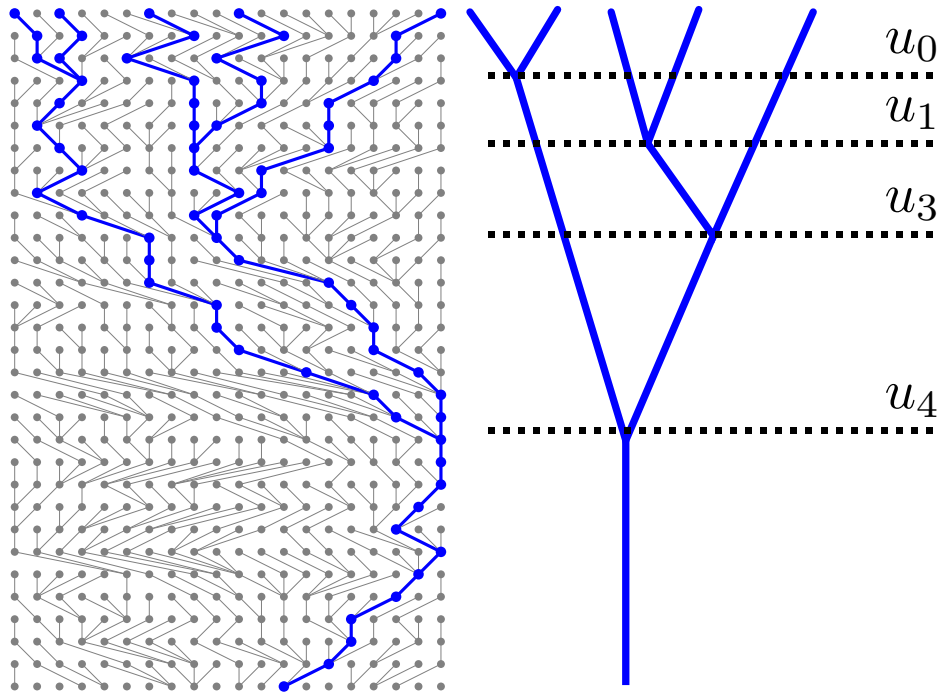


Extension of the basic coalescence



Peter Beerli
Florida State University
#MolEvol2015 MBL Woods Hole

Kingman's coalescent

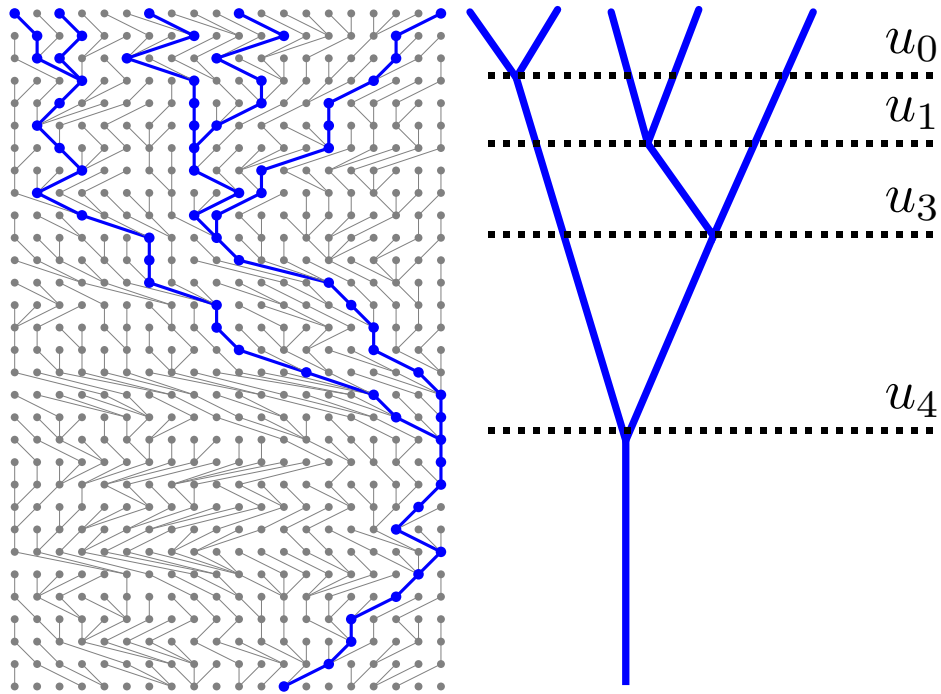


$$P(G|\Theta) = \prod_{j=0}^T e^{-u_j \frac{k_j(k_j-1)}{\Theta}} \frac{2}{\Theta}$$



$$\Theta = 4N_e\mu$$

- ◆ 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



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

 = 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



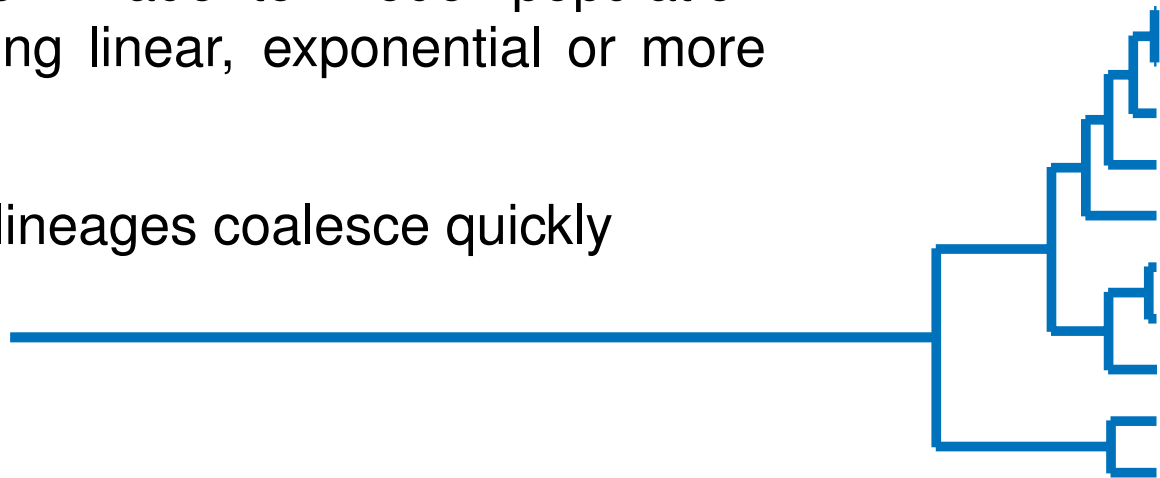
Extensions of the basic coalescence

- ◆ Population growth (2 parameters), fluctuations, bottlenecks
- ◆ Migration among populations (2 to many, potentially thousands, parameters)
- ◆ Population splitting (2 to many parameters)
- ◆ Recombination (2 parameters)

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.

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.

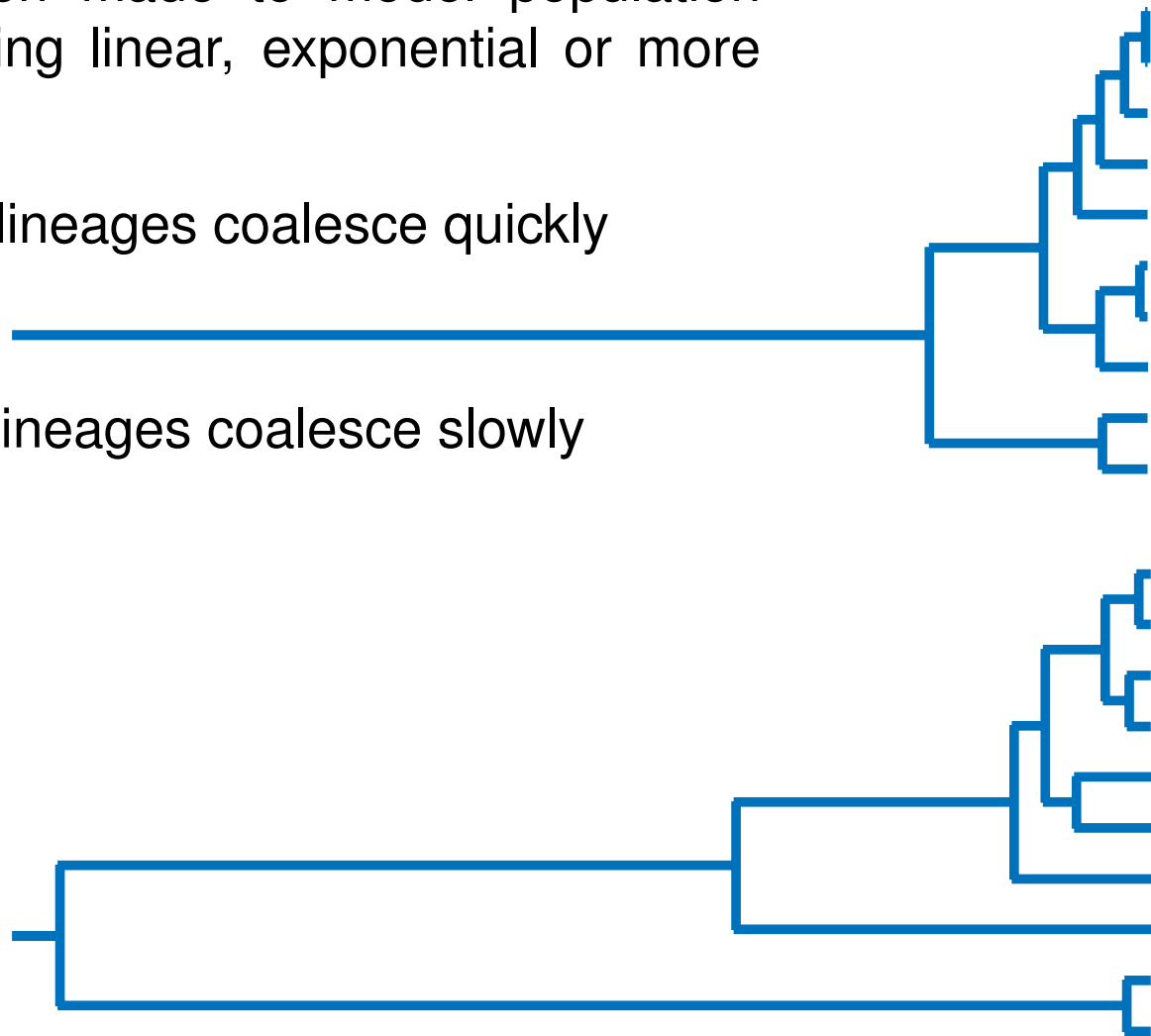
- ◆ In a small population lineages coalesce quickly



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

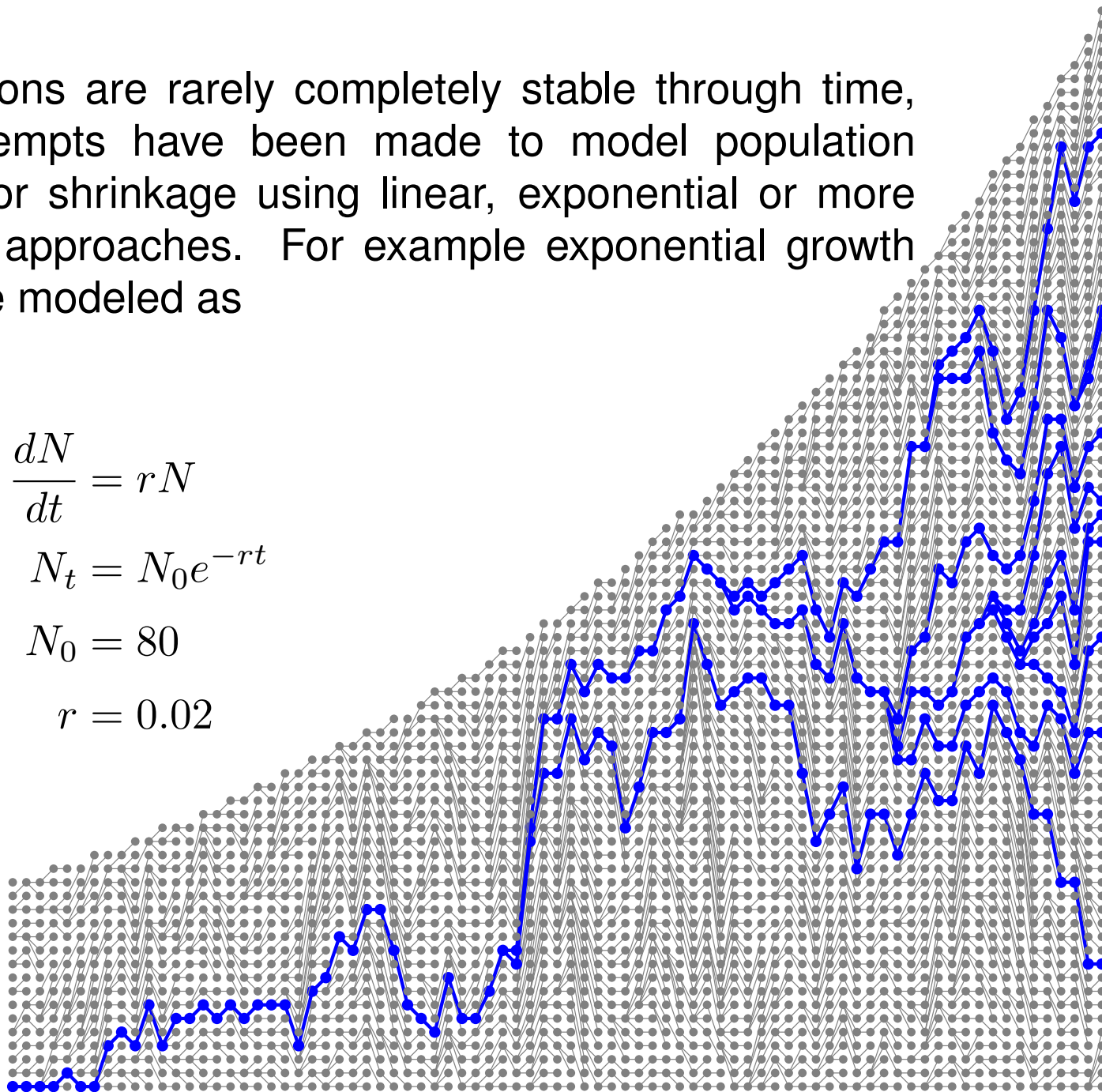
$$\frac{dN}{dt} = rN$$

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

$$N_0 = 80$$

$$r = 0.02$$

Past



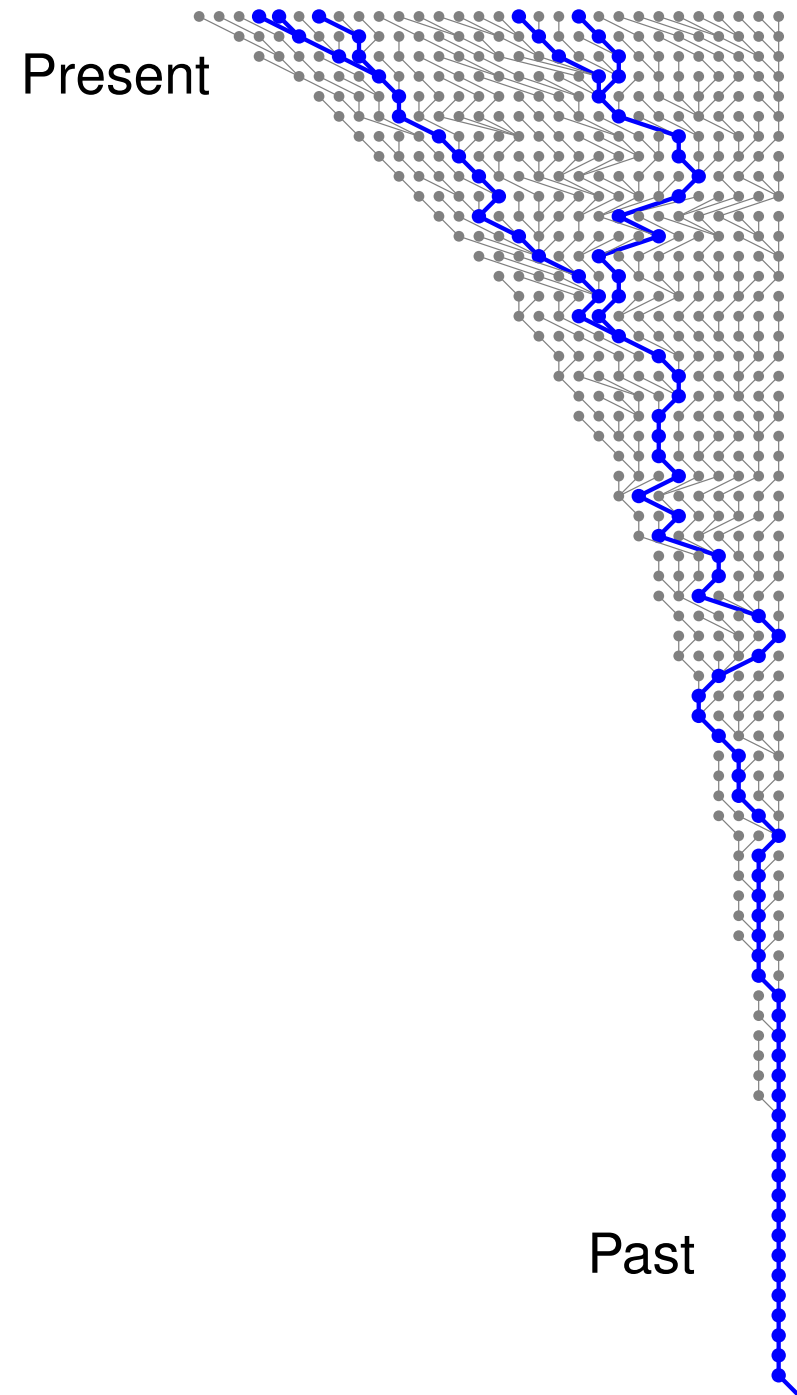
Present

For constant population size we found

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

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

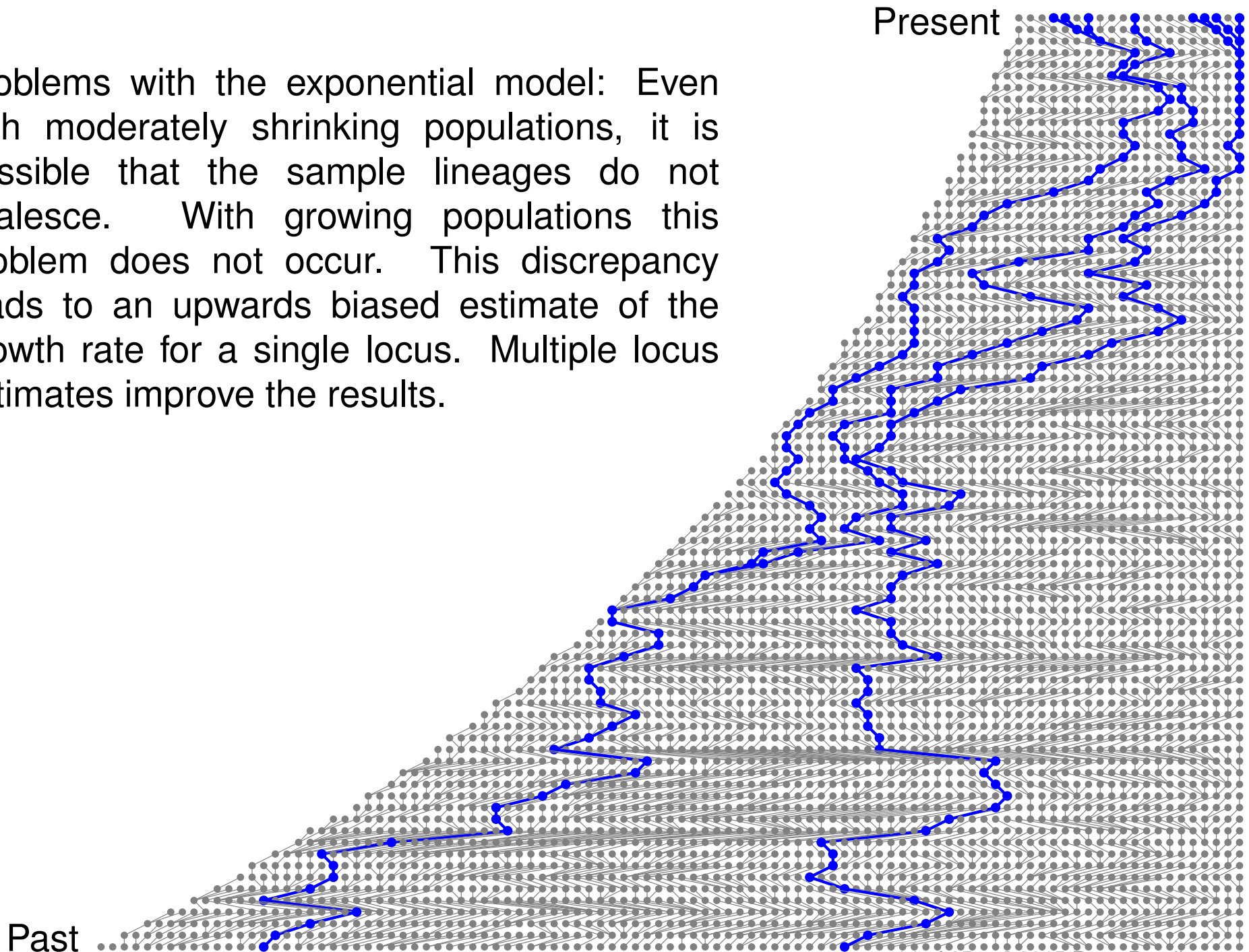
$$p(G|\Theta_0, g) = \prod_j e^{-(t_j - t_{j-1}) \frac{k(k-1)}{\Theta_0 e^{-gt}}} \frac{2}{\Theta_0 e^{-gt}}$$



Extensions of the basic coalescent

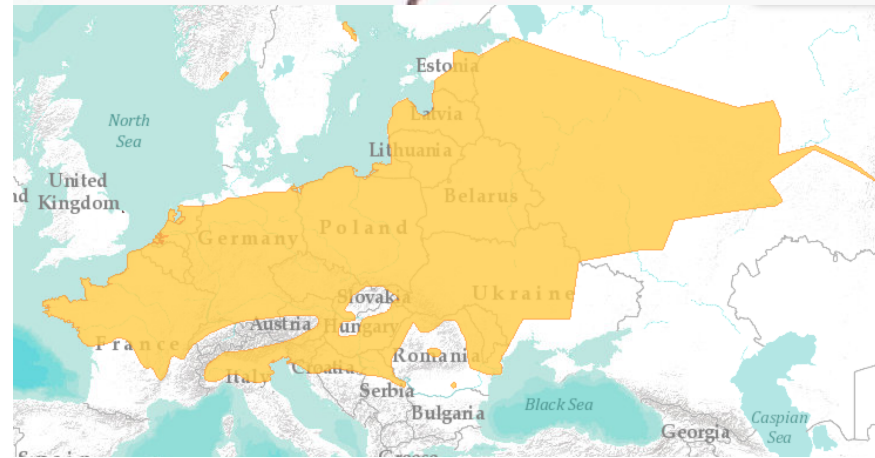
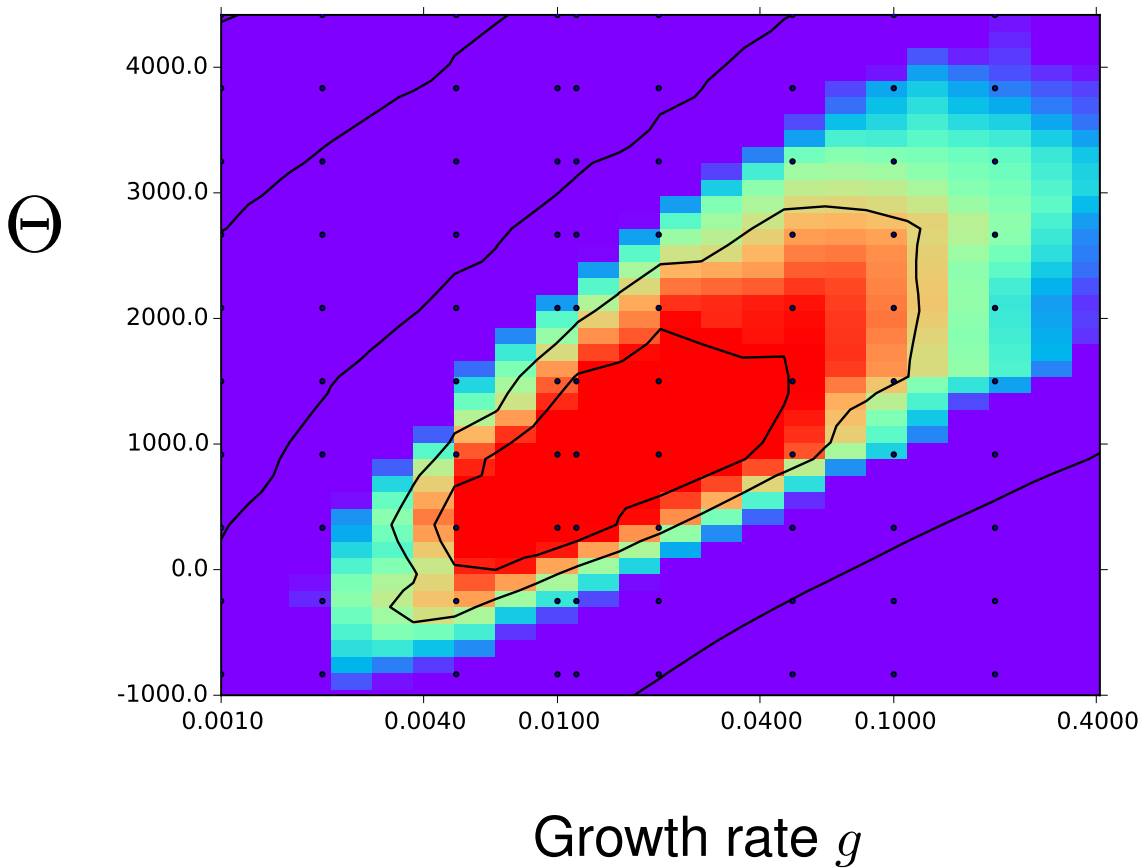
Growth

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.

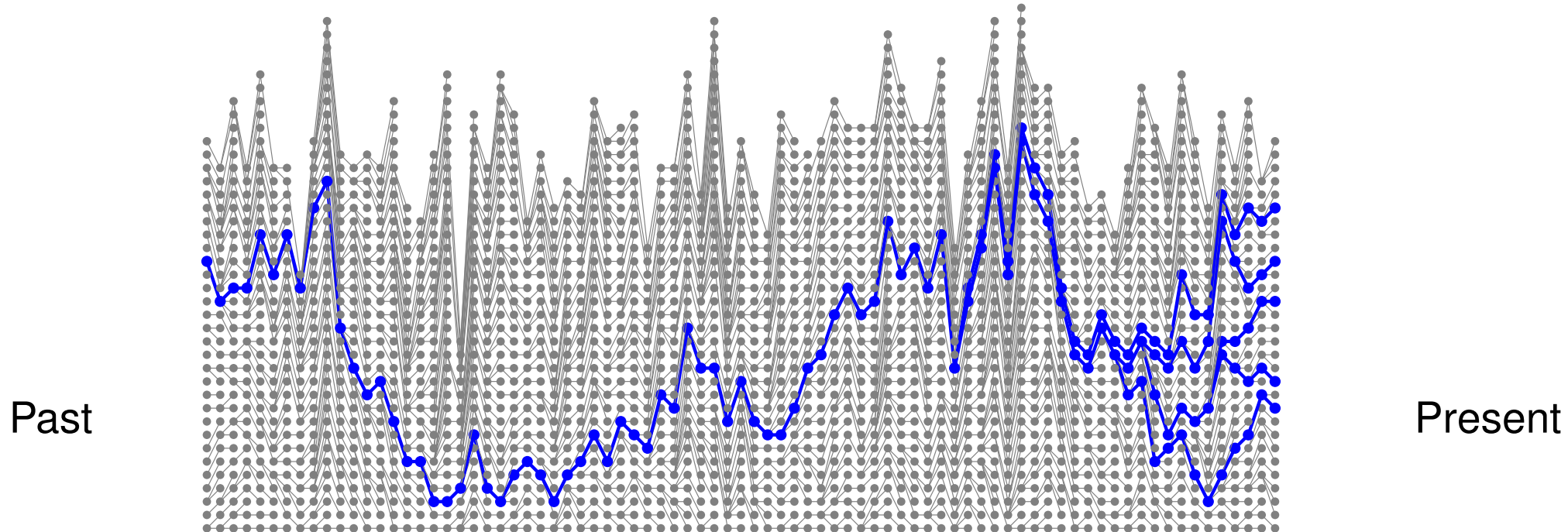




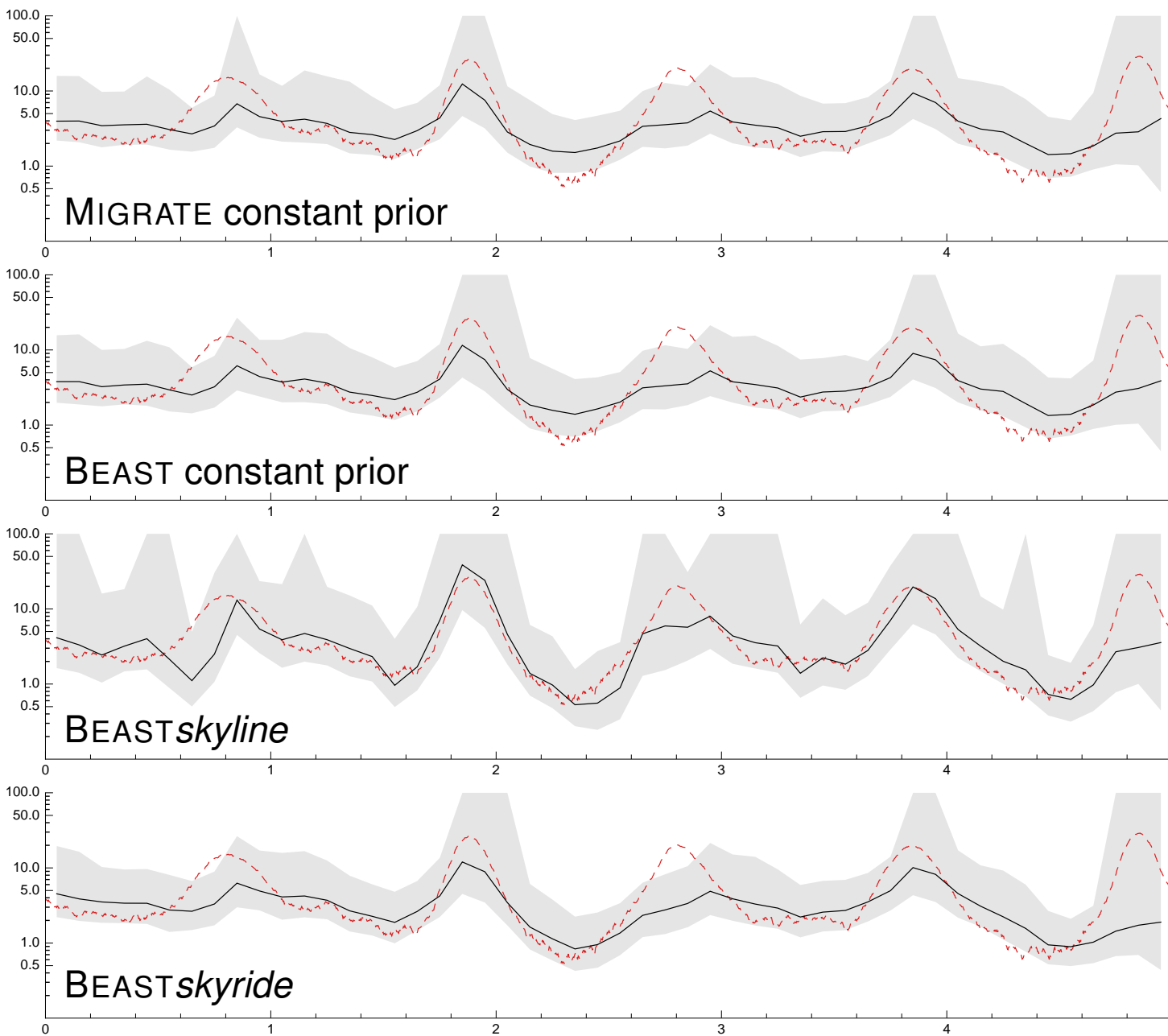
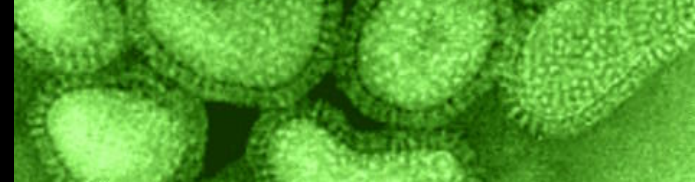
Expansion of *Pelophylax lessonae* in Europe



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.



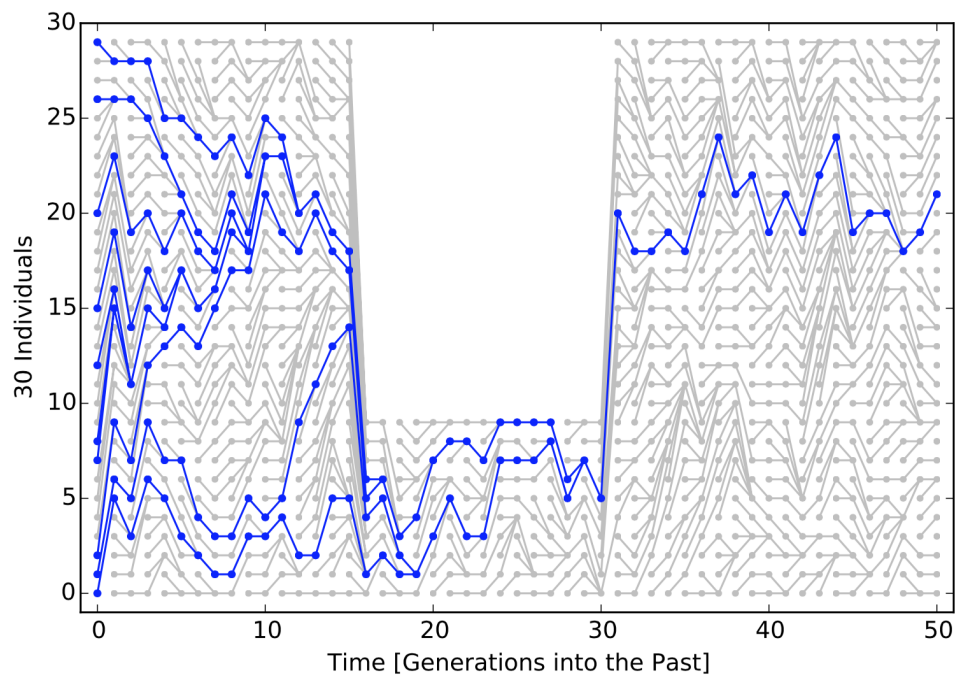
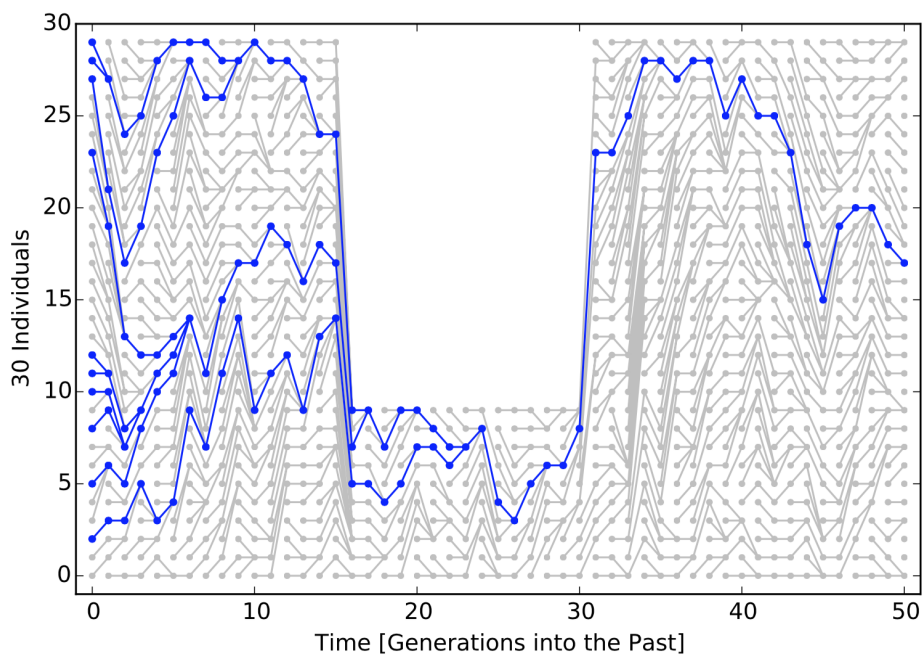
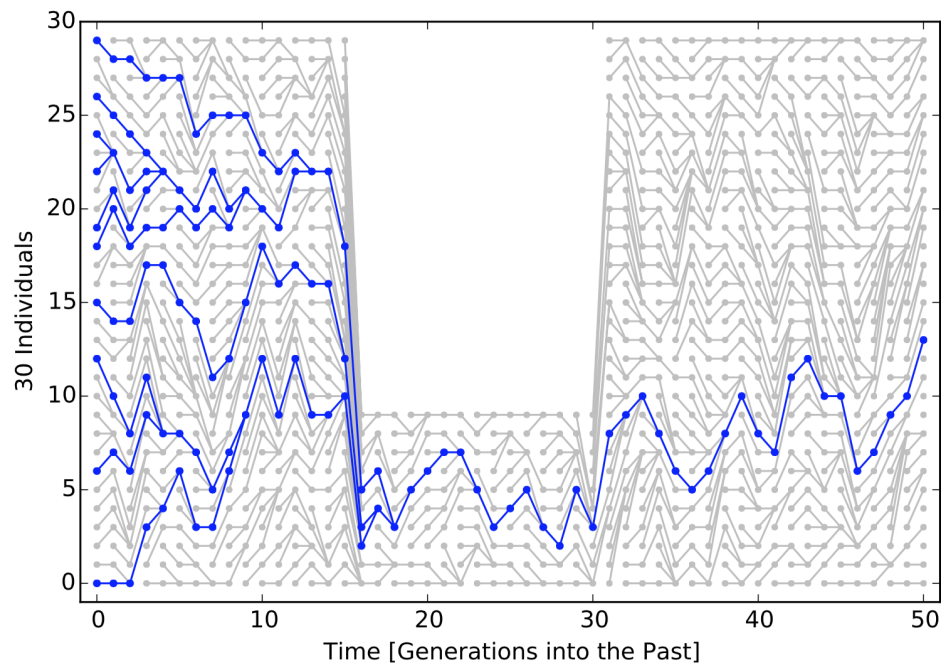
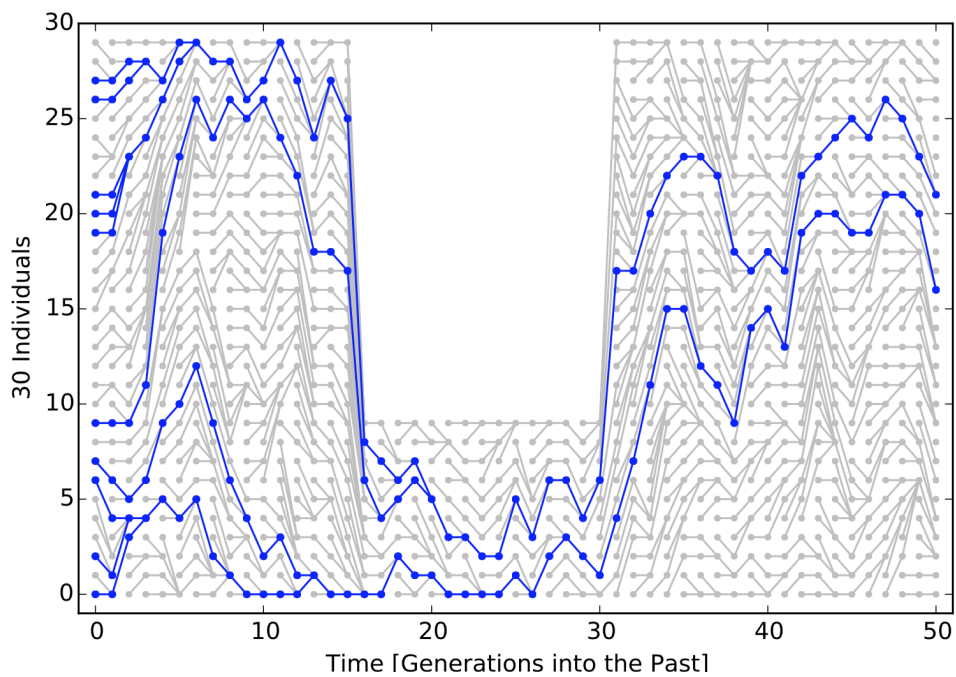
Extensions of the basic coalescent



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.

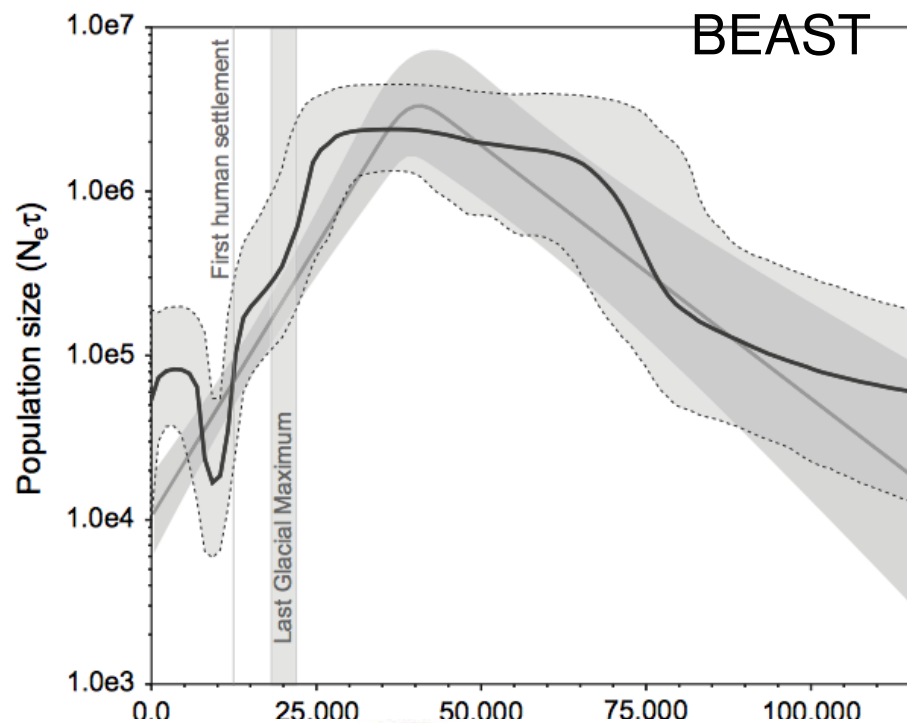
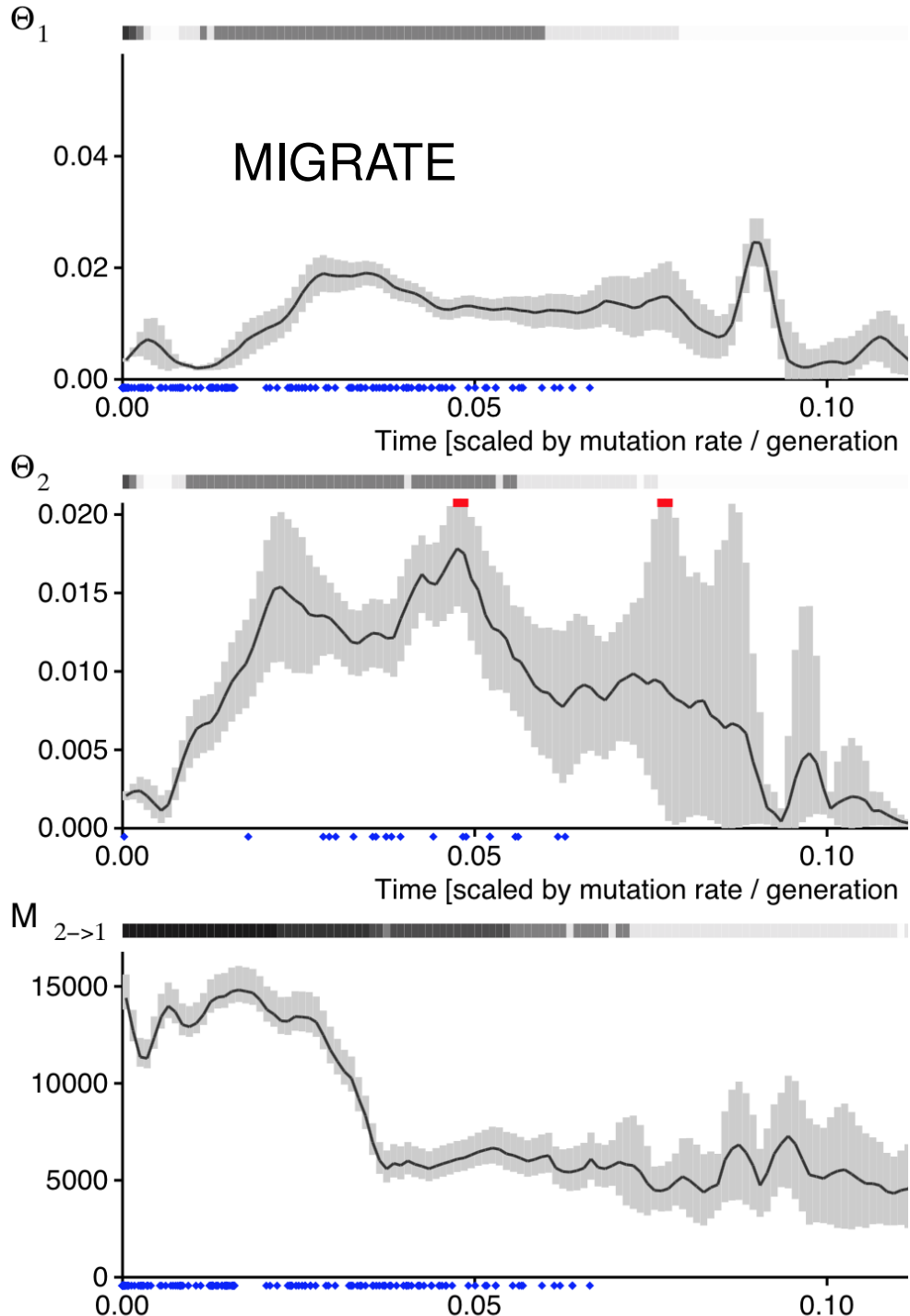
Extensions of the basic coalescent

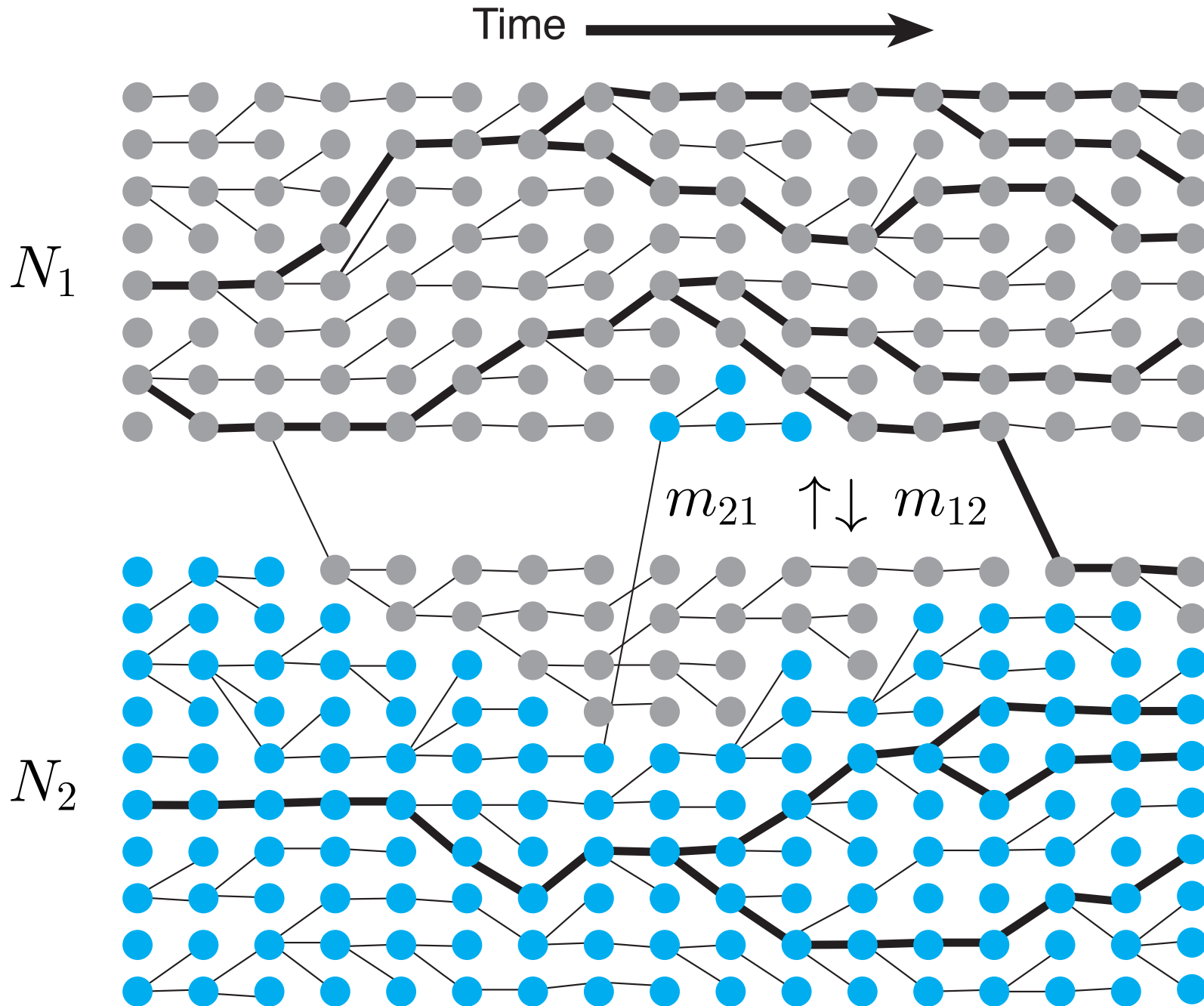
Bottlenecks

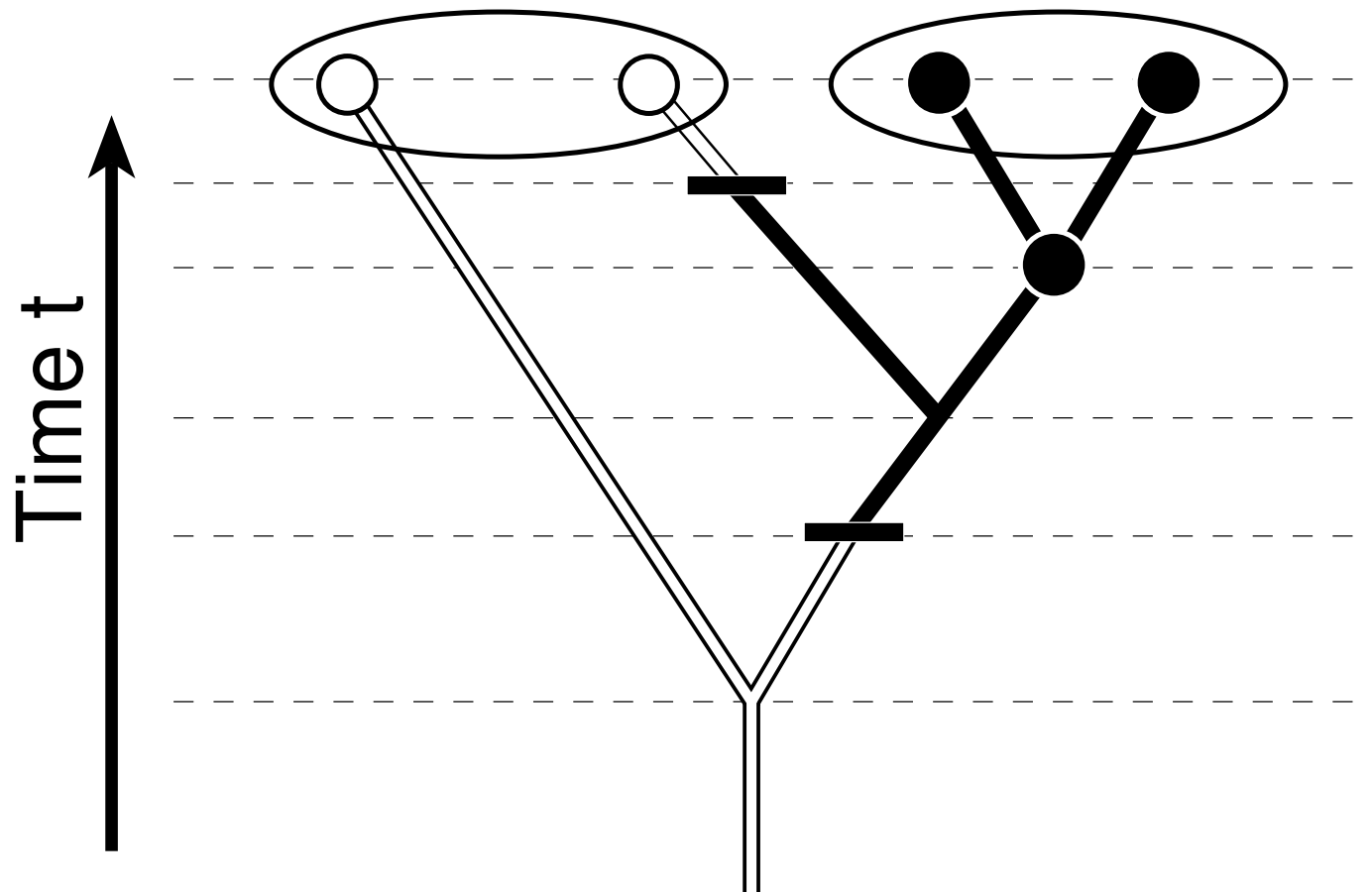


Extensions of the basic coalescent

Skyline plots







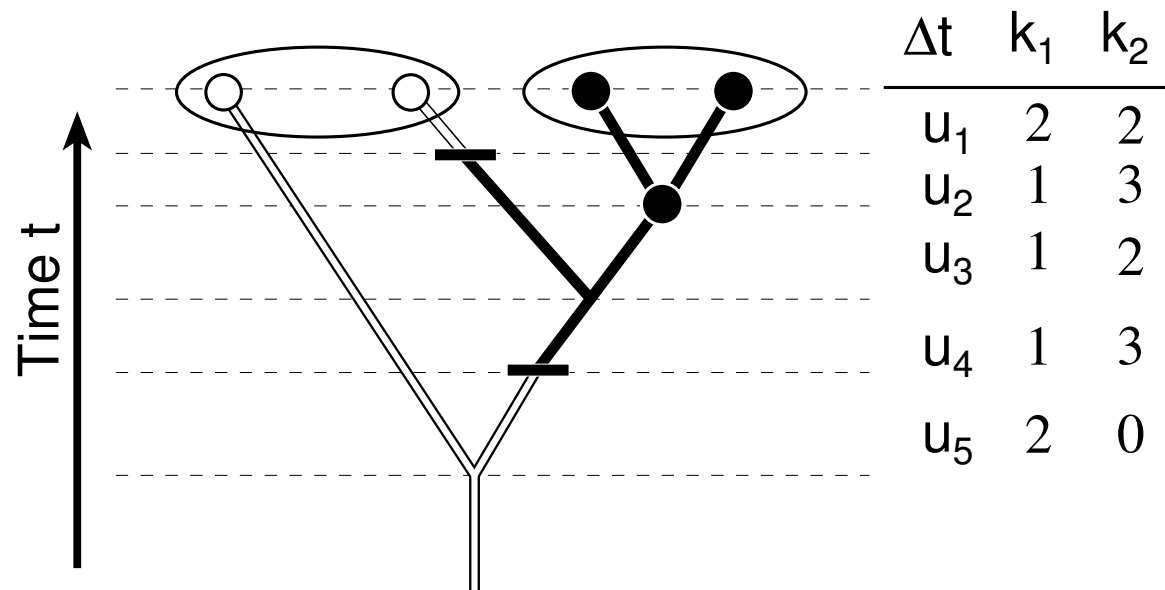
Δt	k_1	k_2
u_1	2	2
u_2	1	3
u_3	1	2
u_4	1	3
u_5	2	0

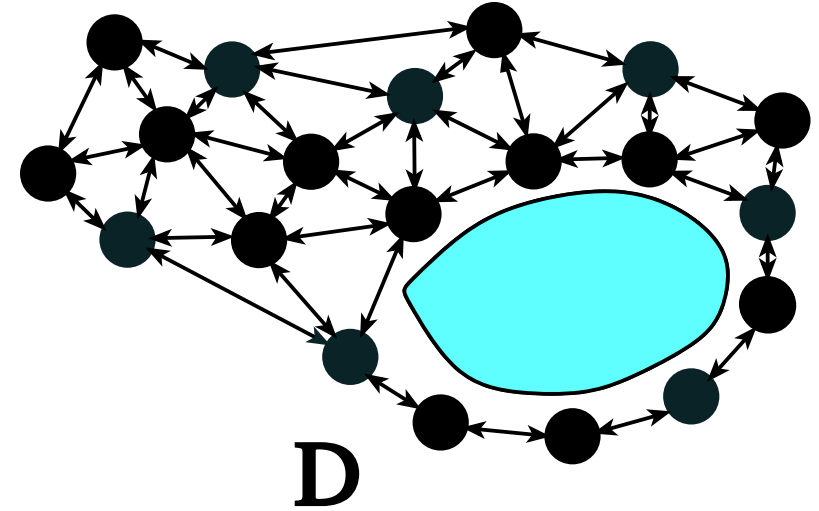
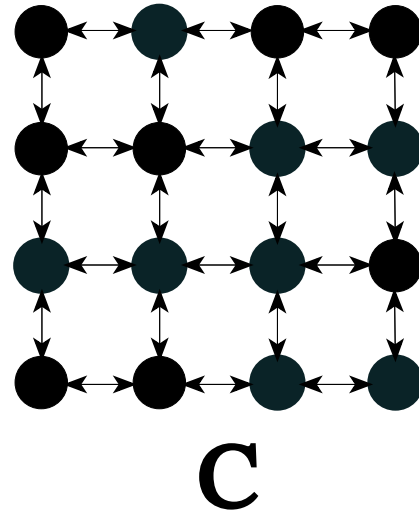
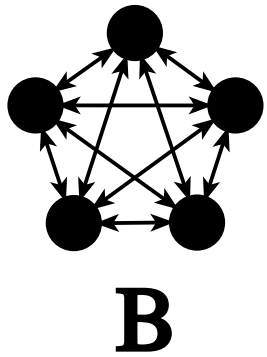
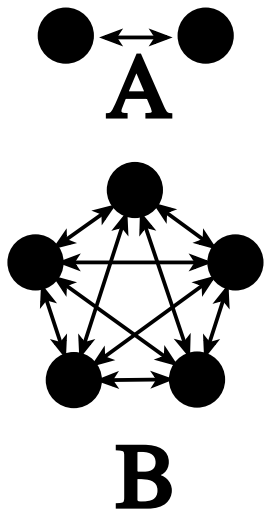
The single population coalescence rate is

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

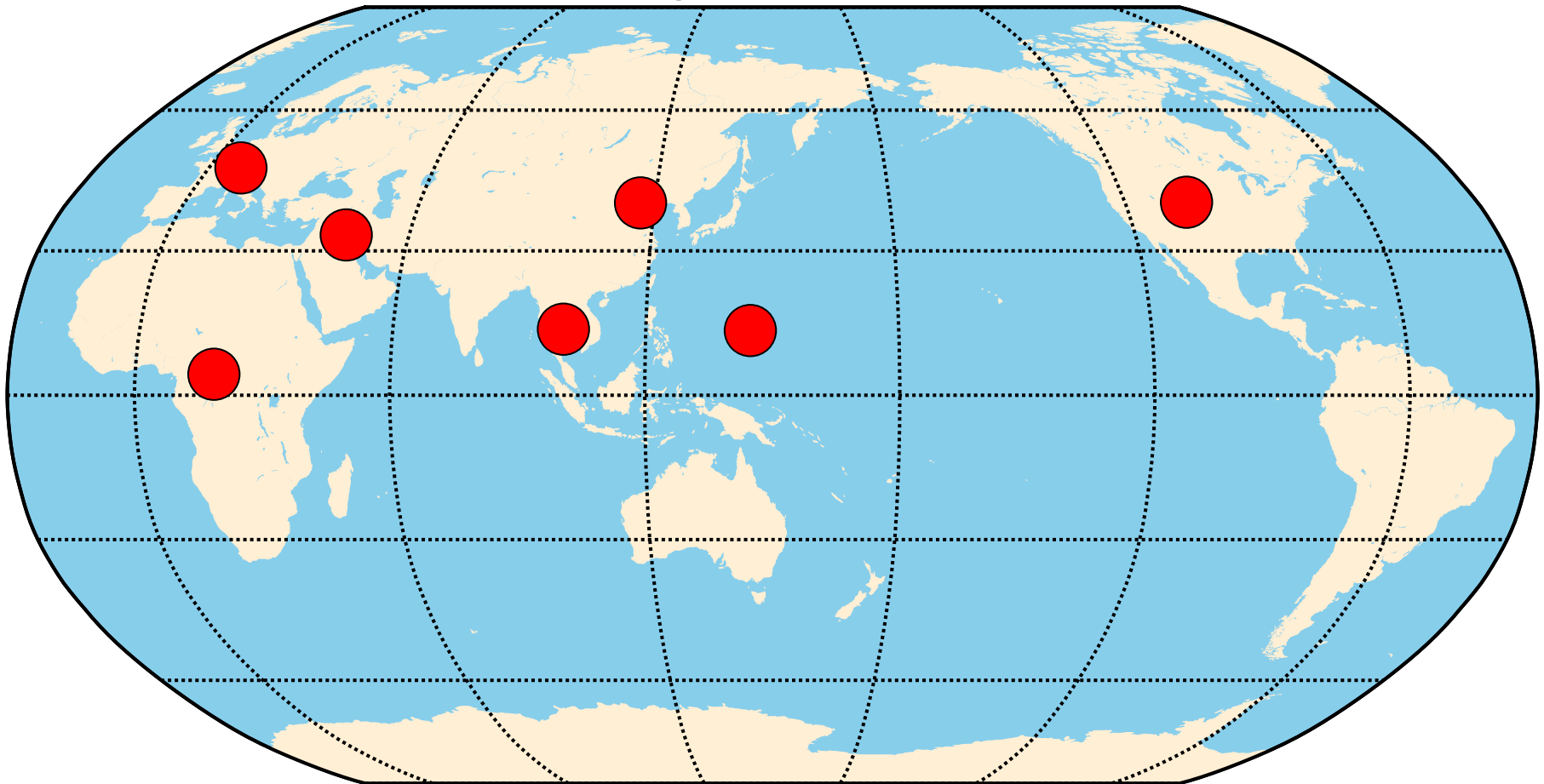
Changes for two populations to

$$\frac{k_1(k_1-1)}{\Theta_1} + \frac{k_2(k_2-1)}{\Theta_2} + k_1M_{2,1} + k_2M_{1,2}$$



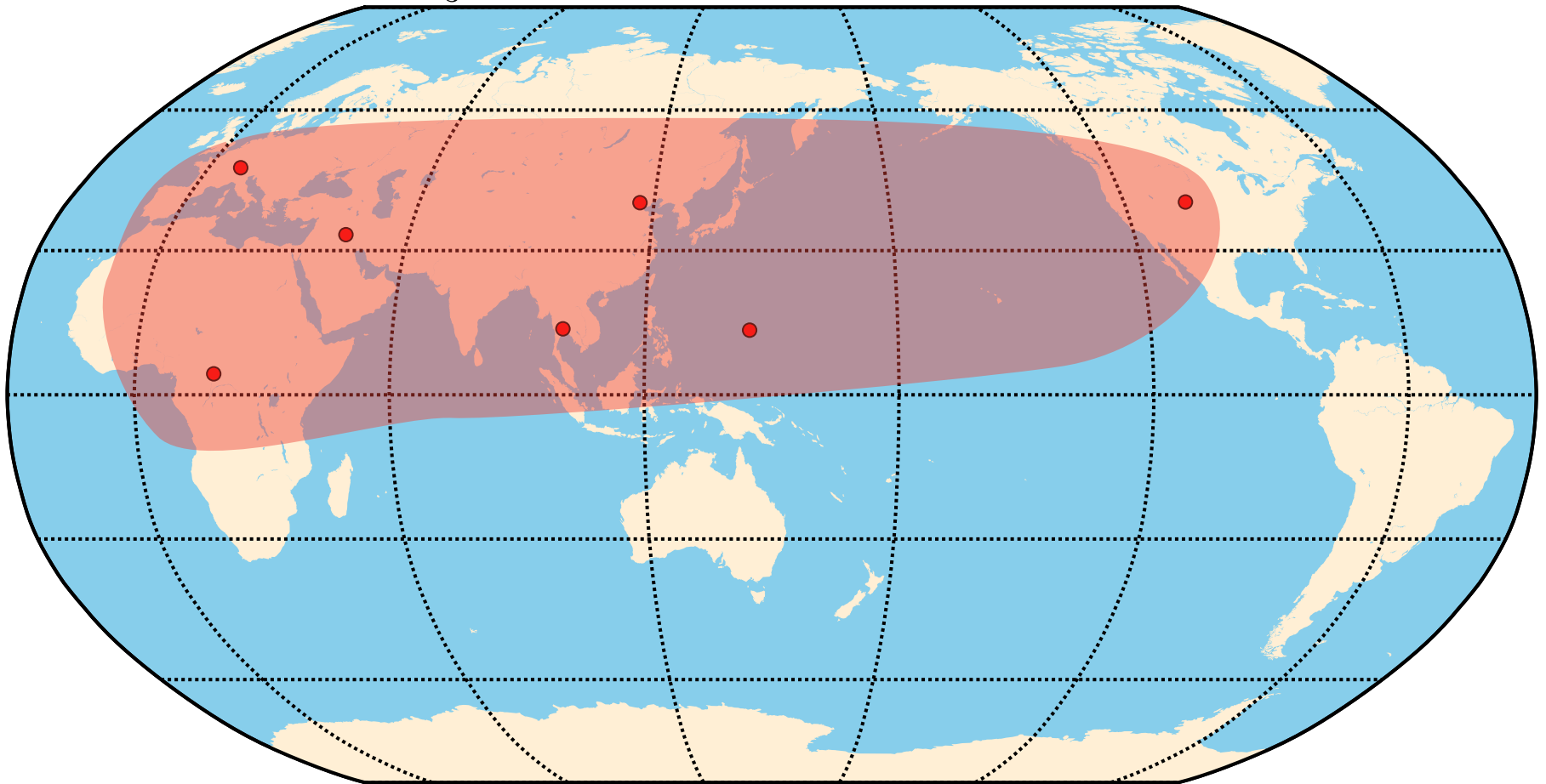


Locations of samples [377 microsatellites]

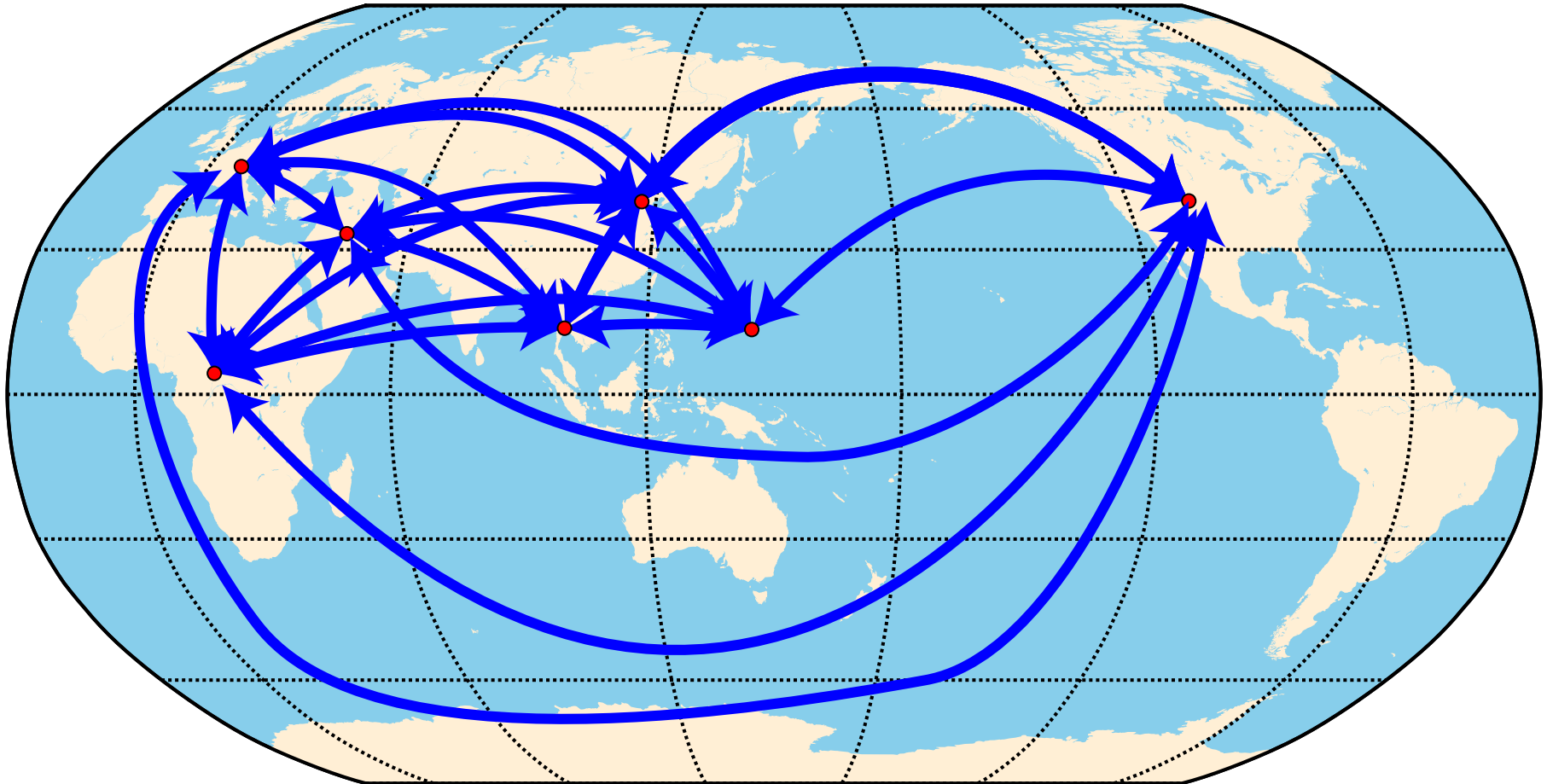


A total of 70 individuals from 7 populations analyzed for 377 microsatellite loci:
Mutation model is Brownian motion approximation to the single-step mutation model

H_3 : One panmictic population

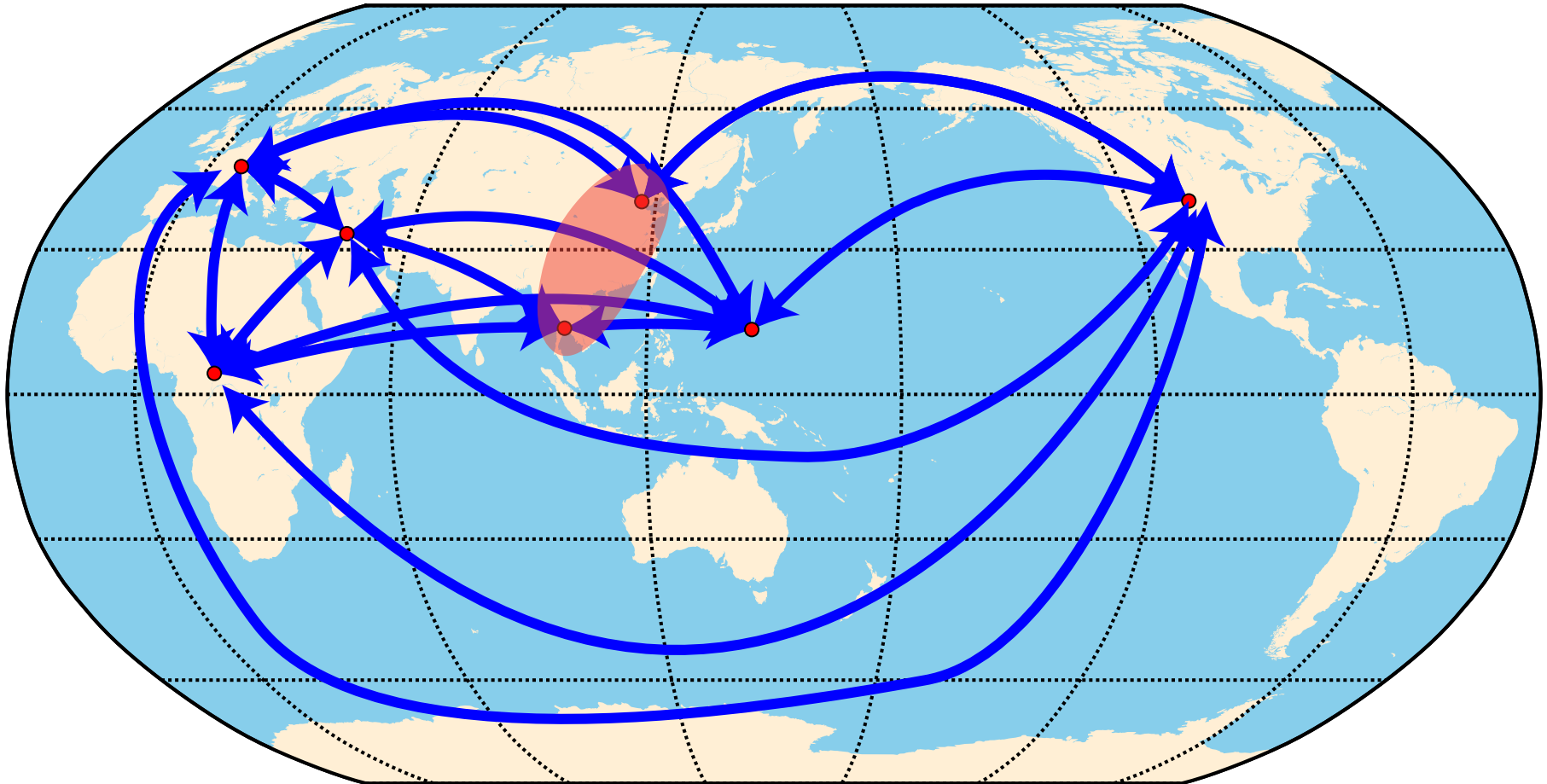


H_2 : Tangled mess

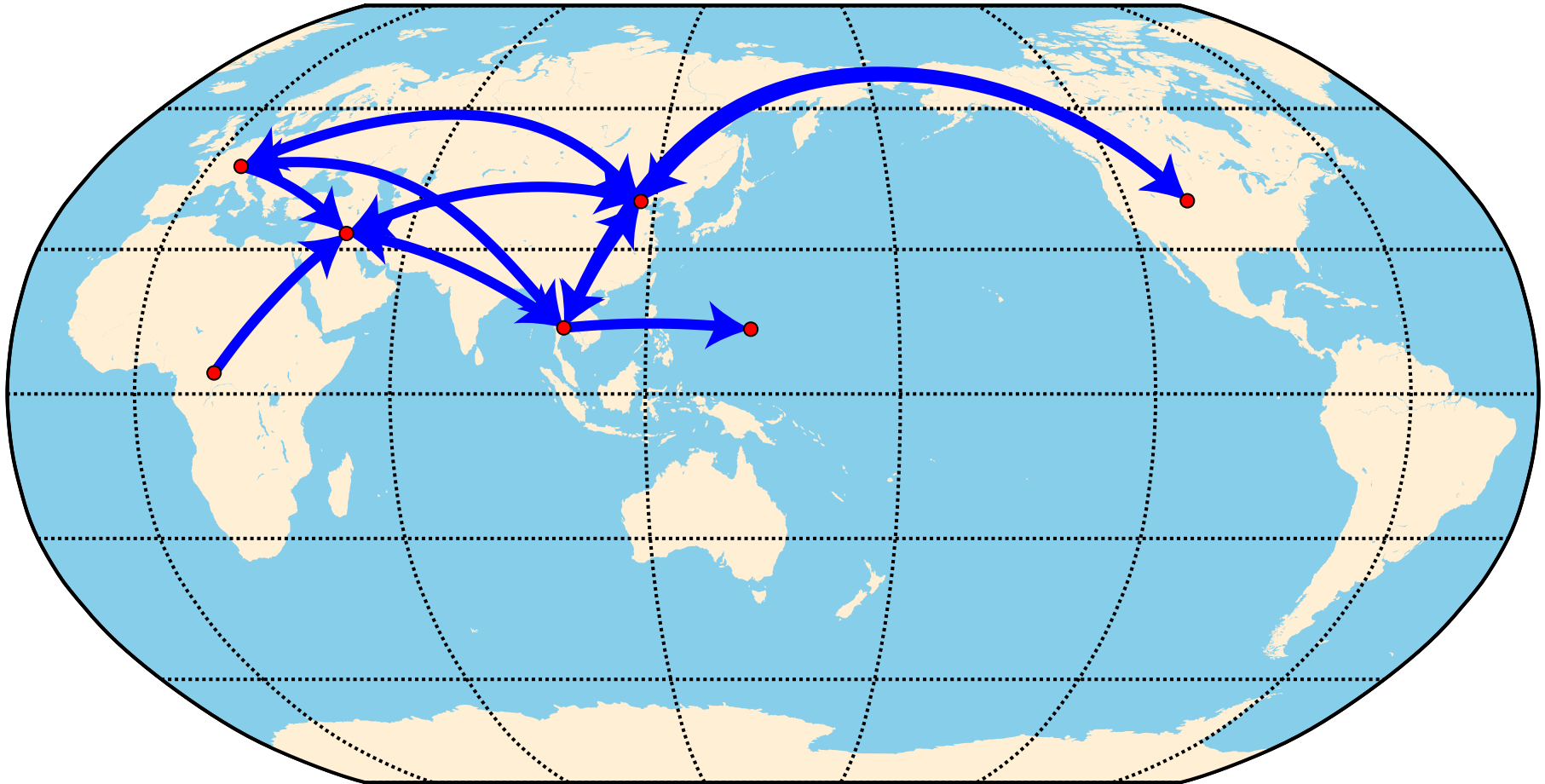


Somewhat less

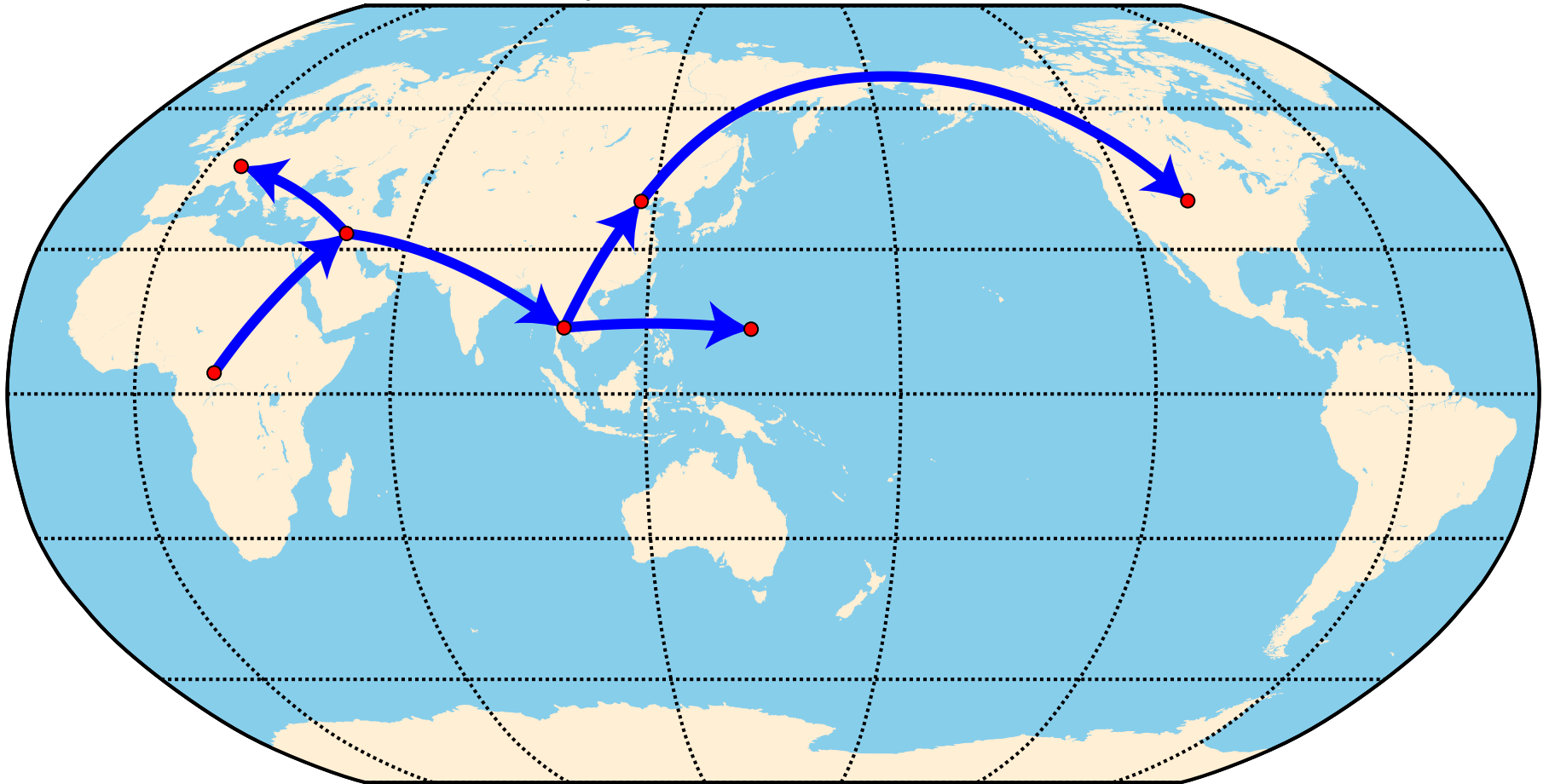
H_4 : $\sqrt{\text{Tangled mess}}$



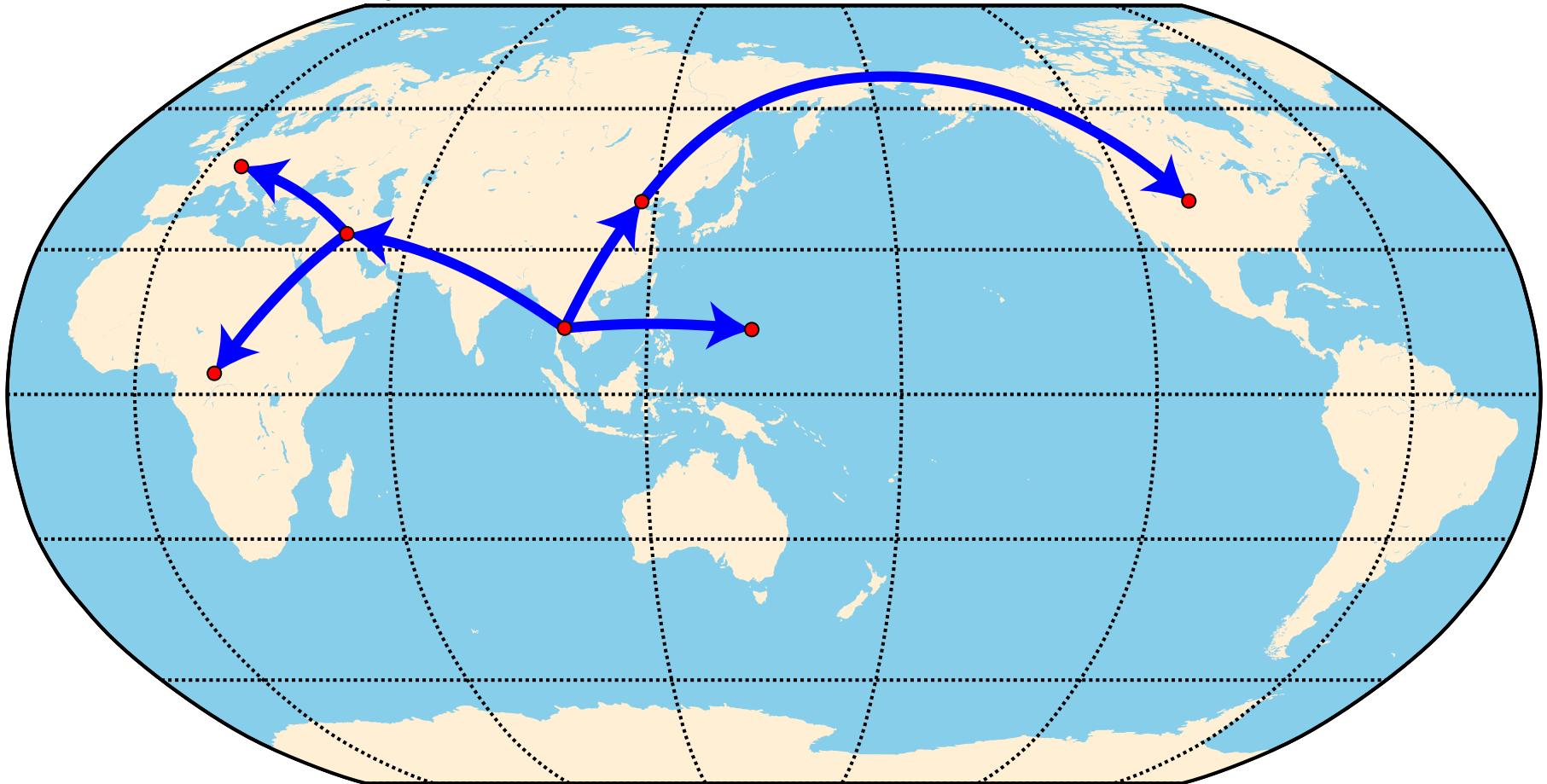
H_1 : Out of Africa, indecision anywhere else



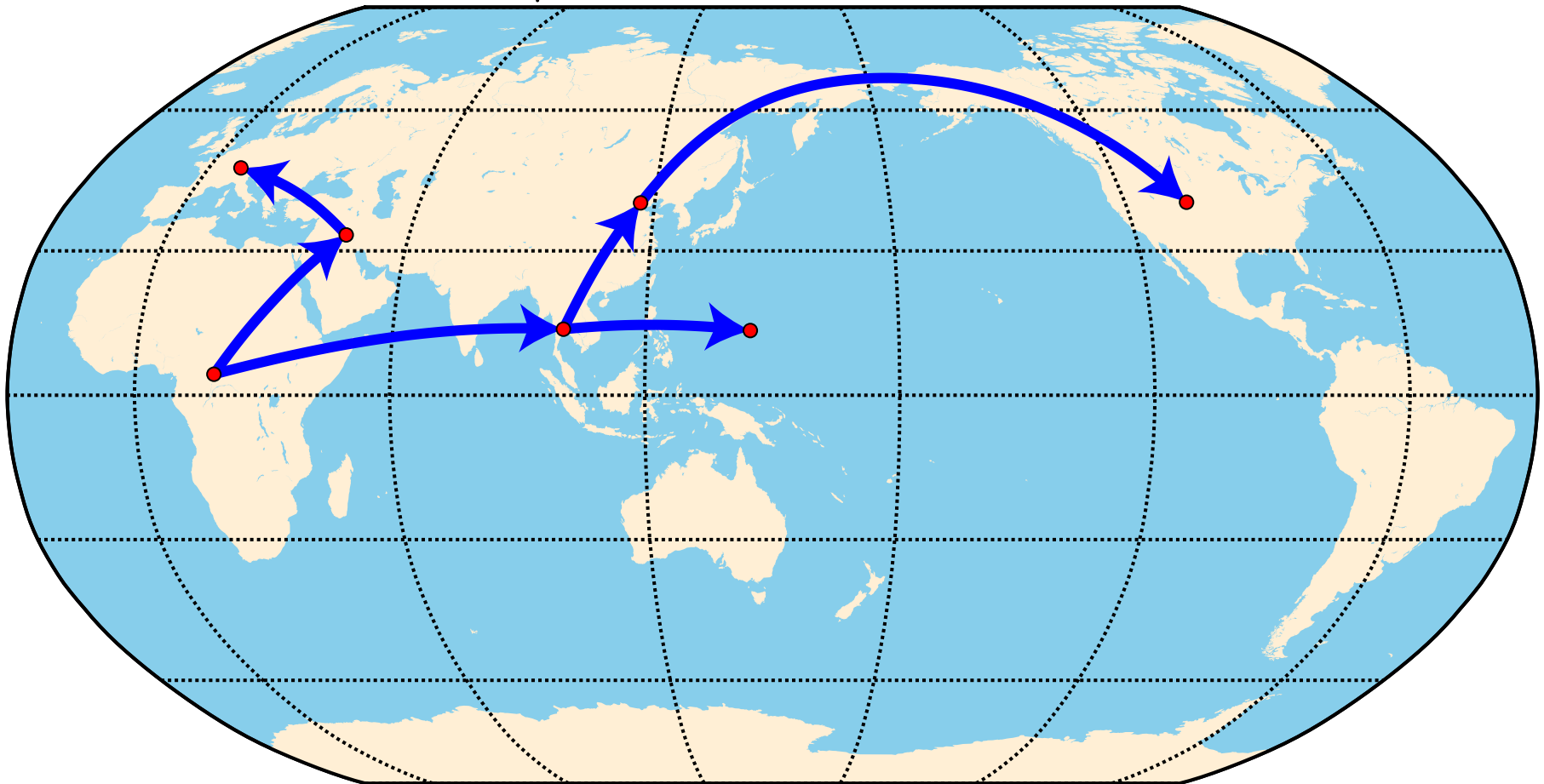
H_5 : Minimal model

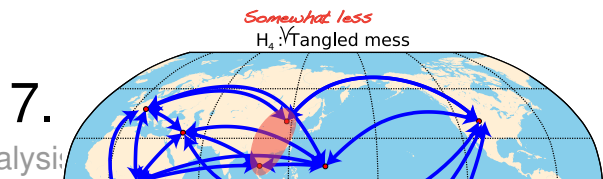
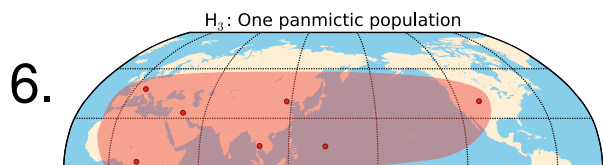
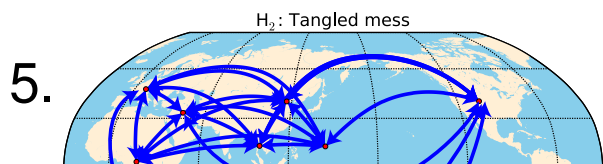
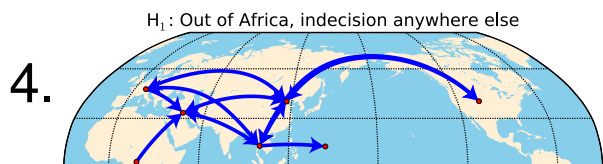
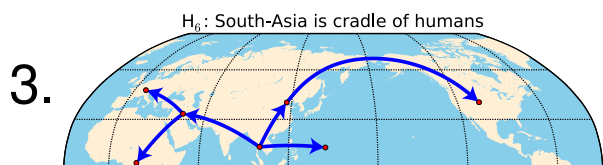
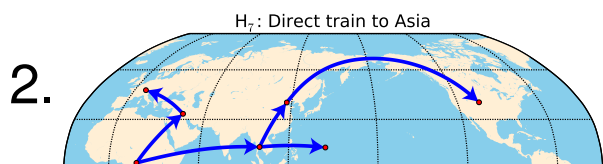
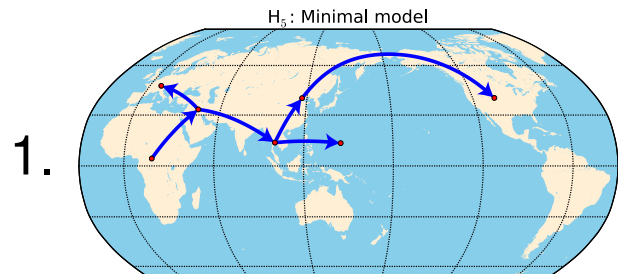


H_6 : South-Asia is cradle of humans



H_7 : Direct train to Asia

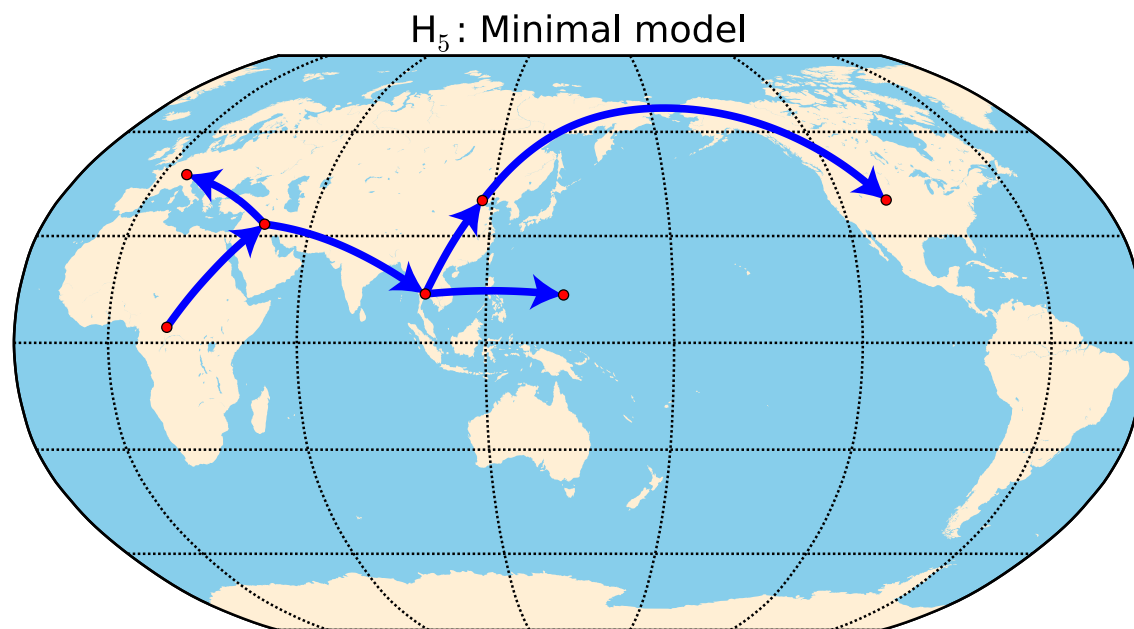




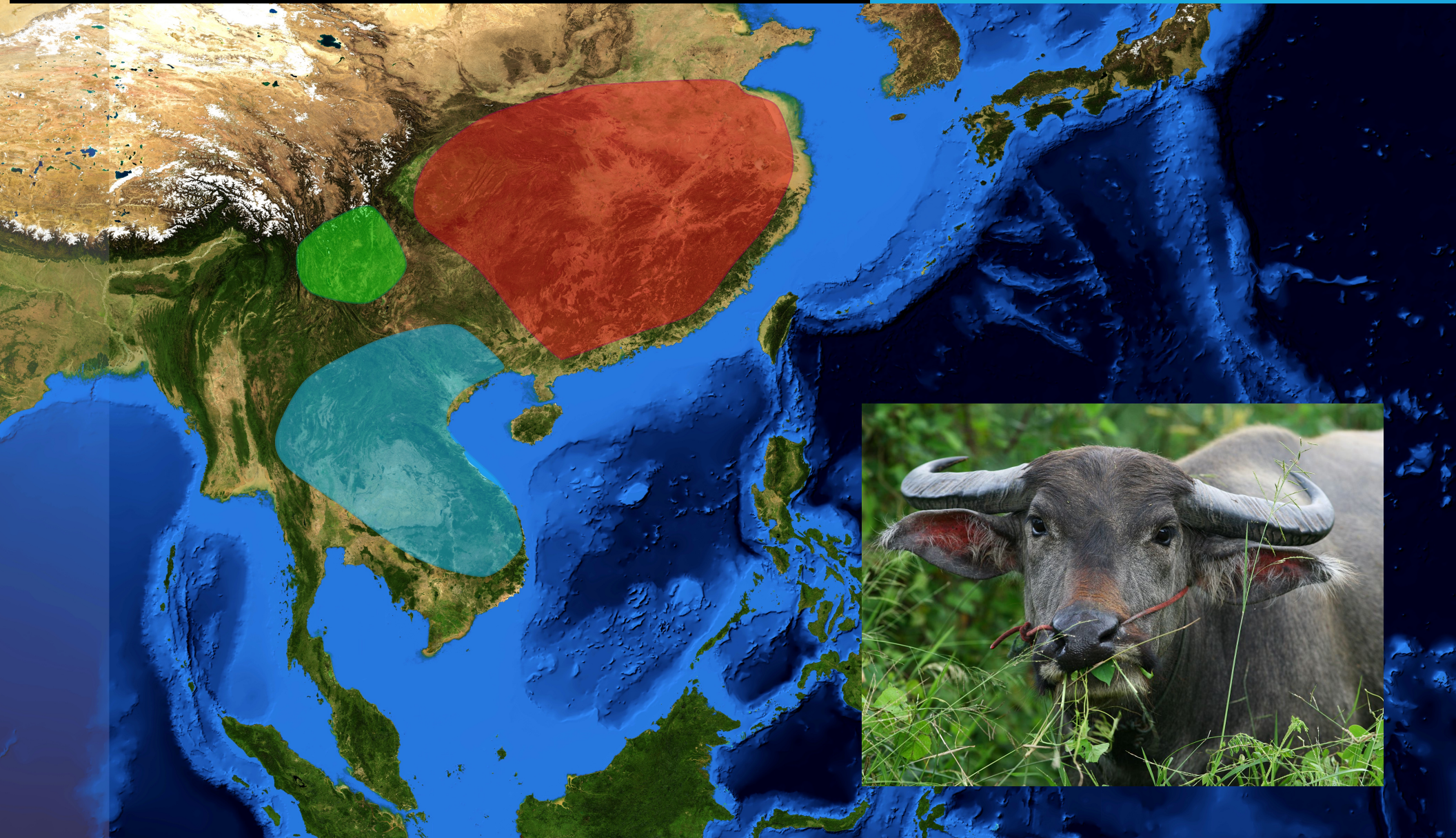
Model order and probability using Bayes factors

all other models: 0.0

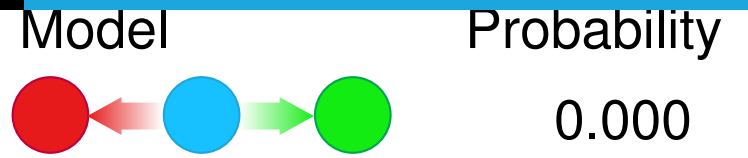
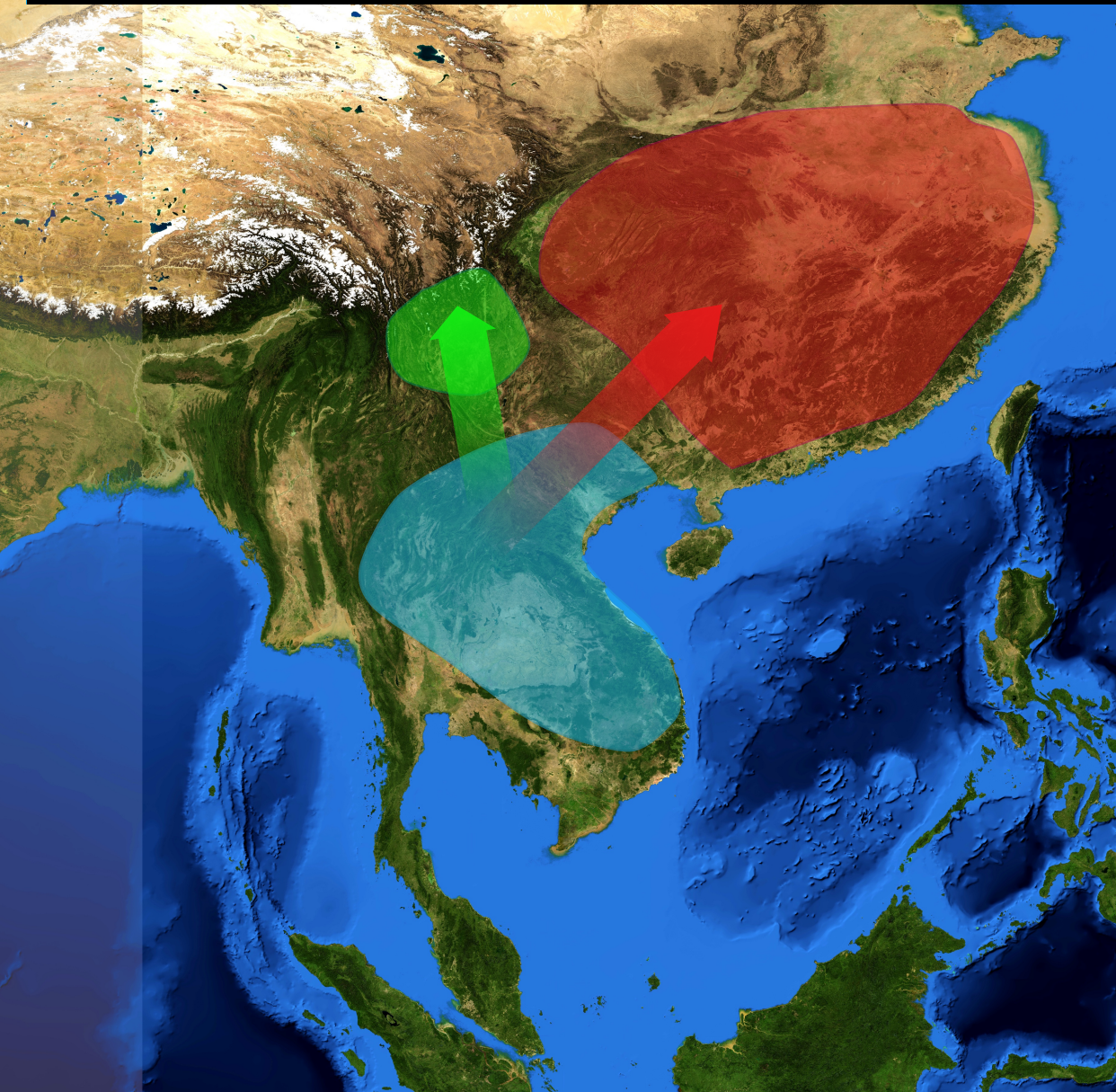
Minimal model 1.0



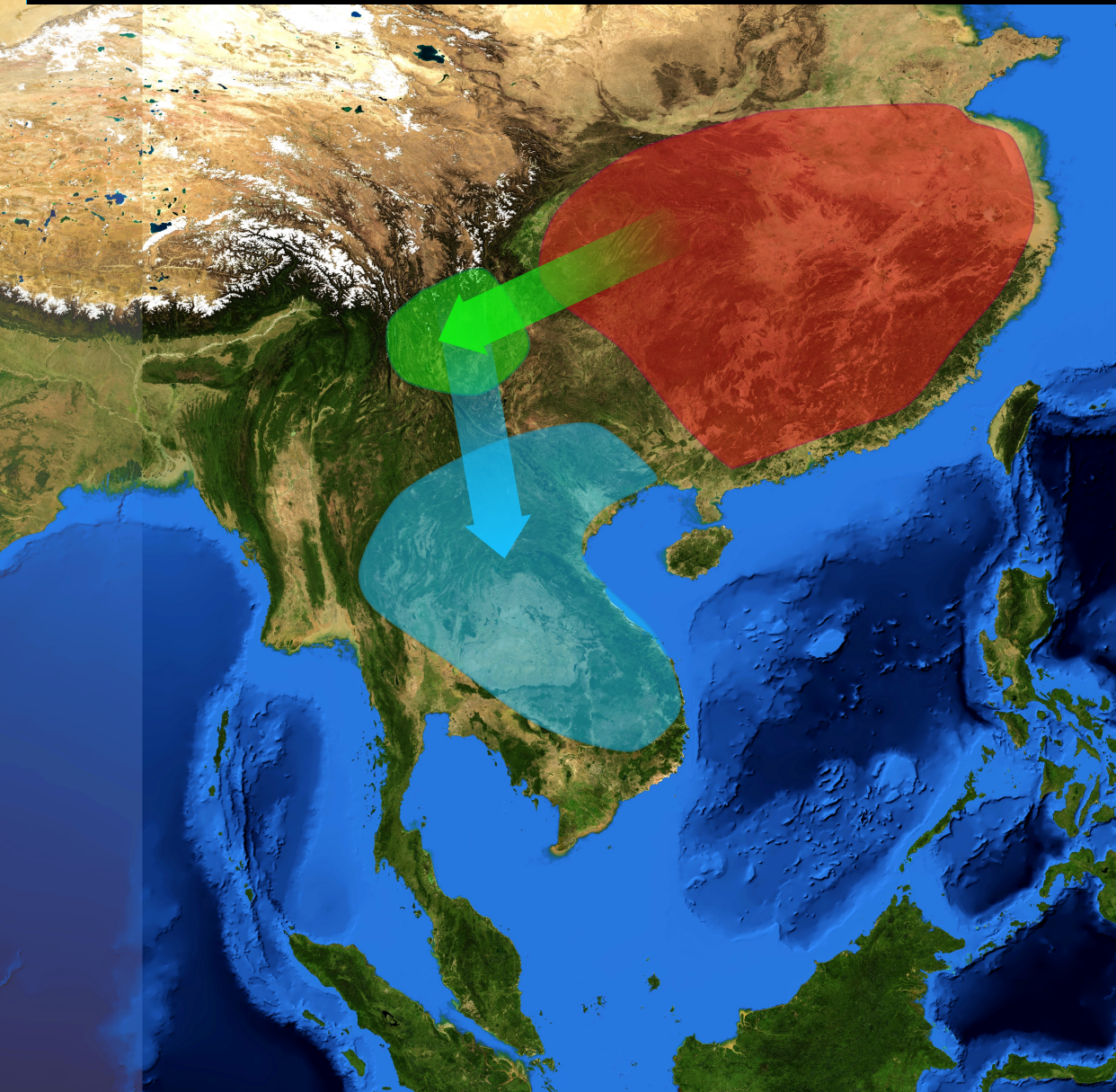
Bayesian Model Comparison









Bayesian Model Comparison



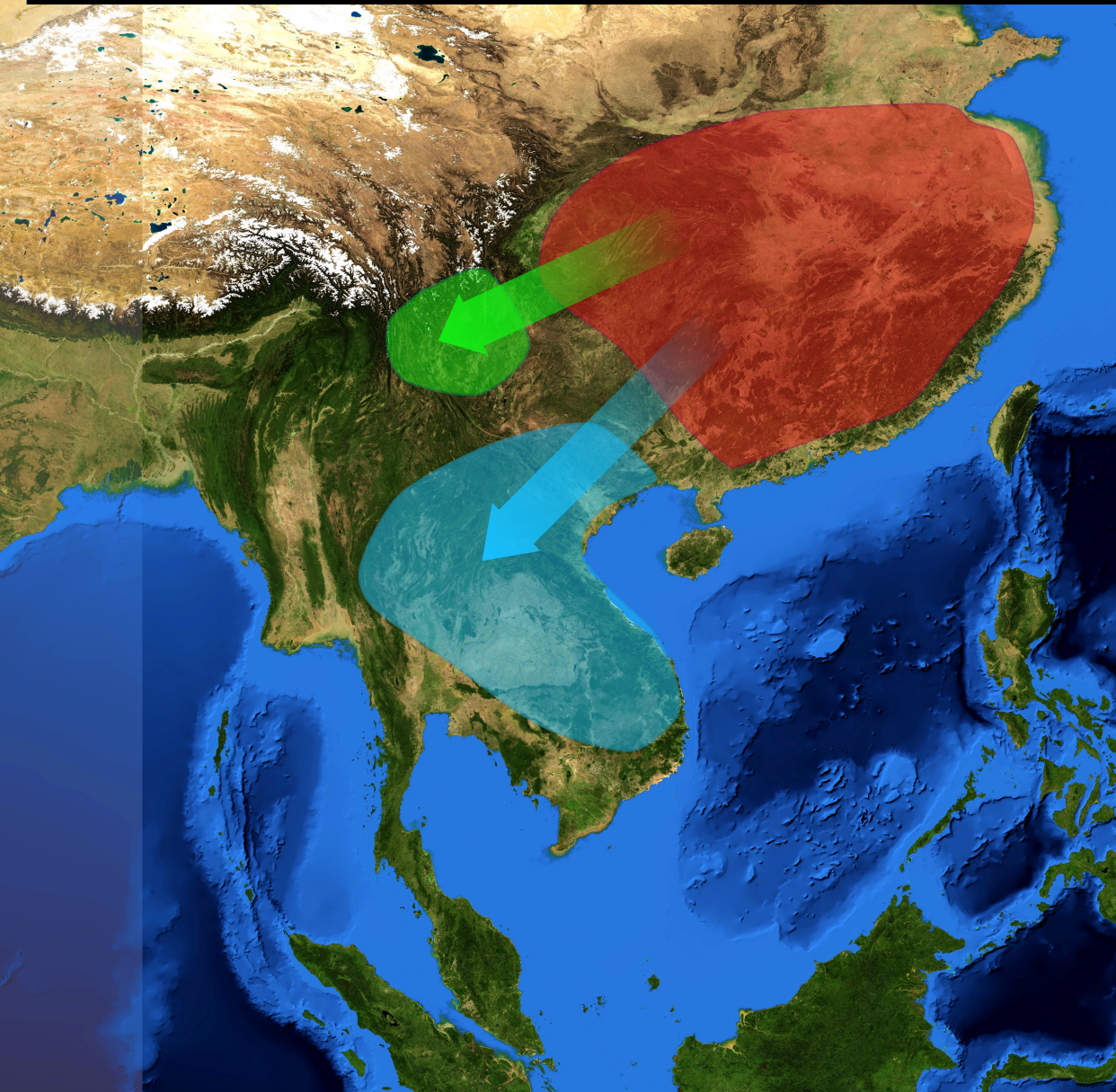
Bayesian Model Comparison












Model	Probability
 ←  → 	0.000
 →  → 	0.002



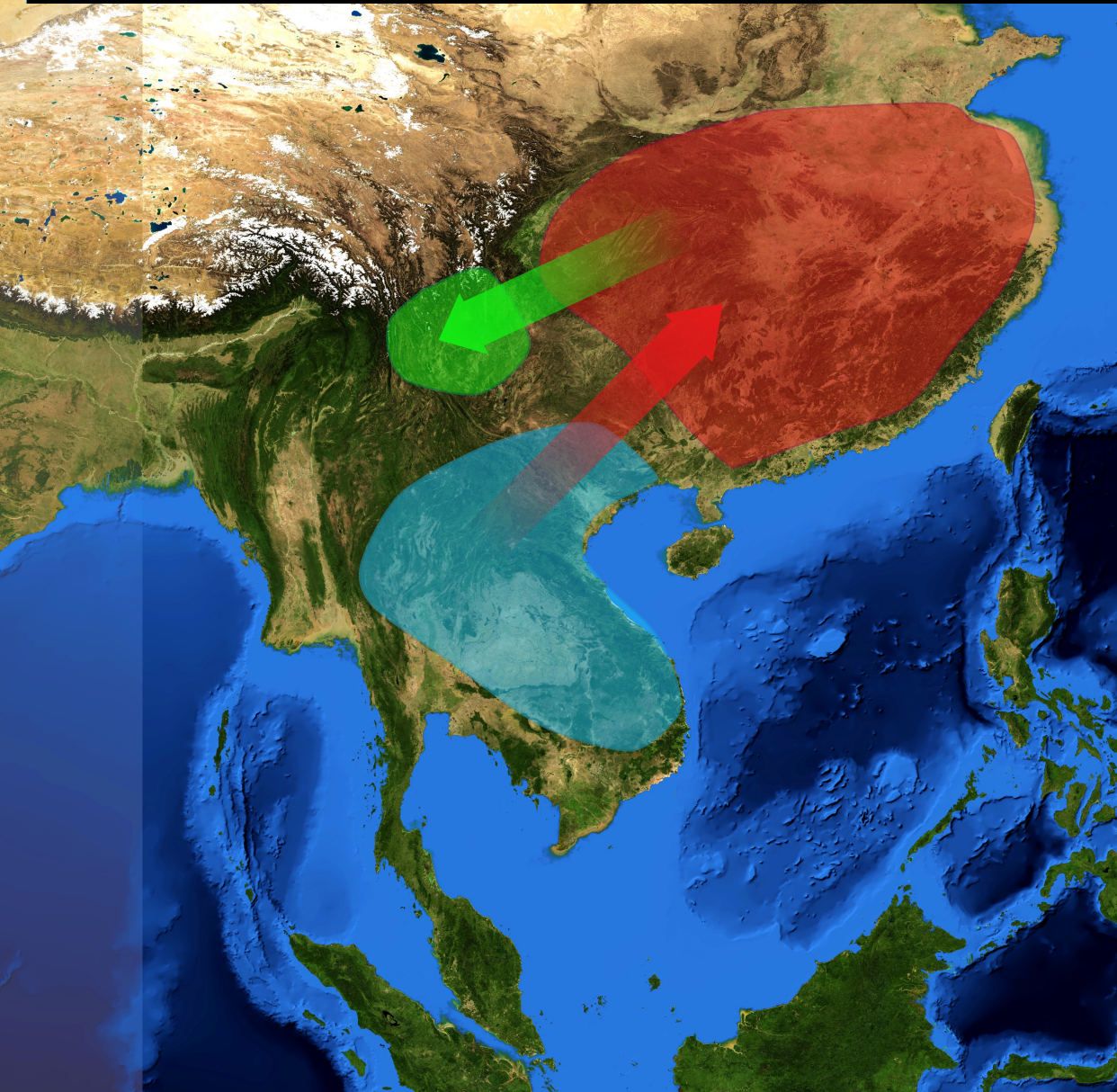
Bayesian Model Comparison




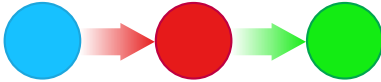


Model	Probability
 ←  → 	0.000
 →  → 	0.002
 ←  → 	0.008

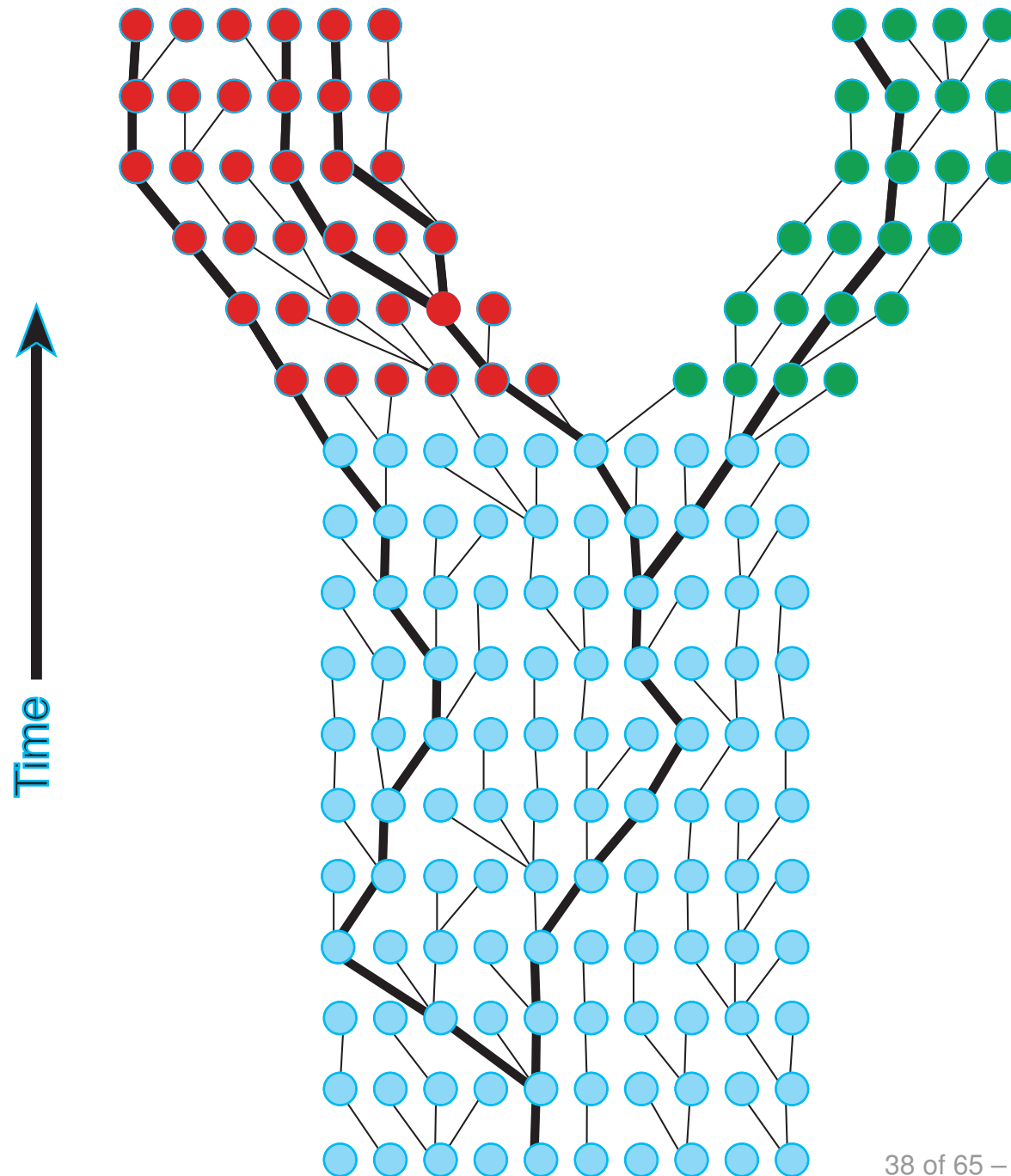


Bayesian Model Comparison

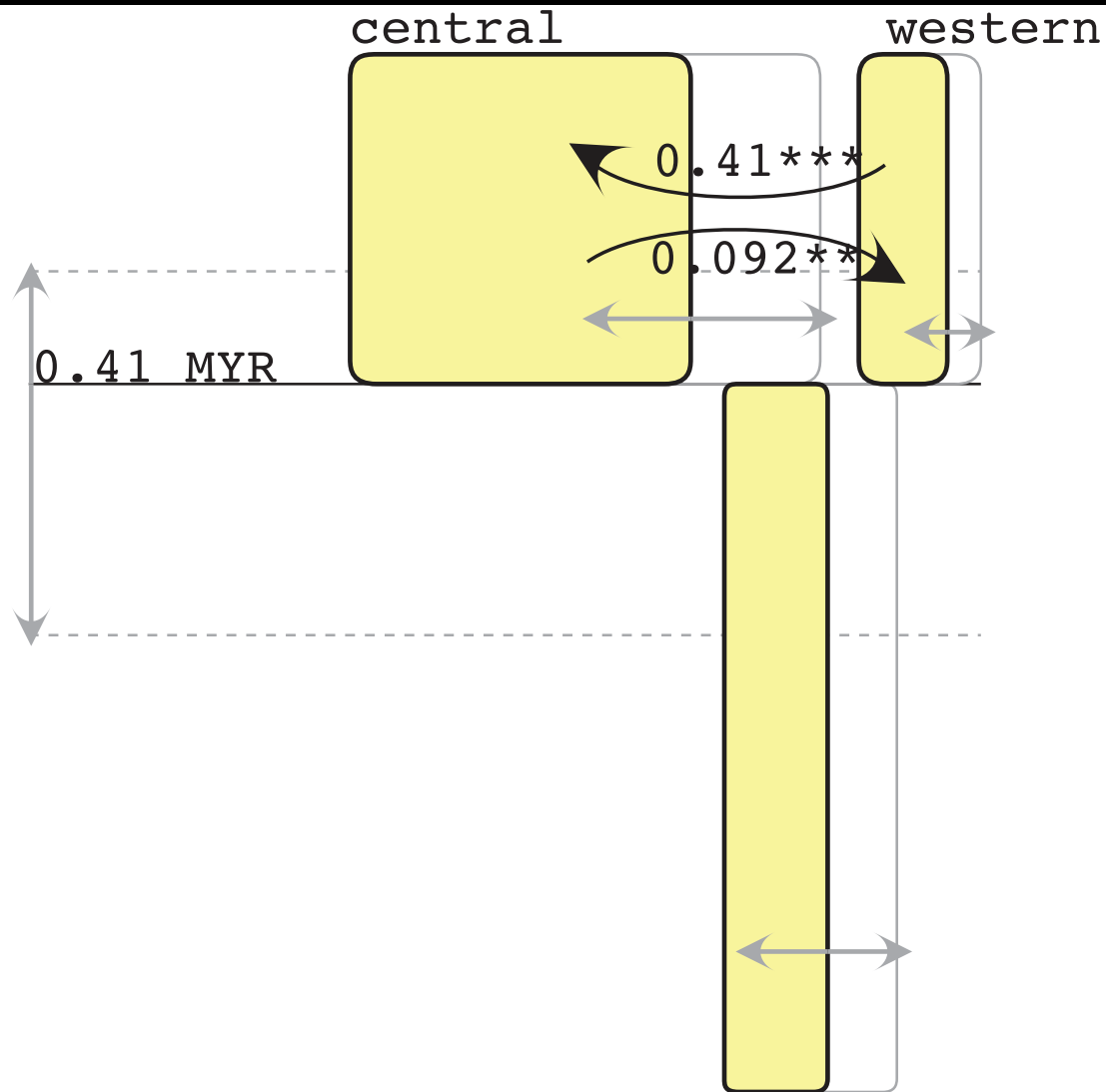


Model	Probability
	0.000
	0.002
	0.008
	0.990





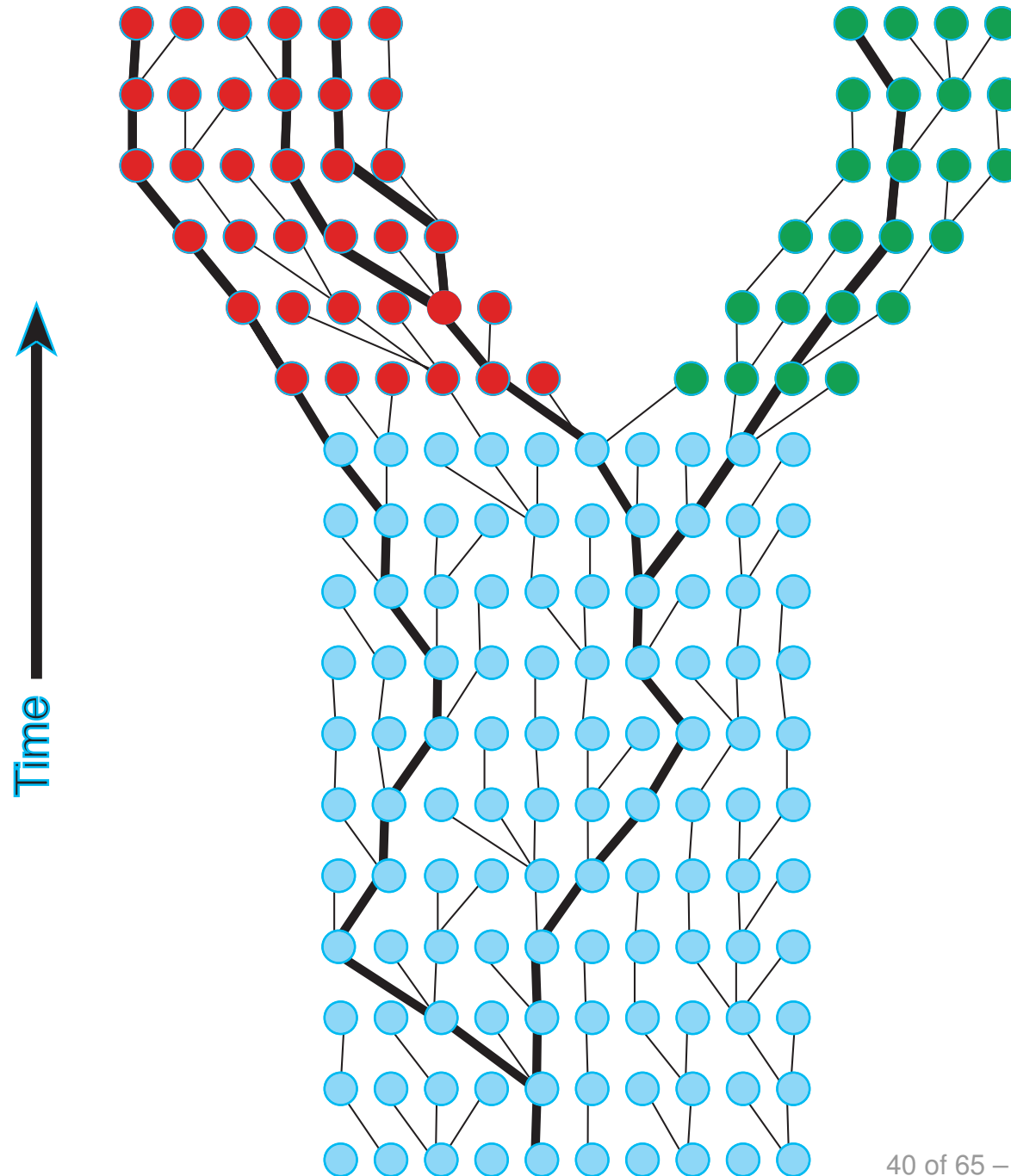
Population splitting



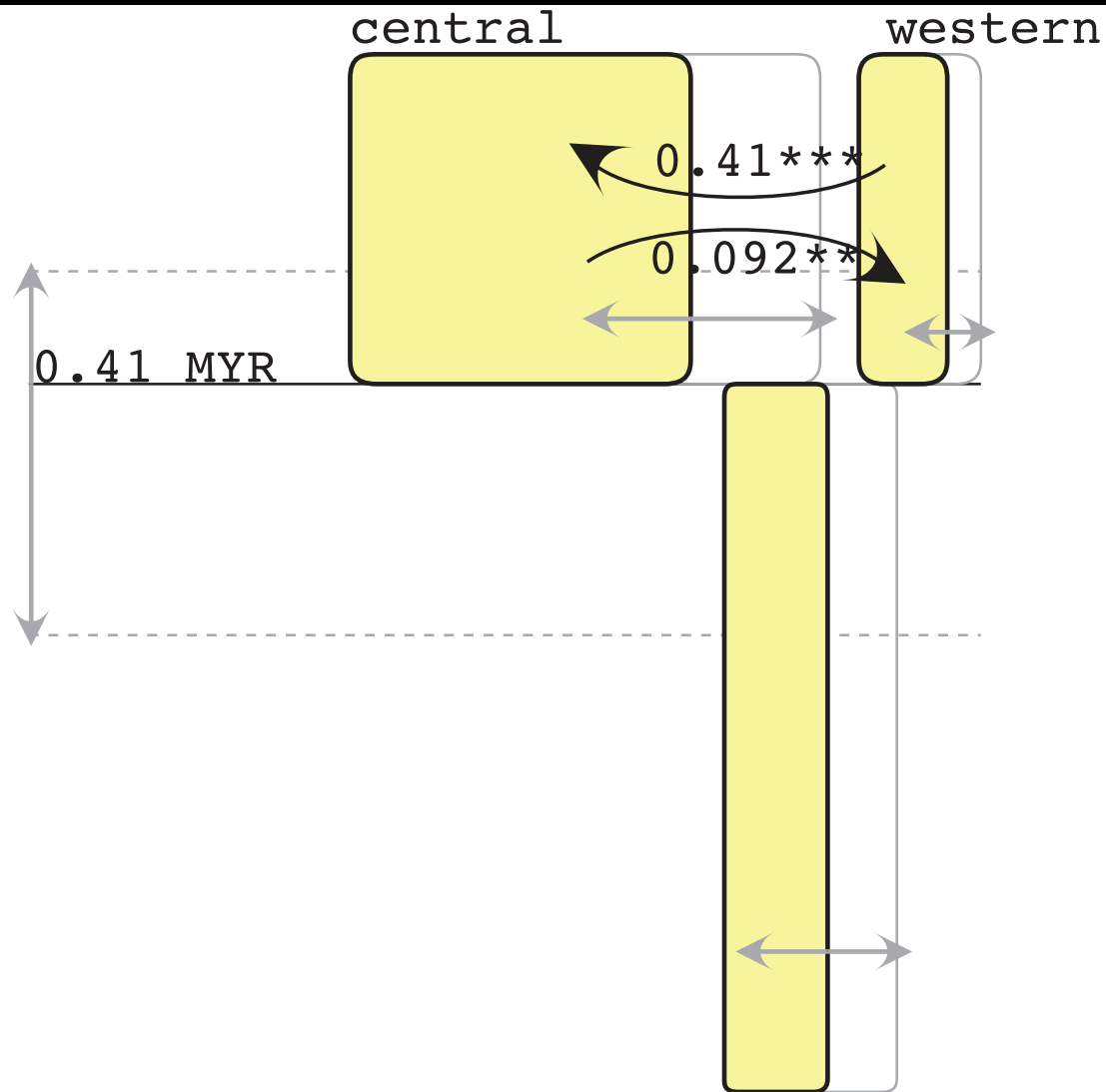
Ancestral N_e (thousands): 8.4



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

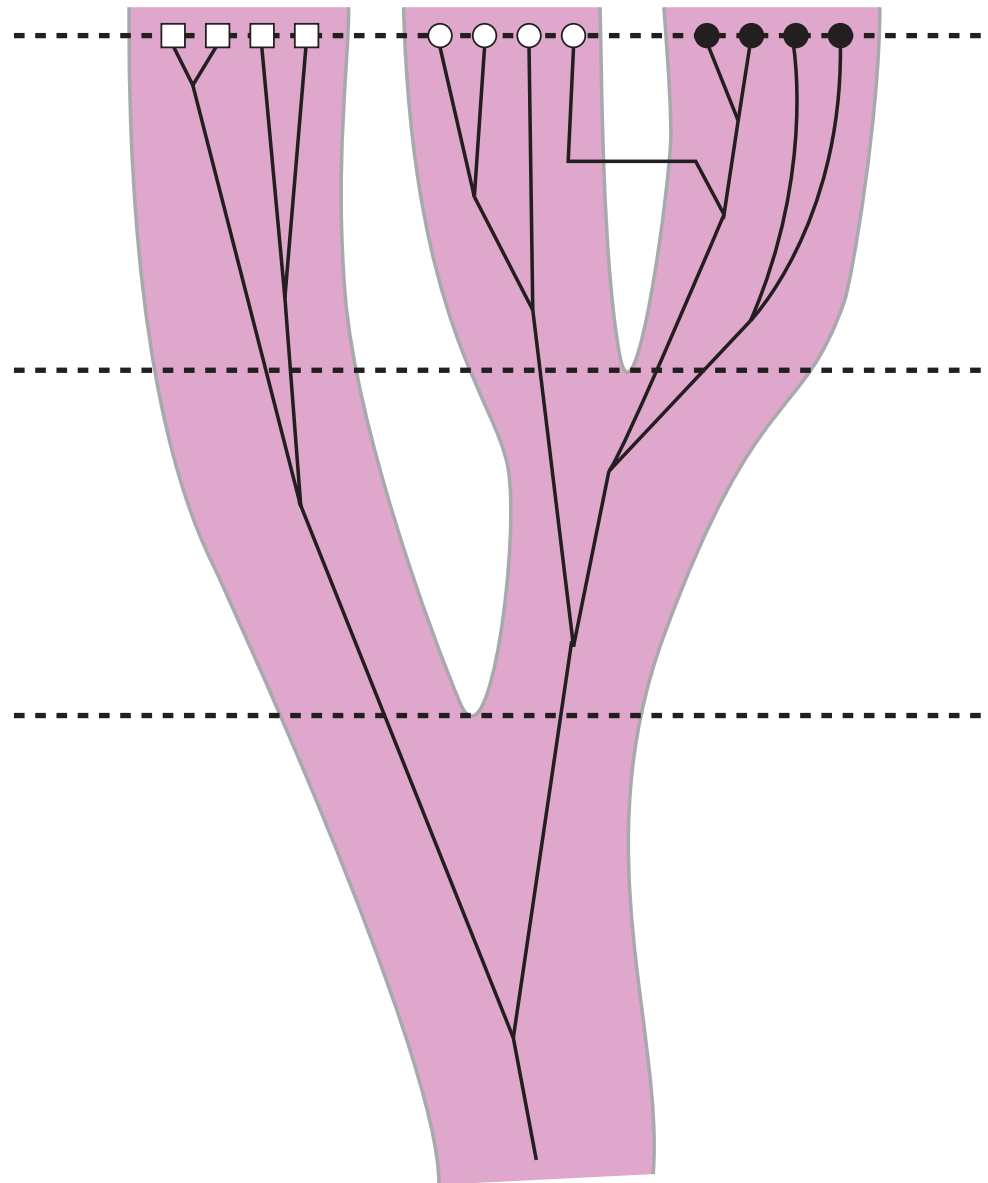


Ancestral N_e (thousands): 8.4



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



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?

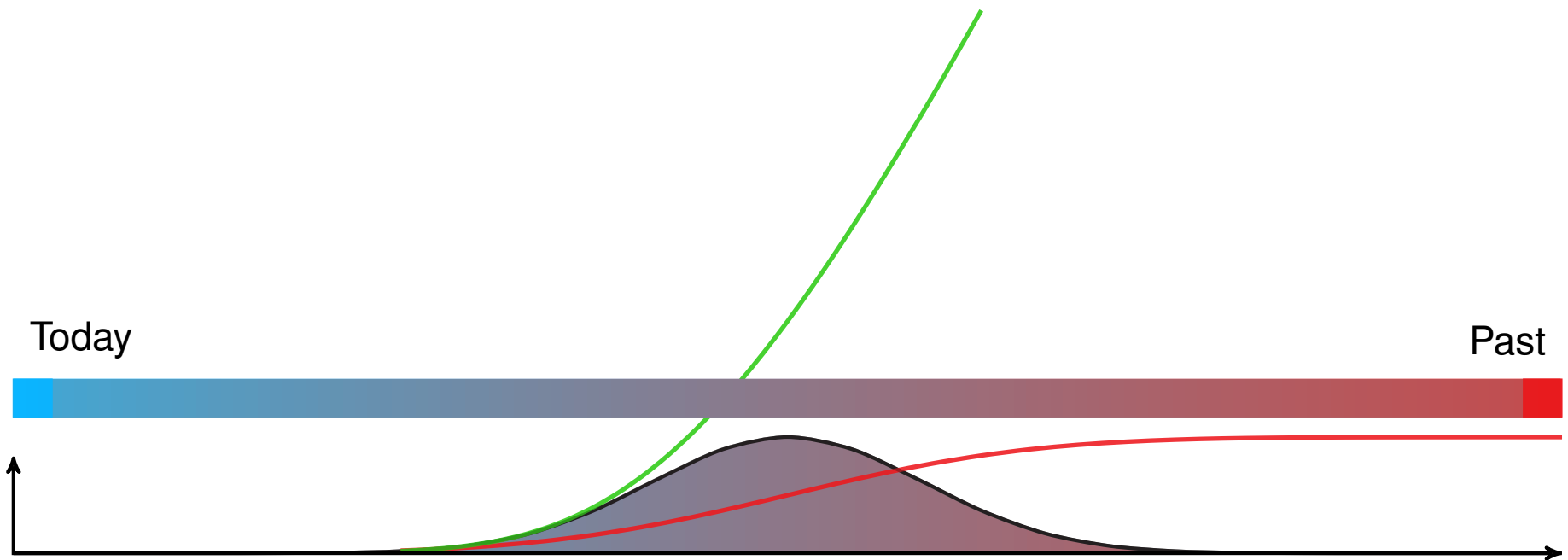
Today

Past

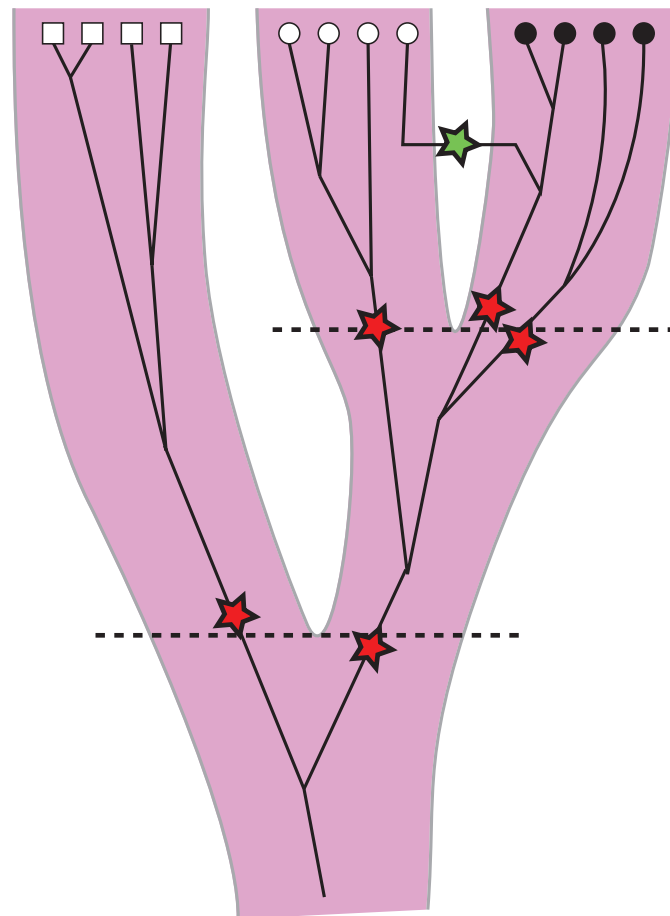


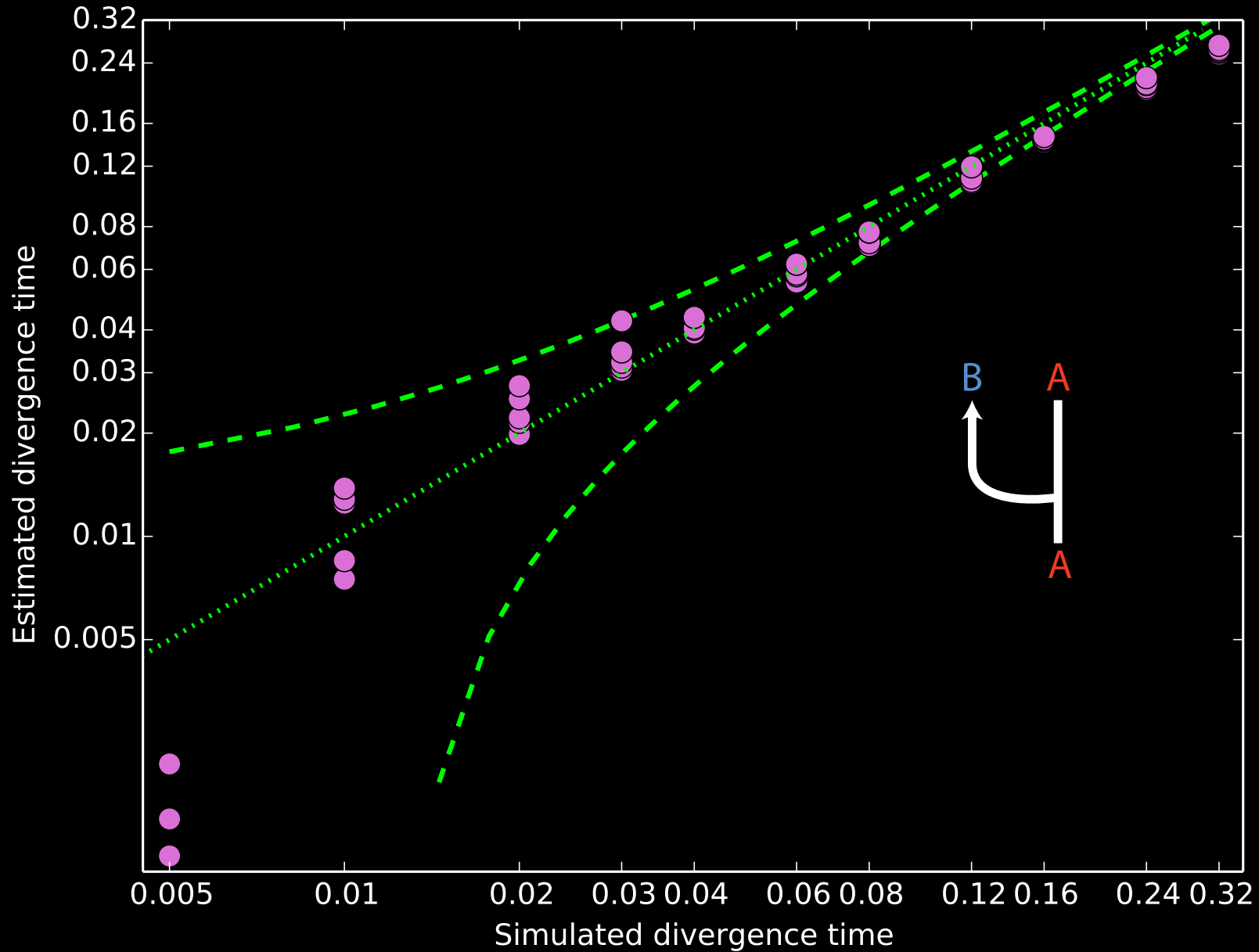
Population splitting

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.



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.

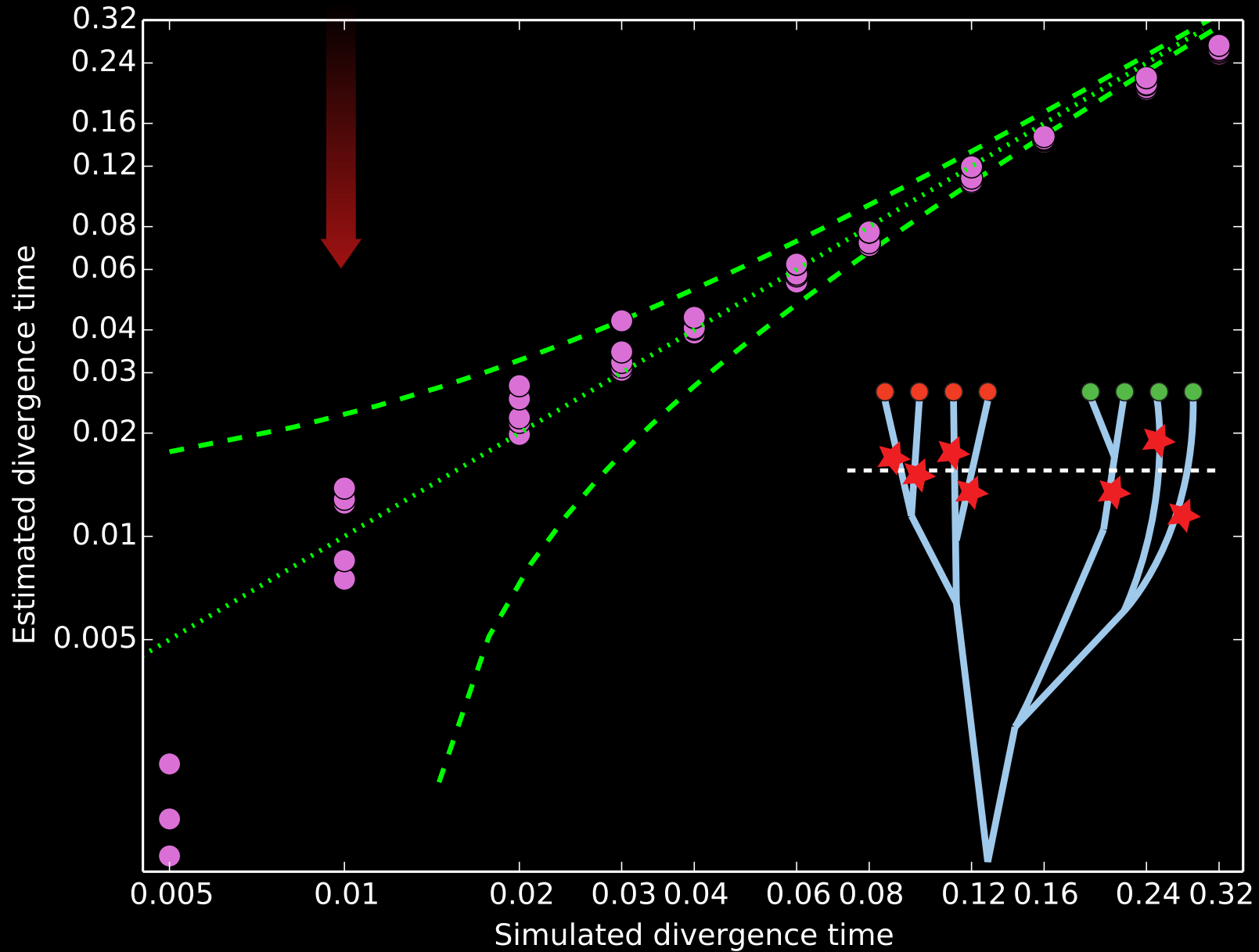




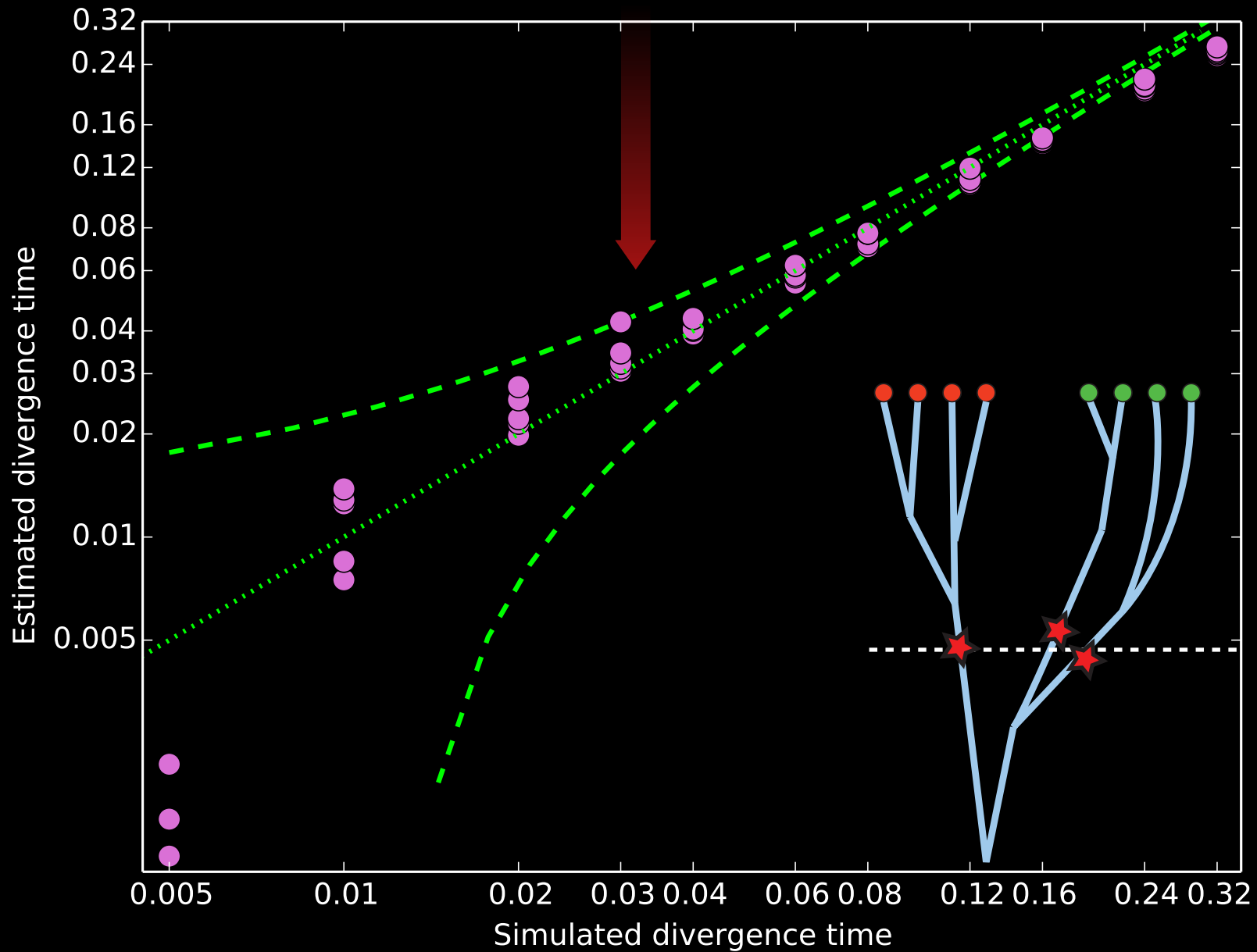
(Palczewski, Ashki, and Beerli [in prep.] An alternative population fission model to the isolation with migration model.)

Population splitting

0.0



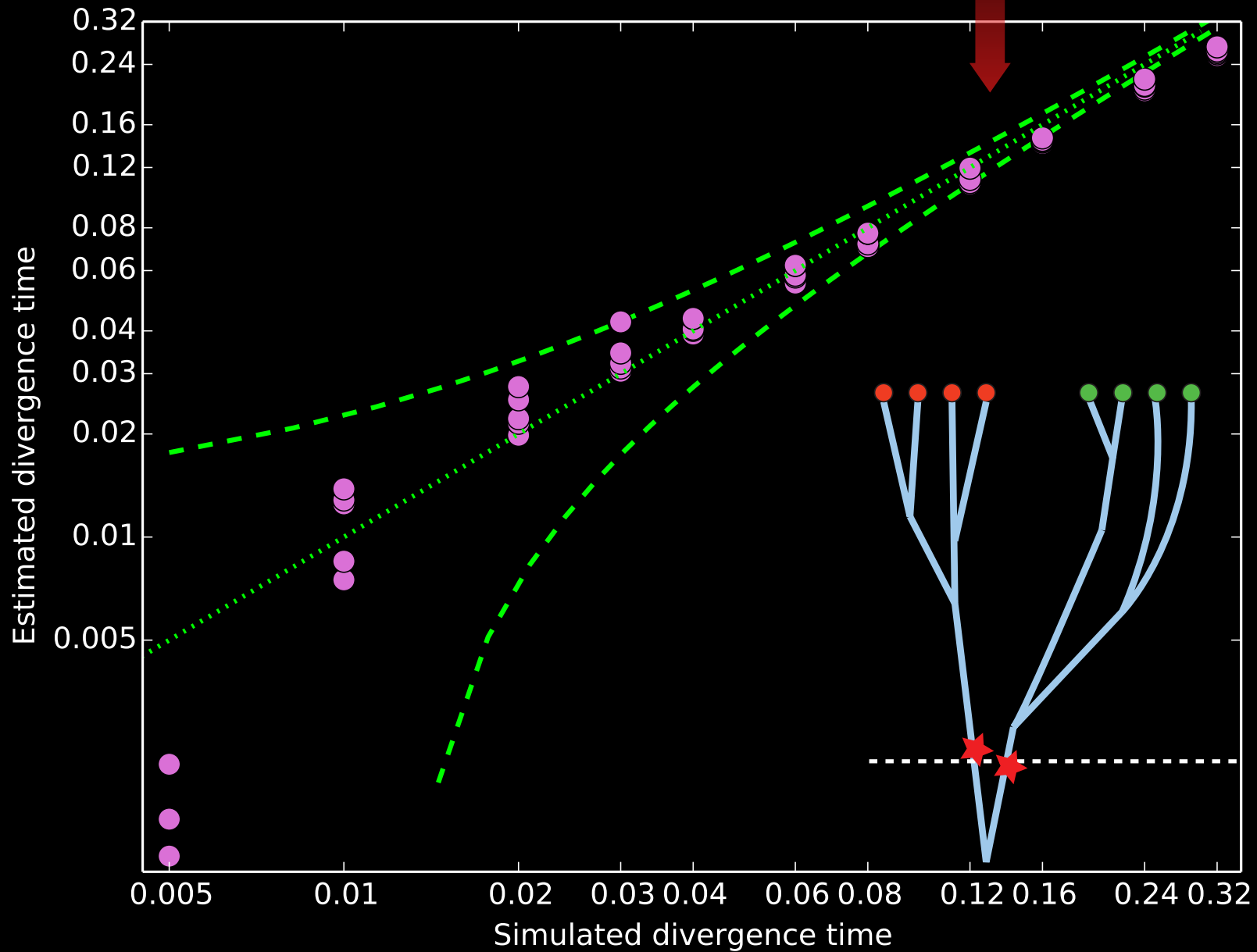
(Palczewski, Ashki, and Beerli [in prep.] An alternative population fission model to the isolation with migration model.)



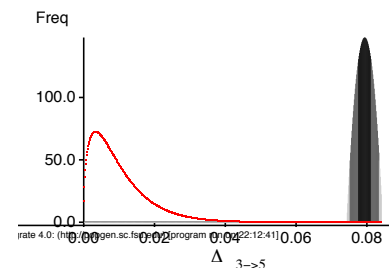
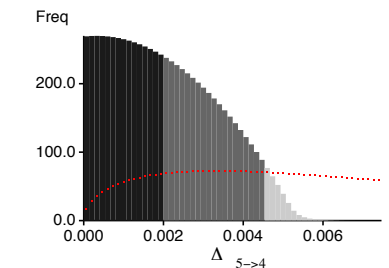
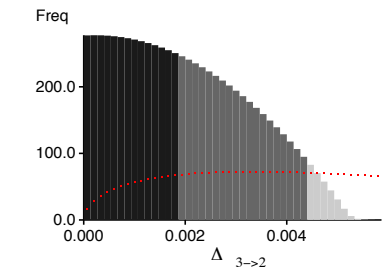
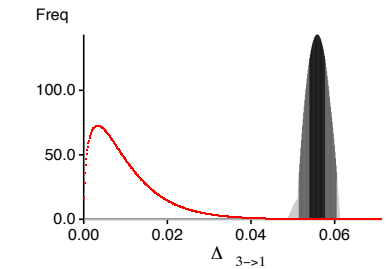
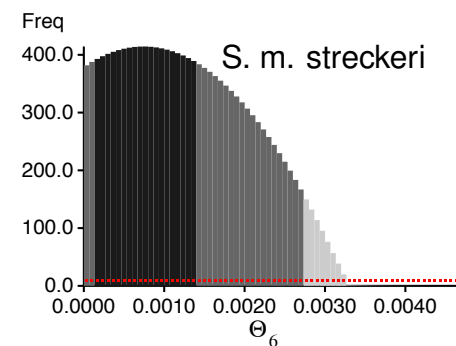
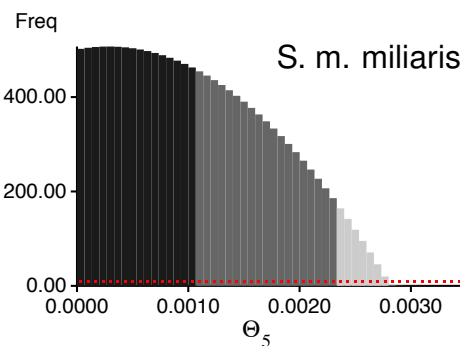
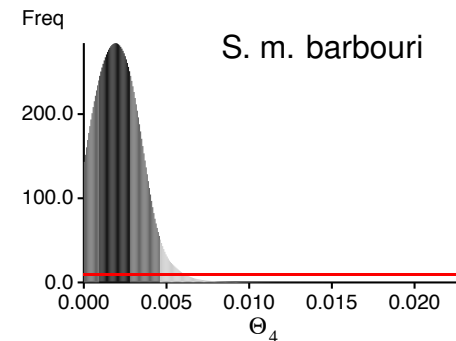
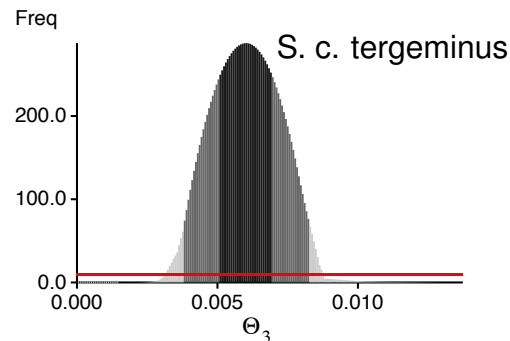
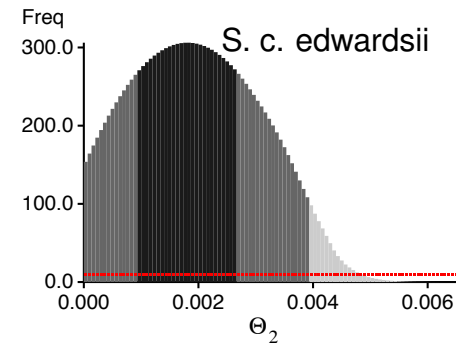
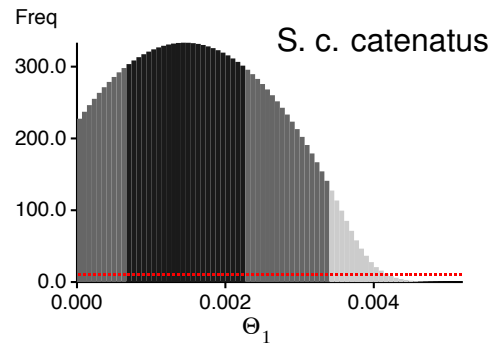
(Palczewski, Ashki, and Beerli [in prep.] An alternative population fission model to the isolation with migration model.)

Population splitting

0.0



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



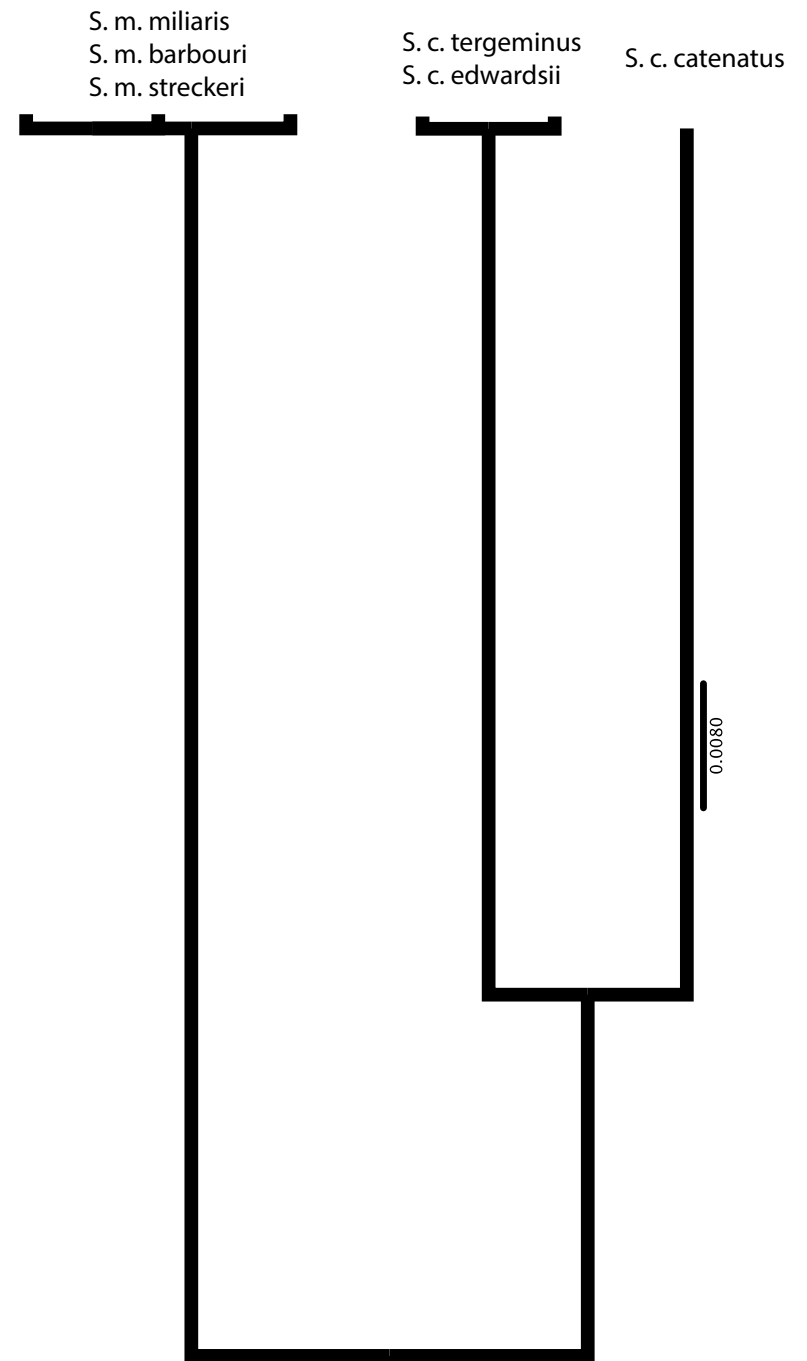
Estimation of splitting dates of 6 subspecies of pygmy rattlesnakes using MIGRATE (data from Kubatko et al. 2011)

Population splitting

Pygmy rattle snakes



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



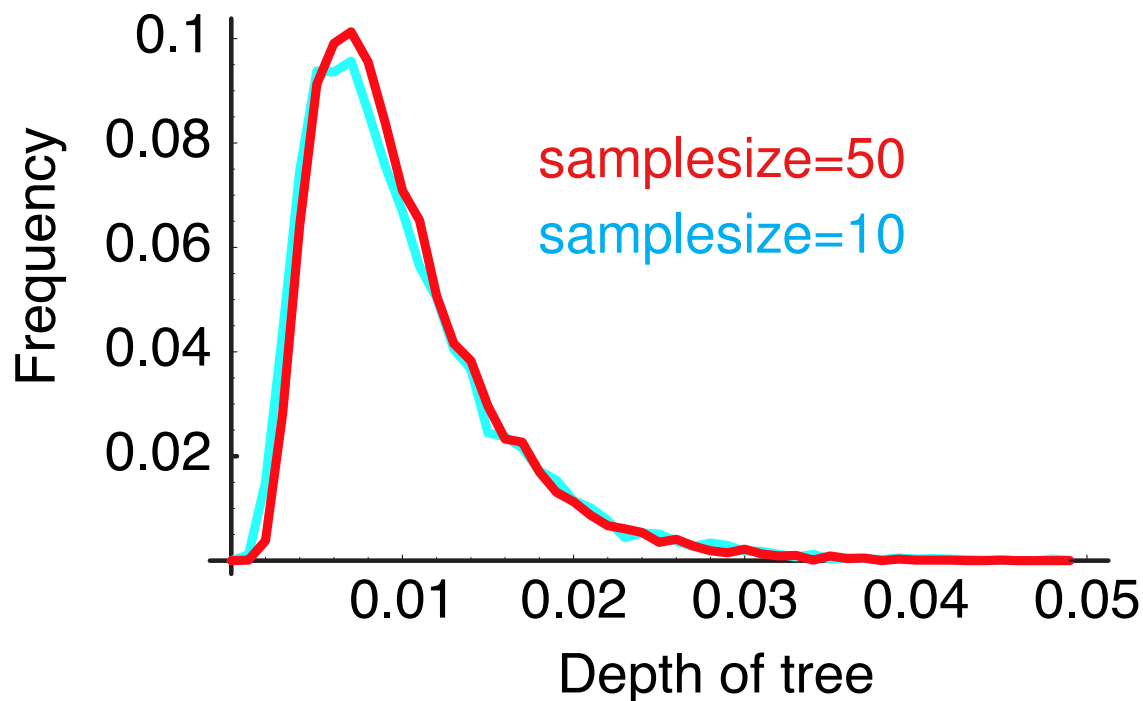


Violating assumptions

The evil reviewer says: *“You shall not use method/program X because your data does not fit the assumptions for...”*

- ◆ Required samples
- ◆ Recombination
- ◆ Population size fluctuation
- ◆ Divergence

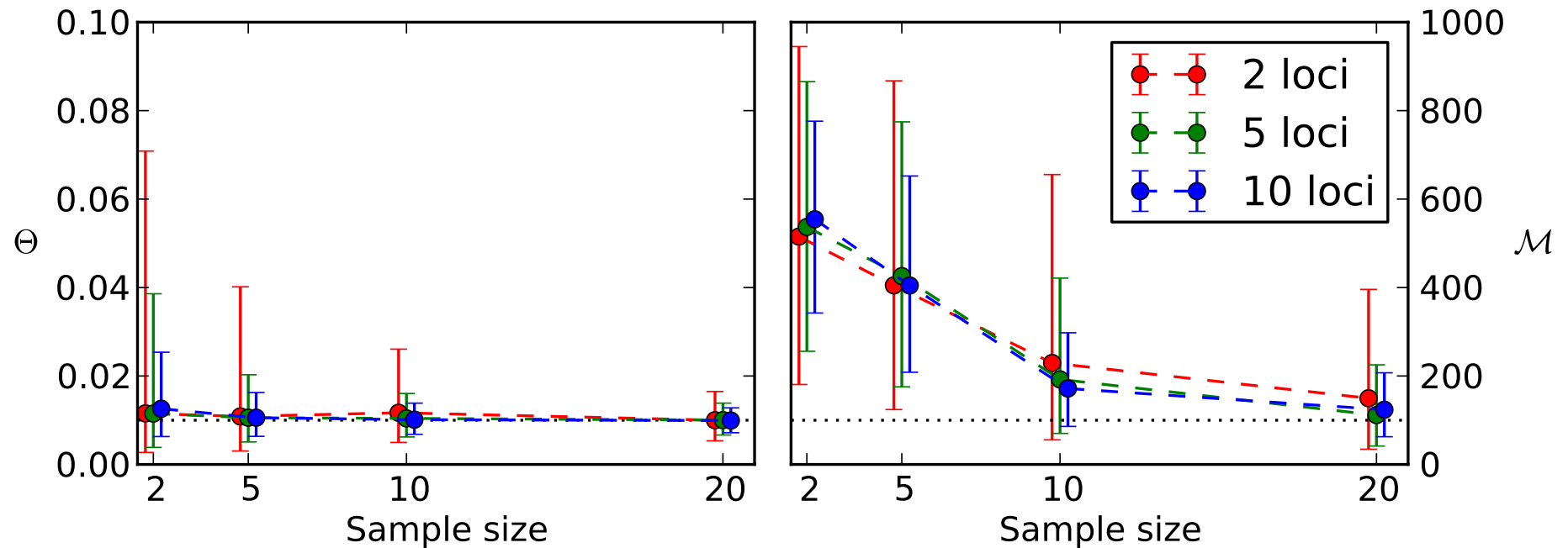
- ◆ The time to the most recent common ancestor is robust to different sample sizes.
- ◆ Simulated sequence data from a single population have shown that after 8 individuals you should better add another locus than more individuals.



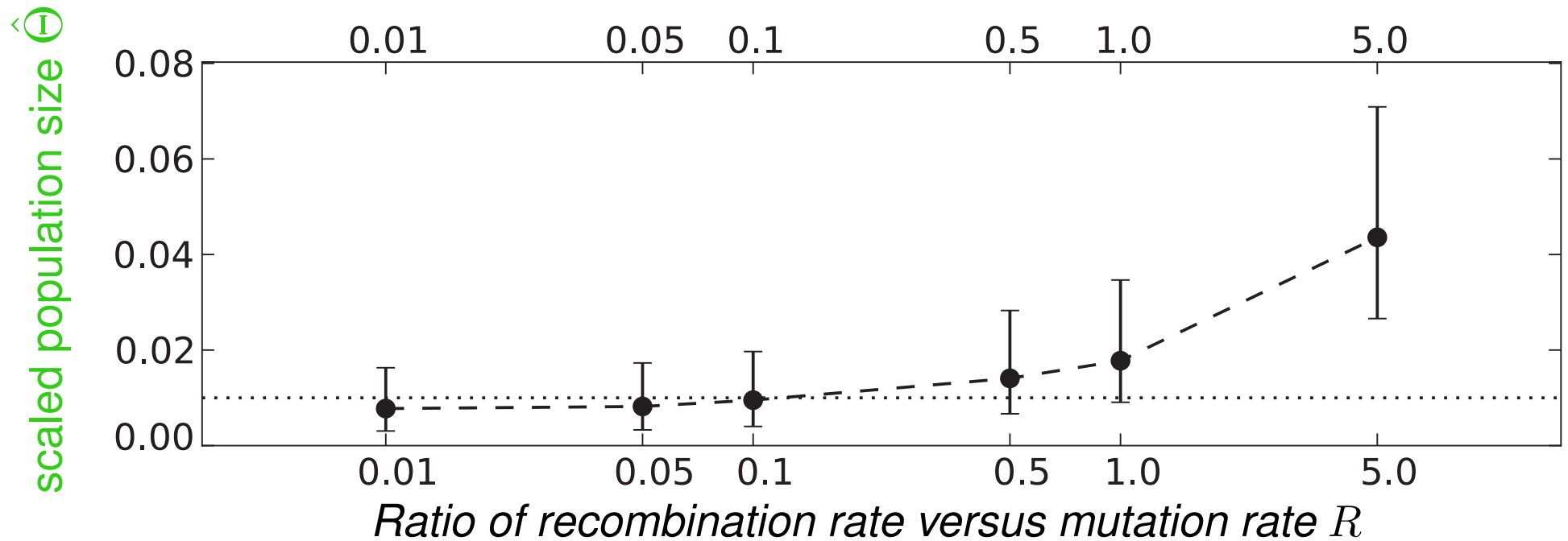
Felsenstein (2005)
Pluzhnikov and Donnelly
(1996)

Required number of samples is small

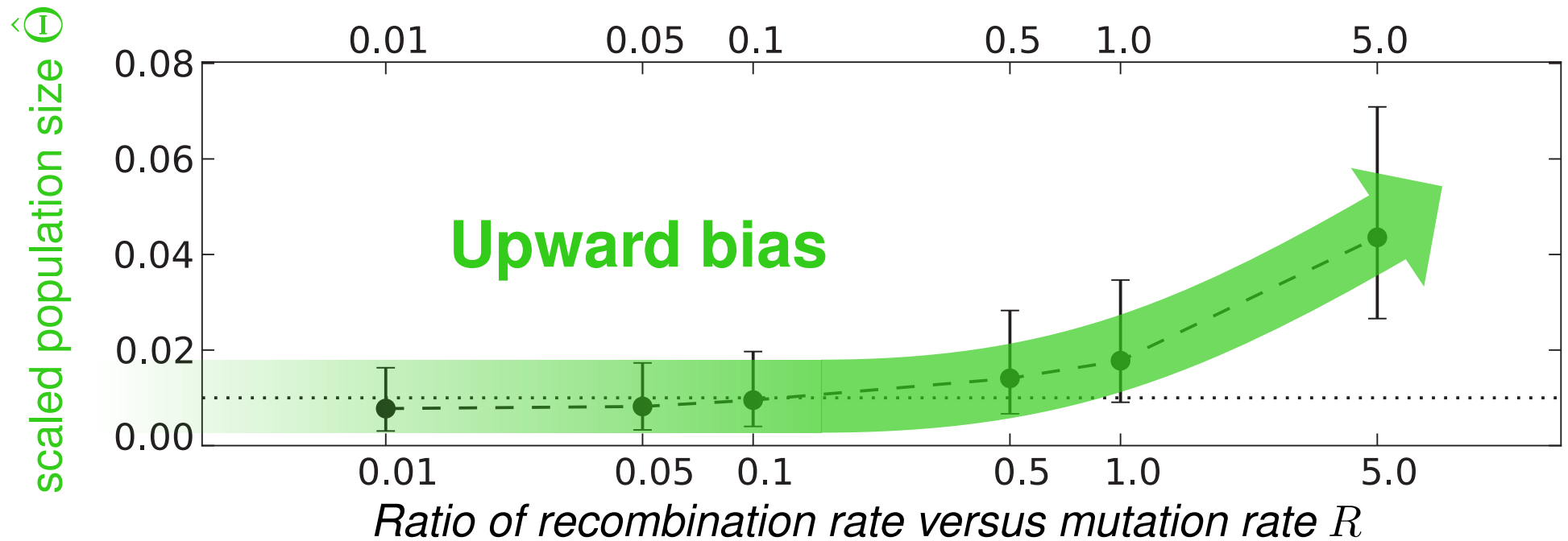
Multiple populations



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; $M = 100$

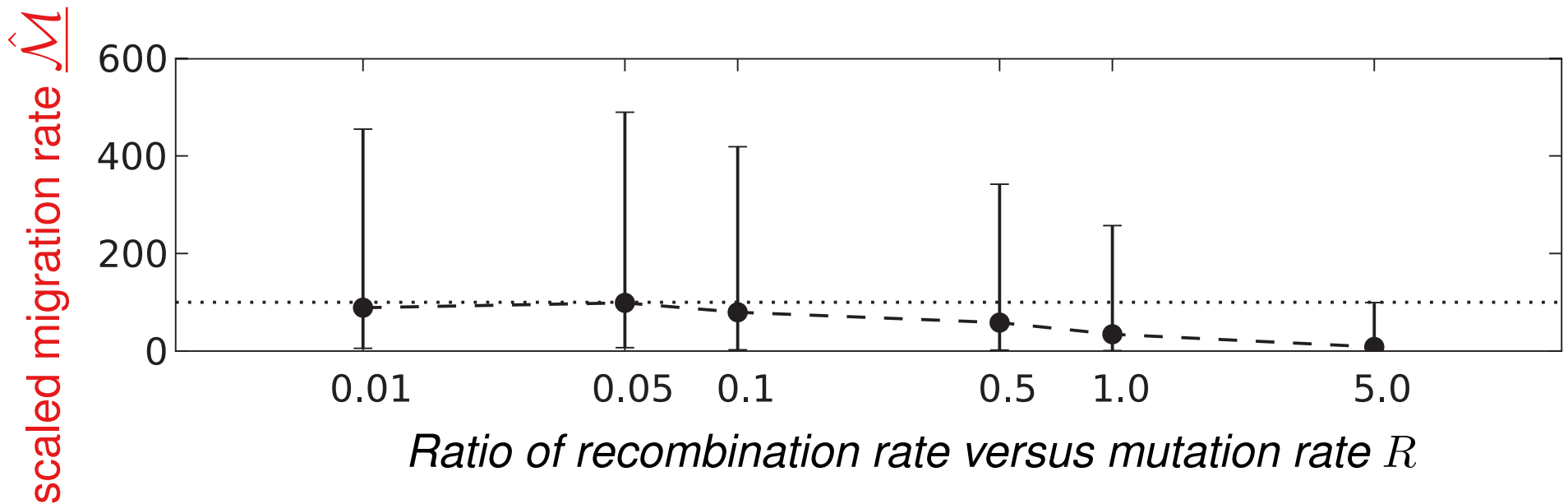


Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.



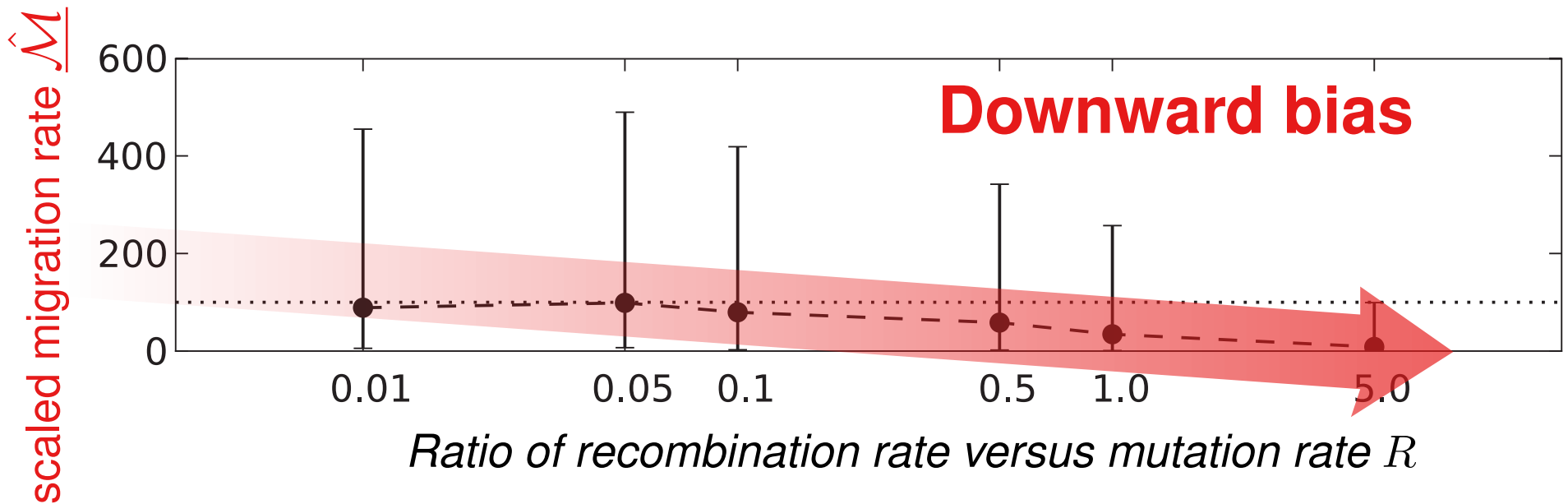
Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

~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.

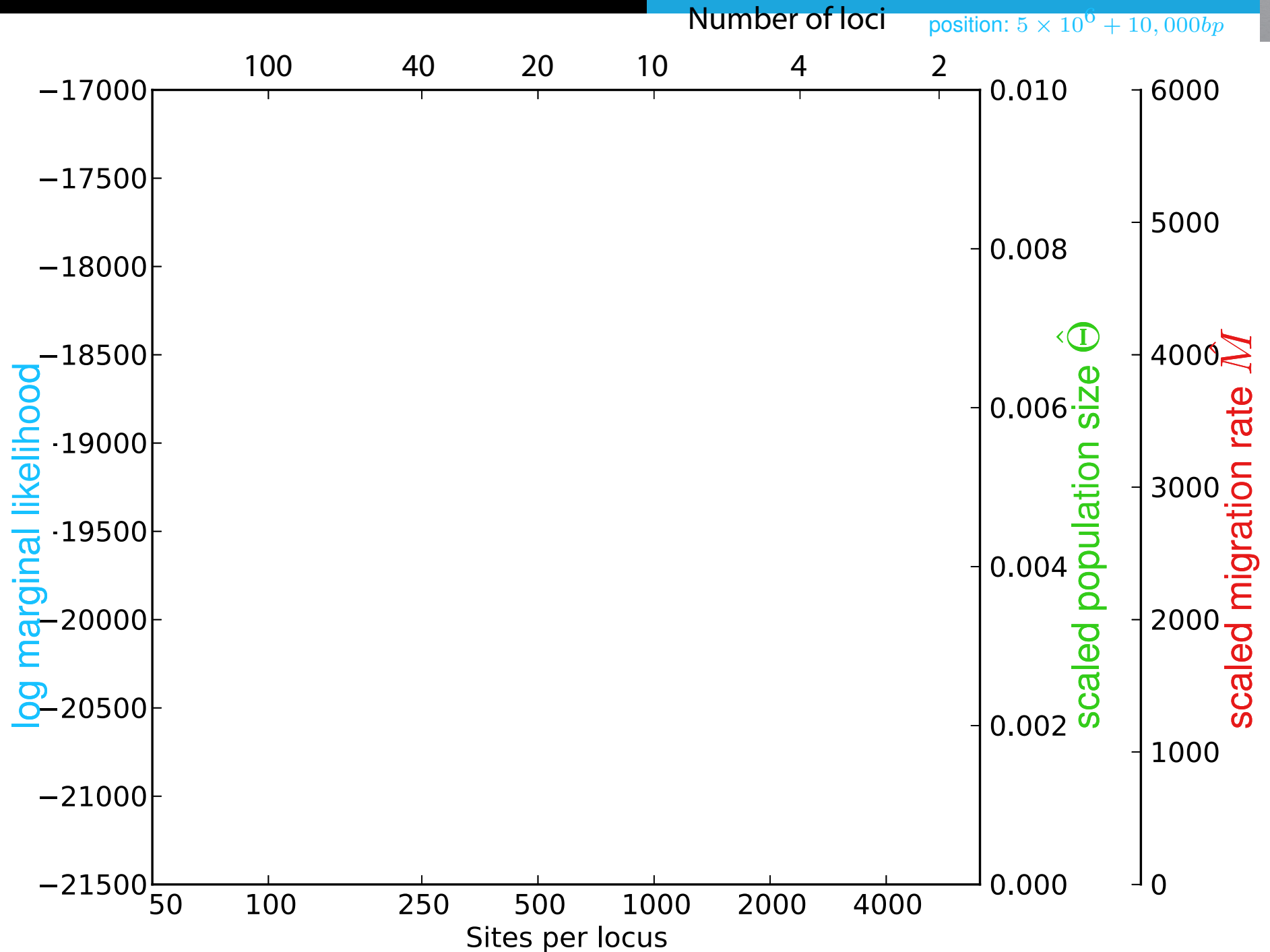
~500 simulated datasets



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

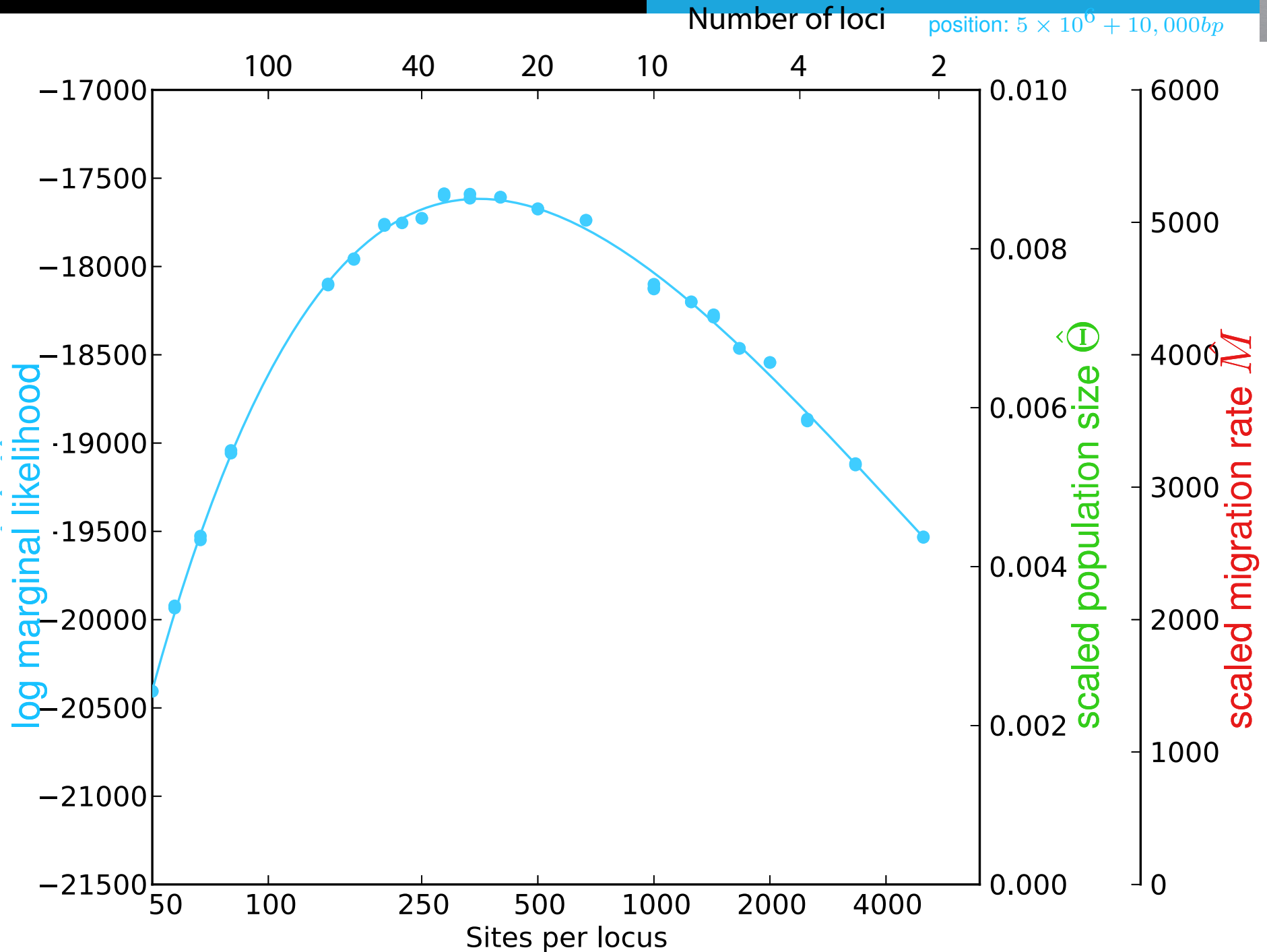
Chopping a real dataset

D. melanogaster Chr2L
position: $5 \times 10^6 + 10,000bp$



Chopping a real dataset

D. melanogaster Chr2L
position: $5 \times 10^6 + 10,000bp$

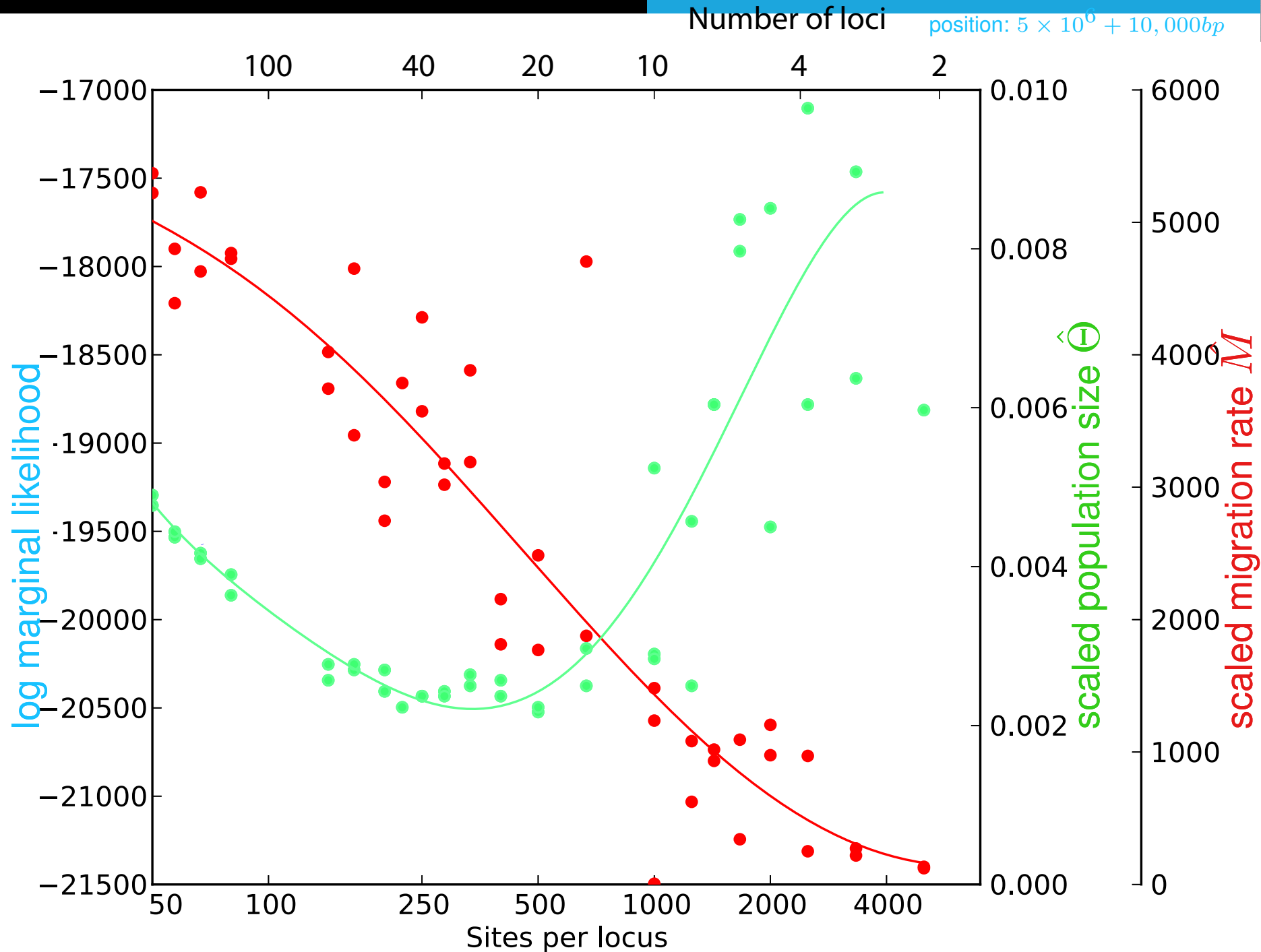


Chopping a real dataset

D. melanogaster Chr2L

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

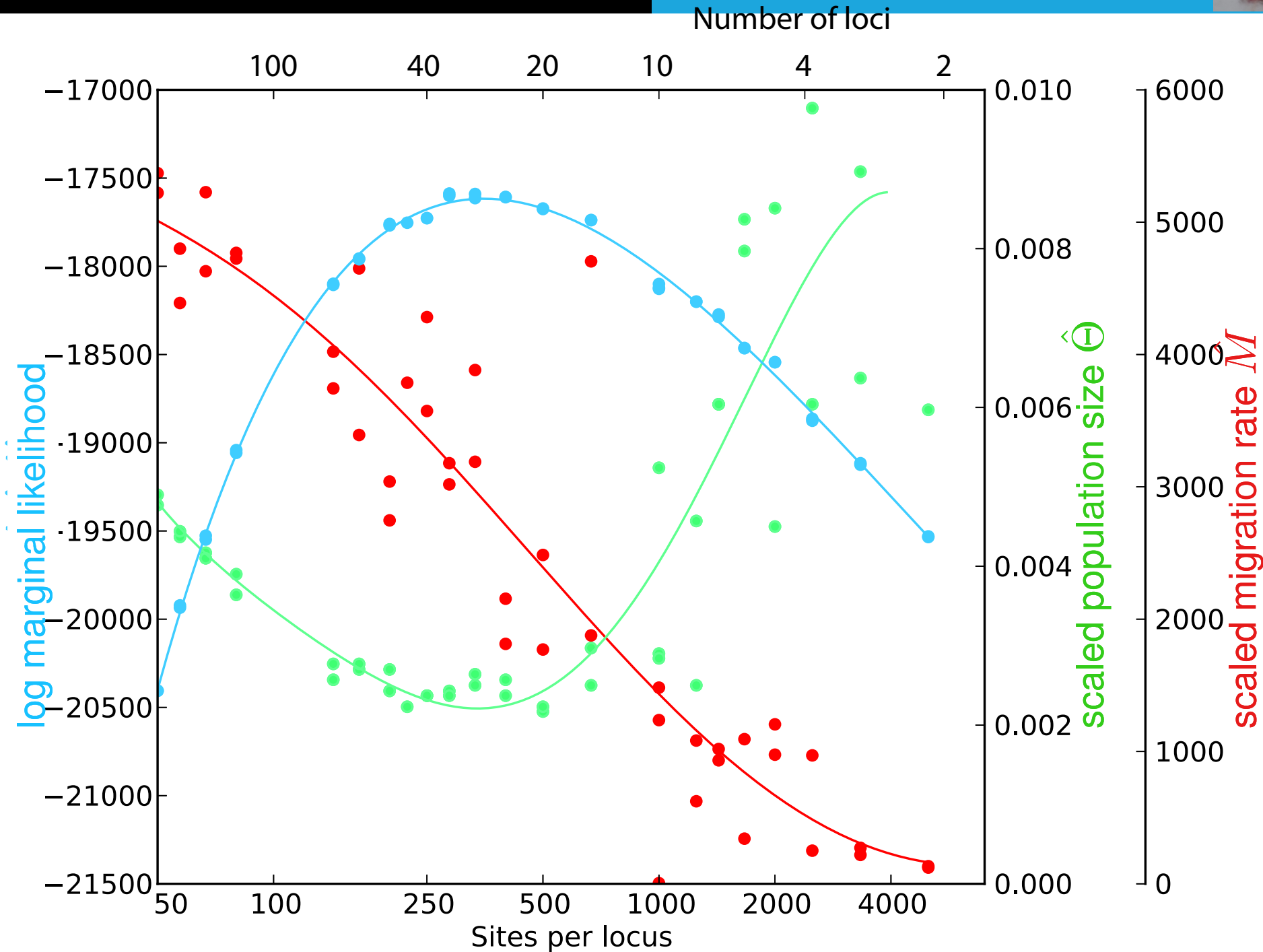
0.0



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 7 simple population models that include speciation with and without migration.
- ◆ (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])

