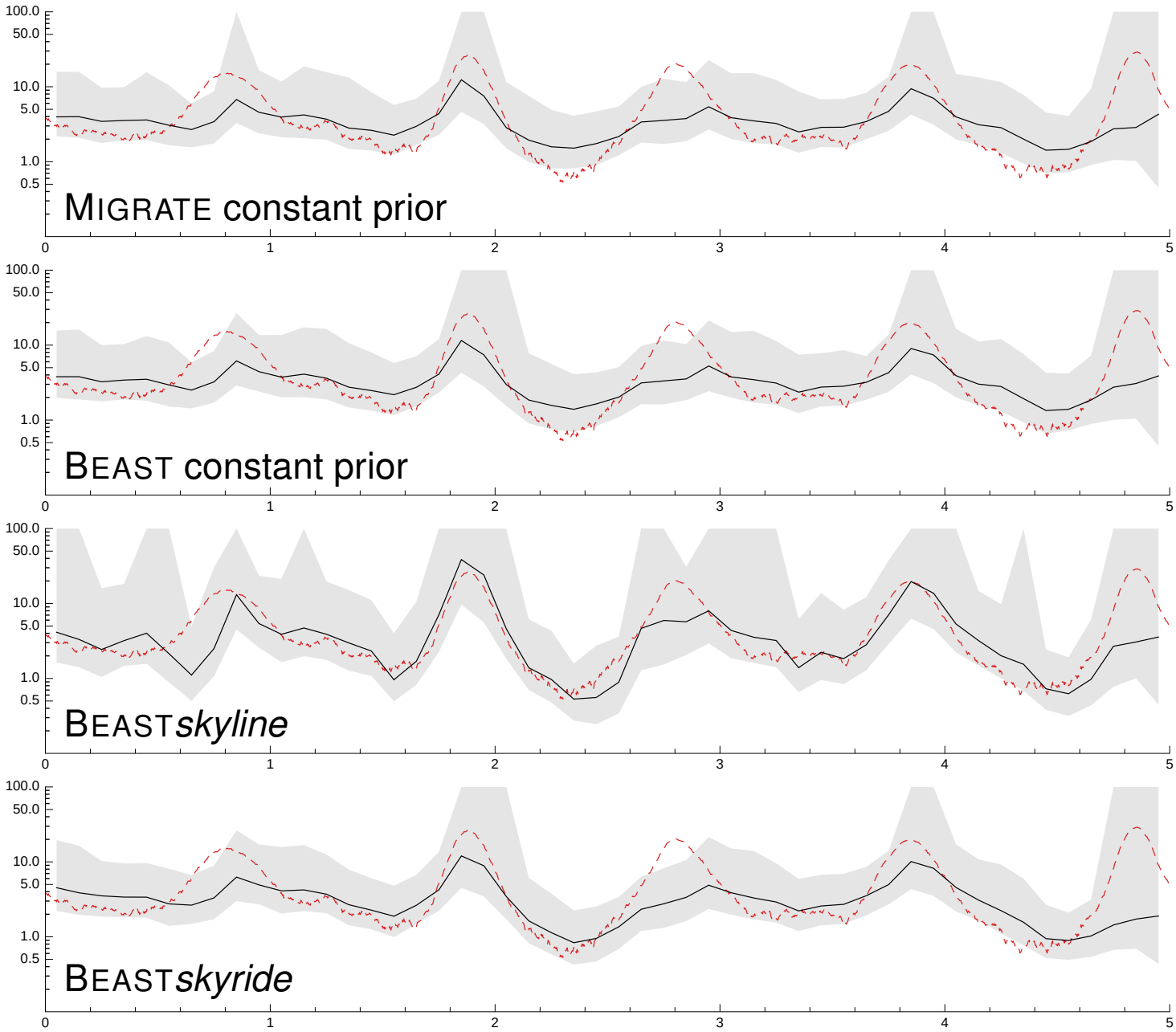




Random fluctuations of the population size are most often ignored. BEAST (and to some extent MIGRATE) can handle such scenarios. BEAST is using a full parametric approach (skyride, skyline) whereas MIGRATE uses a non-parametric approach for its skyline plots that has the tendency to smooth the fluctuations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



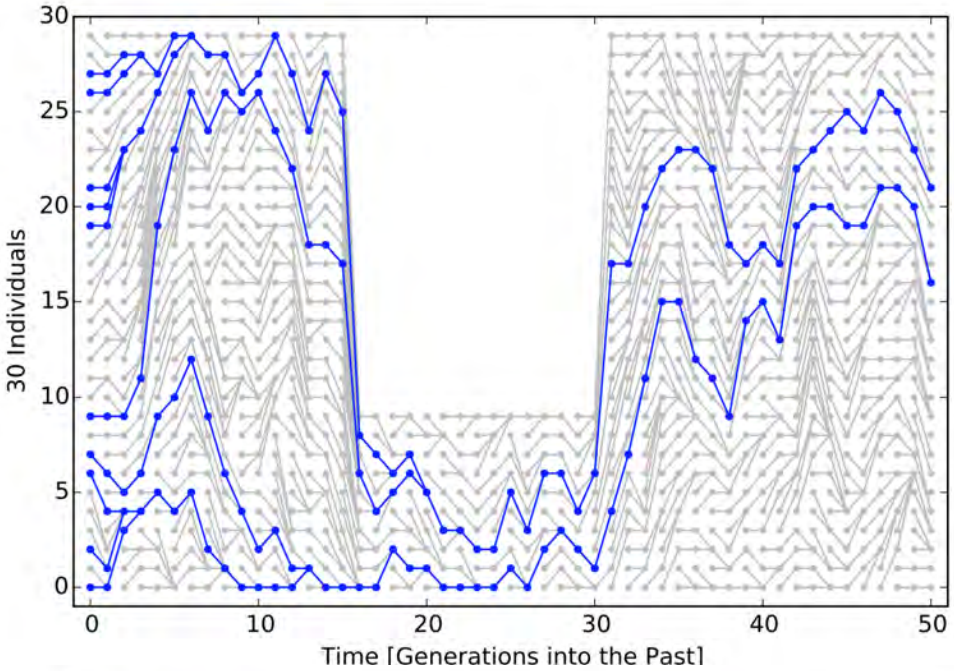
Extensions of the basic coalescent

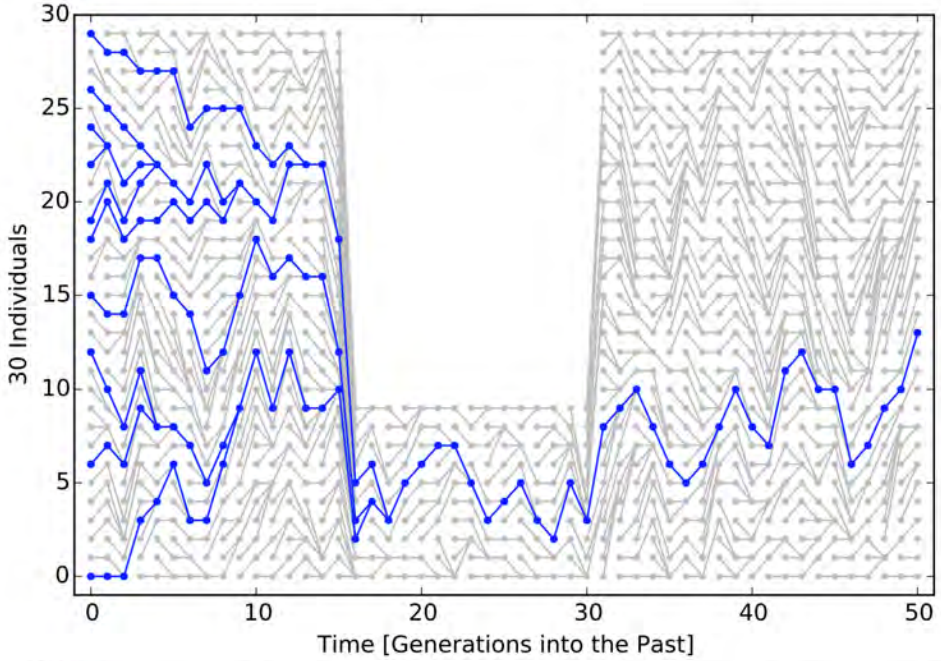


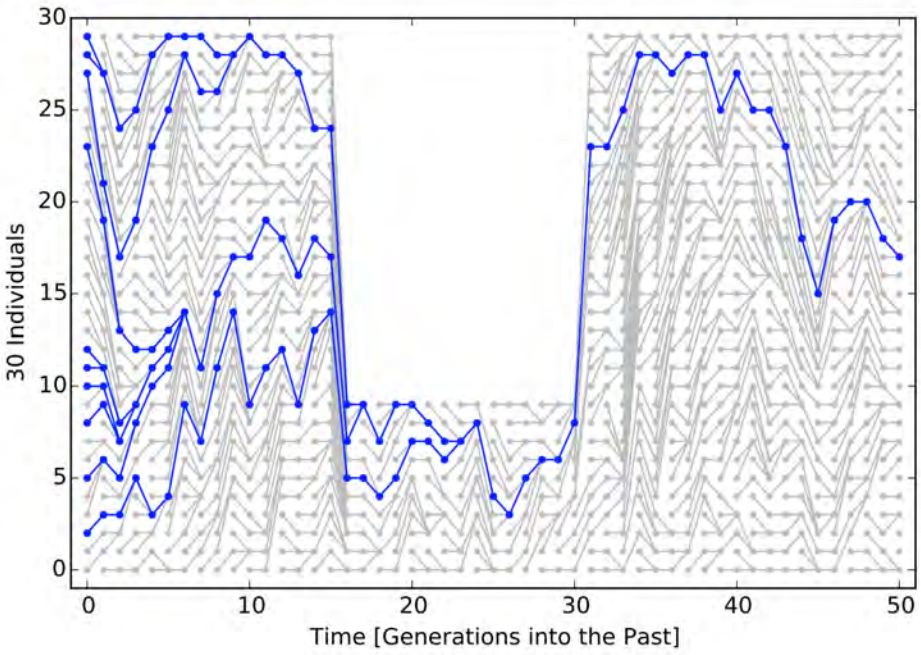
Comparison of the skyline plots of simulated influenza dynamics analyzed by MIGRATE and BEAST. The x-axis is the time in years and the y-axis is effective population size. The data are sequences from 250 individuals sampled at regular intervals over 5 years. The dashed curve is the actual population 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.

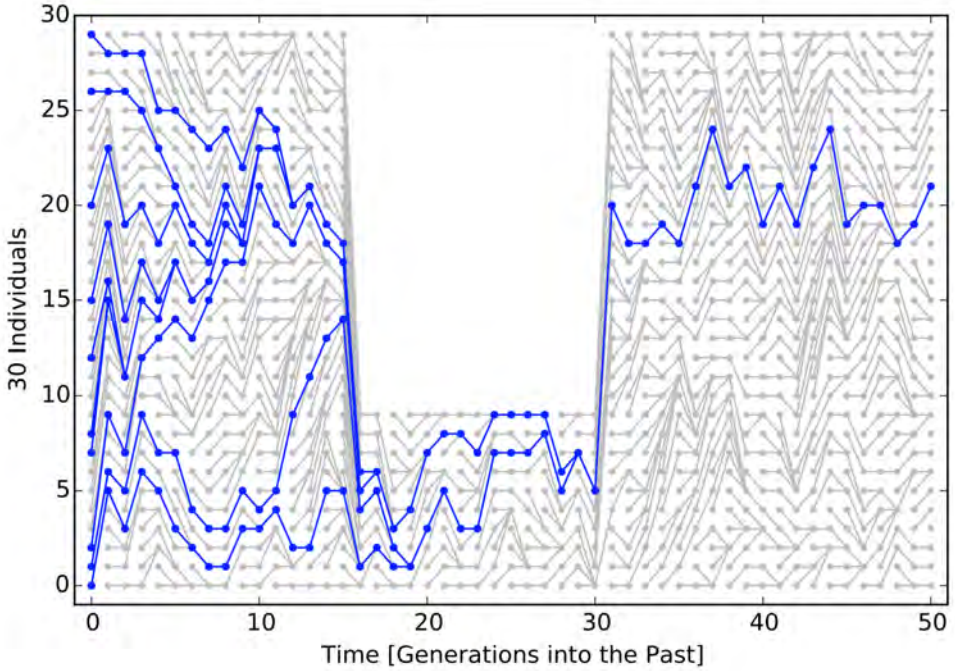
Extensions of the basic coalescent

Bottlenecks



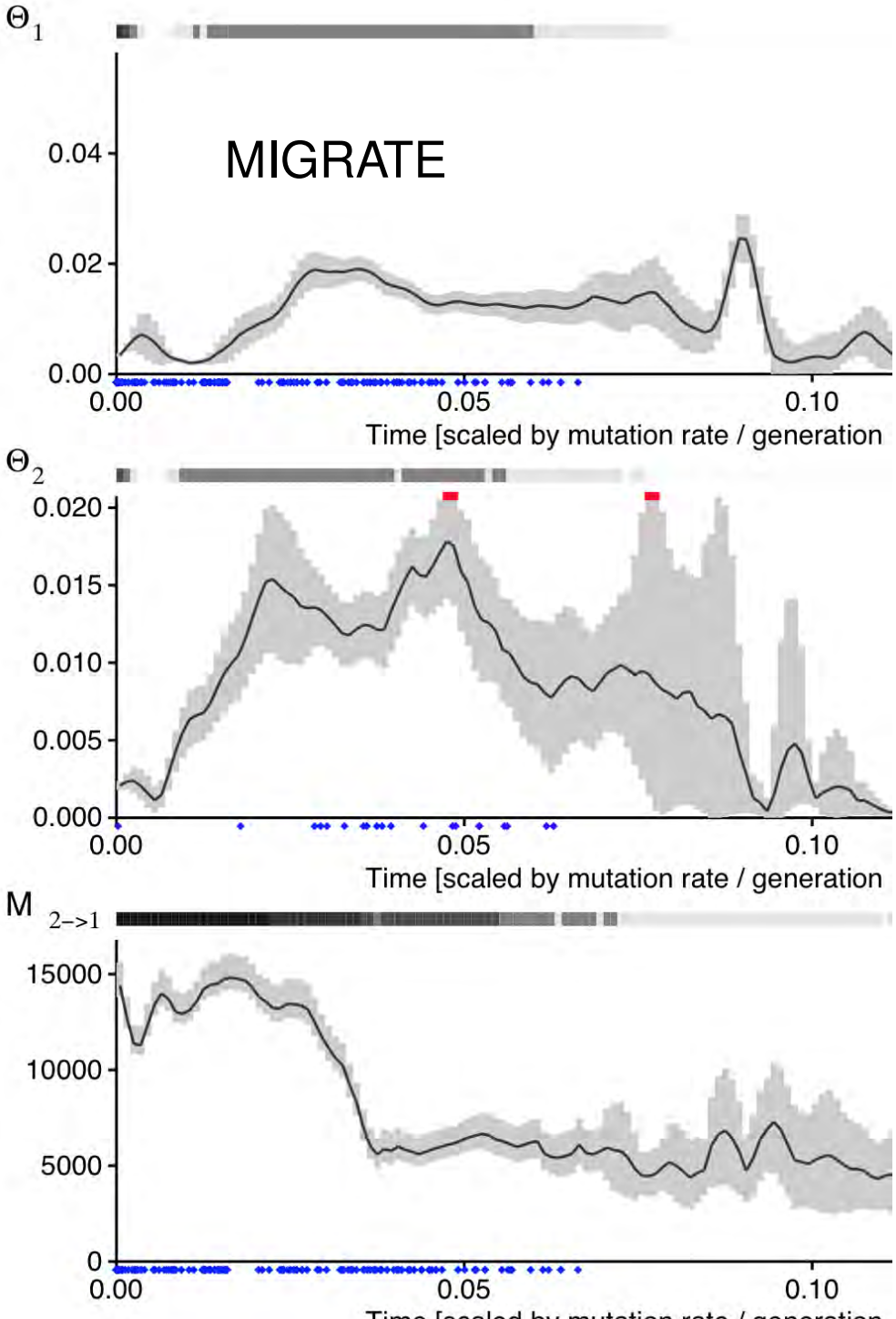


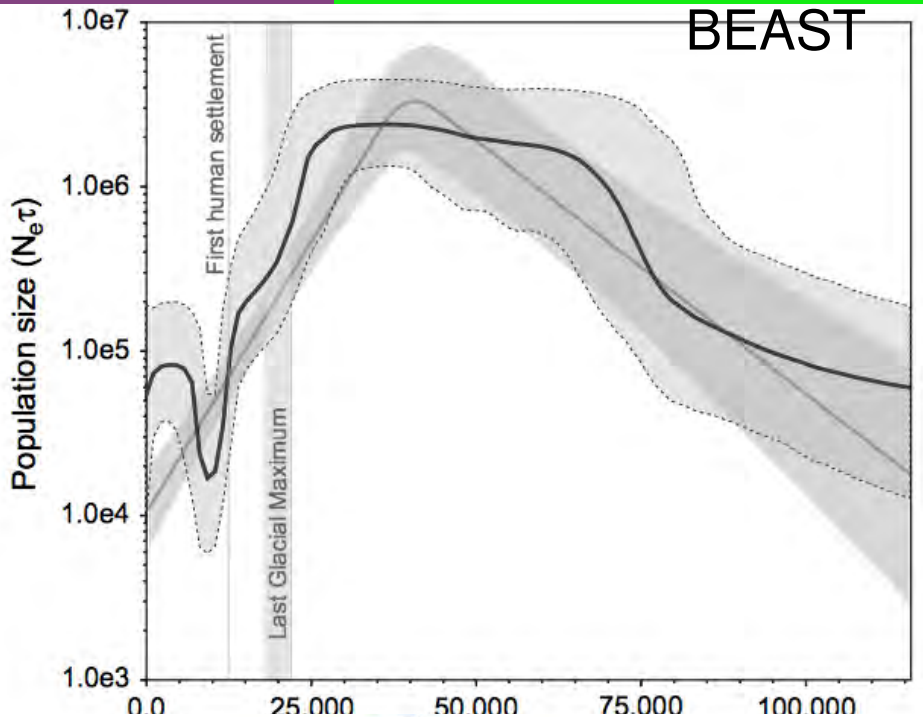




Extensions of the basic coalescent

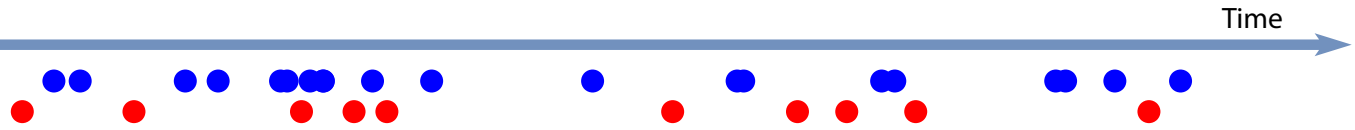
Skyline plots

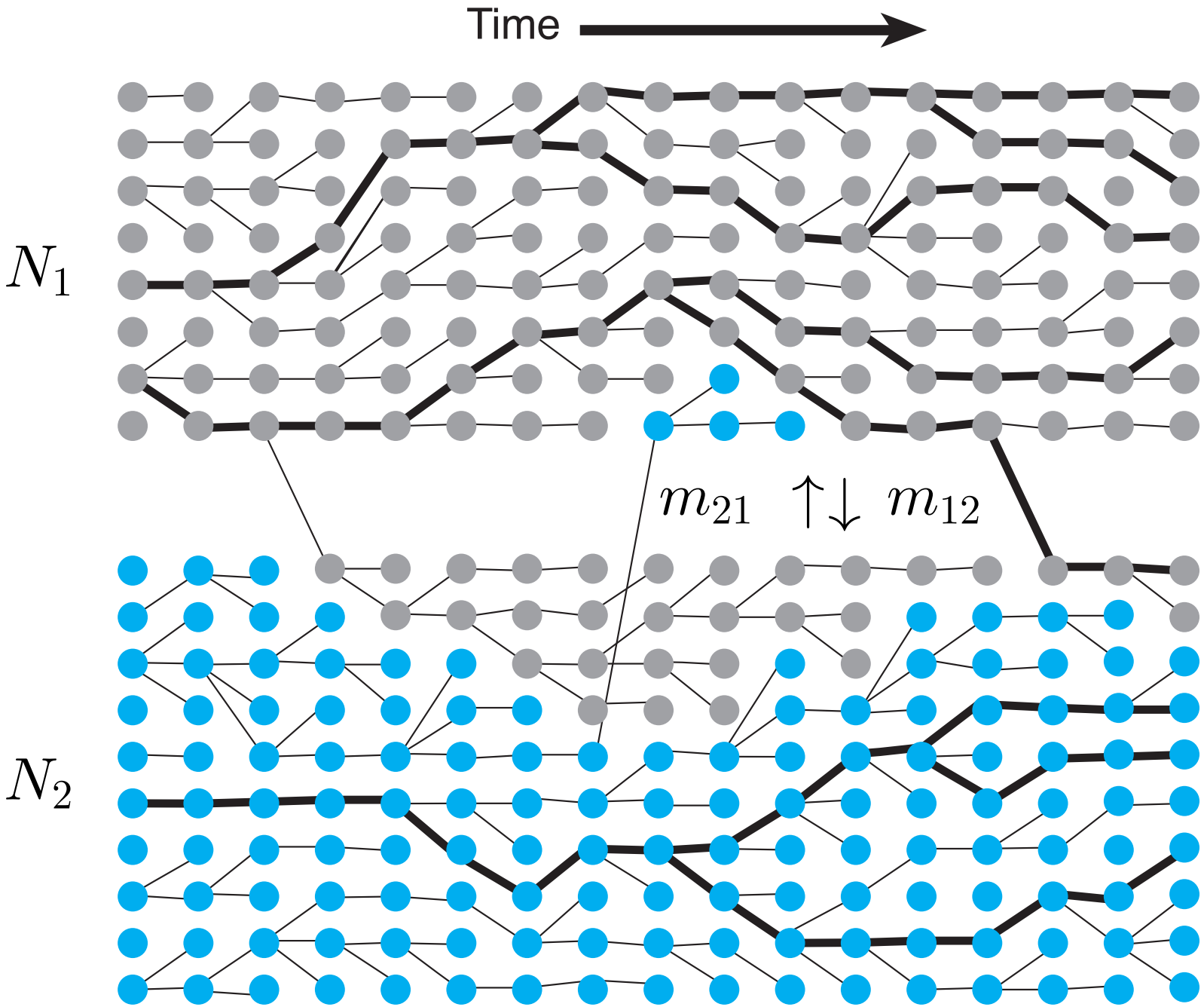


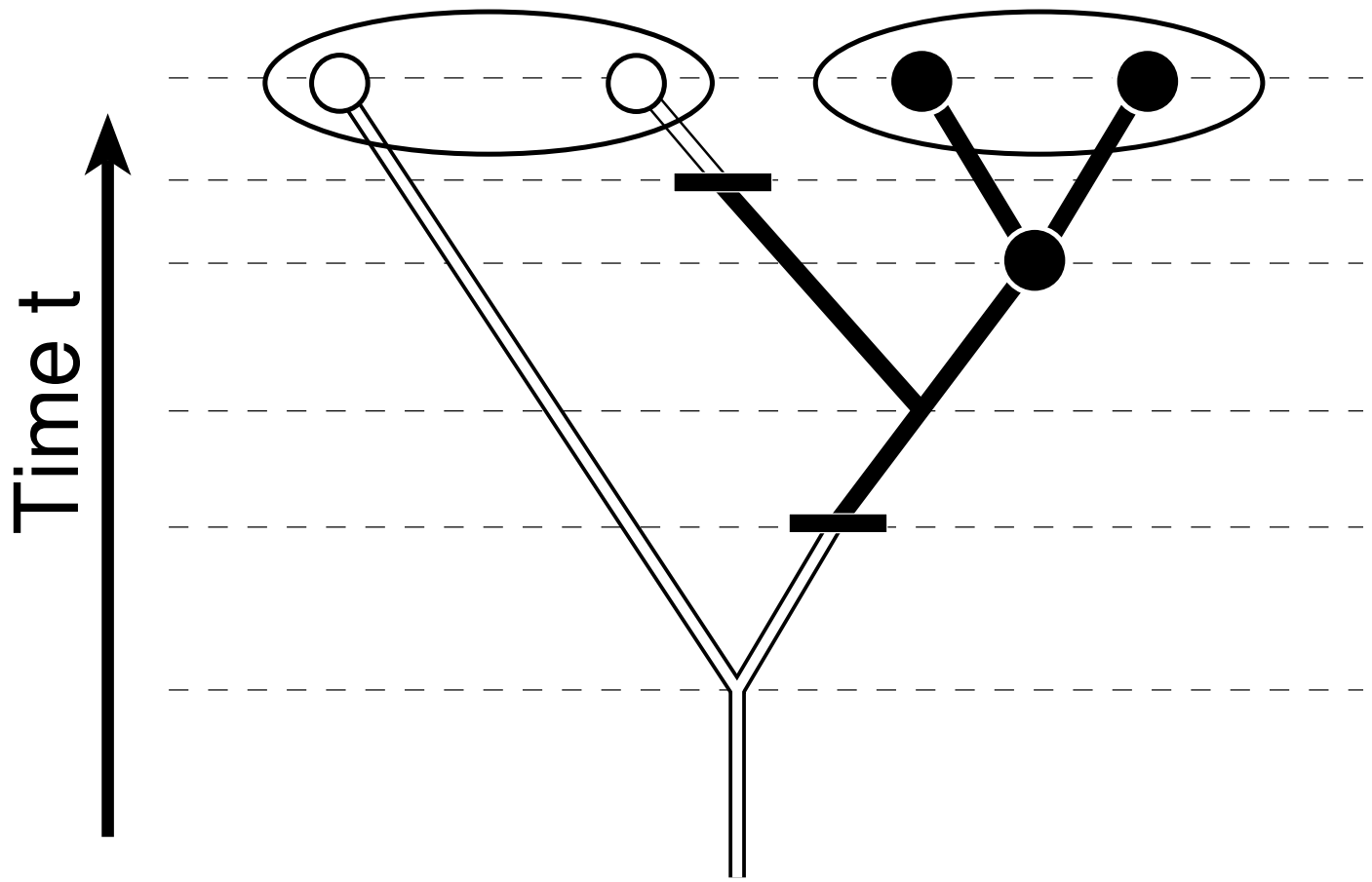


Accommodating more events

An analogy




Time
N_1
m_{21} ↑↓ m_{12}
N_2



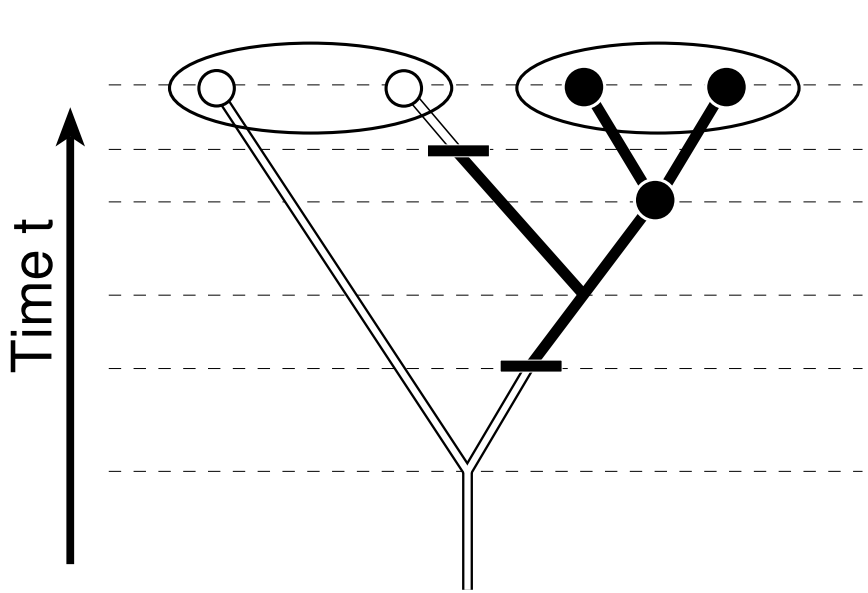
	Δt	k_1	k_2
u_1		2	2
u_2		1	3
u_3		1	2
u_4		1	3
u_5		2	0

The single population coalescence rate is

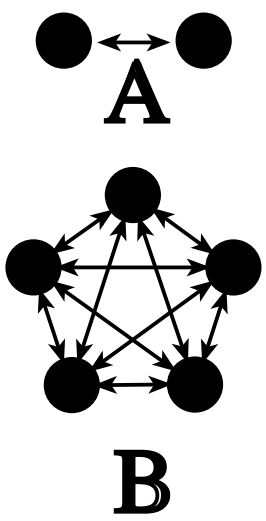
$$\frac{k(k-1)}{4N}.$$

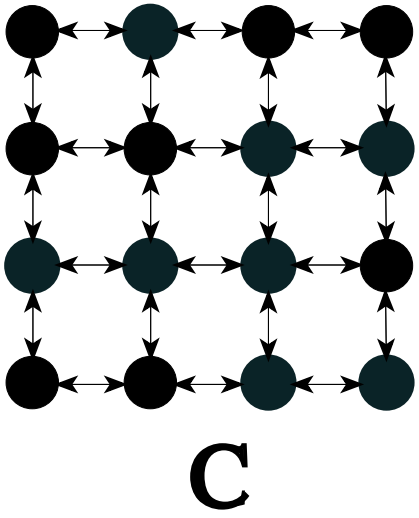
Changes for two populations to

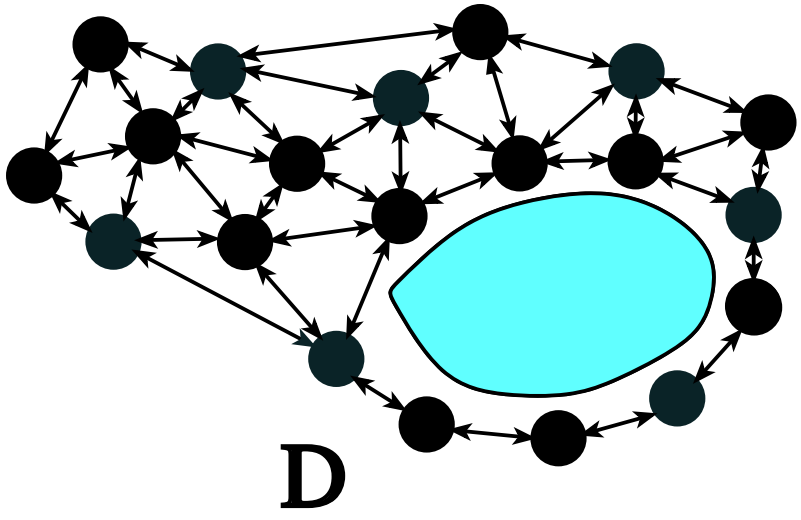
$$\frac{k_1(k_1-1)}{\Theta_1} + \frac{k_2(k_2-1)}{\Theta_2} + k_1 M_{2,1} + k_2 M_{1,2}$$

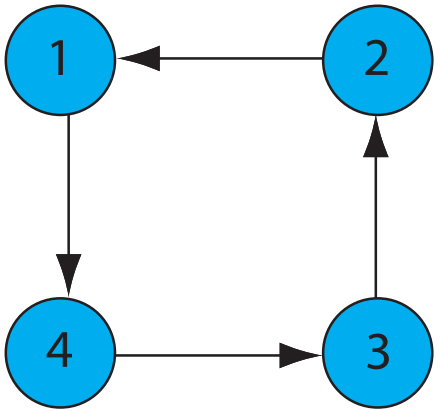


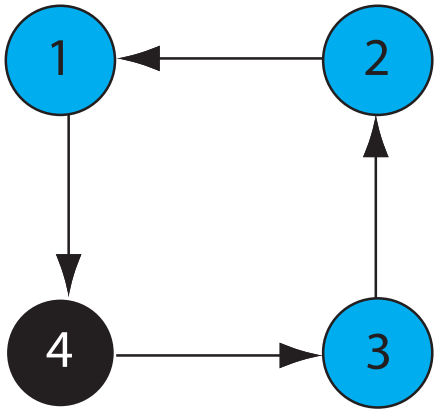
Δt	k_1	k_2
u_1	2	2
u_2	1	3
u_3	1	2
u_4	1	3
u_5	2	0

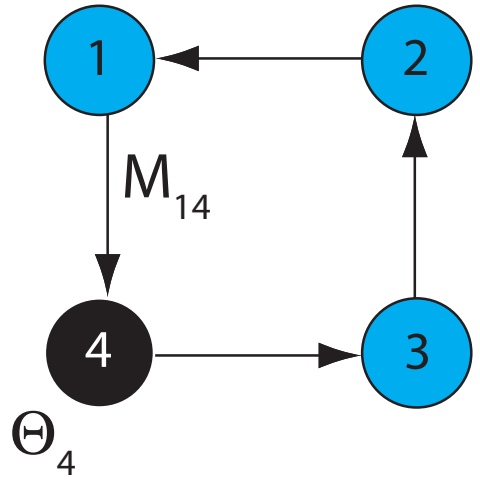

A
B


C

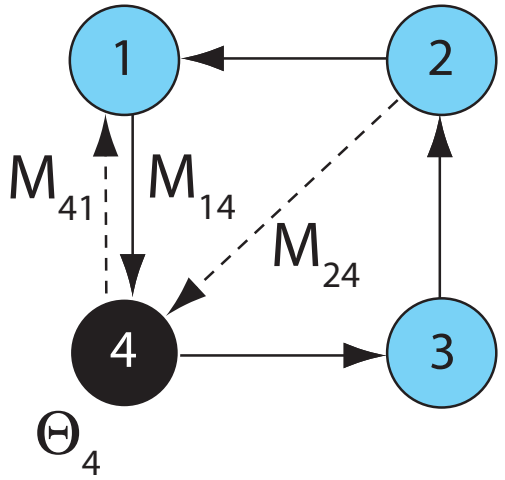

D


1
2
4
3


1
2
4
3

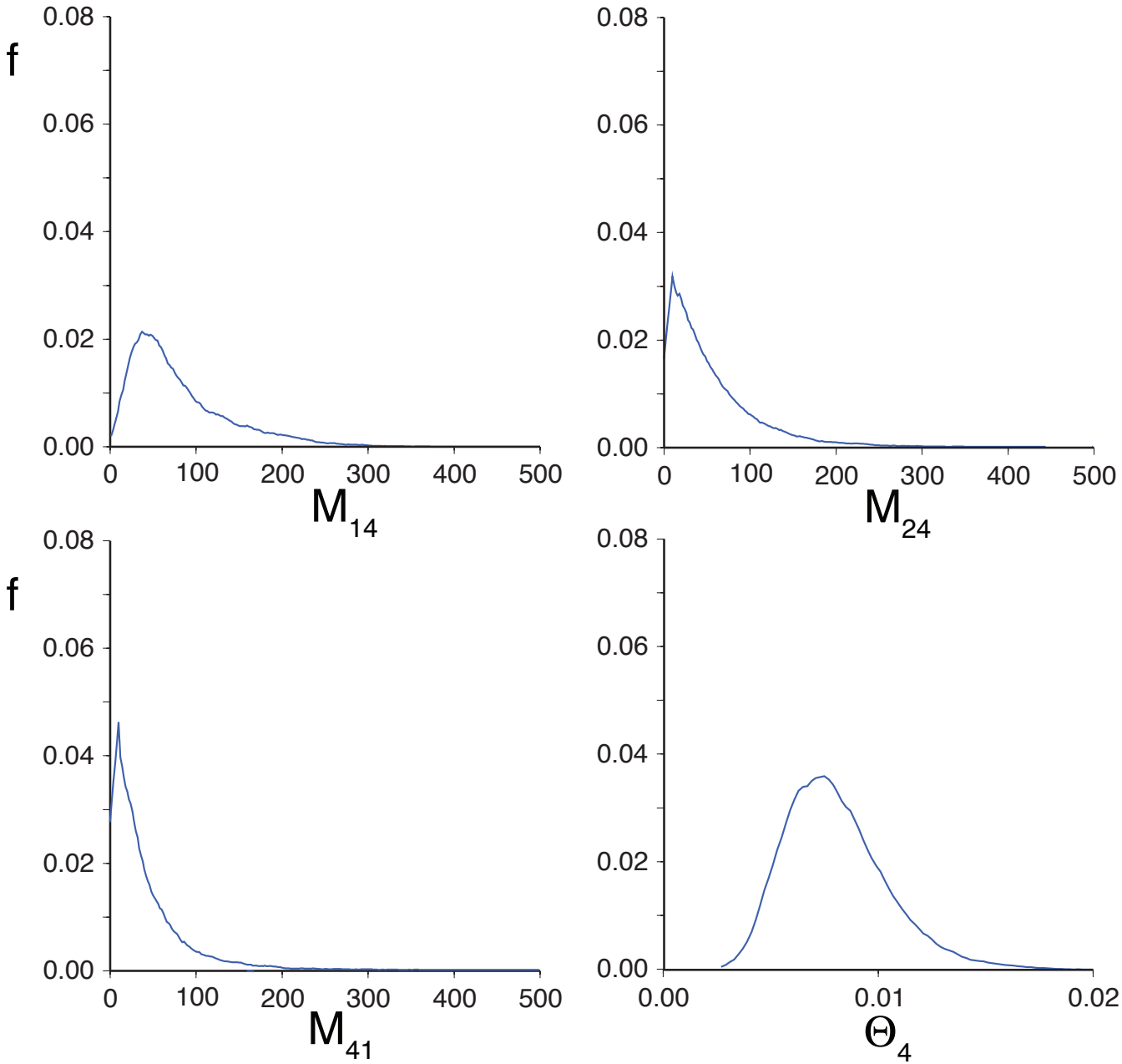


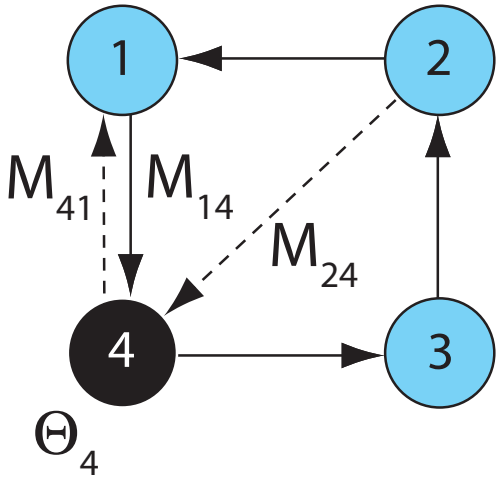
Beerli (2006) Comparison of Bayesian and maximum likelihood inference of population genetic parameters. Bioinformatics.


1
2
M41
M14
M24
4
3
Θ4

Bayesian inference

Synthetic data

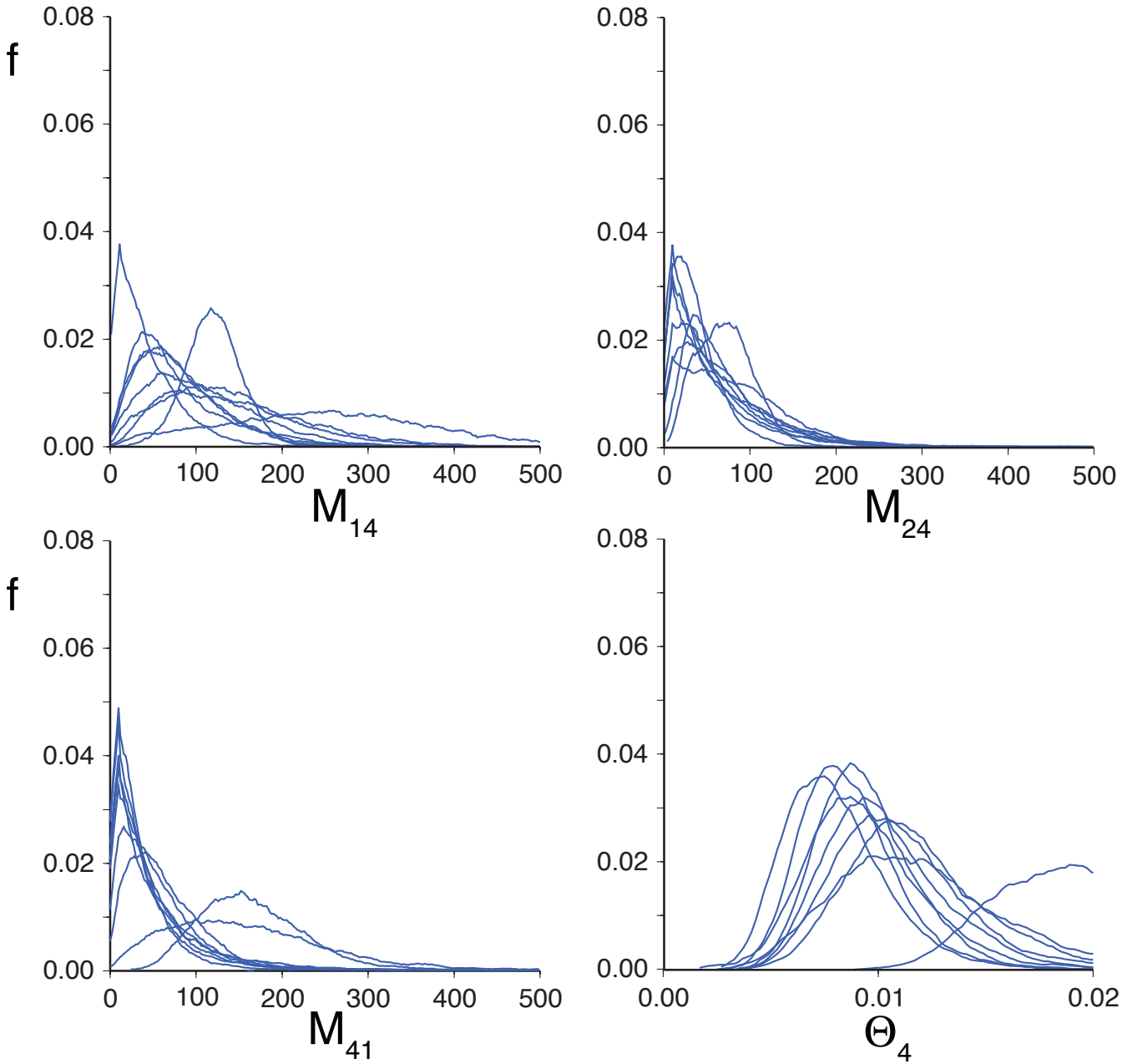


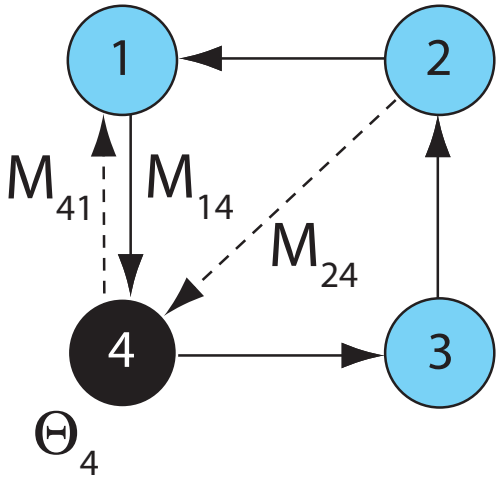


Beerli (2006) Comparison of Bayesian and maximum likelihood inference of population genetic parameters. Bioinformatics.

Bayesian inference

Synthetic data

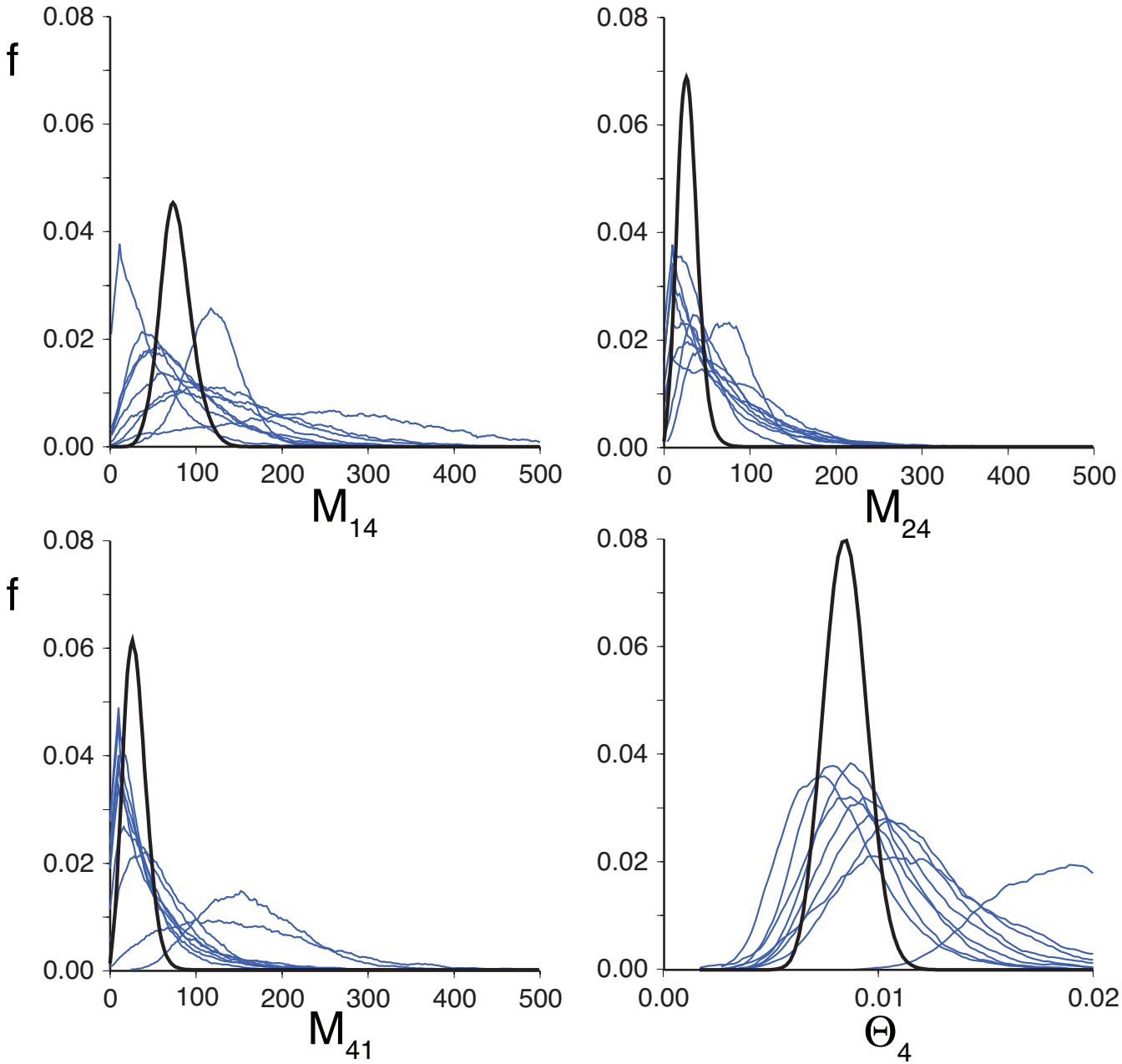


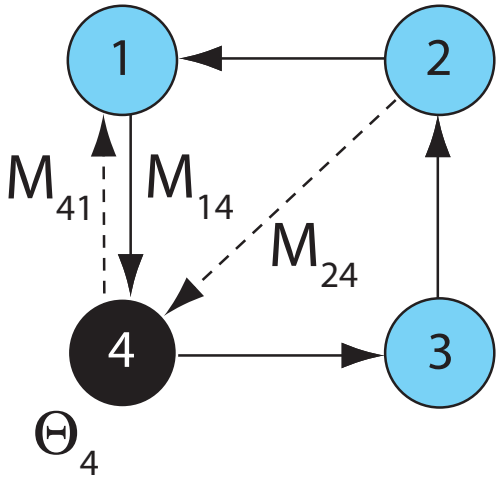


Beerli (2006) Comparison of Bayesian and maximum likelihood inference of population genetic parameters. Bioinformatics.

Bayesian inference

Synthetic data

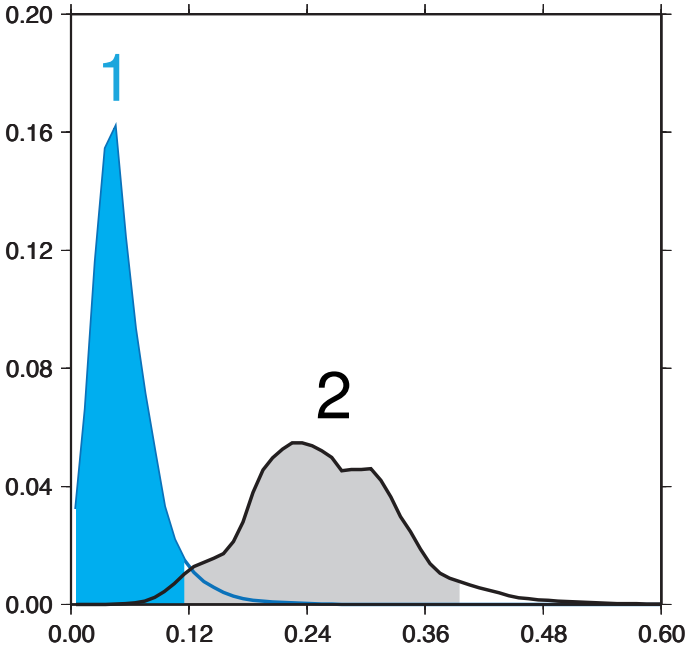




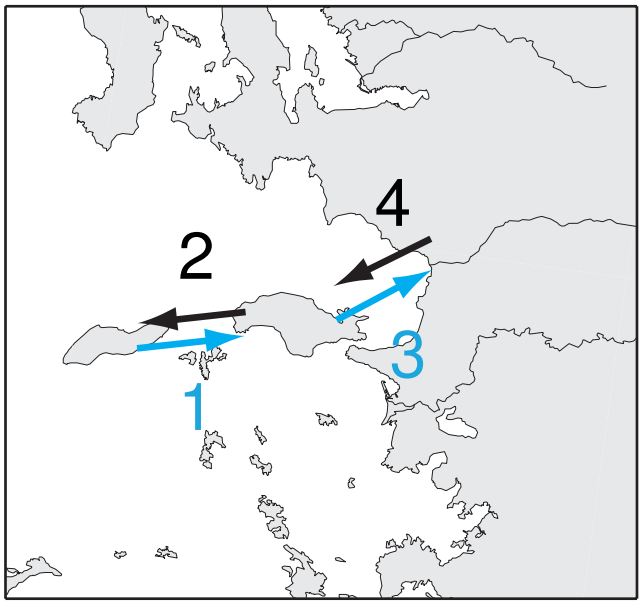
Beerli (2006) Comparison of Bayesian and maximum likelihood inference of population genetic parameters. Bioinformatics.

Obvious migration pattern

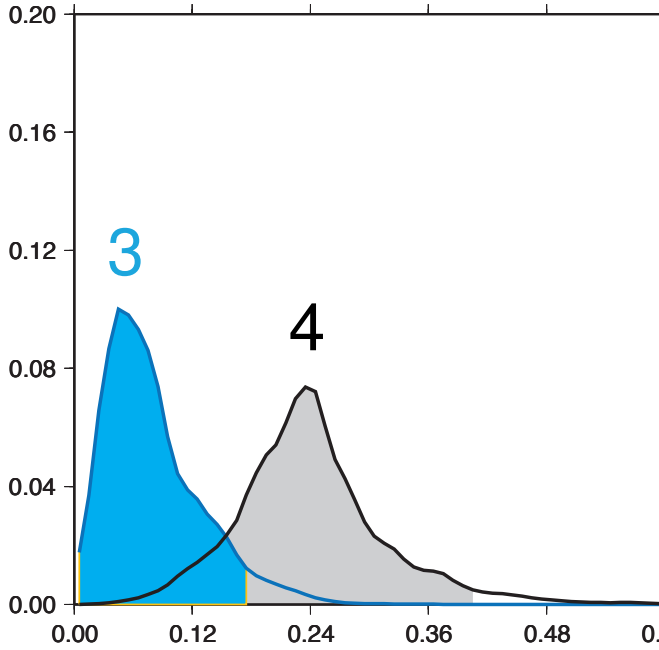
$p(\mathcal{M}|D)$



scaled migration rate

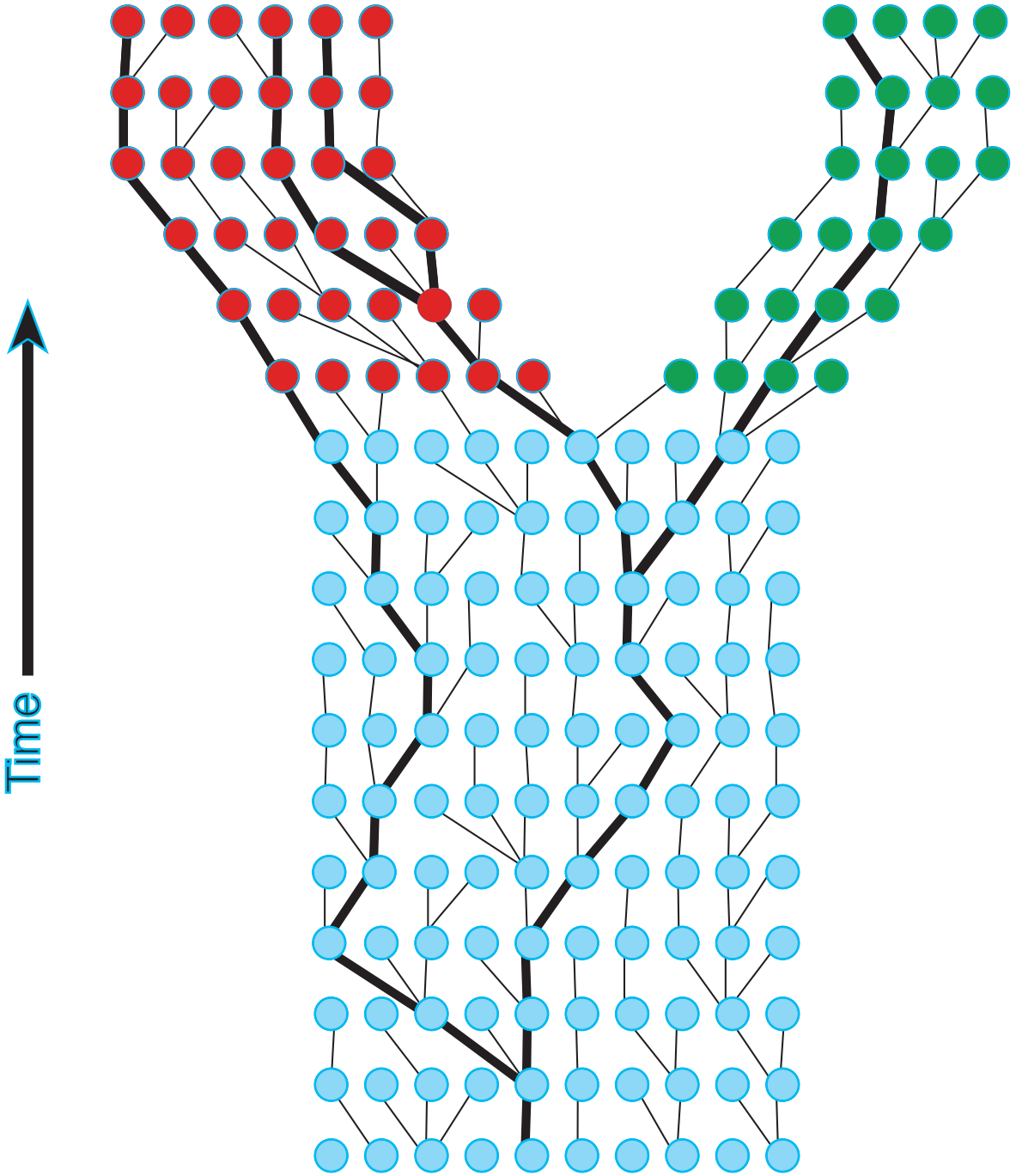


$p(\mathcal{M}|D)$

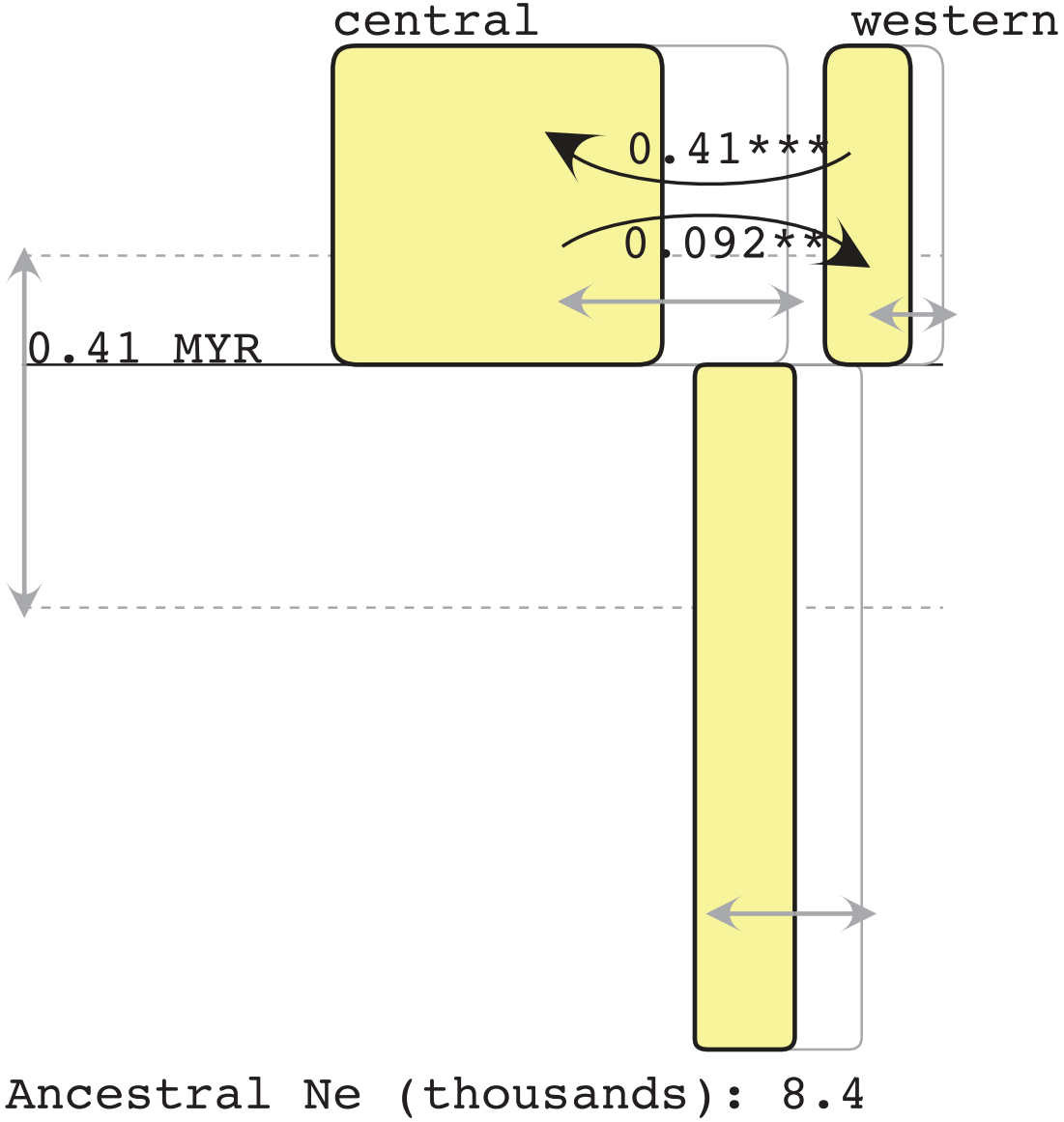


scaled migration rate

Beerli, P. (2009). How to use MIGRATE or why are Markov chain Monte Carlo programs difficult to use.

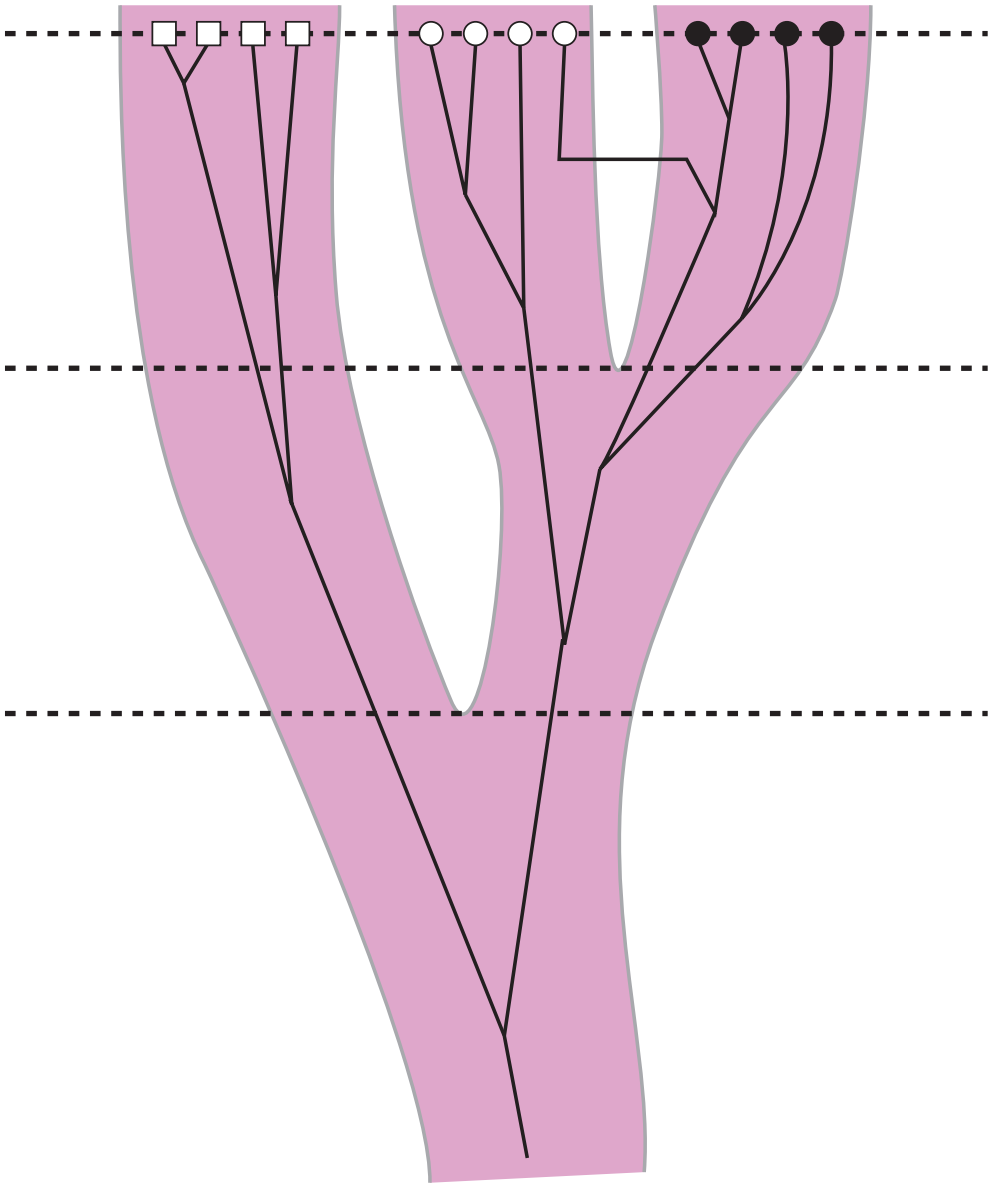

Time

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



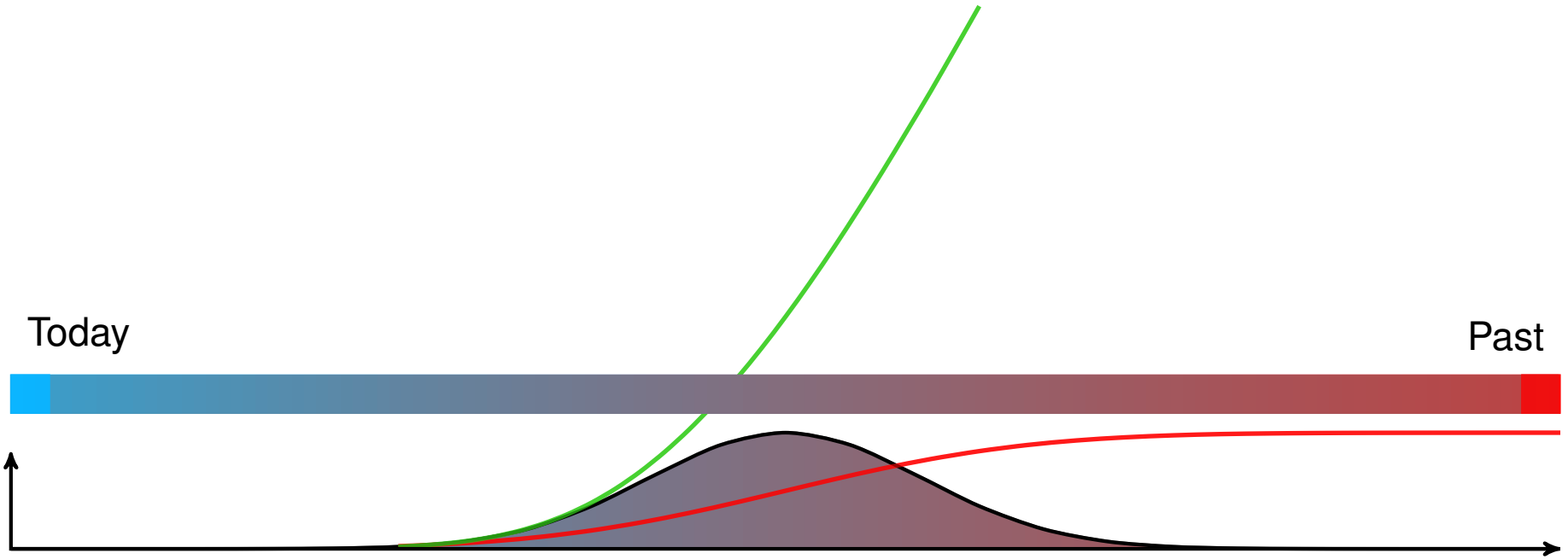
Population splitting

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

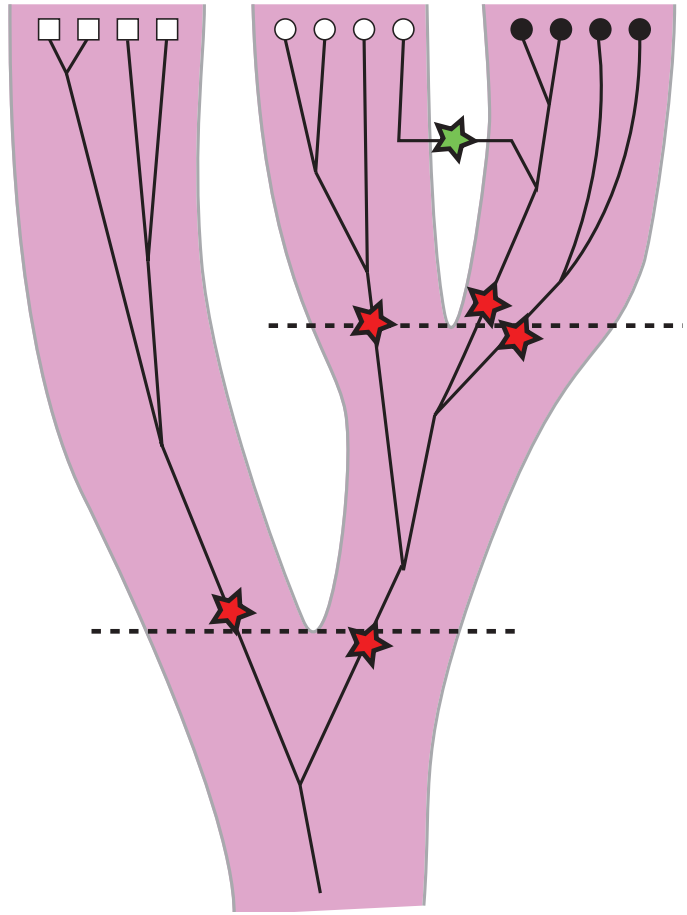
Looking backwards in time we could think about the risk of **A** turning into **B** which becomes larger and larger the further back in time the lineage goes. In the coalescence framework we are well accustomed to that thinking: we use the risk of a coalescent or the risk of a migration event. This risk 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

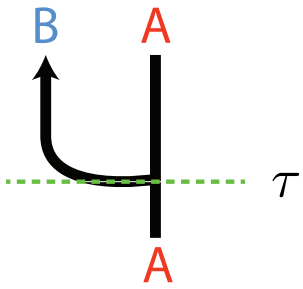
One lineage is easy, but what about the genealogy? Each lineage is at risk of being in the ancestral population, thus we need to consider coalescences, migration events, and population label changing events. This results in genealogies that are realizations of migration and population splitting events.



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

Population splitting

Comparison of estimated versus simulated divergence times for different number of loci



Model, τ → Genealogy

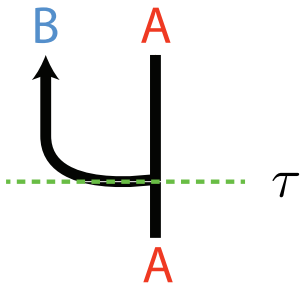
Genealogy → Sequence data

Sequence Data → Model → $\hat{\tau}$

(Beerli, Ashki, Mashayekhi, and Palczewski (still in review) Population divergence estimation using individual lineage label switching.)

Population splitting

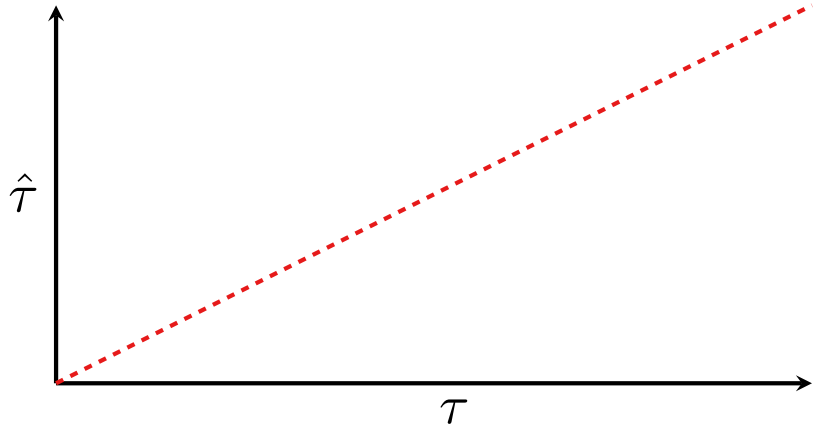
Comparison of estimated versus simulated divergence times for different number of loci



Model, τ → Genealogy

Genealogy → Sequence data

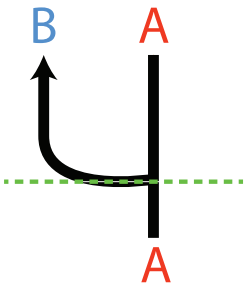
Sequence Data → Model → $\hat{\tau}$

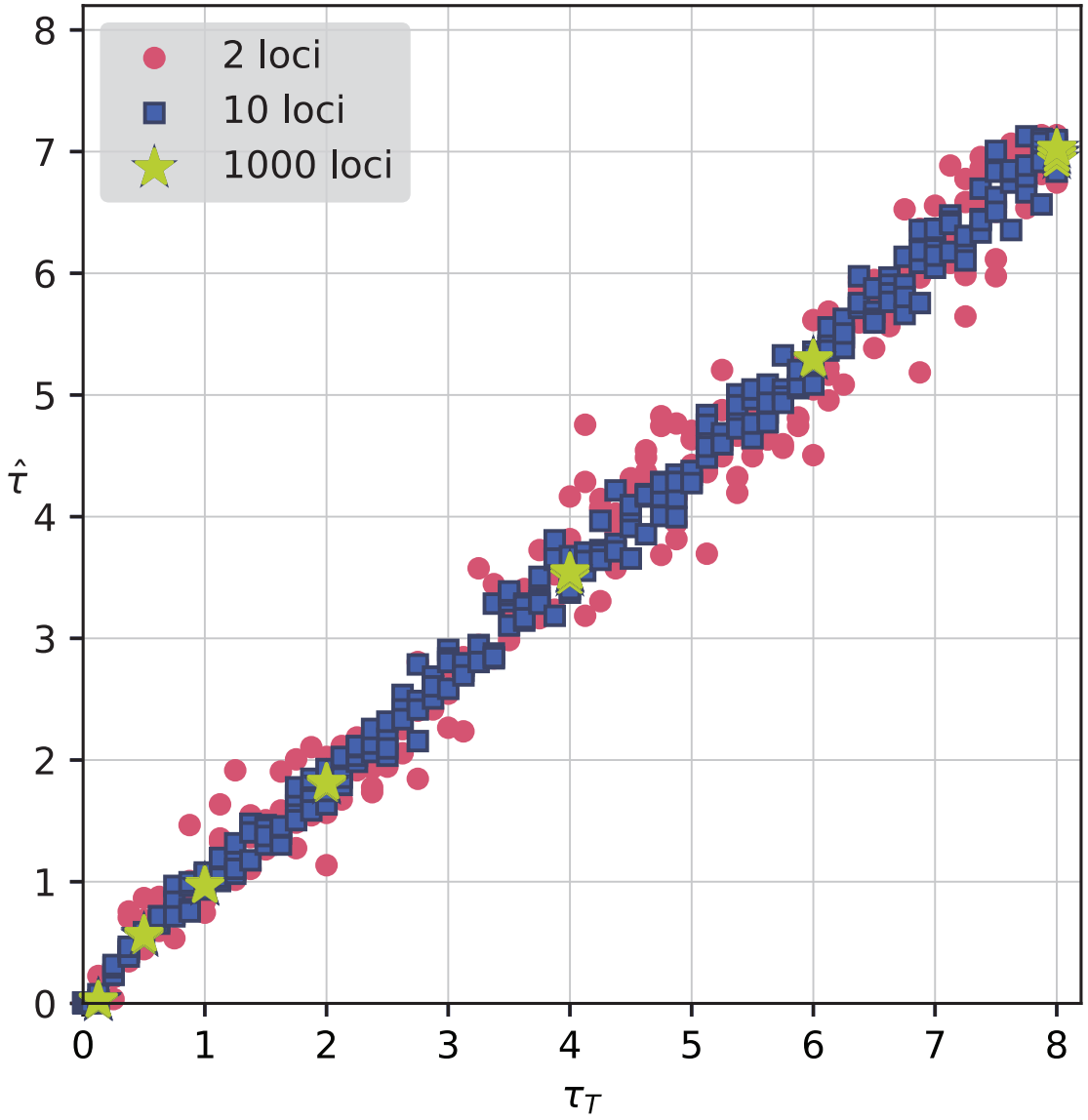


(Beerli, Ashki, Mashayekhi, and Palczewski (biorxiv) Population divergence estimation using individual lineage label switching.)

Population splitting

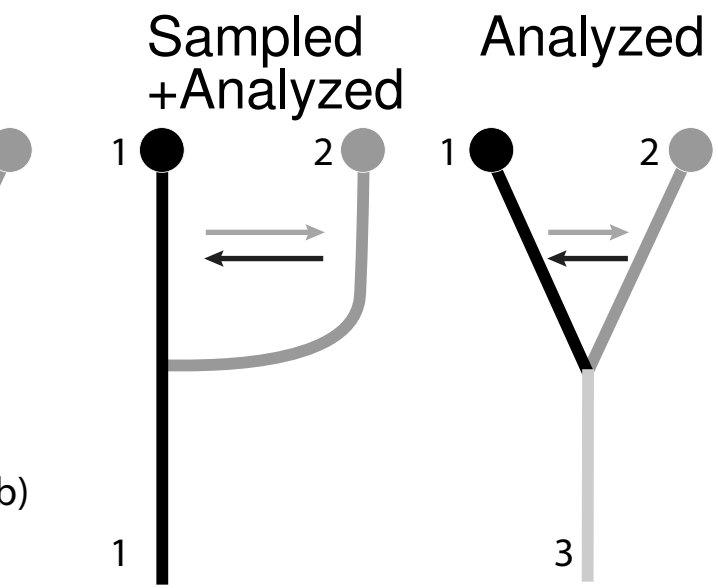
Comparison of estimated versus simulated divergence times for different number of loci

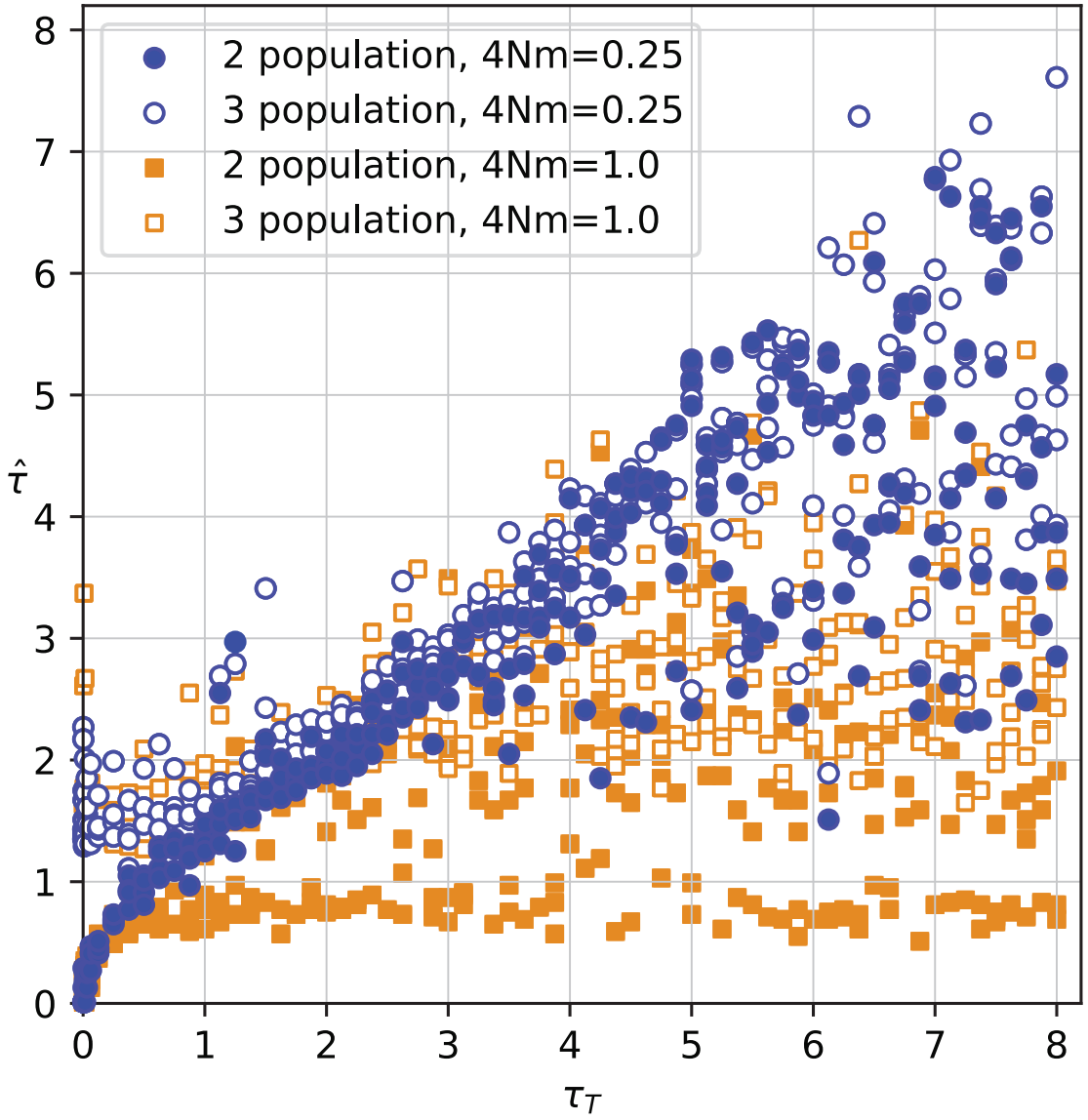




Beerli, Ashki, Mashayekhi, and Palczewski (Biorxiv) Population divergence estimation using individual lineage label switching.

Population splitting



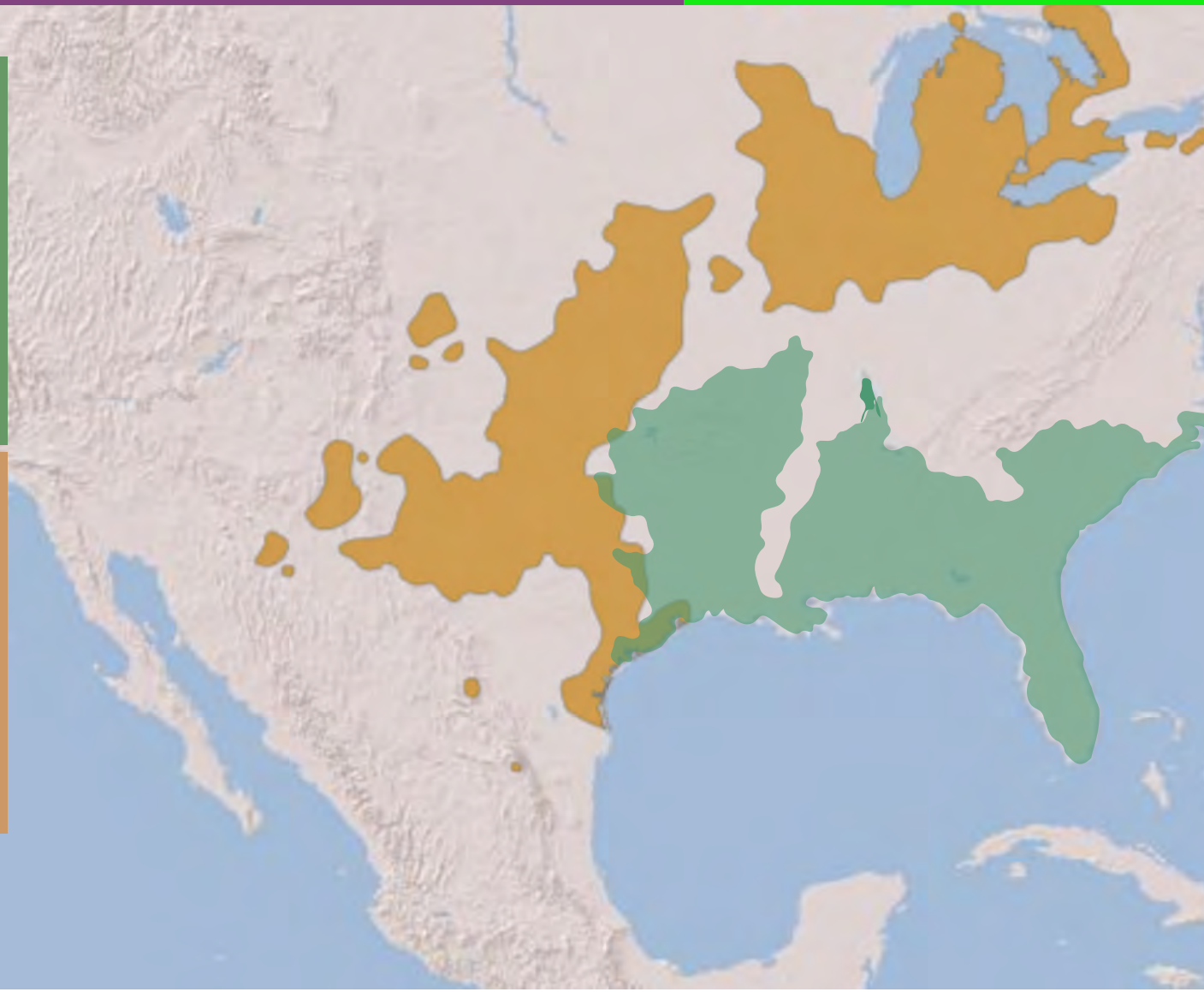


Beerli, Ashki, Mashayekhi, and Palczewski (biorxiv) Population divergence estimation using individual lineage label switching.

Population splitting

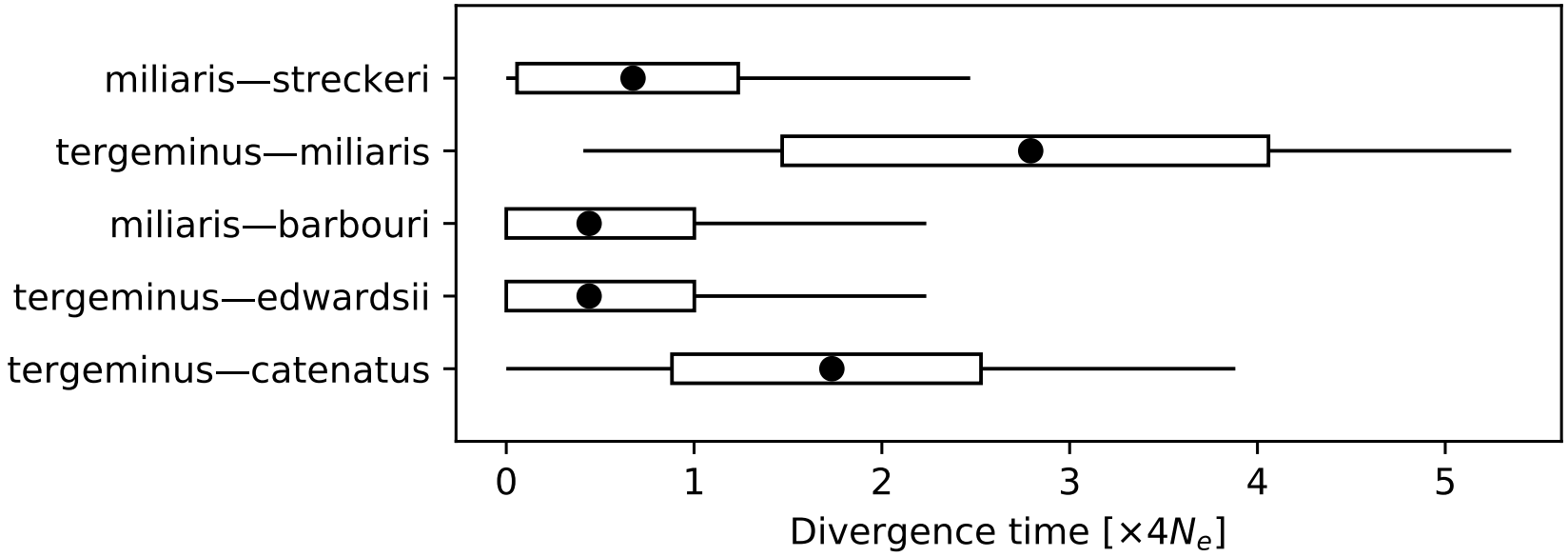
Lisle Gibbs, Ohio (Kubatko et al. 2011)

Phylogenetics of pygmy rattle snakes



Population splitting

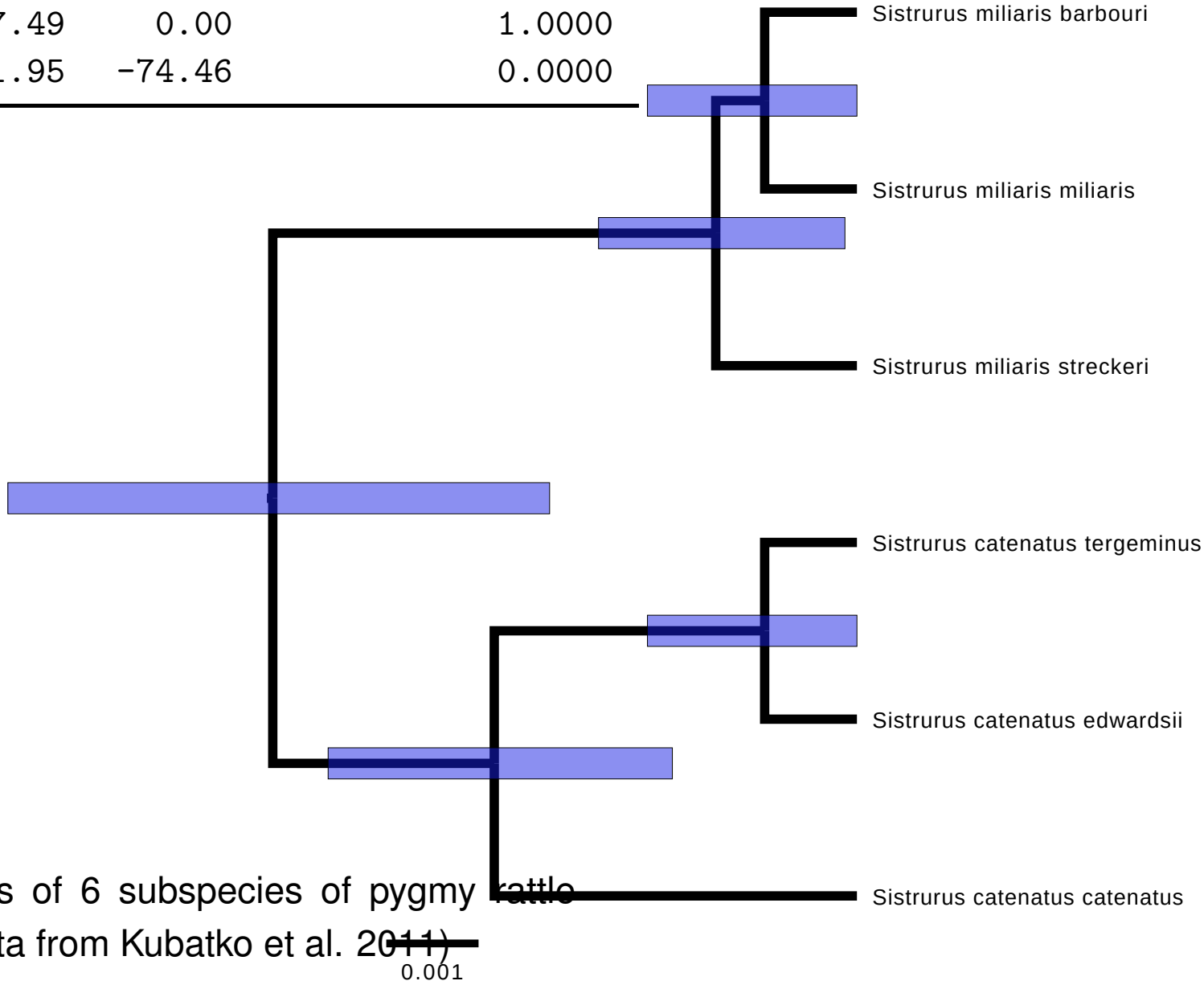
Pygmy rattle snakes



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

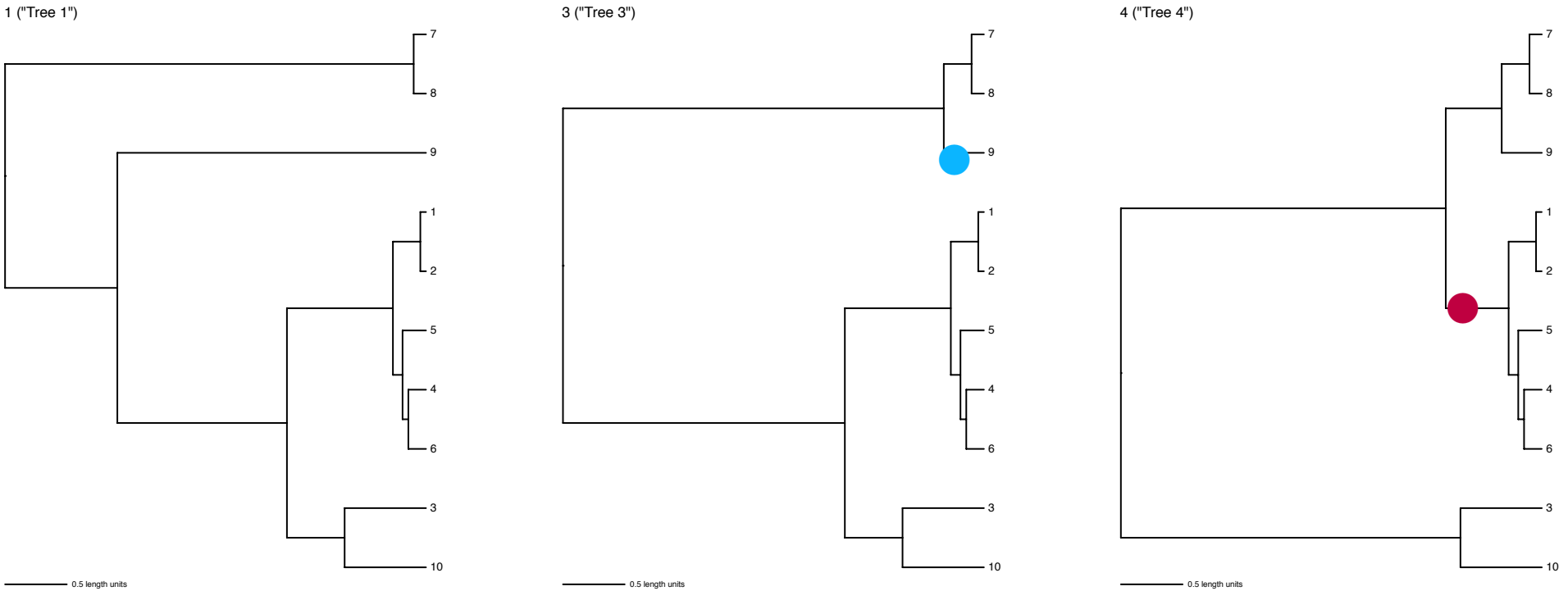
Population splitting: Pygmy rattle snakes

Model	Log(mL)	LBF	Model-probability
1: 3 species:	-15887.49	0.00	1.0000
2: 6 species:	-15961.95	-74.46	0.0000



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

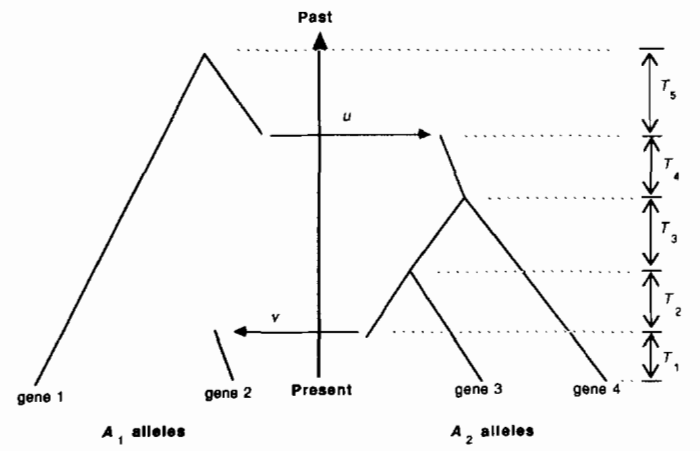
Coalescent and Recombination



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

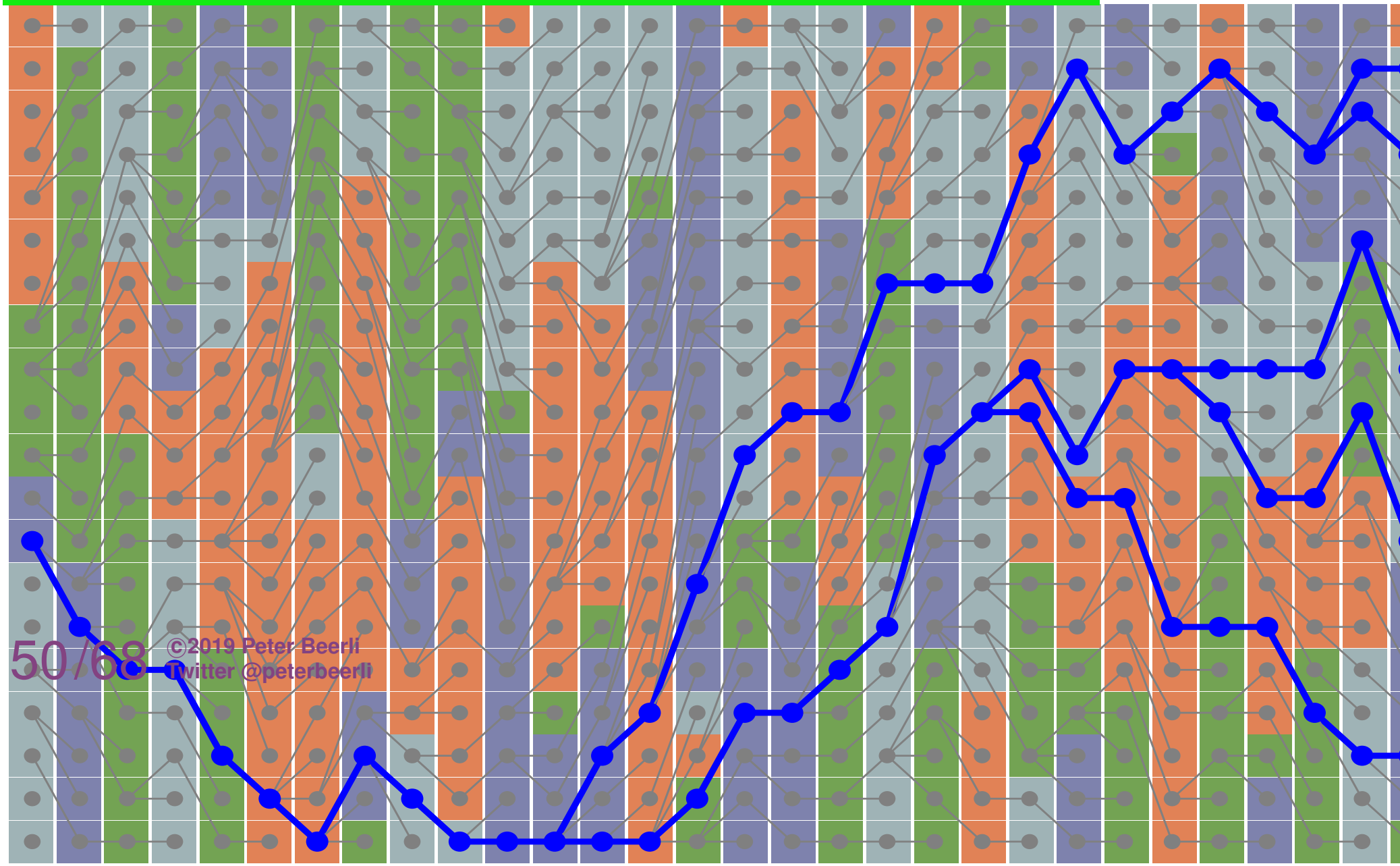
Coalescent and Selection

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)



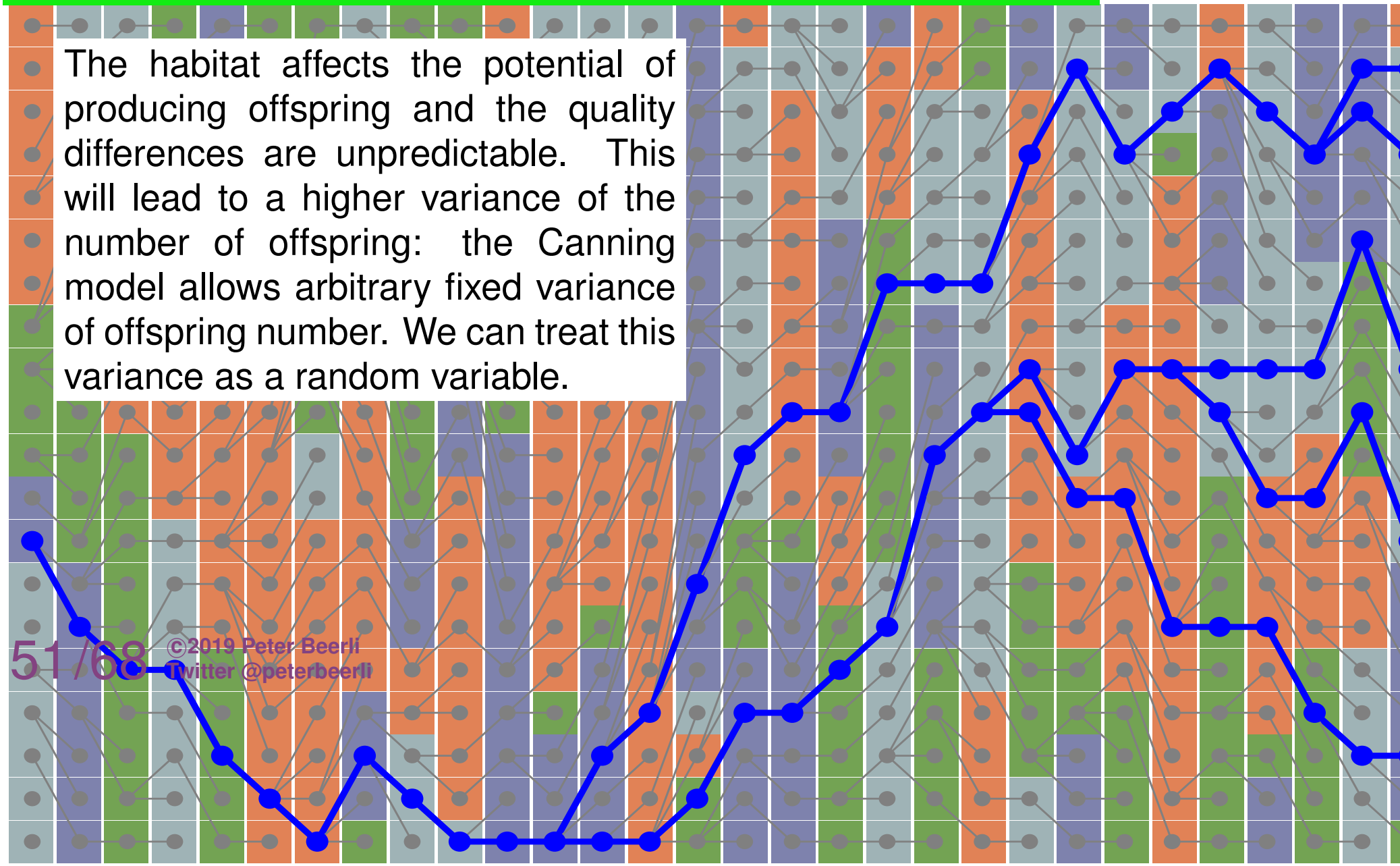
positive selection:

Offspring number is a random variable

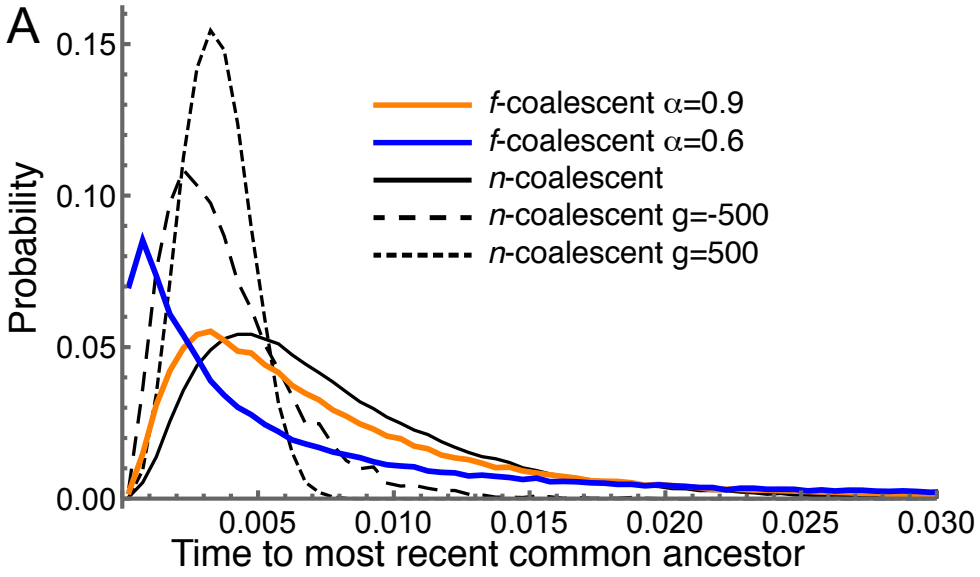


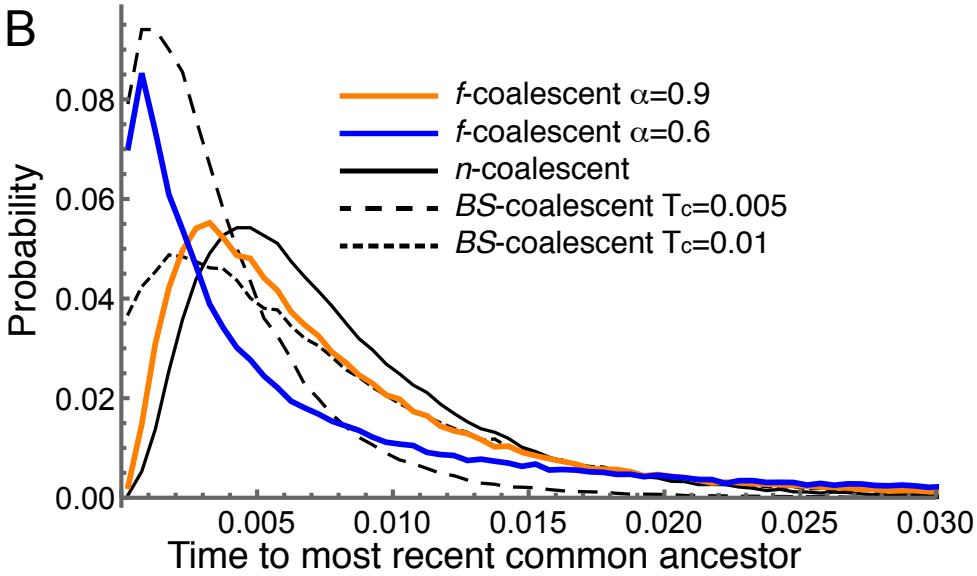
Offspring number is a random variable

The habitat affects the potential of producing offspring and the quality differences are unpredictable. This will lead to a higher variance of the number of offspring: the Canning model allows arbitrary fixed variance of offspring number. We can treat this variance as a random variable.



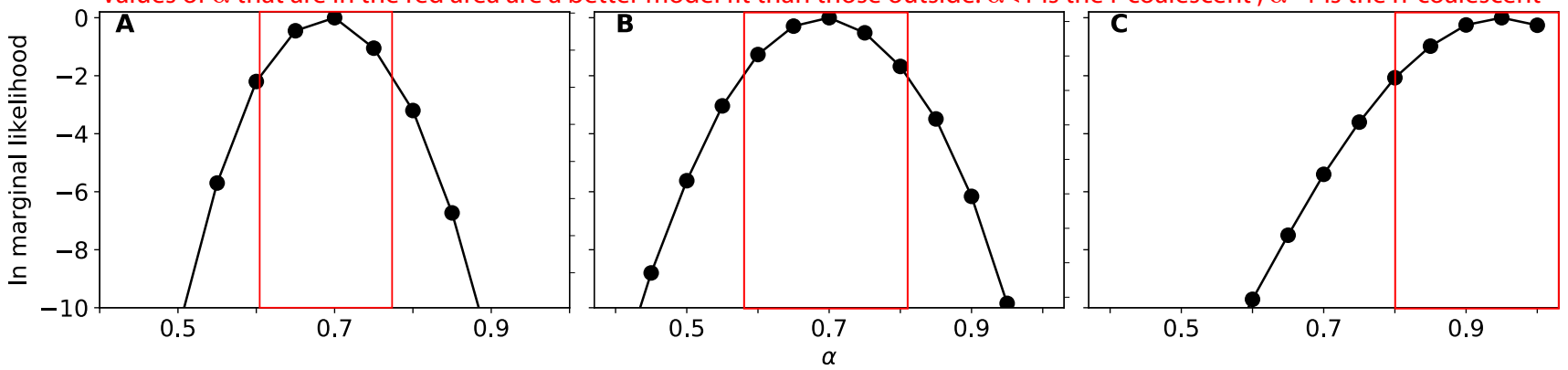
Extensions of Coalescence theory

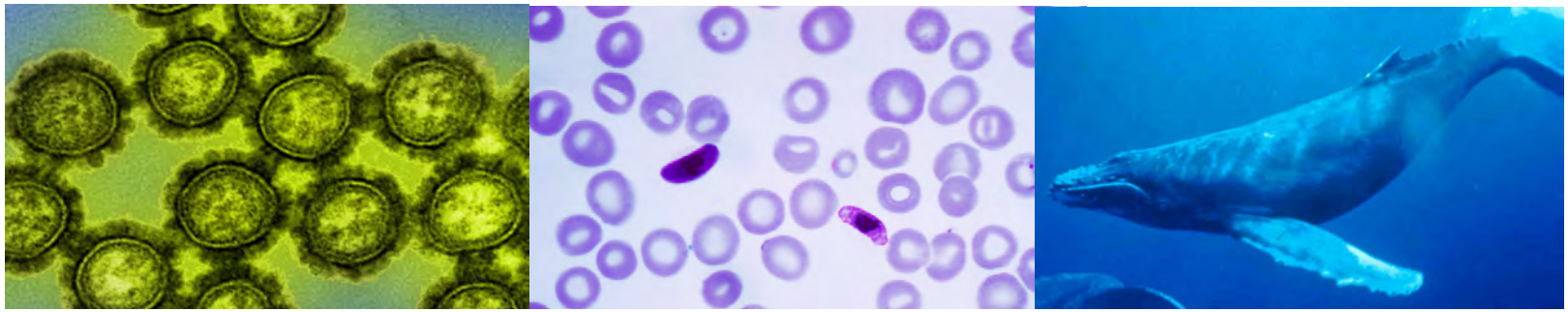




Different α: model comparison with real data

Values of α that are in the red area are a better model fit than those outside. $\alpha<1$ is the f-coalescent , $\alpha=1$ is the n-coalescent





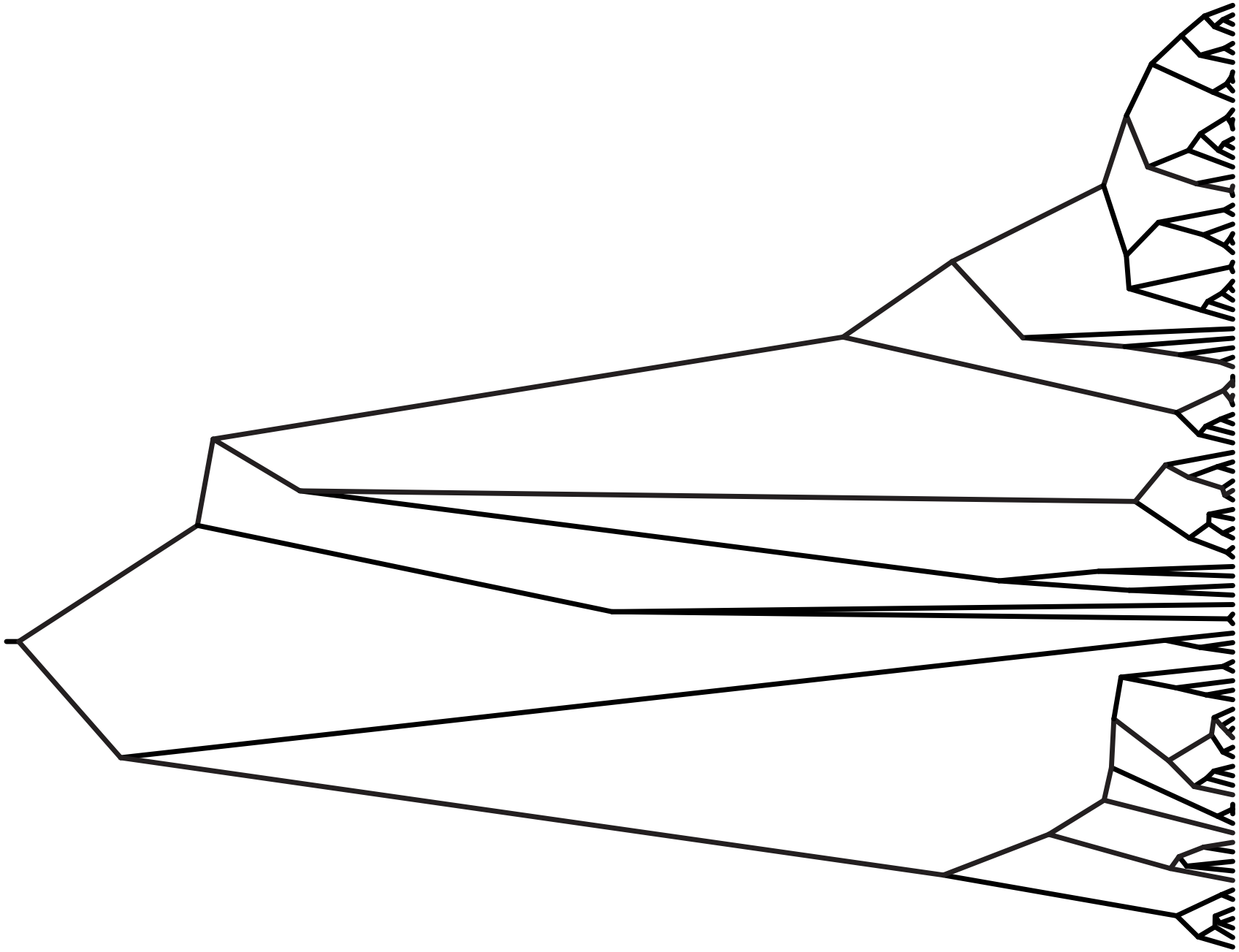
Model selection using relative marginal likelihoods of DNA sequence from the flu (H1N1), Malaria parasites, Humpback whales.

Robustness of the coalescence

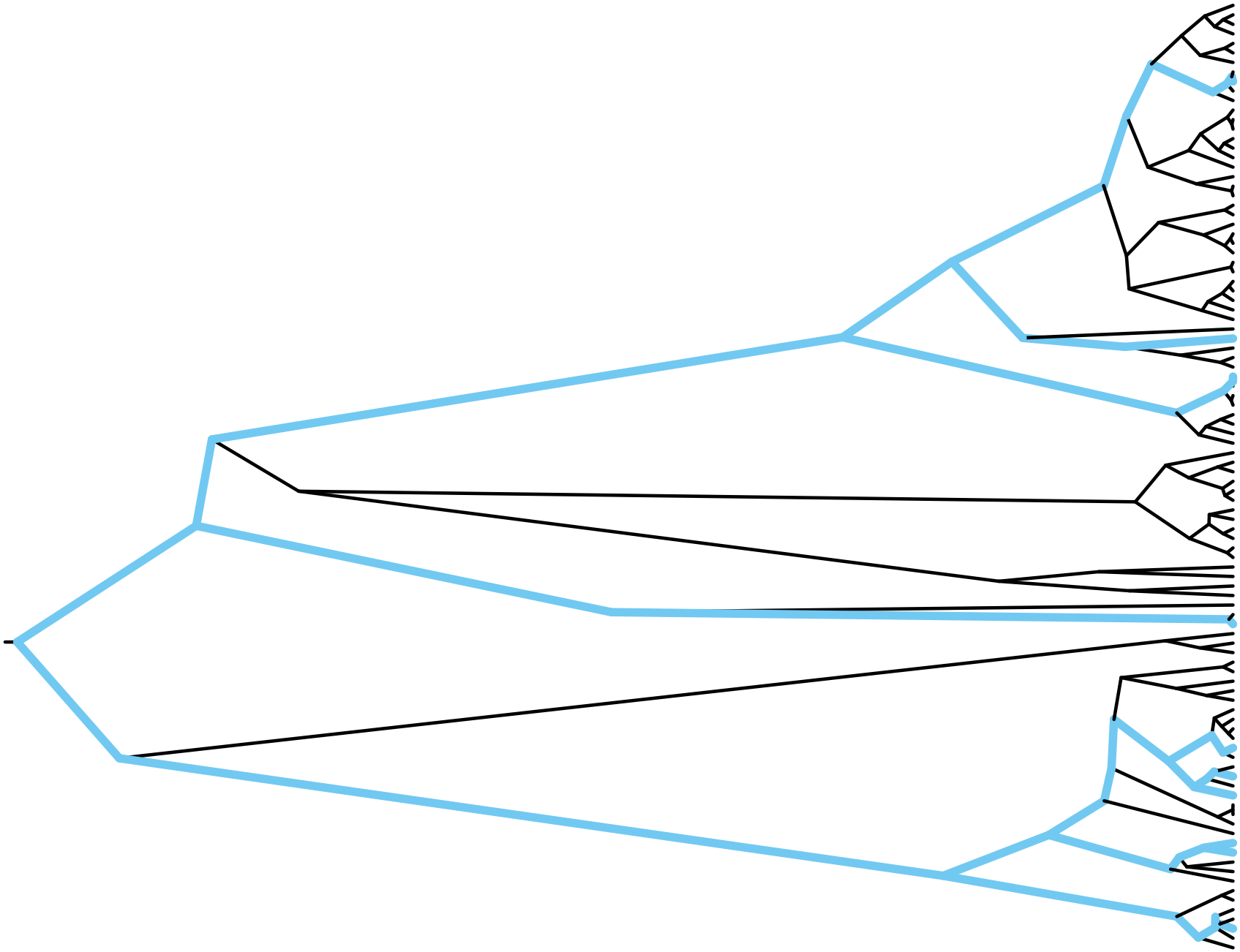
Violating assumptions

- Required samples (small samples/ deep coalescence)
- Average over long time
- Recombination

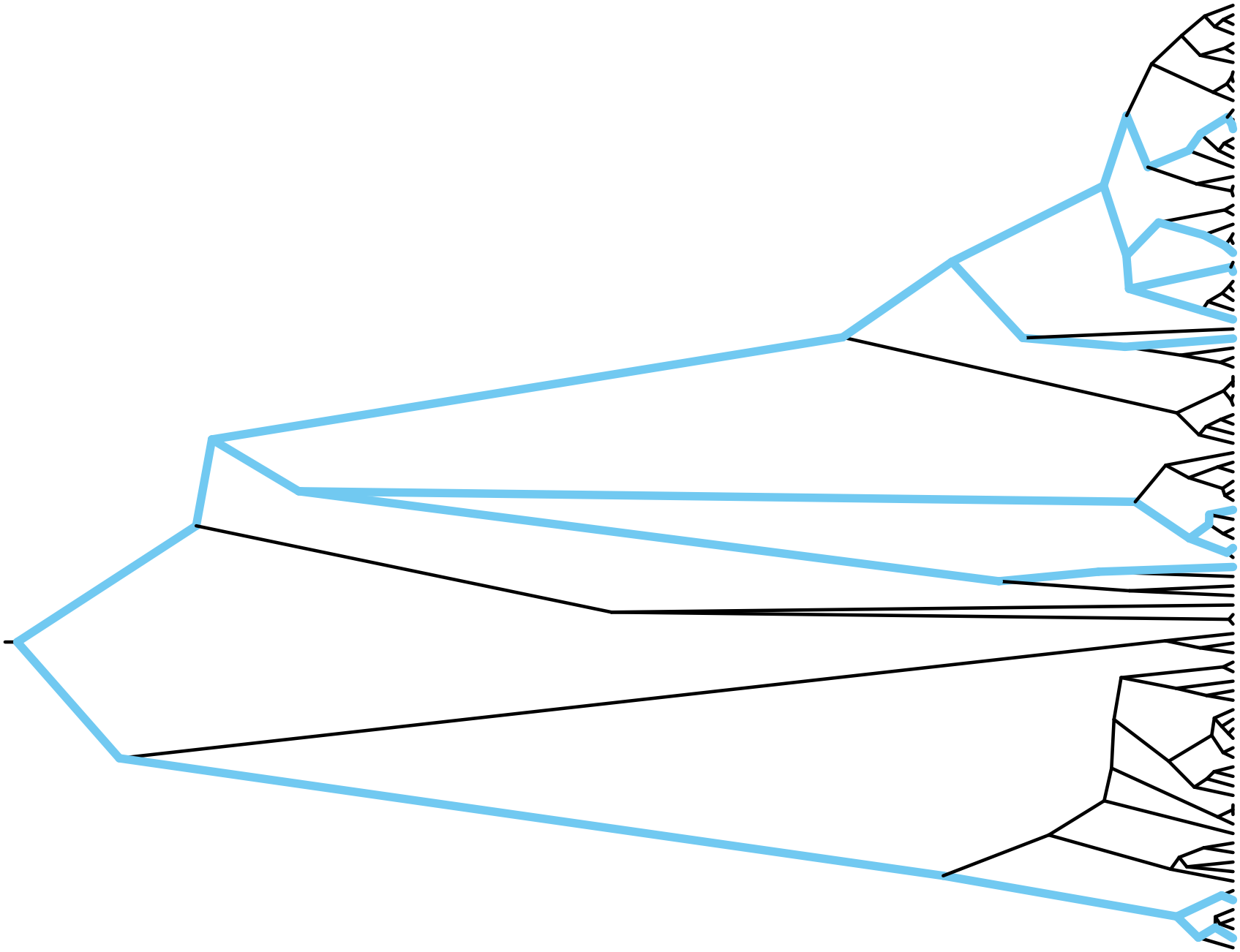
Required samples is small



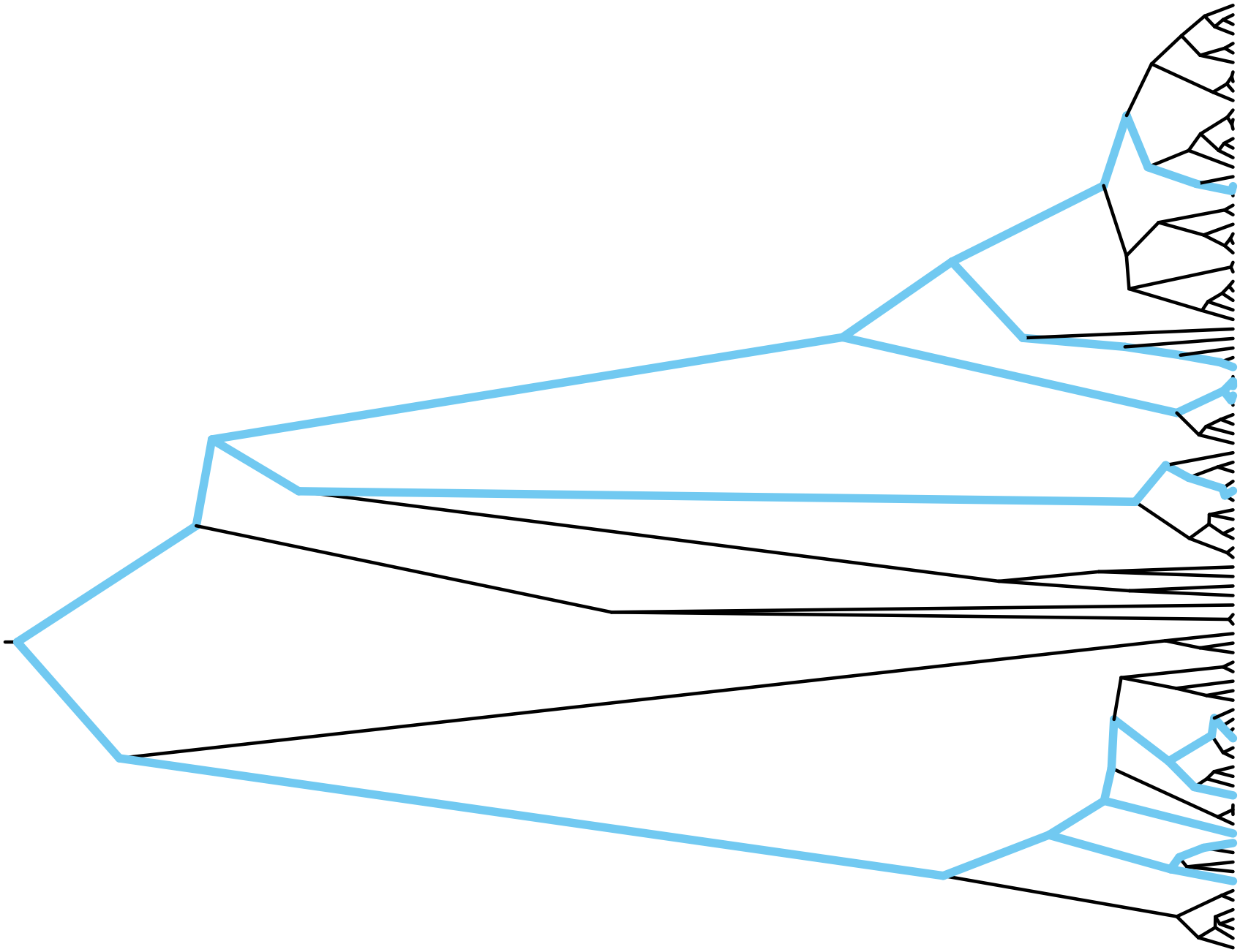
Required samples is small



Required samples is small

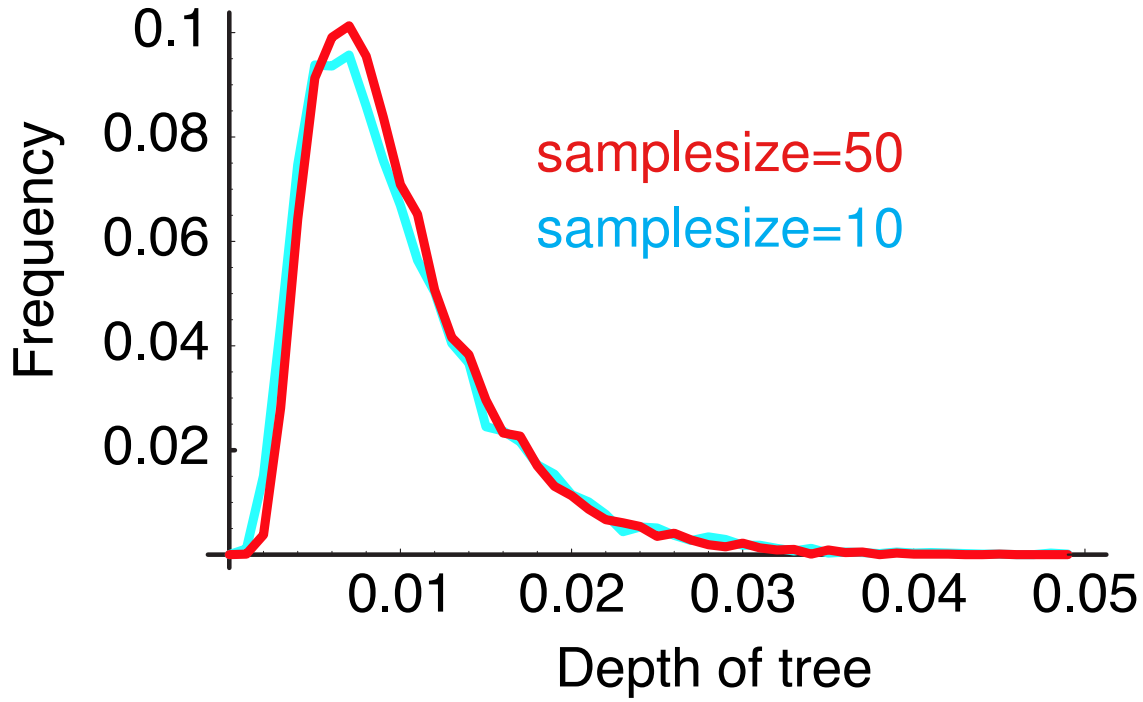


Required samples is small



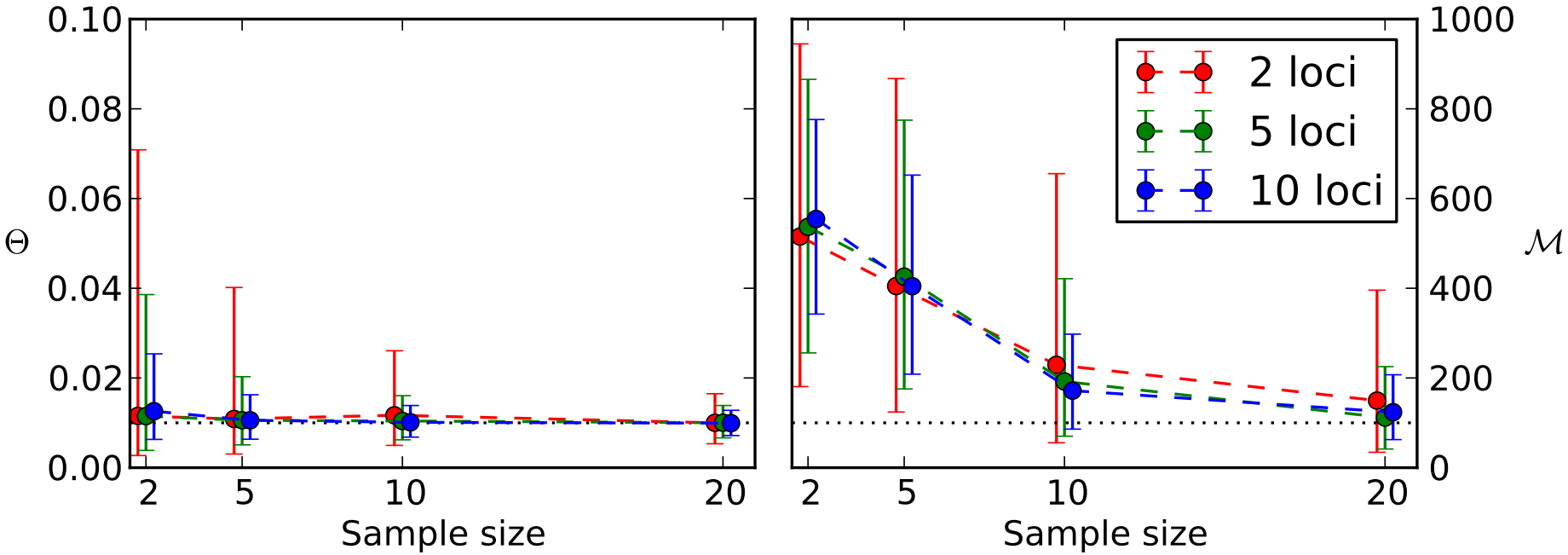
Required samples is small

- 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

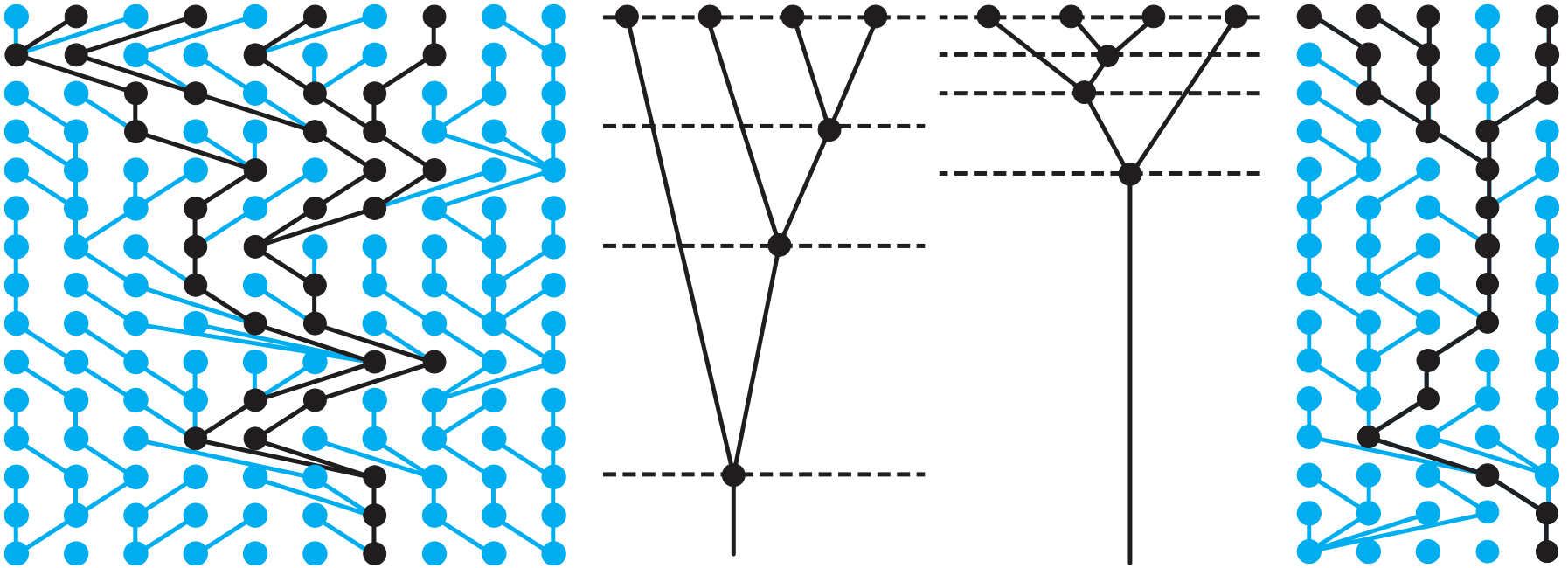


Medium variability DNA dataset: Mutation-scaled population size Θ and mutation-scaled migration rate M versus sample size for 2, 5, and 10 loci. The true $\Theta_T = 0.01$ is marked with the dotted gray line; $M = 100$

Average of parameters over long time

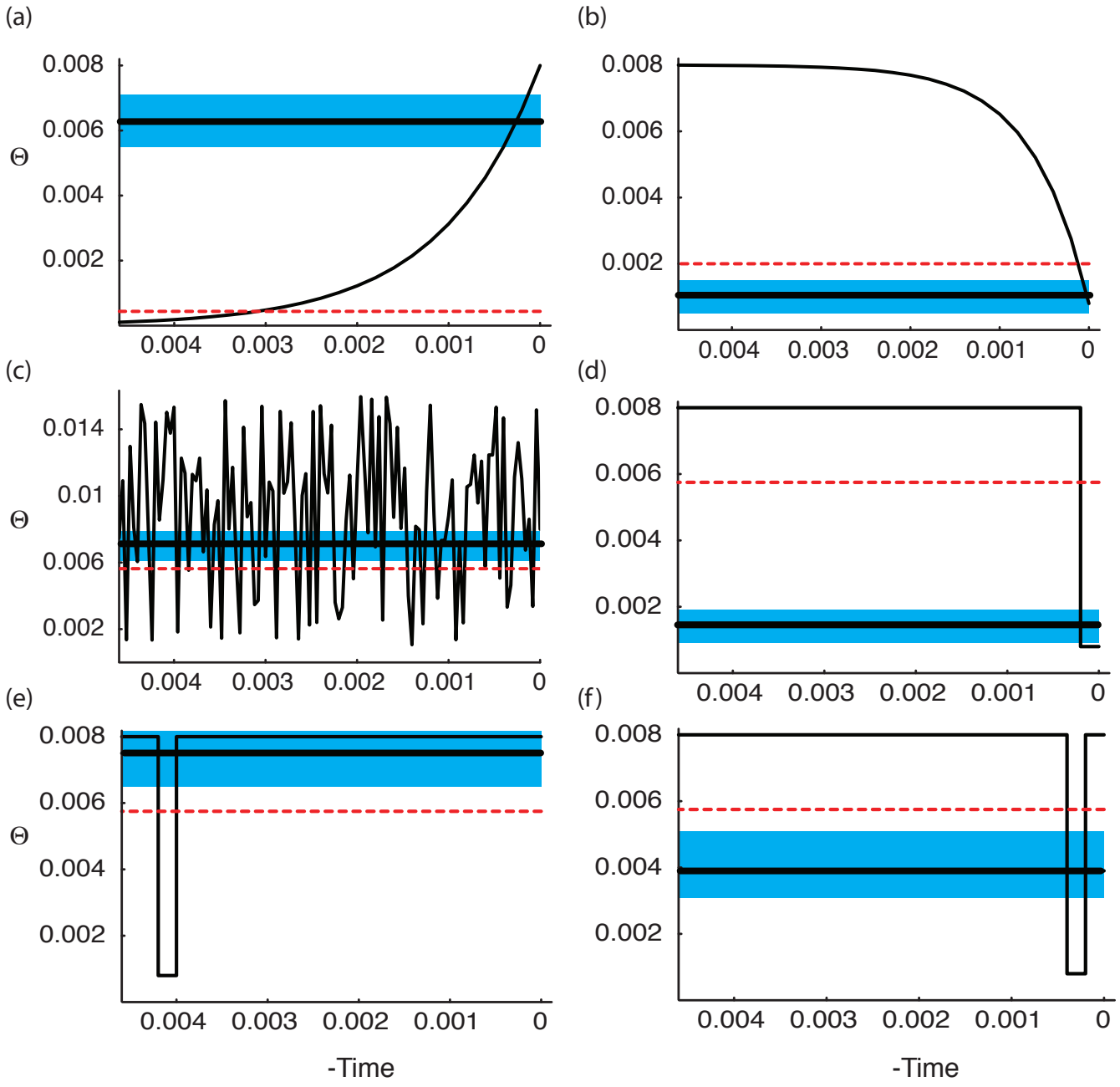
Researchers from the frequency-based camp claim that the coalescence-based methods are working on an evolutionary time-scale and therefore are not really usable in a conservation genetics or management context.

There is some truth to this claim because the time scale for the genealogies is in generations and with large populations such genealogies are deep, but ...



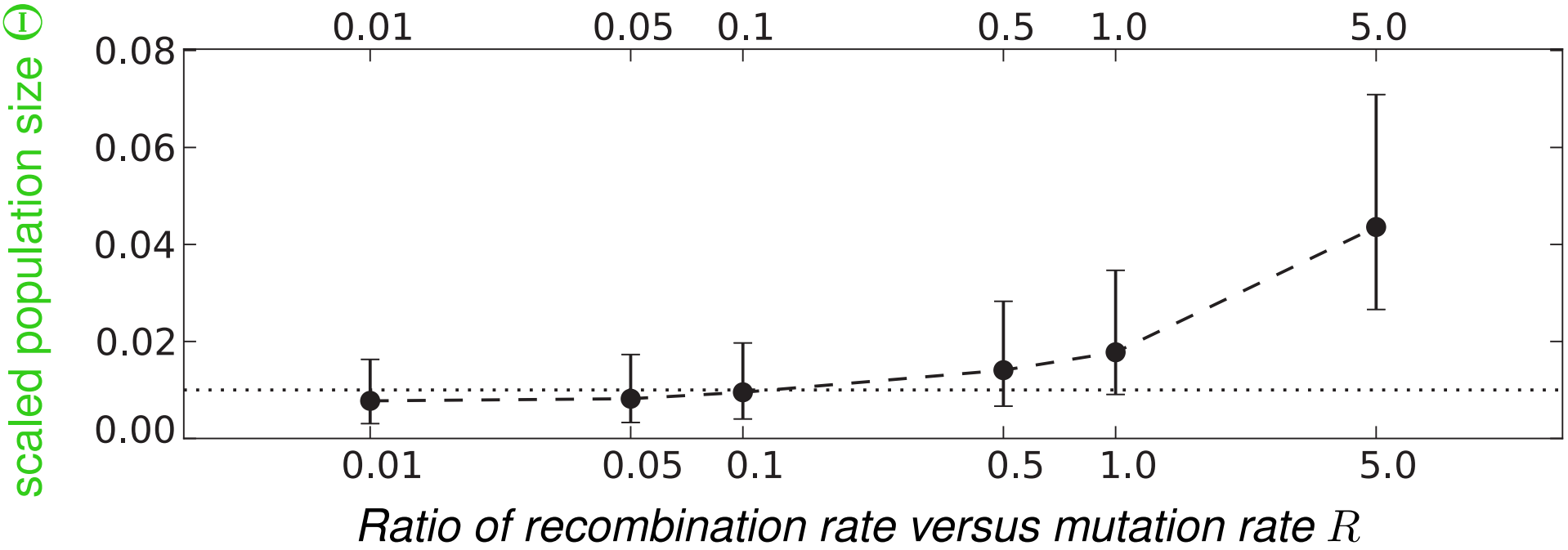
Average of parameters over long time

True value
MIGRATE estimate
Support interval
Harmonic mean



Ignoring recombination

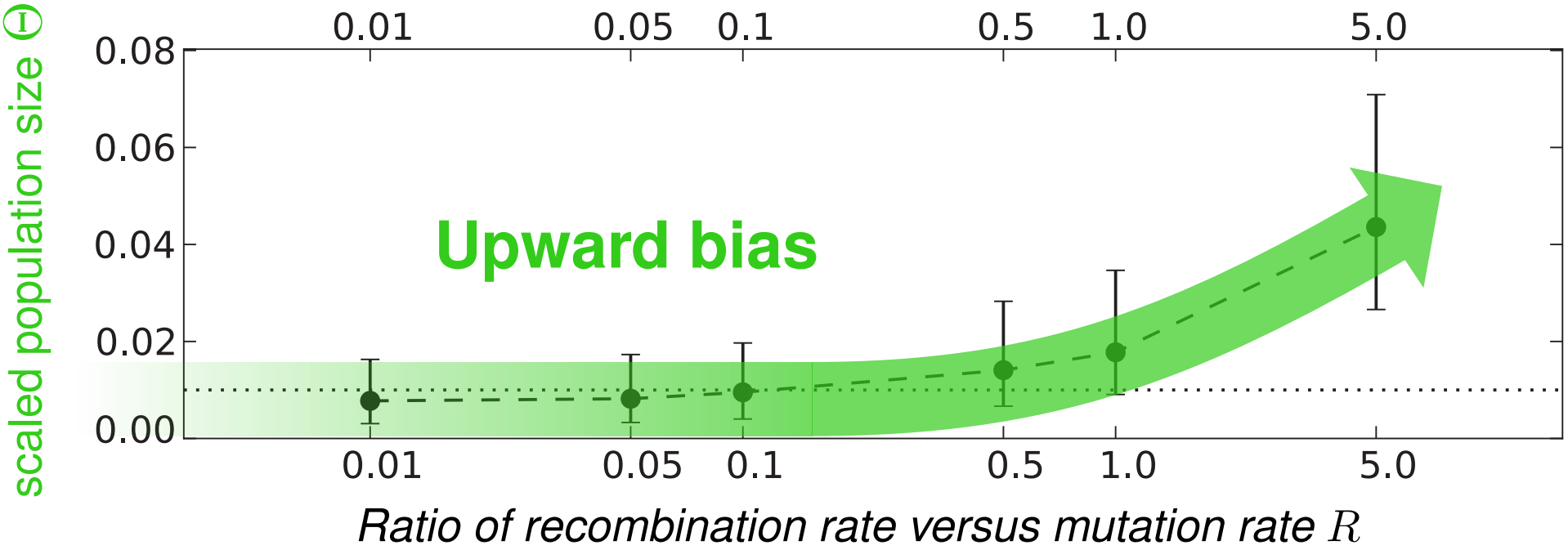
∼500 simulated datasets



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

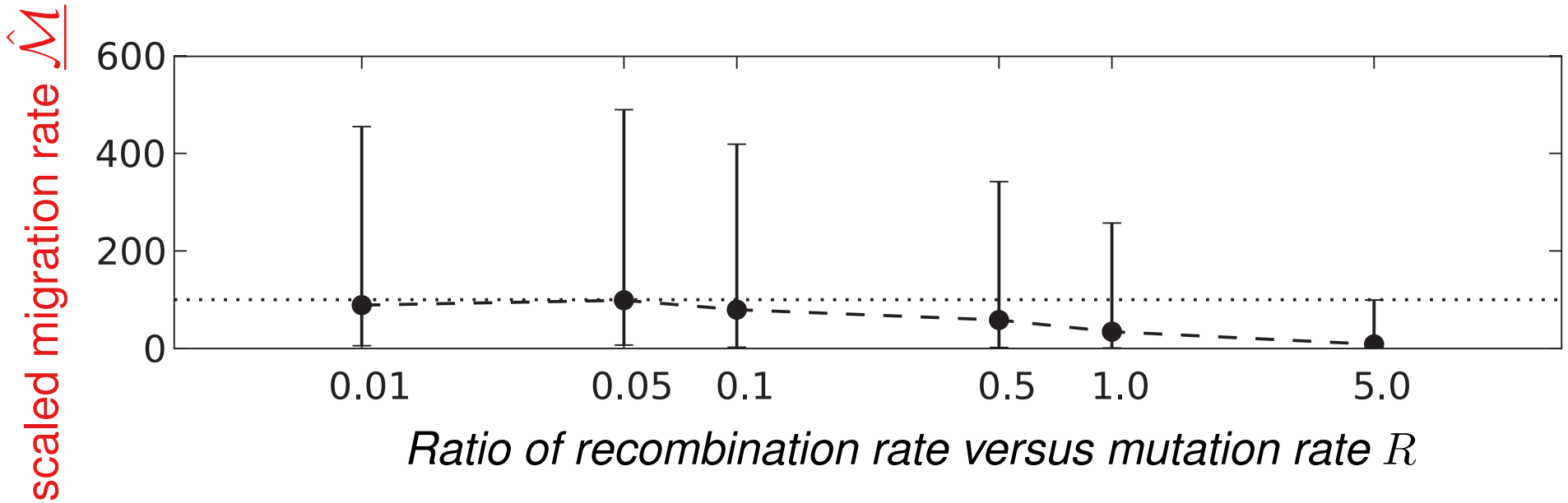
~500 simulated datasets



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

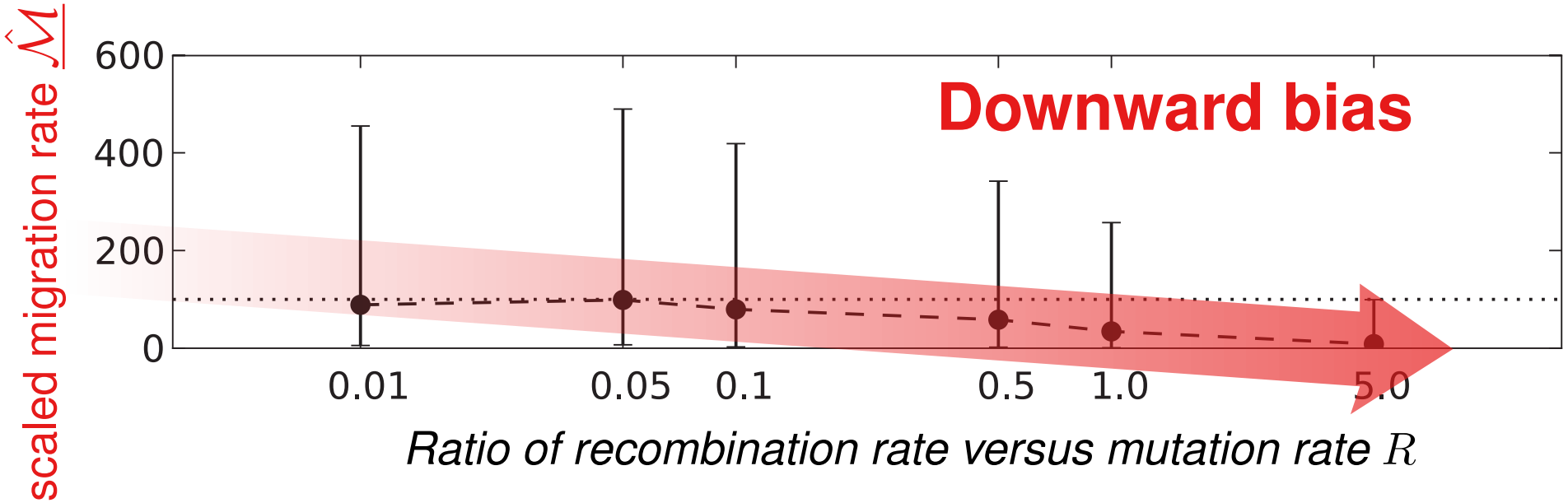
∼500 simulated datasets



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.

Outlook

- We will have a lab this afternoon where you will learn about the basics of MIGRATE using some simple urchin and bird datasets,

- Tomorrow we will learn about Bayesian model selection and then will do a lab where we 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.