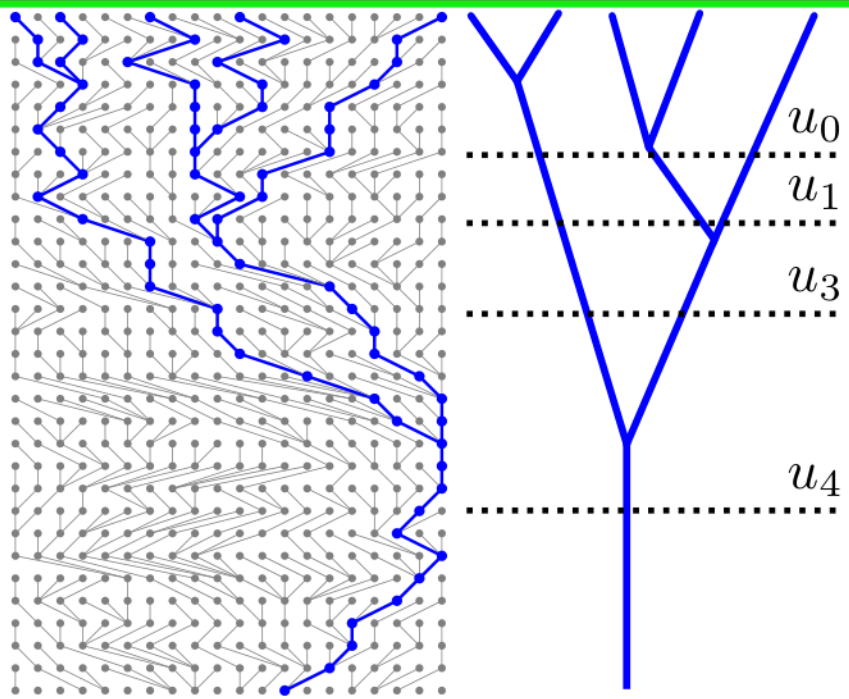
The Coalescent:

Inference using trees of individuals

Peter Beerli
Scientific Computing, Florida State University @peterbeerli

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

Analogy #3

Extensions of the basic coalescence

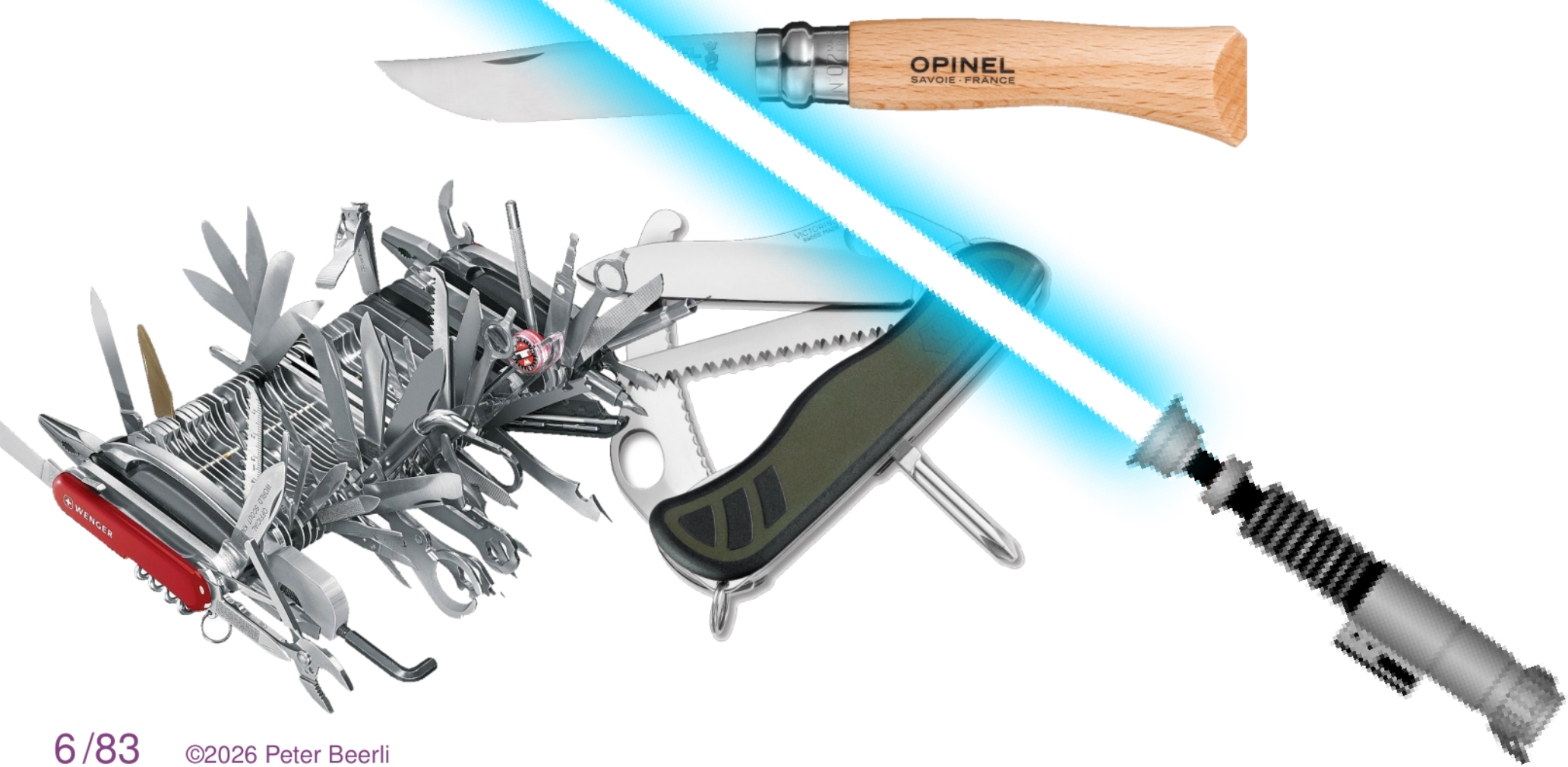
Analogy #3

Extensions of the basic coalescence

Analogy #3

Extensions of the basic coalescence

Analogy #3



Extensions of the basic coalescence

- Population growth (two parameters), fluctuations, bottlenecks
- Migration among populations (potentially thousands, parameters)
- Population splitting (many parameters)
- SNPs, site frequency spectra, mutation models
- Effect of assumption violation

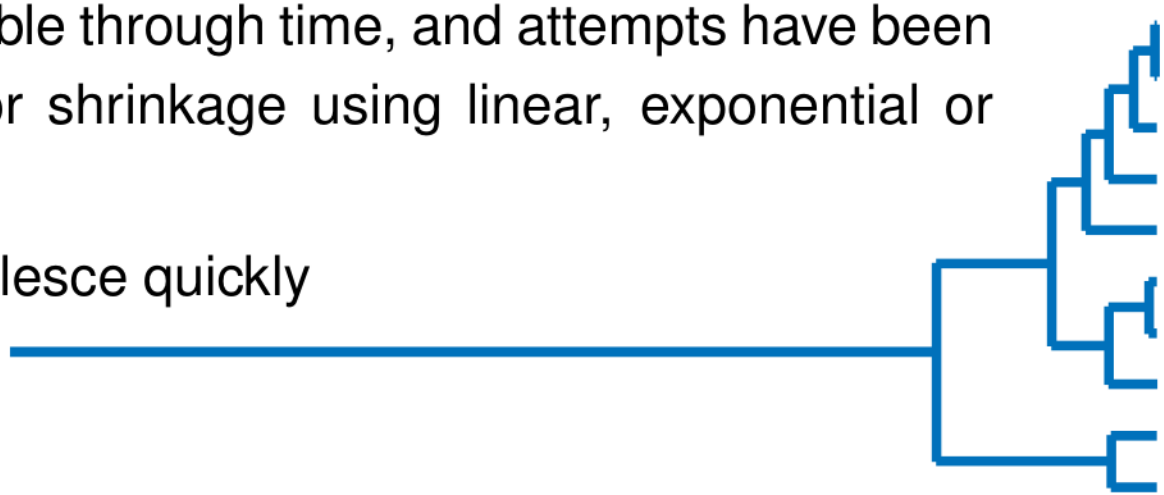
Extensions of the basic coalescent

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.

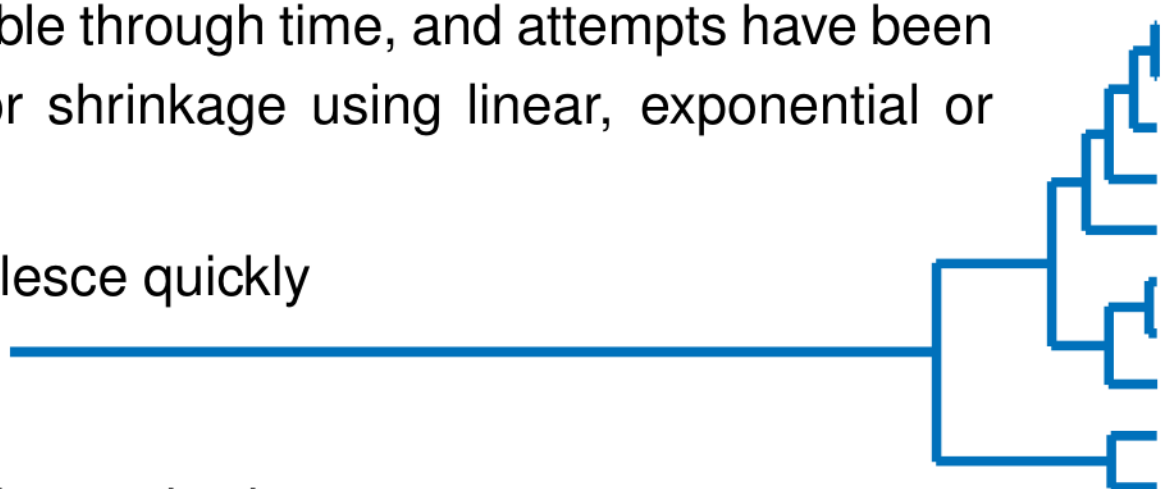
- In a small population lineages coalesce quickly



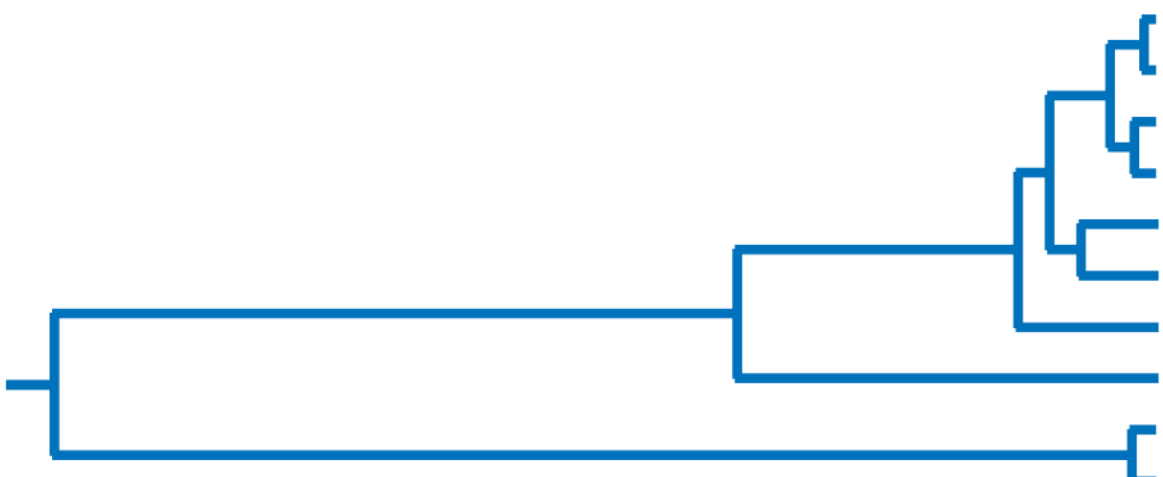
Extensions of the basic coalescent

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



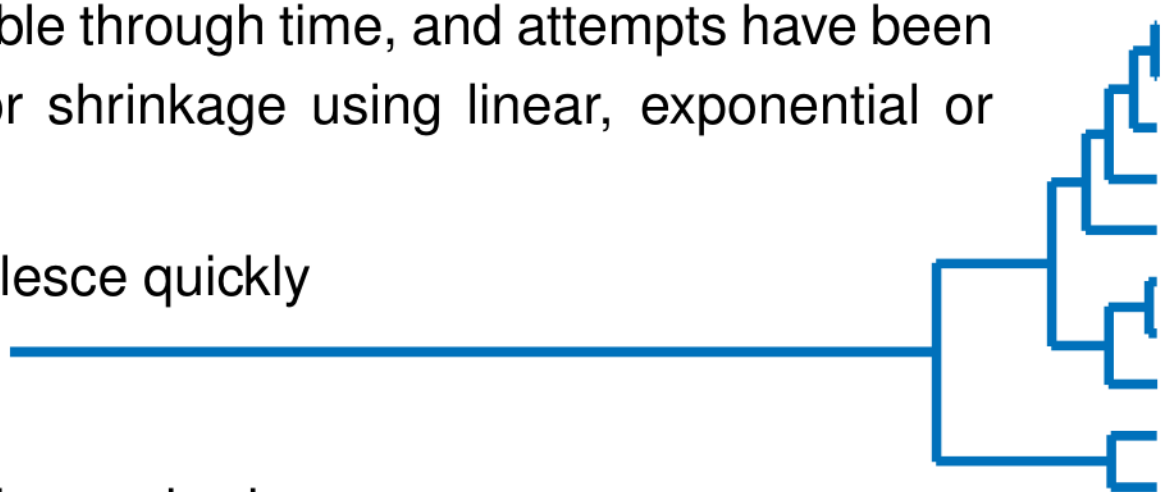
- In a large population lineages coalesce slowly



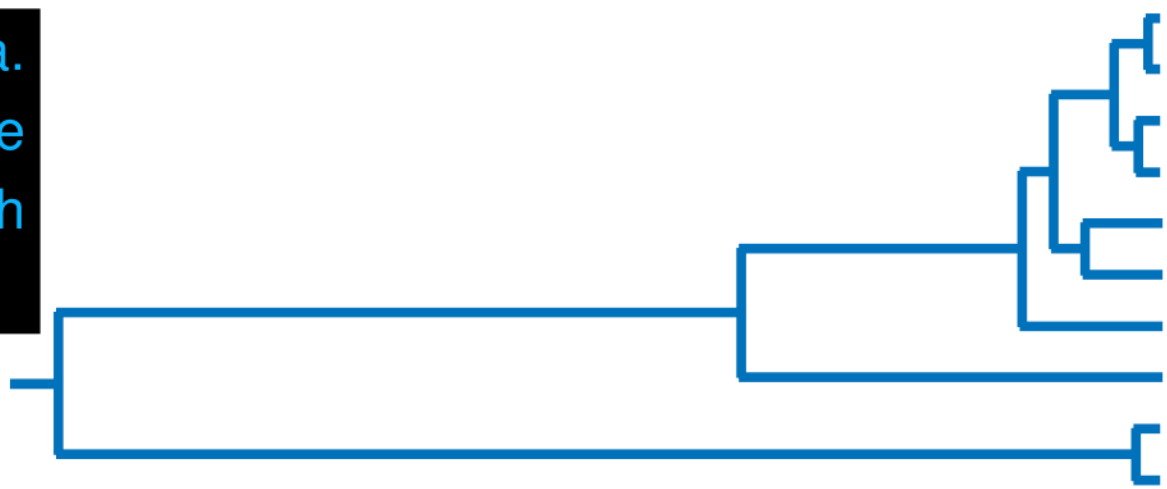
Extensions of the basic coalescent

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

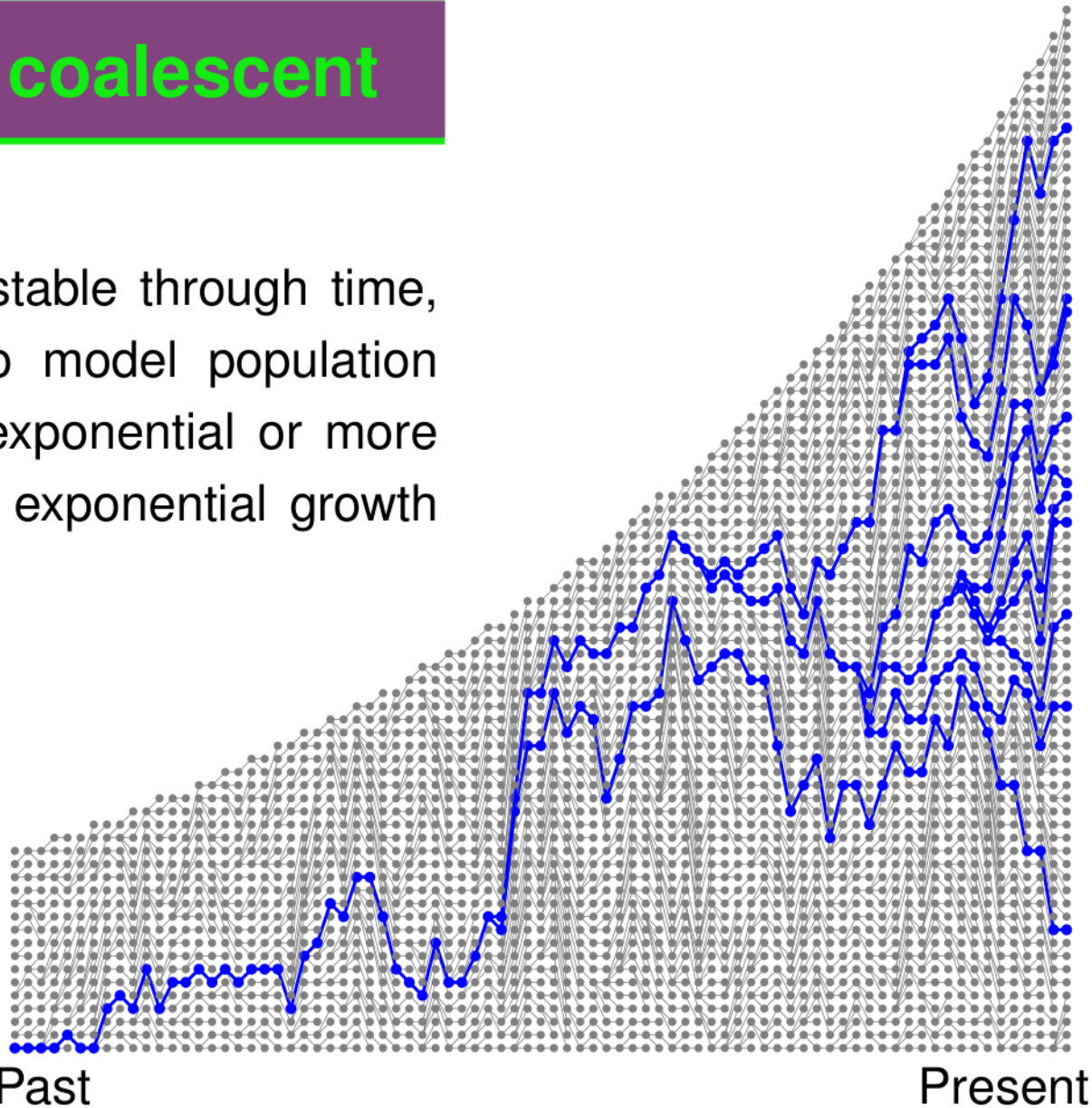
Populations are rarely completely stable through time, and attempts have been made to model population growth or shrinkage using linear, exponential or more general approaches. For example exponential growth could be modeled as

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

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

$$N_0 = 80$$

$$r = 0.02$$



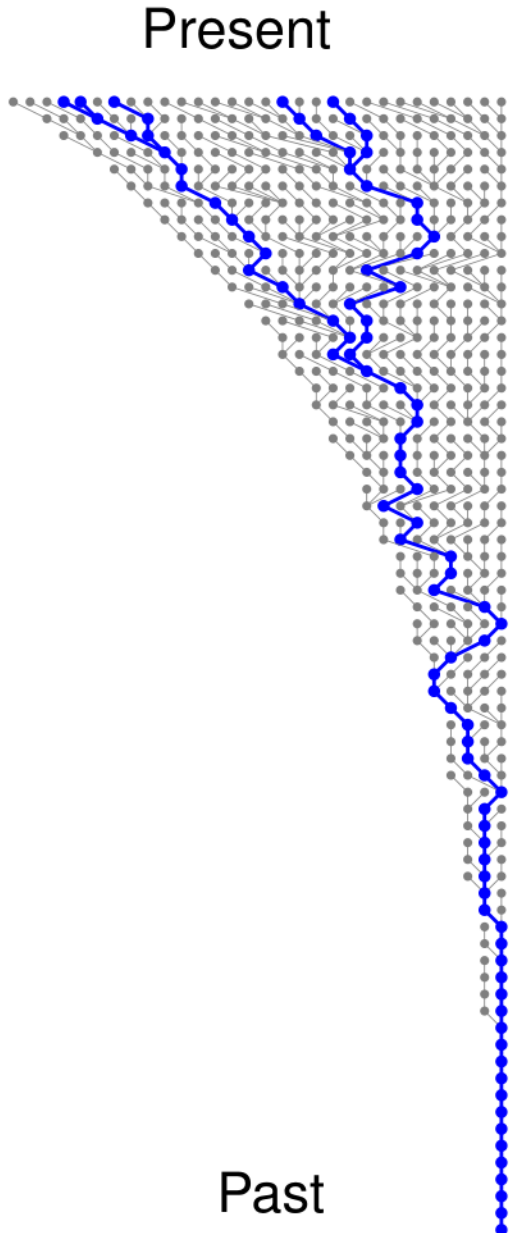
Extensions of the basic coalescent

For constant population size we found

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

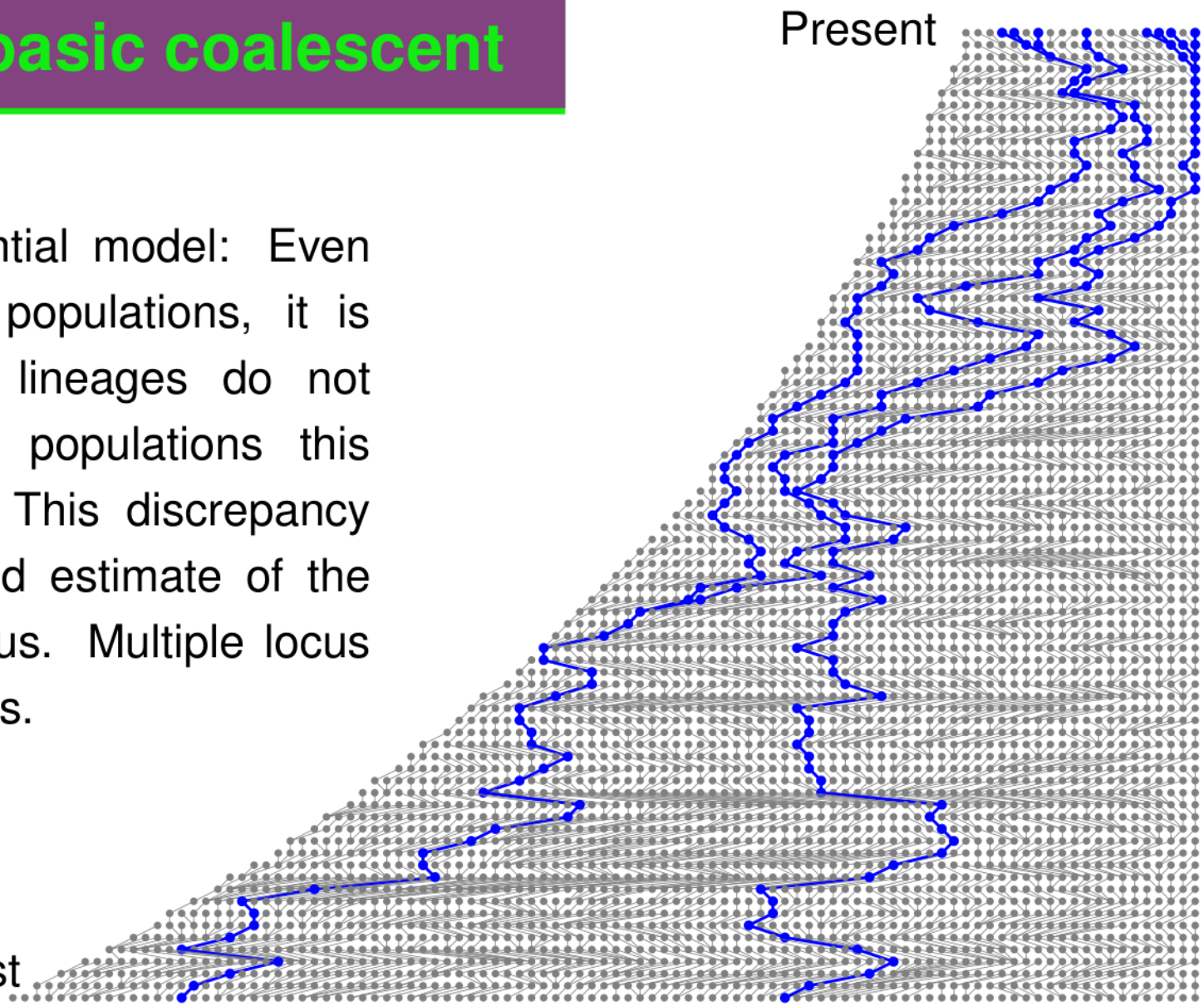
Relaxing the constant size to exponential growth and using $g = r/\mu$ leads to

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

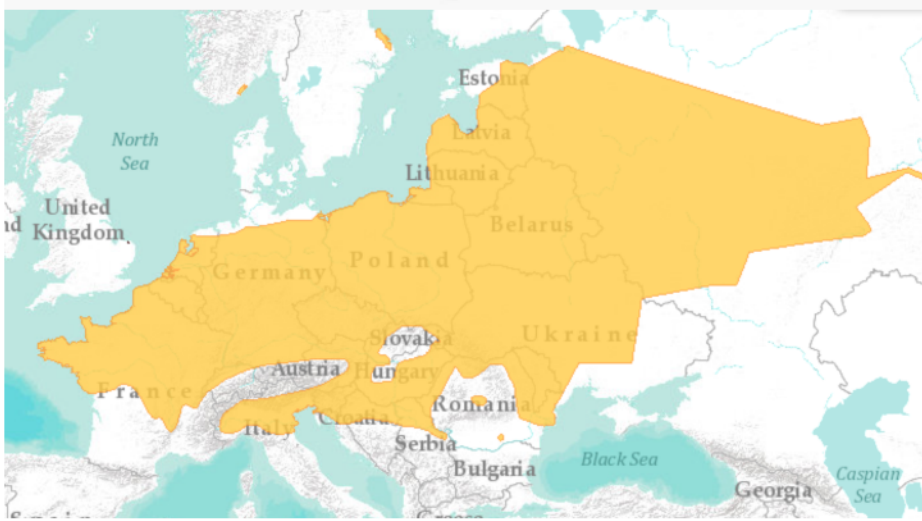


Extensions of the basic coalescent

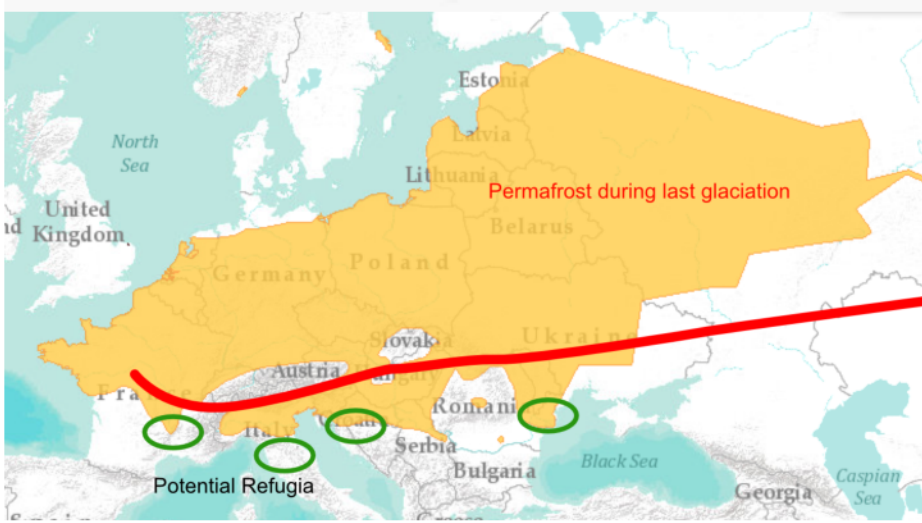
Problems with the exponential model: Even with moderately shrinking populations, it is possible that the sample lineages do not coalesce. With growing populations this problem does not occur. This discrepancy leads to an upwards biased estimate of the growth rate for a single locus. Multiple locus estimates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



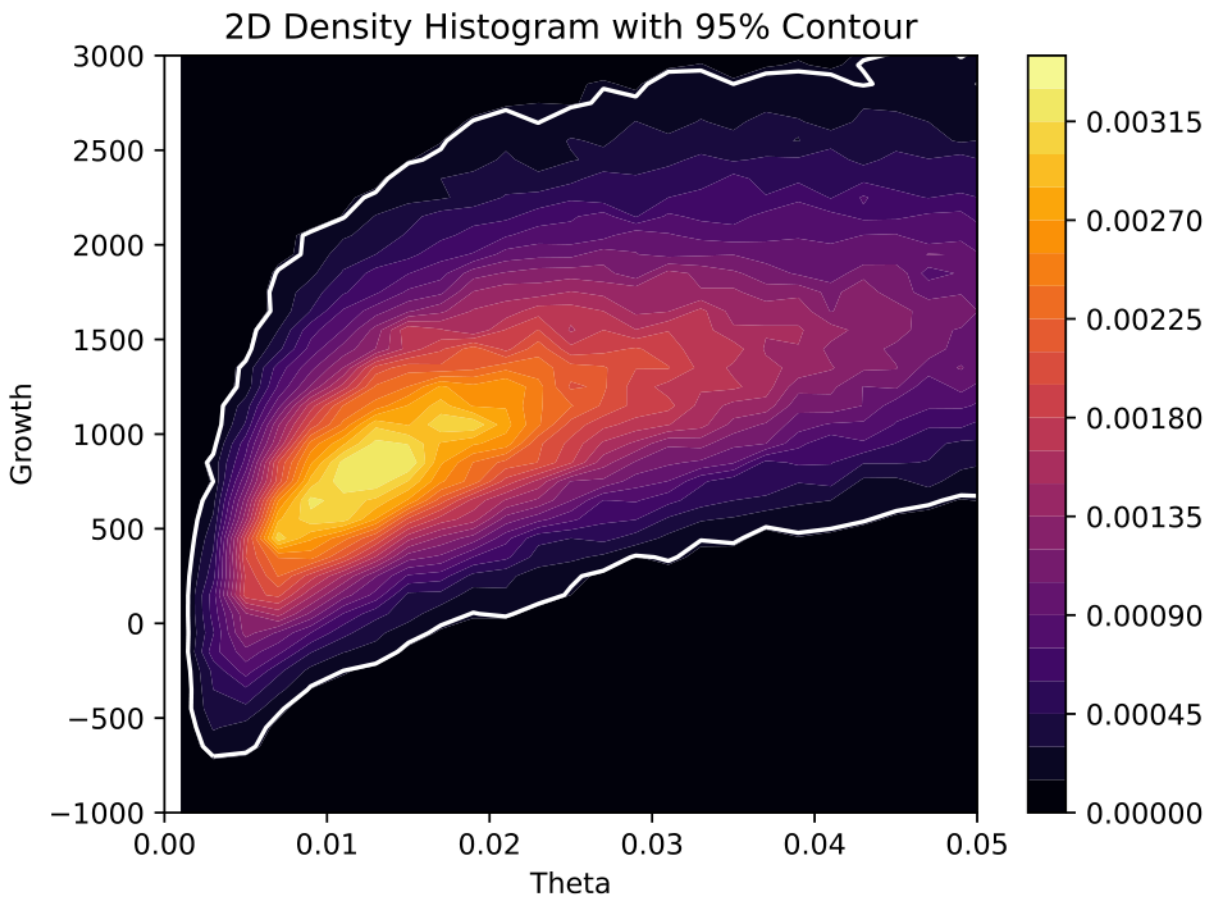
Grow-A-Frog



Grow-A-Frog

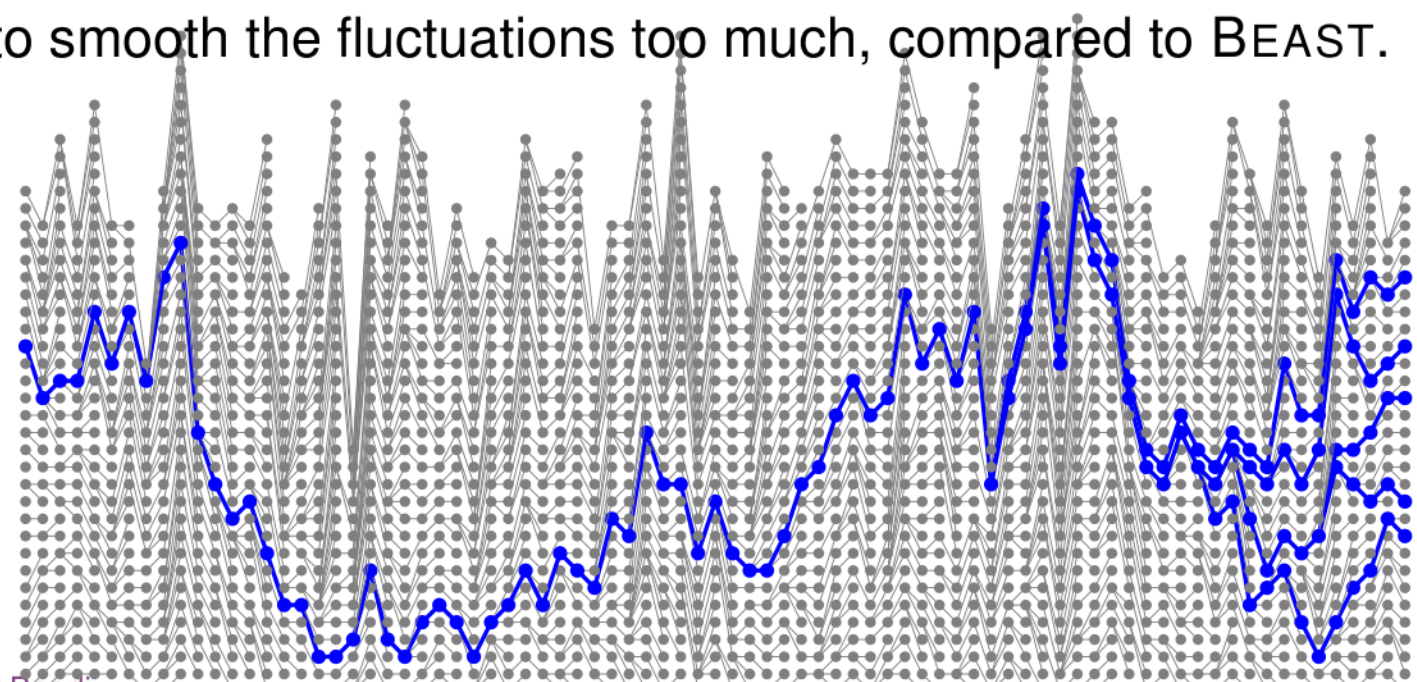


Grow-A-Frog



Extensions of the basic coalescent

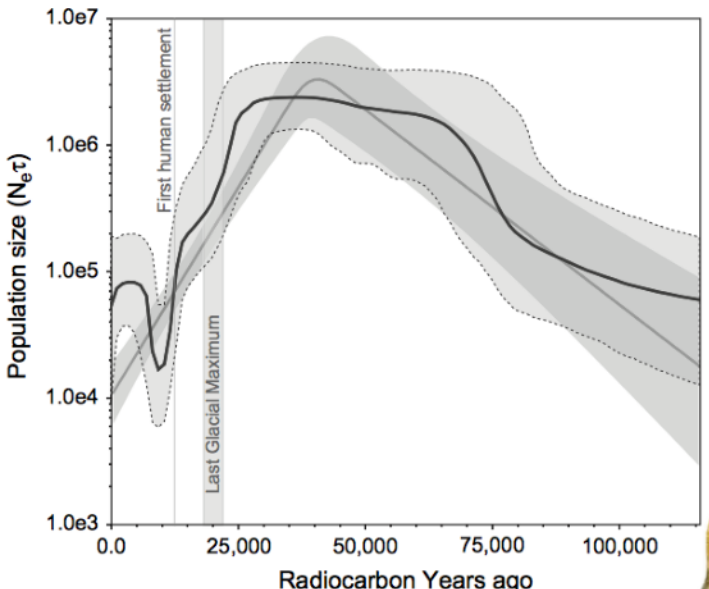
Random fluctuations of the population size are most often ignored. BEAST and REVBAYES (and to some extent MIGRATE) can handle such scenarios. BEAST and REVBAYES are using a full parametric approach (skyride, skyline, skyfish) whereas MIGRATE uses a non-parametric approach for its skyline plots that has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



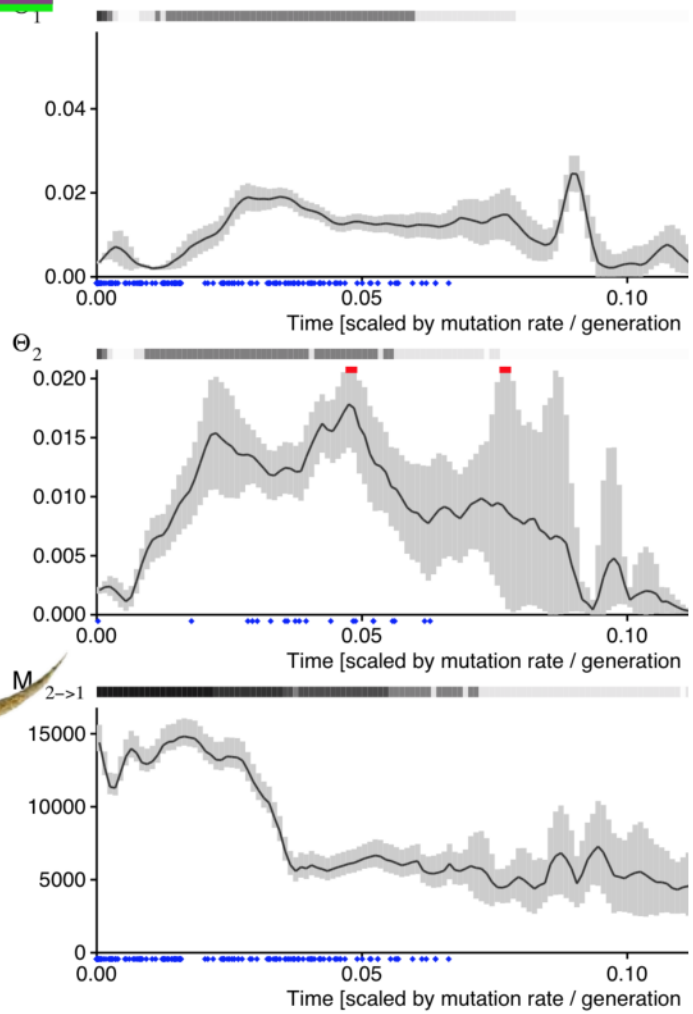
Past Present

Rise and Fall of Steppe Bisons

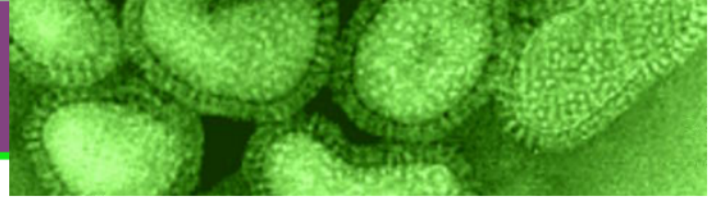
BEAST



MIGRATE

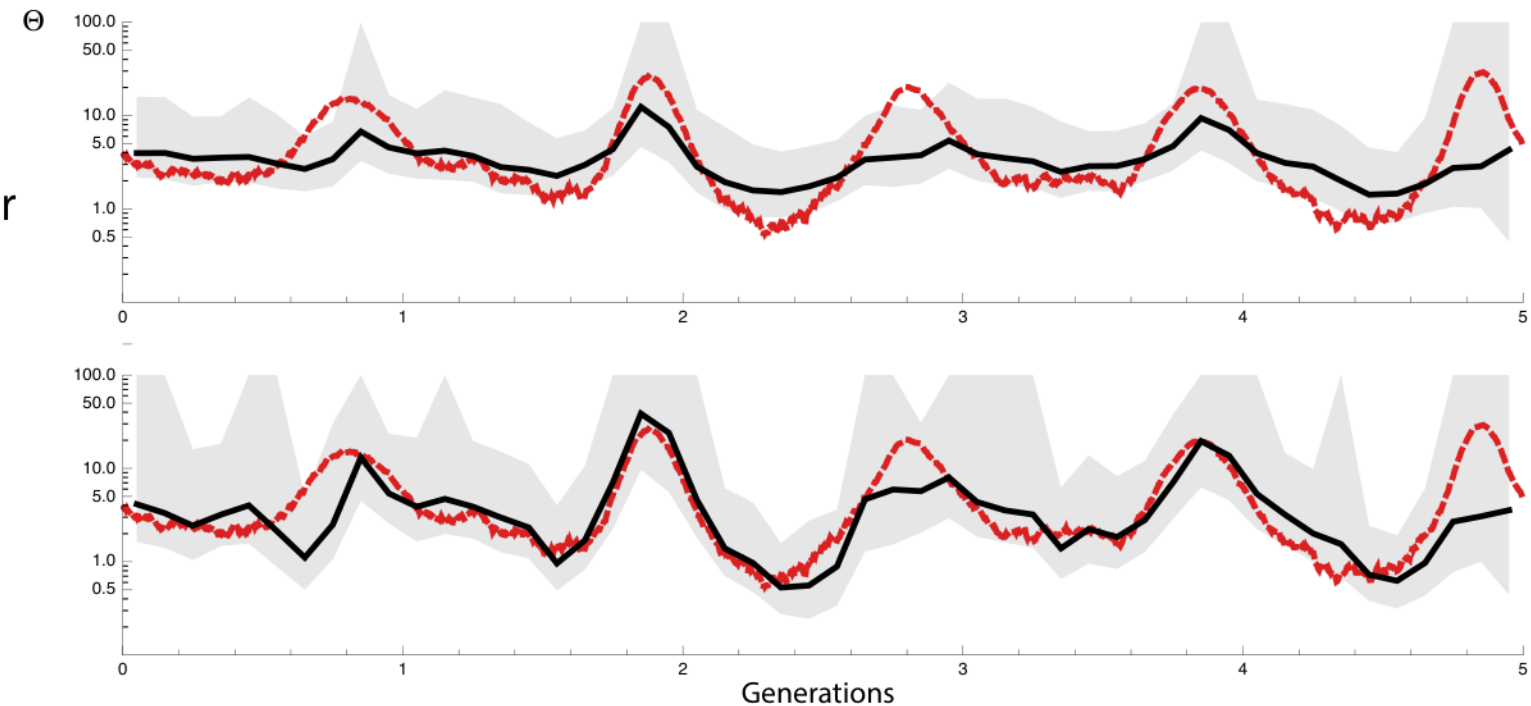


Extensions of the basic coalescent



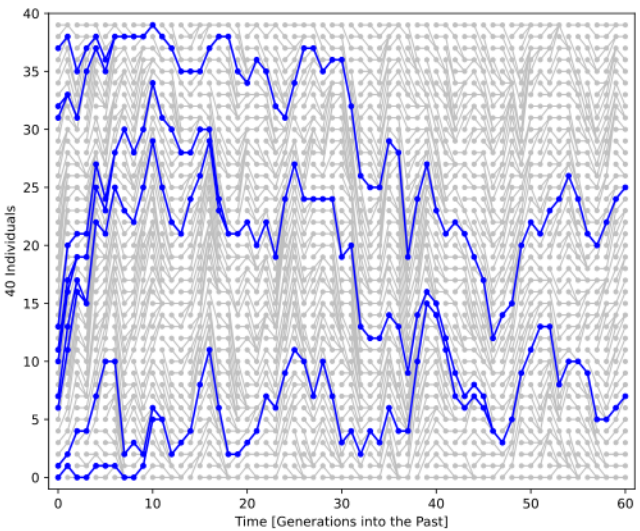
MIGRATE constant prior

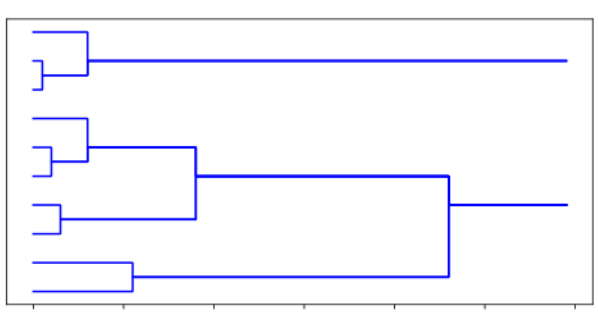
BEAST *skyline*



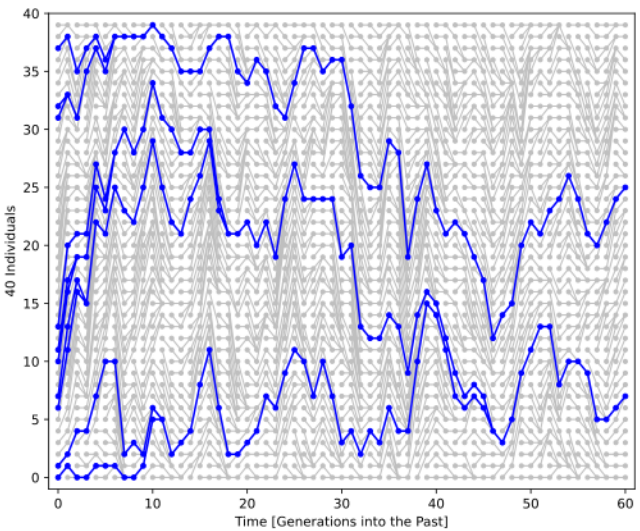
Comparison of the skyline plots of simulated influenza dynamics analyzed by MIGRATE and BEAST. The x-axis is the time in years and the y-axis is effective population size. The data are sequences from 250 individuals sampled at regular intervals over 5 years. The dashed curve is the actual population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

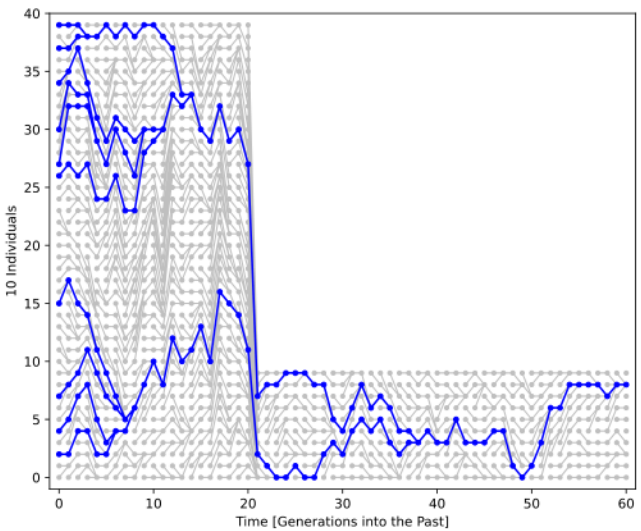
How well can we estimate bottlenecks?

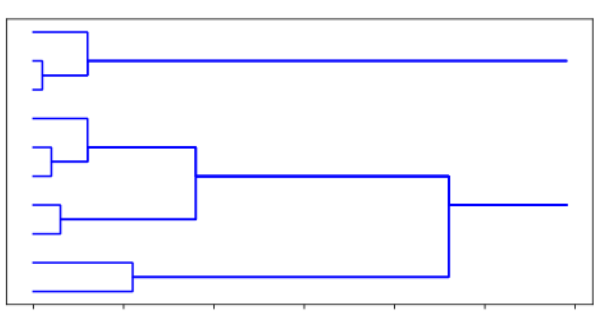


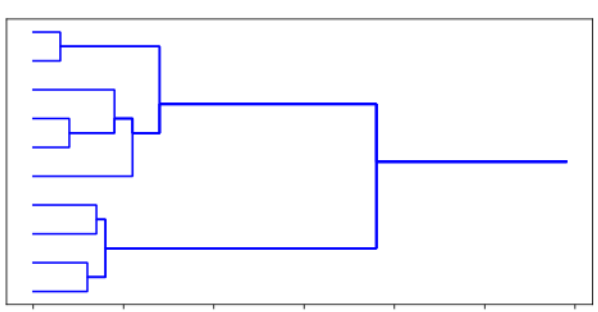


How well can we estimate bottlenecks?

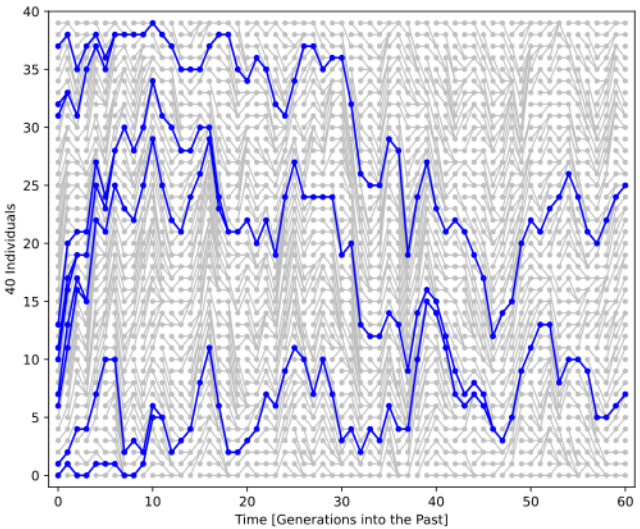


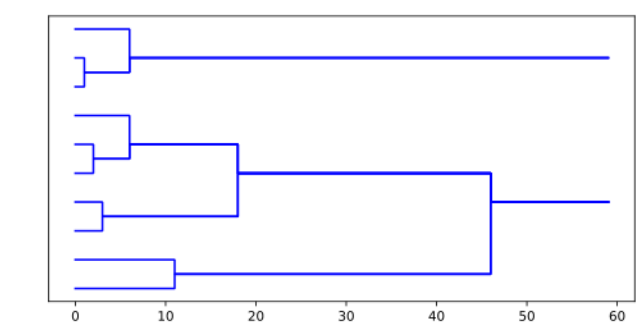


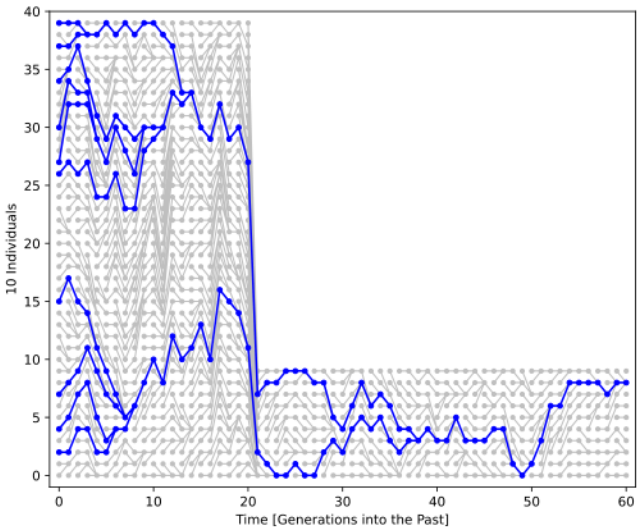


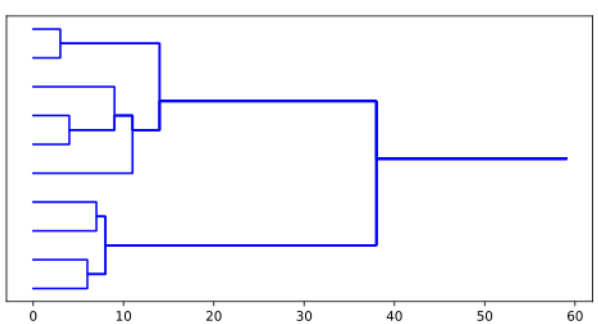


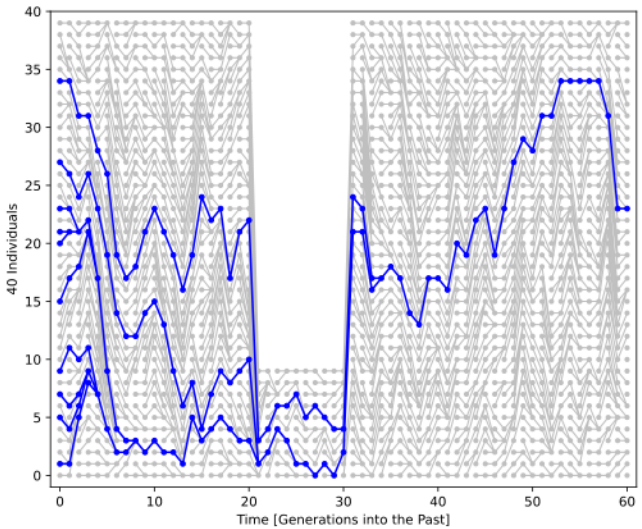
How well can we estimate bottlenecks?

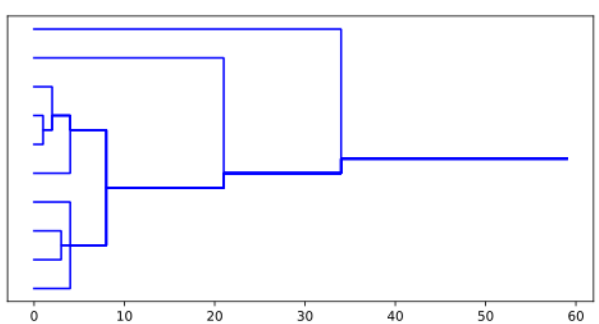




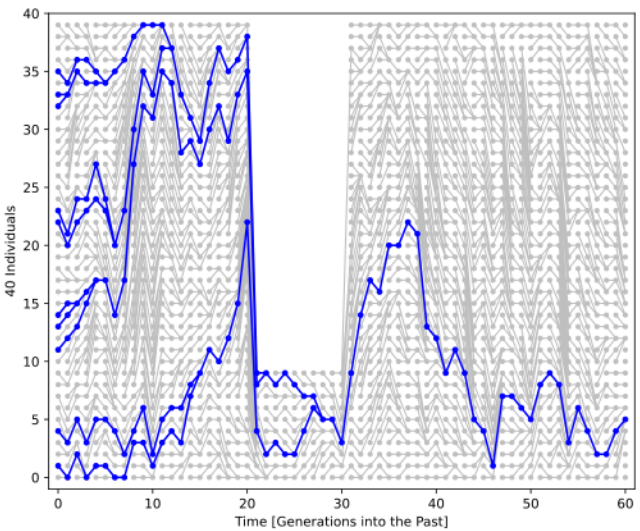


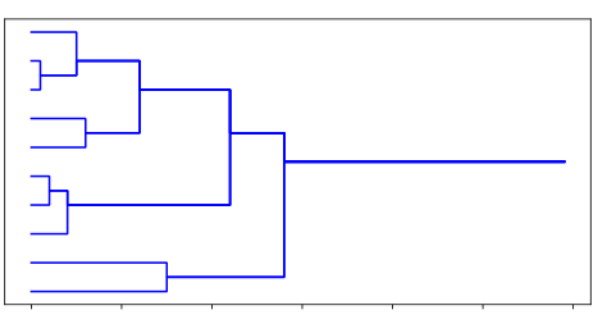


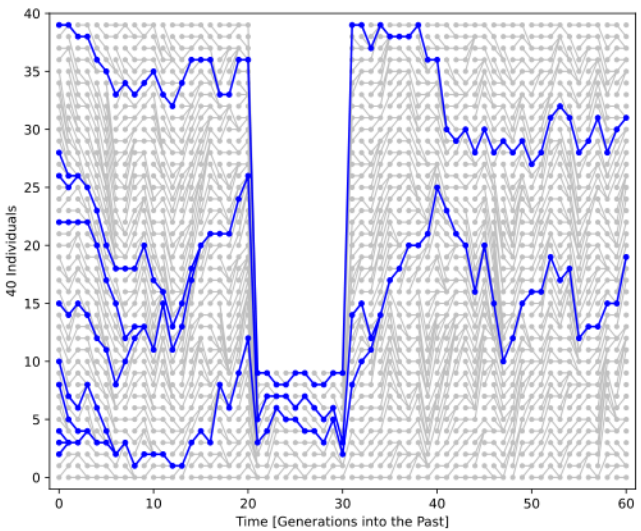


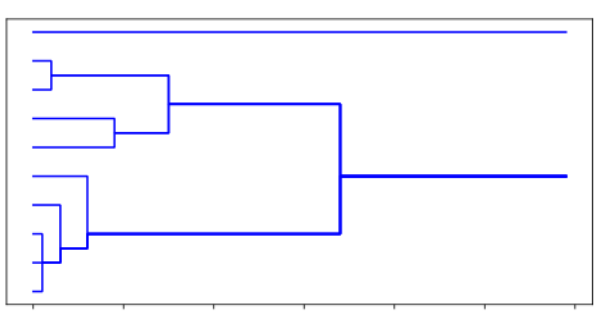


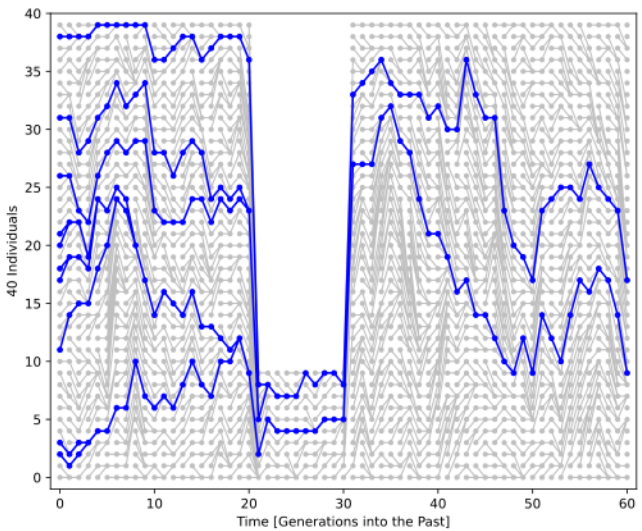
How well can we estimate bottlenecks?

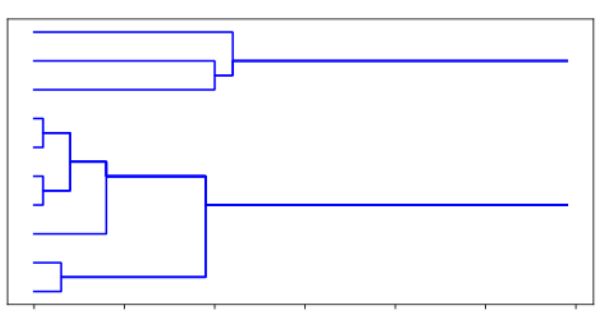




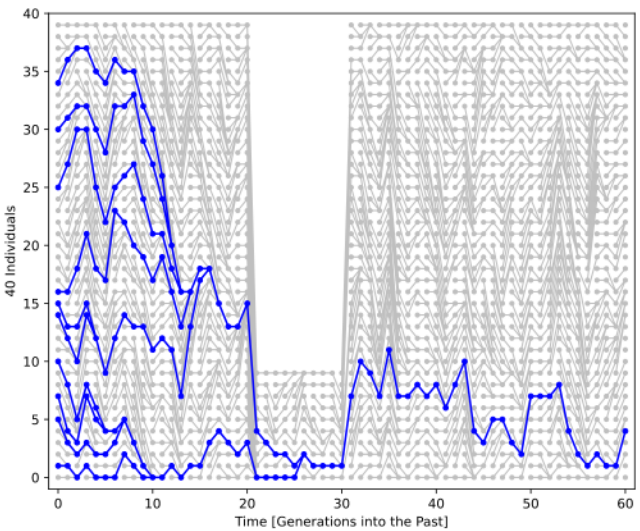


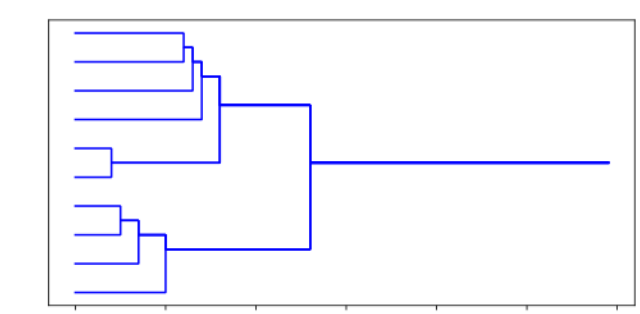


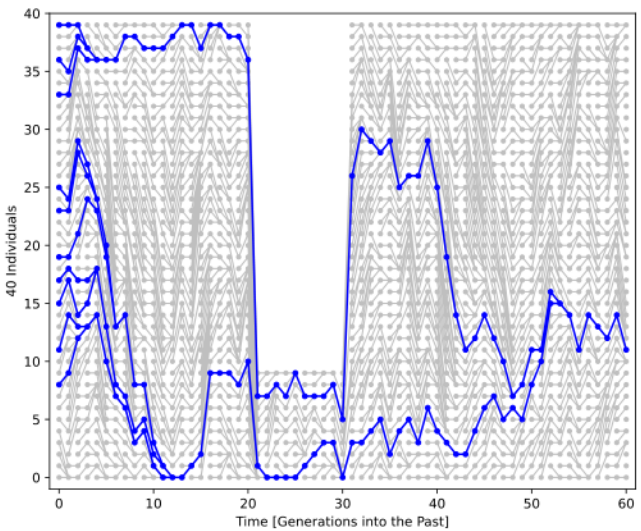


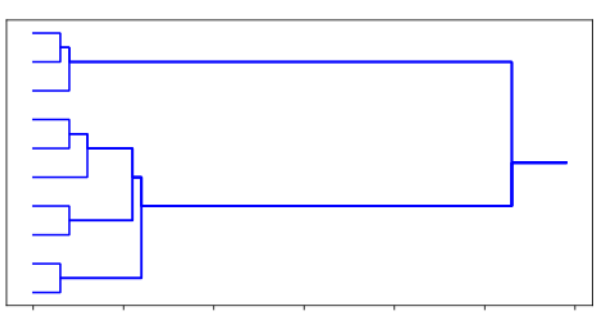


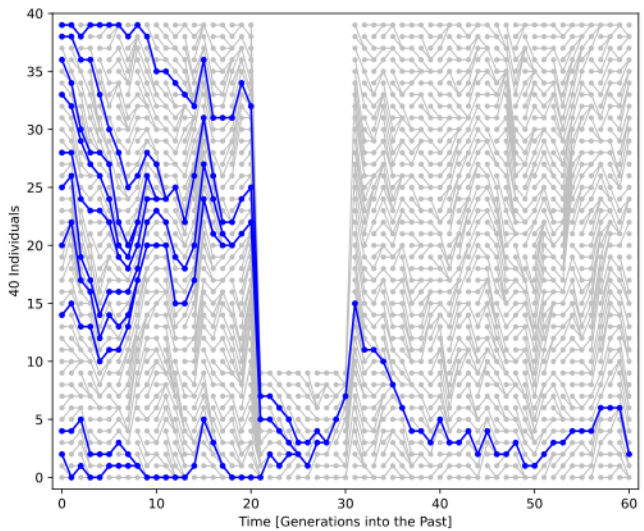
How well can we estimate bottlenecks?

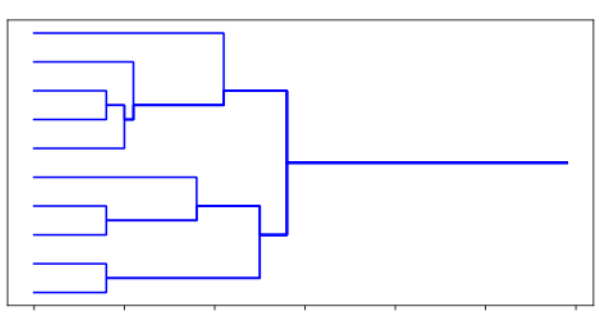




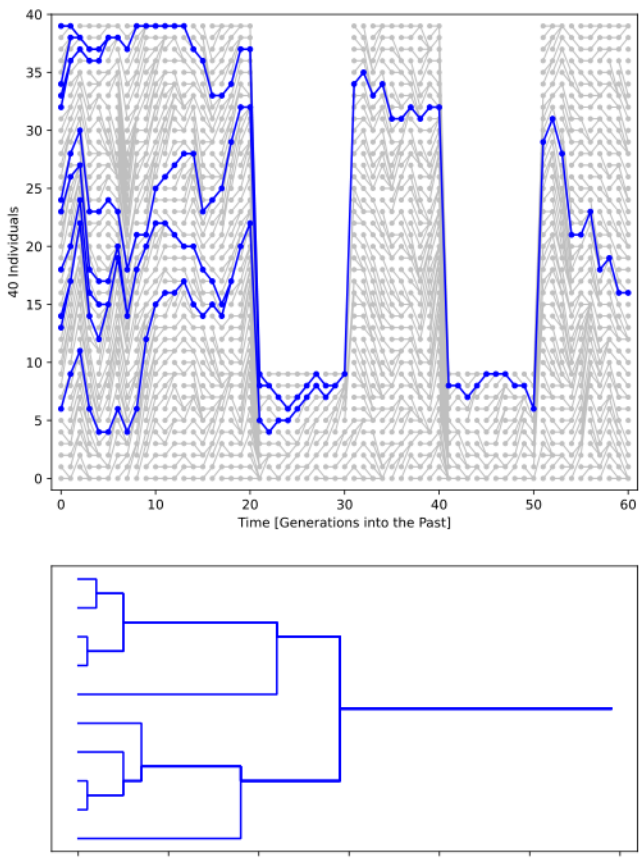








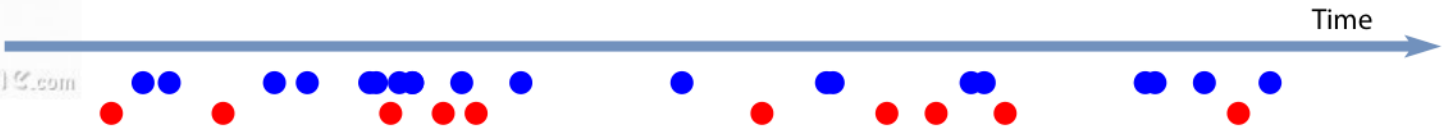
Two or more bottlenecks



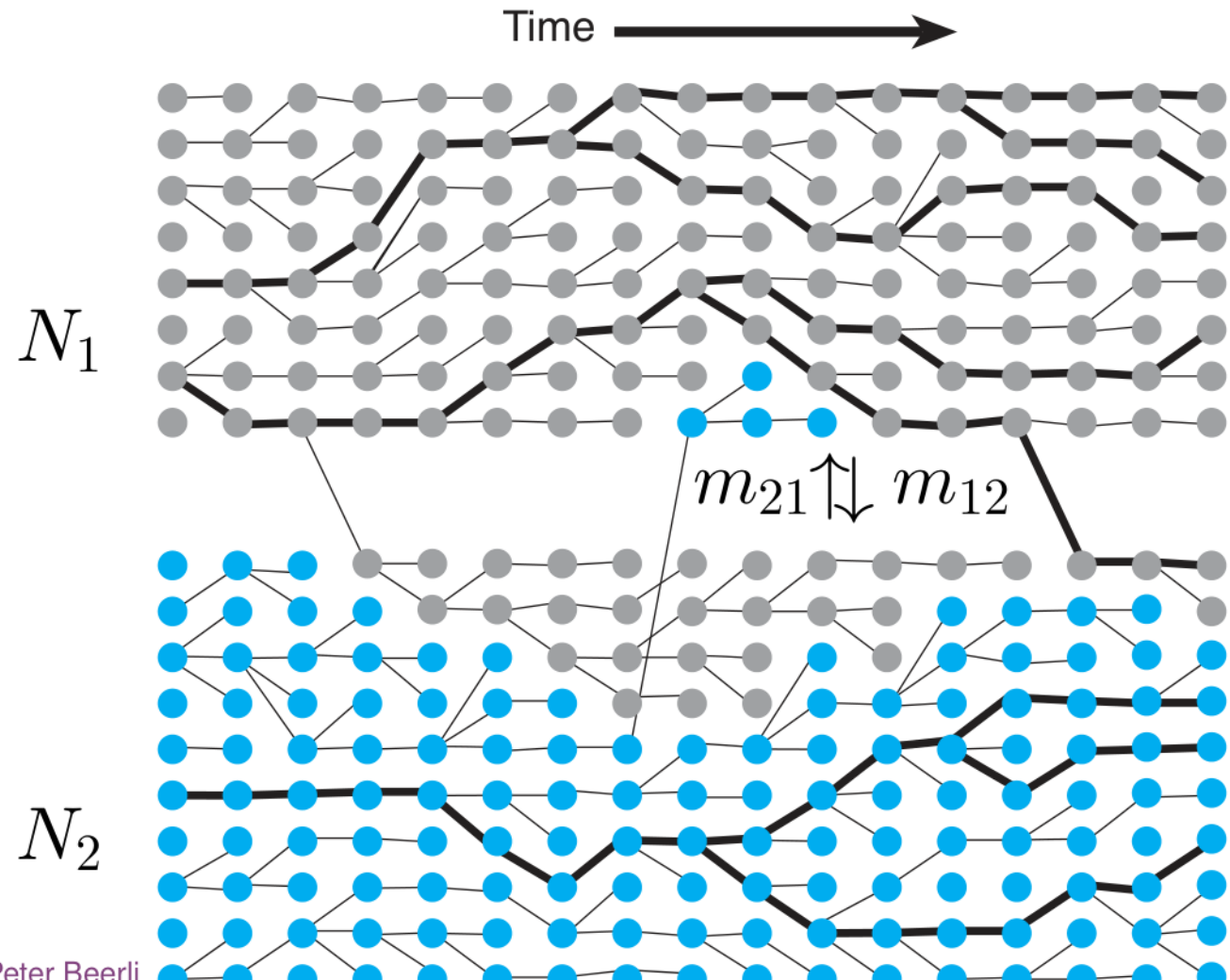
Software: popvizard at https://github.com/pbeerli/popvizard.git

Accommodating more events

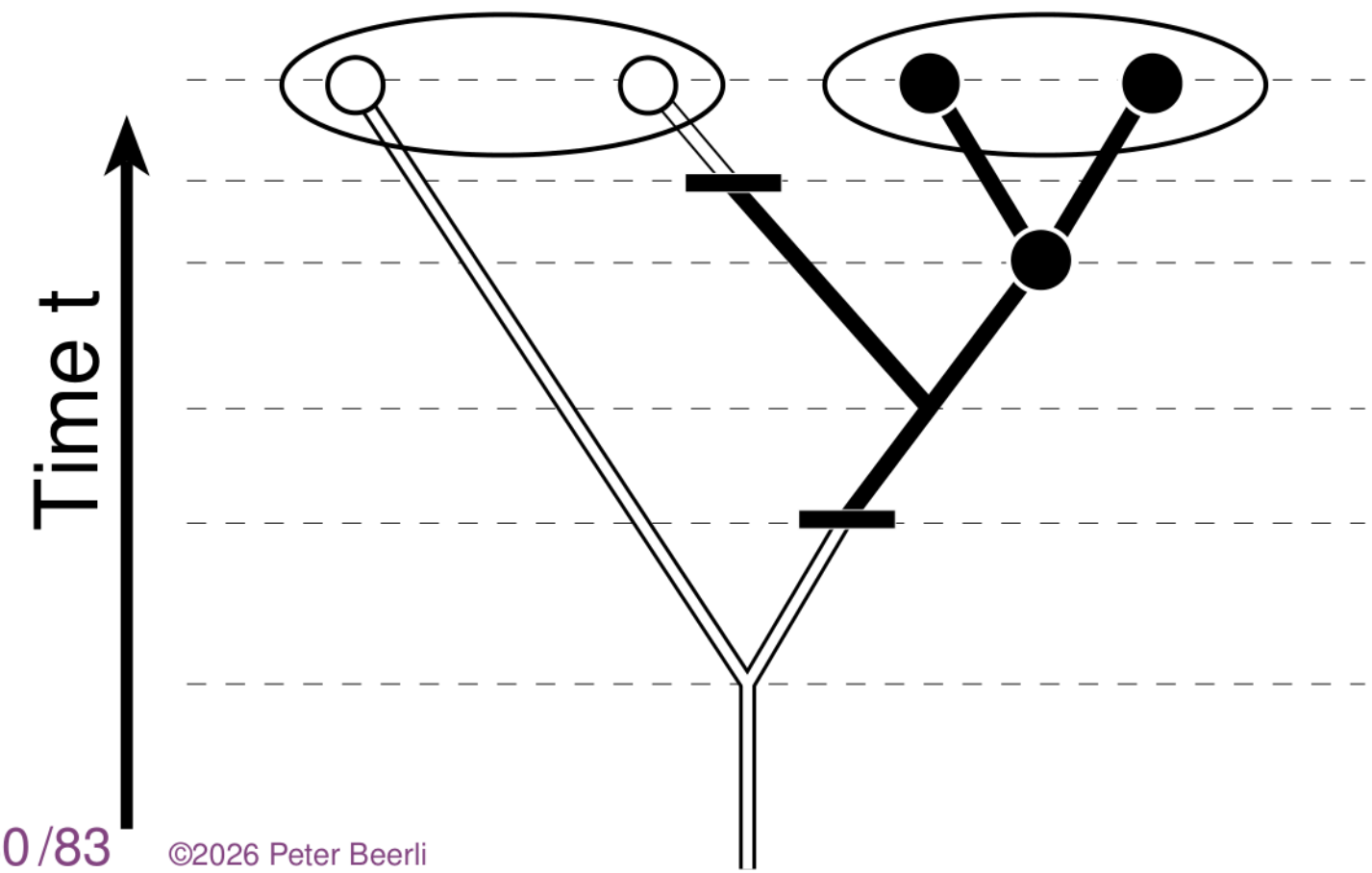
An analogy



Extensions of the basic coalescent



Extensions of the basic coalescent

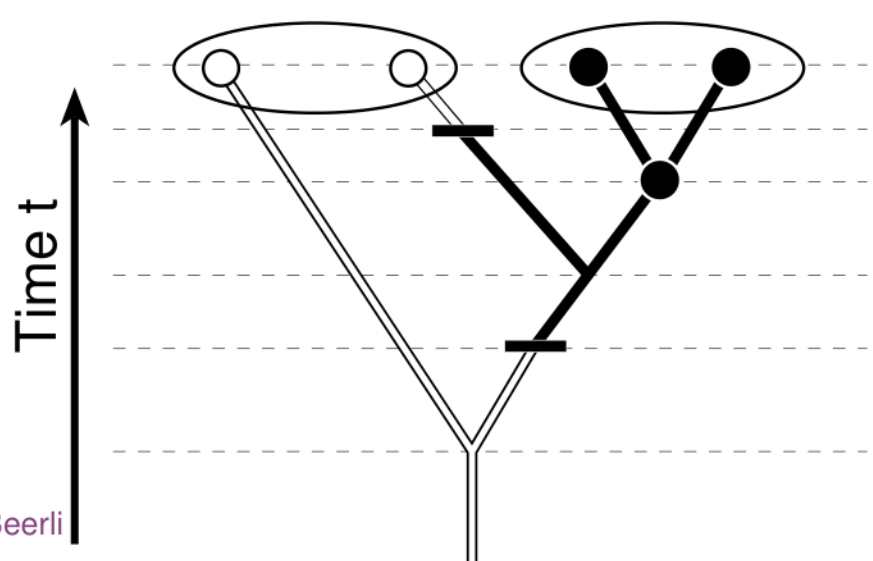


Δt	k_1	k_2
u_1	2	2
u_2	1	3
u_3	1	2
u_4	1	3
u_5	2	0

Extensions of the basic coalescent

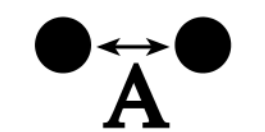
The single population coalescence rate is $\frac{k(k-1)}{4N}$.

Changes for two populations to $\frac{k_1(k_1-1)}{\Theta_1} + \frac{k_2(k_2-1)}{\Theta_2} + k_1 M_{2,1} + k_2 M_{1,2}$

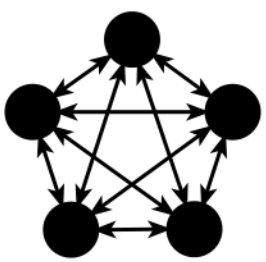


Δt	k_1	k_2
u_1	2	2
u_2	1	3
u_3	1	2
u_4	1	3
u_5	2	0

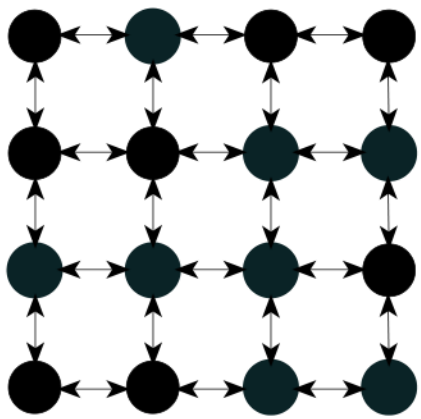
Structured populations



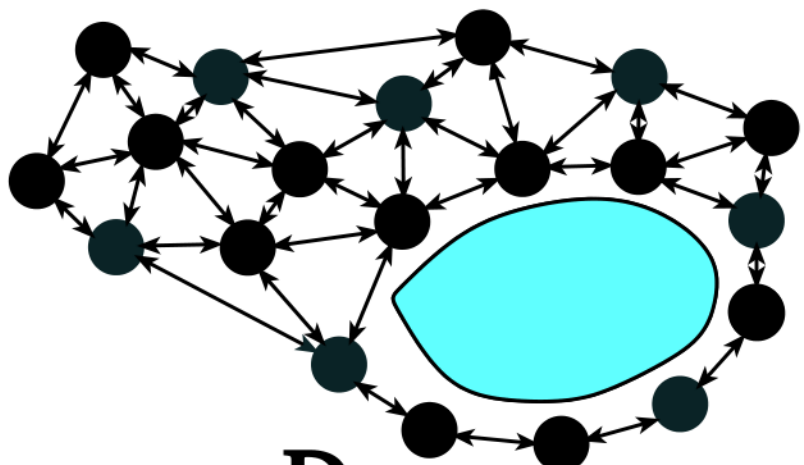
A



B



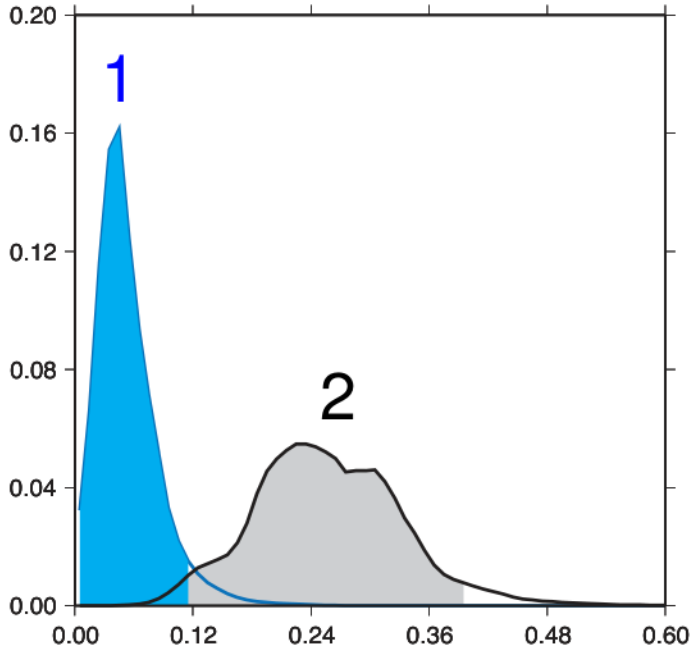
C



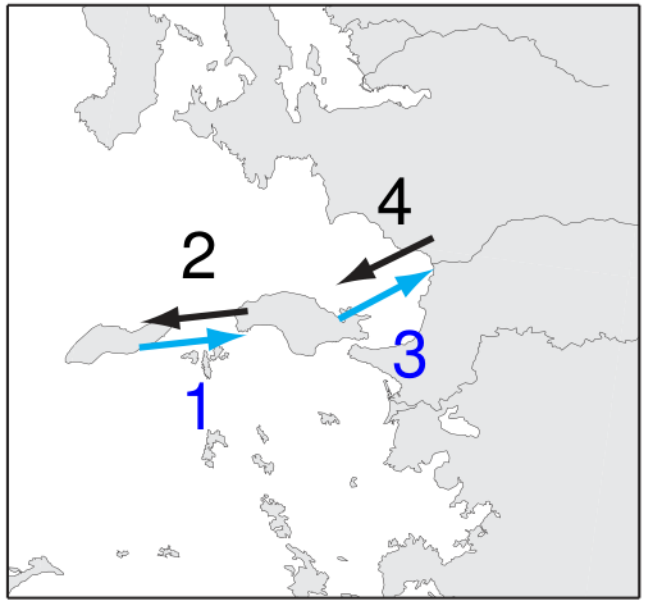
D

Obvious migration pattern

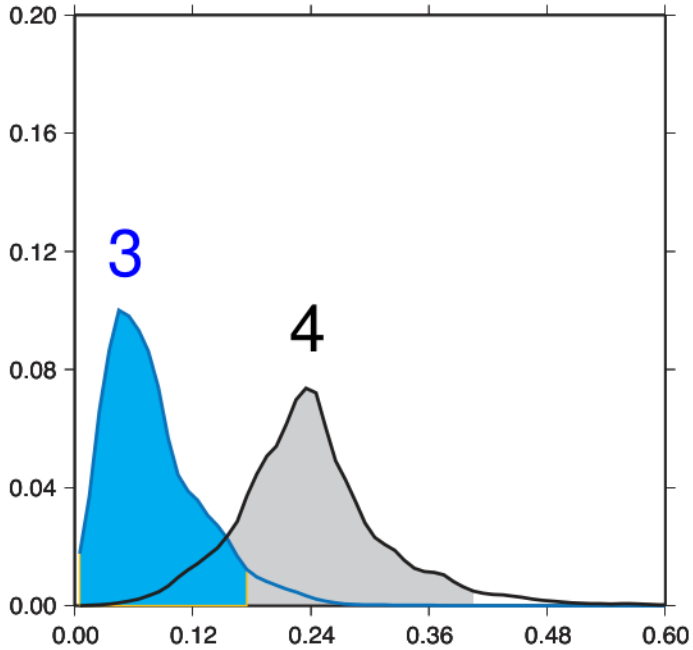
$p(\mathcal{M}|D)$



scaled migration rate

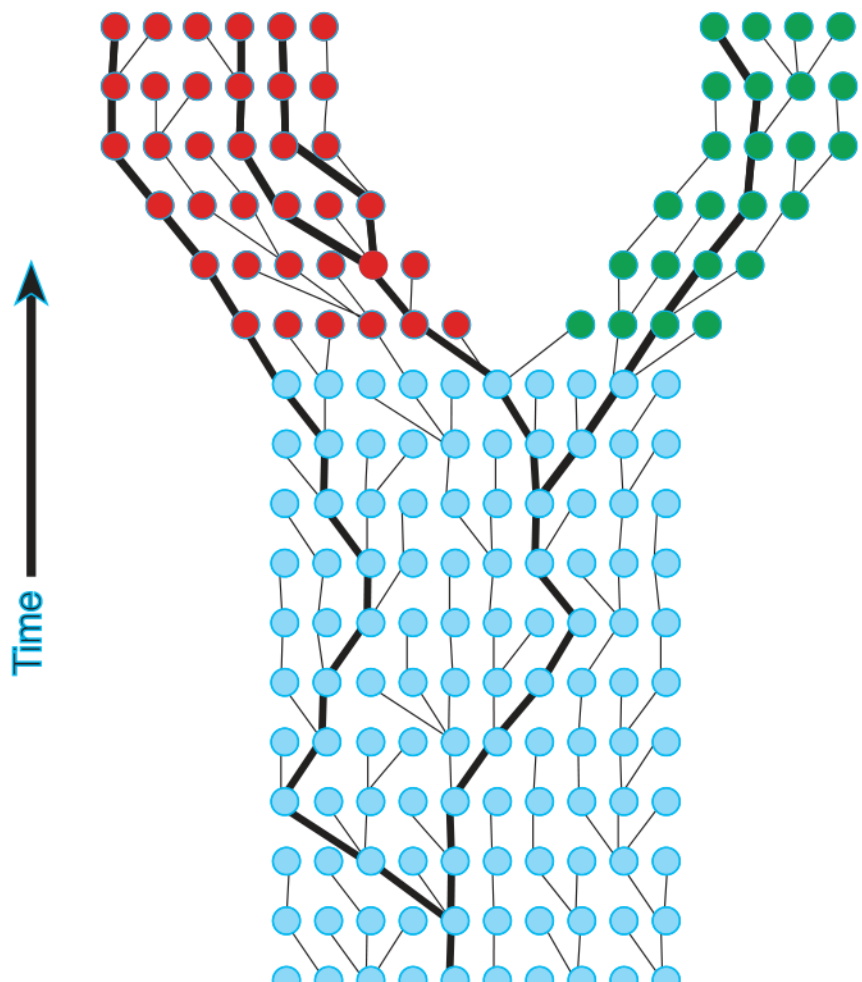


$p(\mathcal{M}|D)$

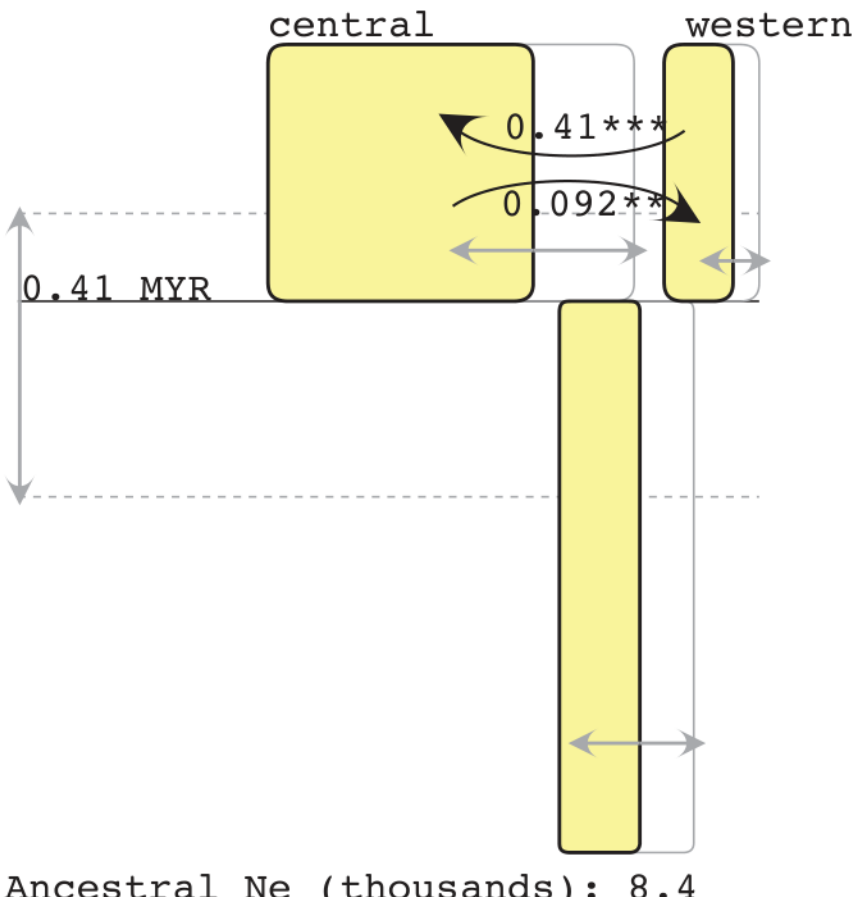


scaled migration rate

Extensions of the basic coalescent



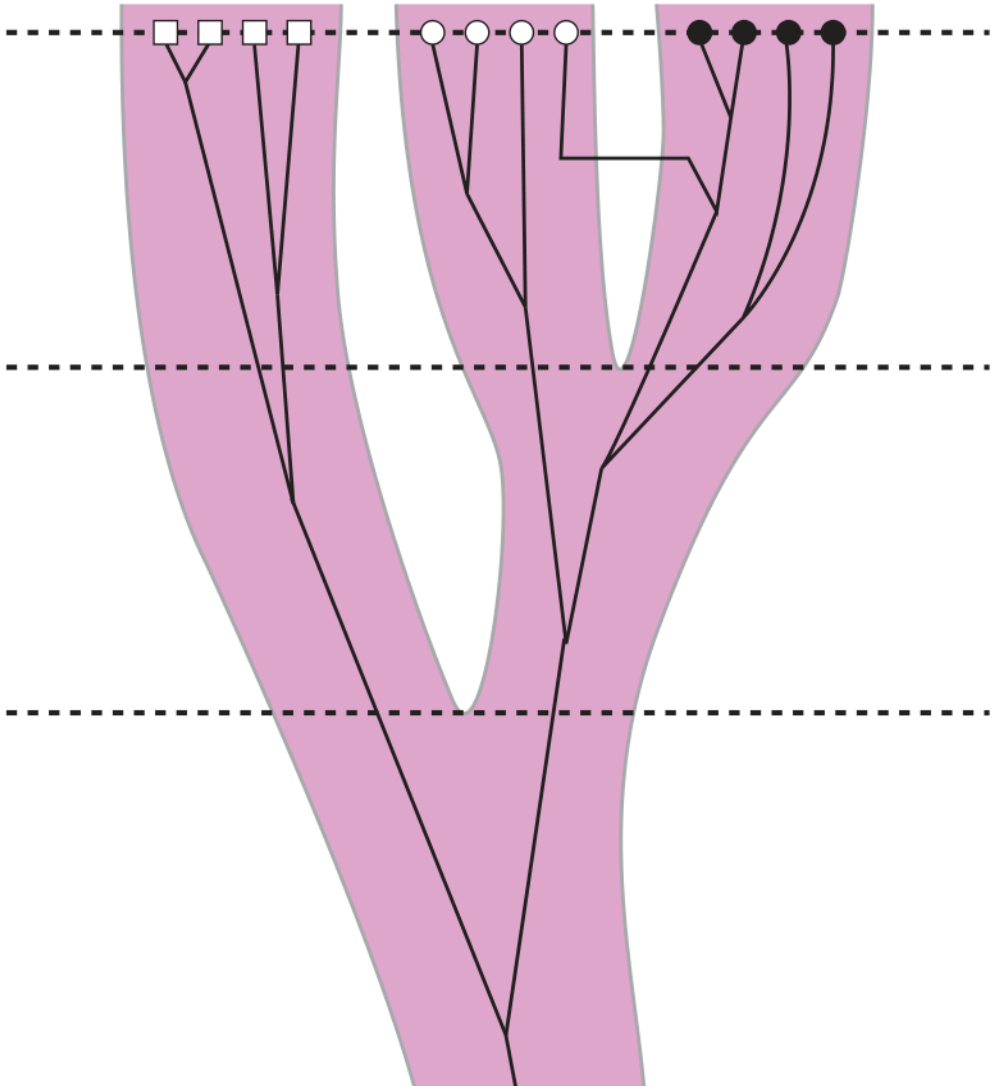
Population splitting



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.

Hey 2010

Population splitting



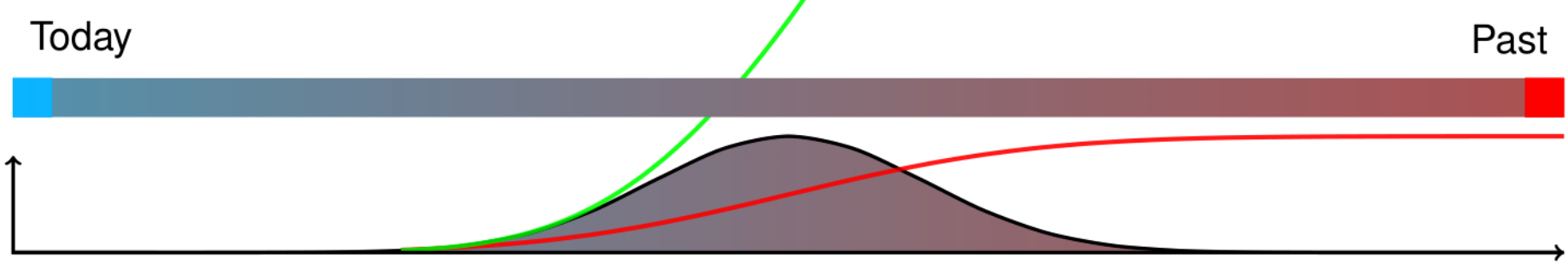
Population splitting

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?

(Beerli P., Ashki H., Mashayekhi S., and Palczewski M. 2022. Population divergence time estimation using individual lineage label switching. G3 Genes – Genomes – Genetics, 12(4),URL https://doi.org/10.1093/g3journal/jkac040.)

Population splitting

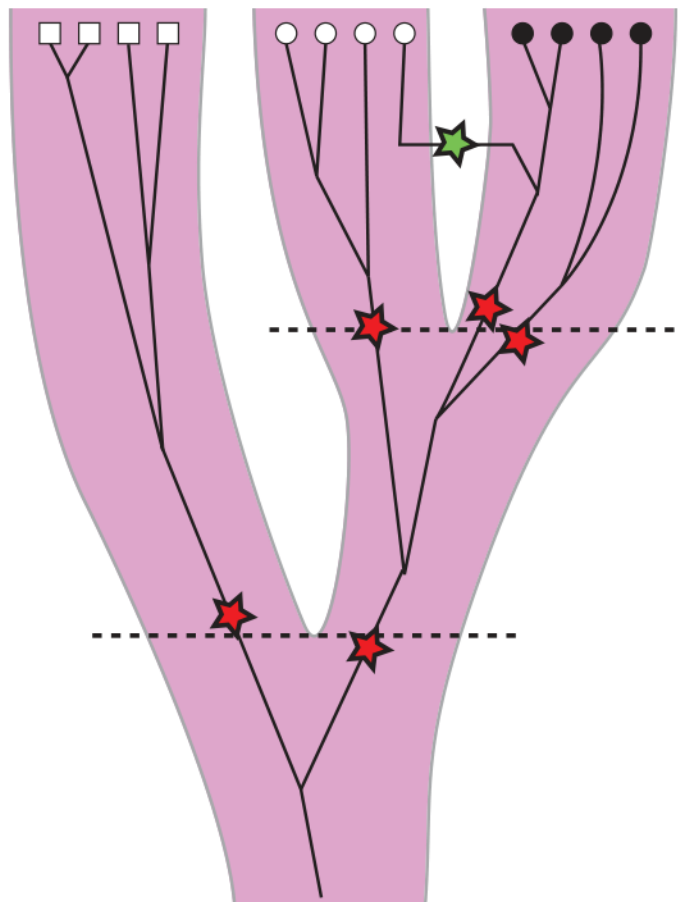
Looking backwards in time we could think about the risk of A turning into B which becomes larger and larger the further back in time the lineage goes. In the coalescence framework we are well accustomed to that thinking: we use the risk of a coalescent or the risk of a migration event. This risk can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



(Beerli P., Ashki H., Mashayekhi S., and Palczewski M. 2022. Population divergence time estimation using individual lineage label switching. G3 Genes – Genomes – Genetics, 12(4),URL https://doi.org/10.1093/g3journal/jkac040.)

Population splitting

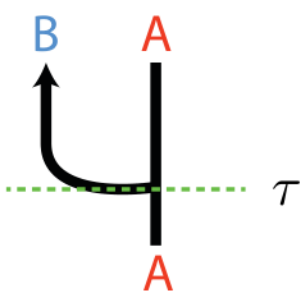
One lineage is easy, but what about the genealogy? Each lineage is at risk of being in the ancestral population, thus we need to consider coalescences, migration events, and population label changing events. This results in genealogies that are realizations of migration and population splitting events.



(Beerli P., Ashki H., Mashayekhi S., and Palczewski M. 2022. Population divergence time

Population splitting

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



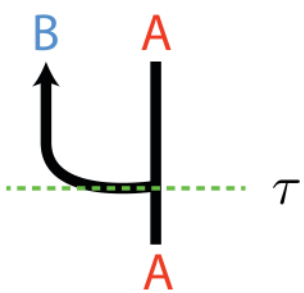
Model, τ → Genealogy

Genealogy → Sequence data

Sequence Data → Model → $\hat{\tau}$

Population splitting

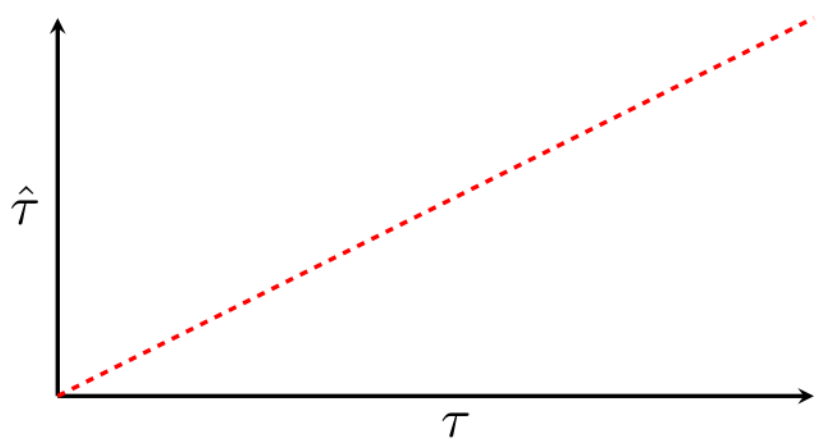
Comparison of estimated versus simulated divergence times for different number of loci



Model, τ → Genealogy

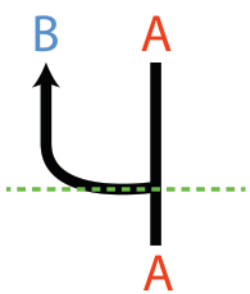
Genealogy → Sequence data

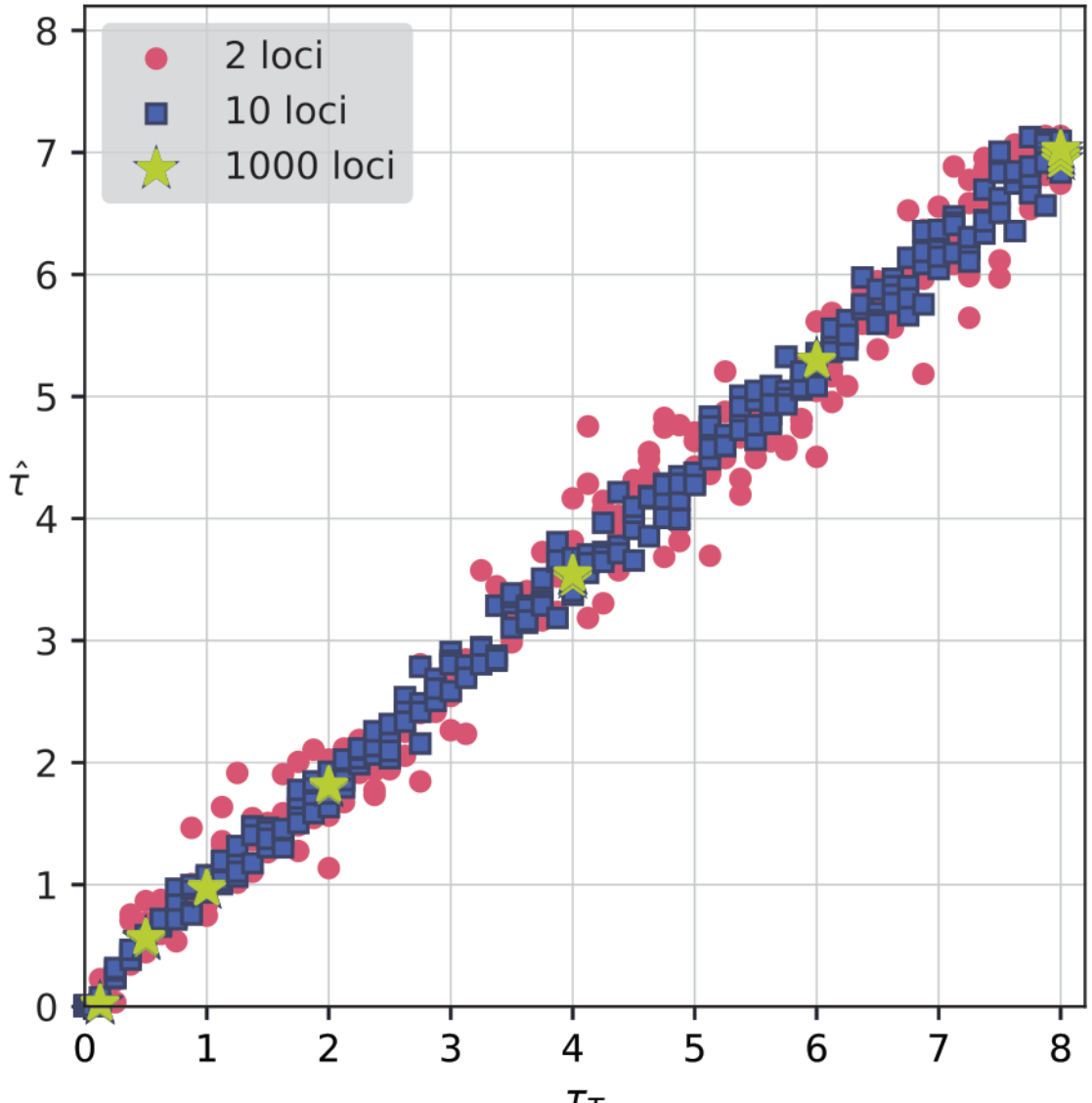
Sequence Data → Model → $\hat{\tau}$



Population splitting

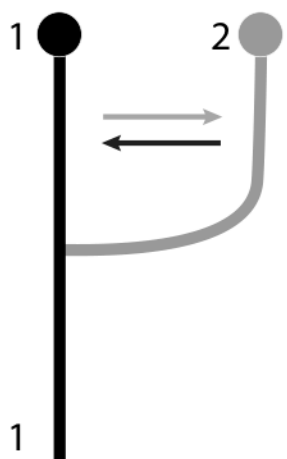
Comparison of estimated versus simulated divergence times for different number of loci

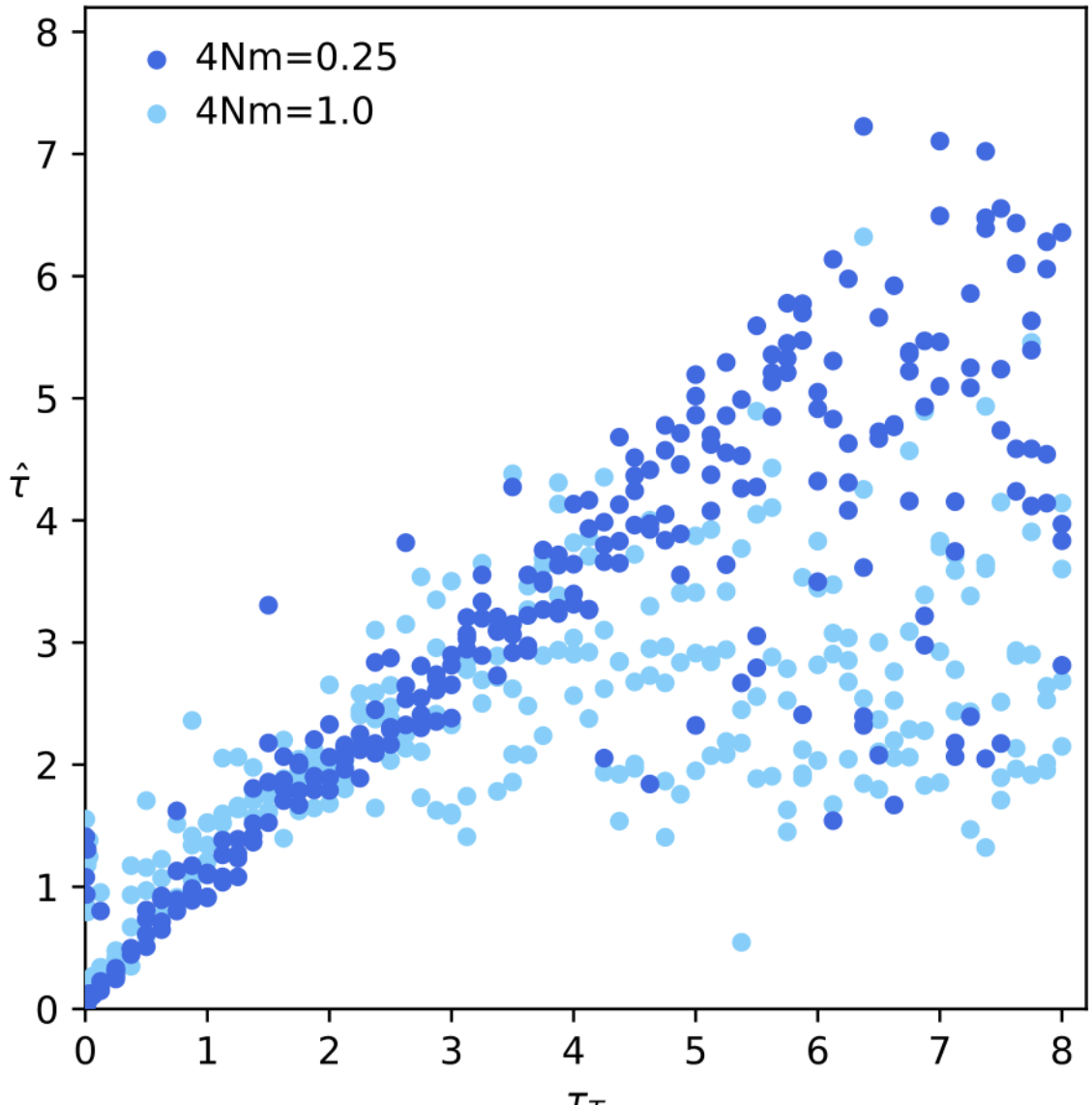




(Beerli P., Ashki H., Mashayekhi S., and Palczewski M. 2022. Population divergence time estimation using indivi

Population splitting



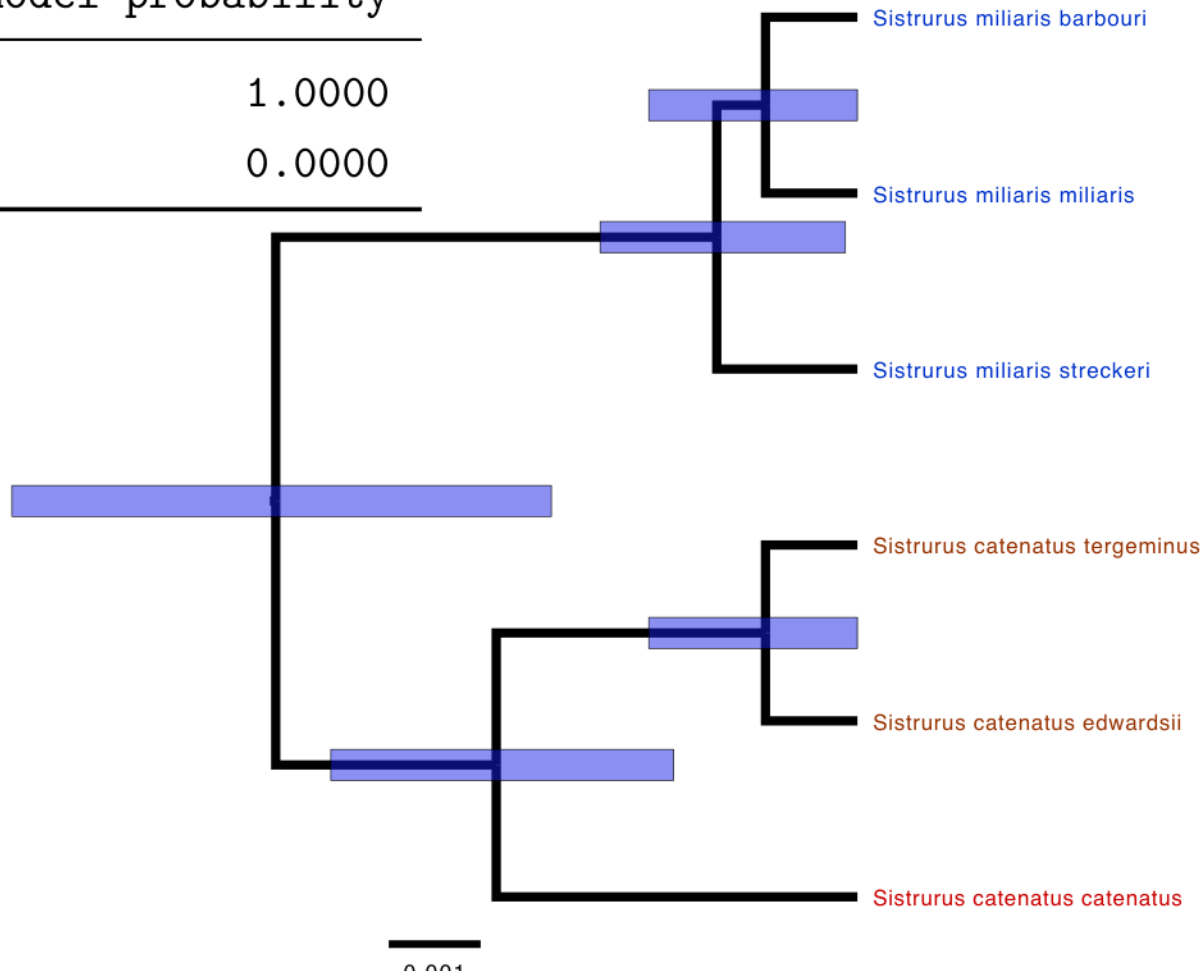


(Beerli P., Ashki H., Mashayekhi S., and Palczewski M. 2022. Population divergence time

Population splitting: Pygmy rattle snakes

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

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

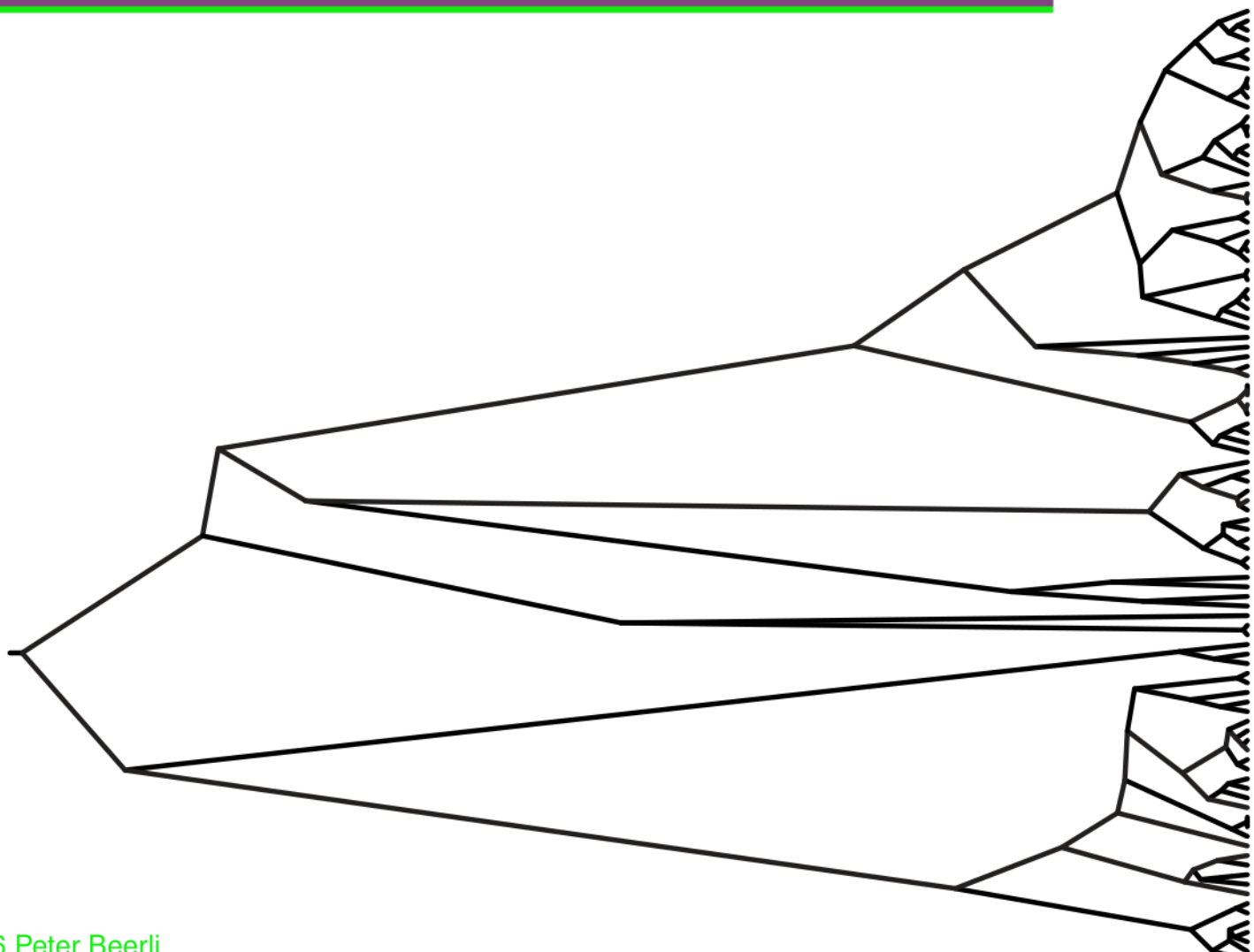


Robustness of the coalescence

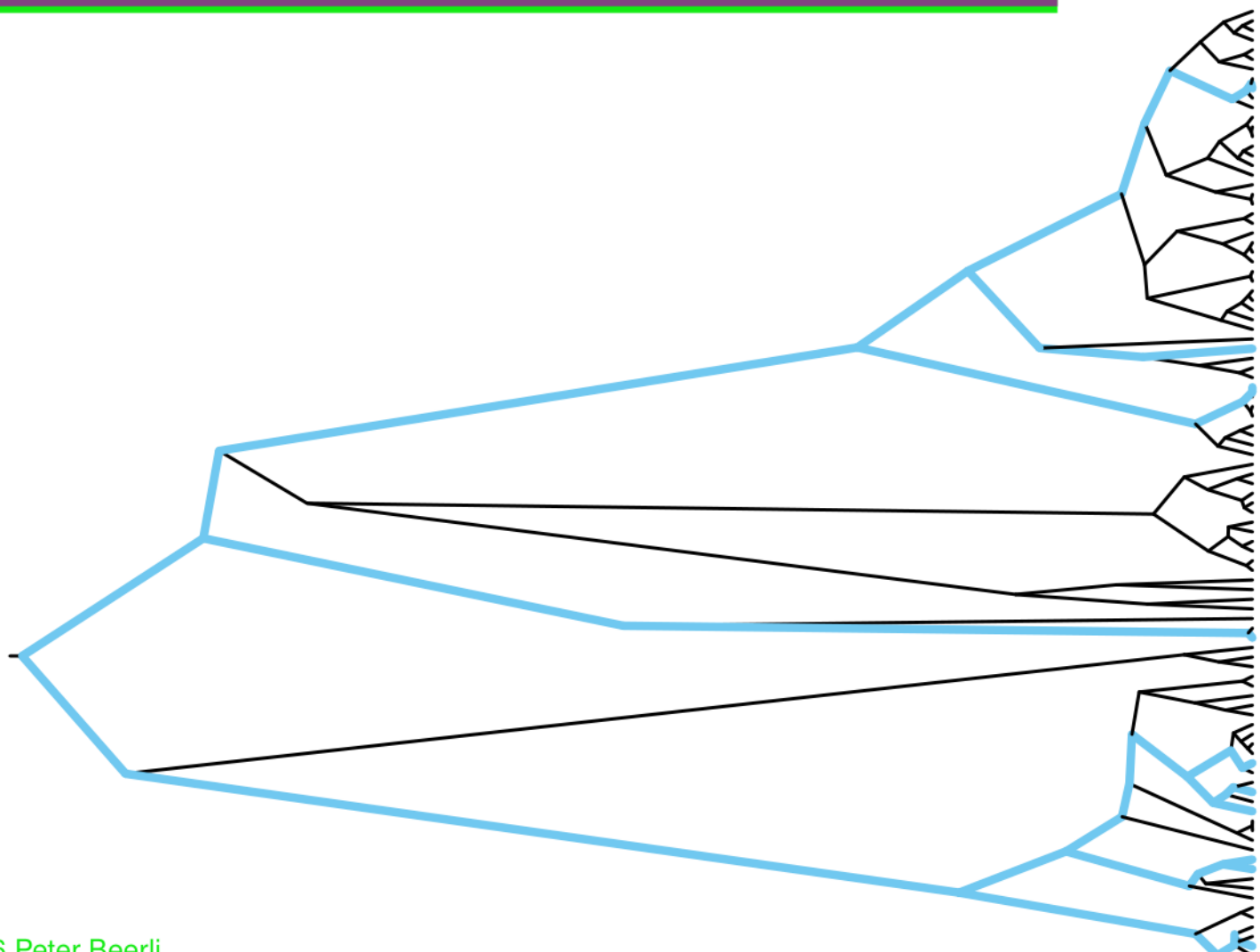
Violating assumptions

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

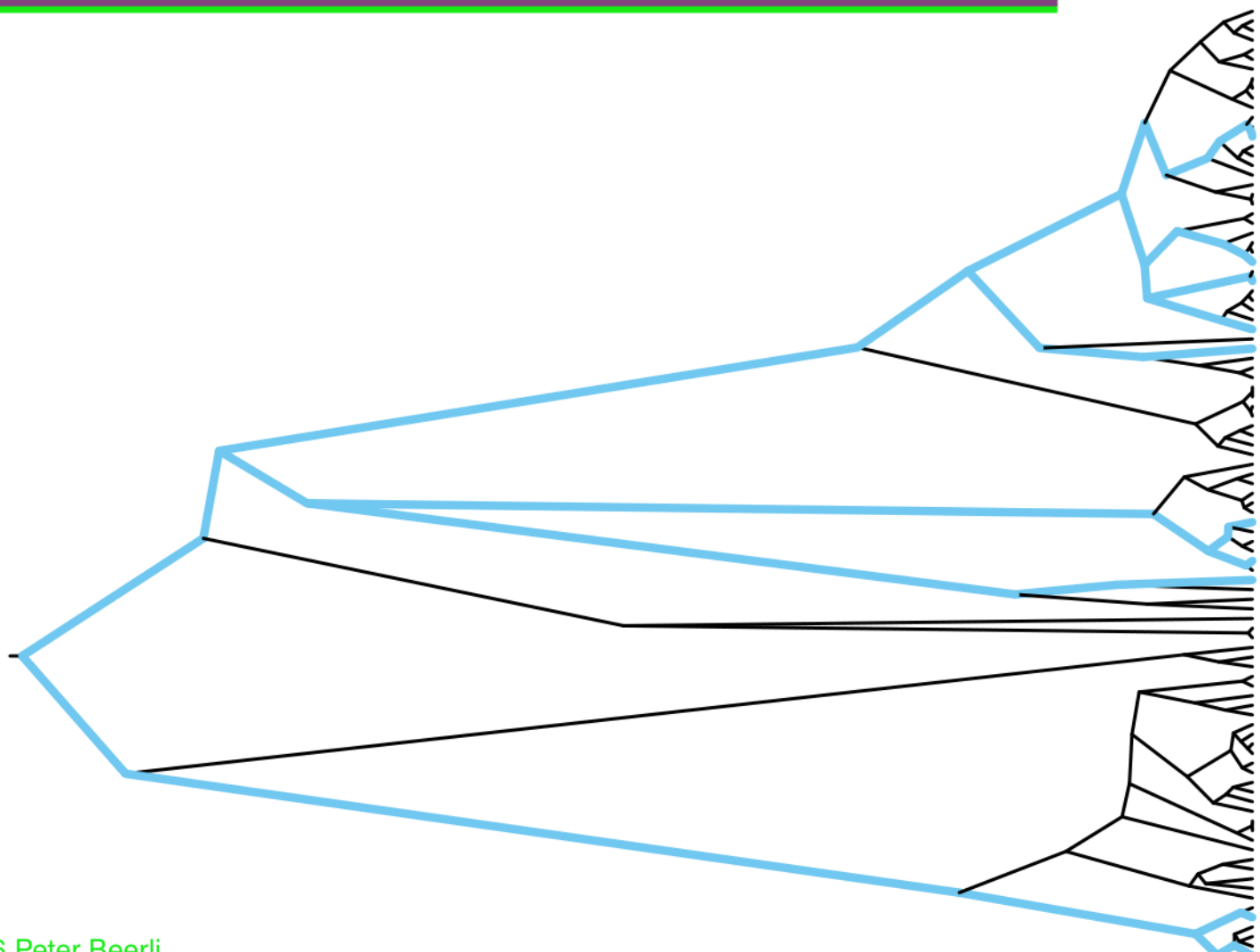
Required samples is small



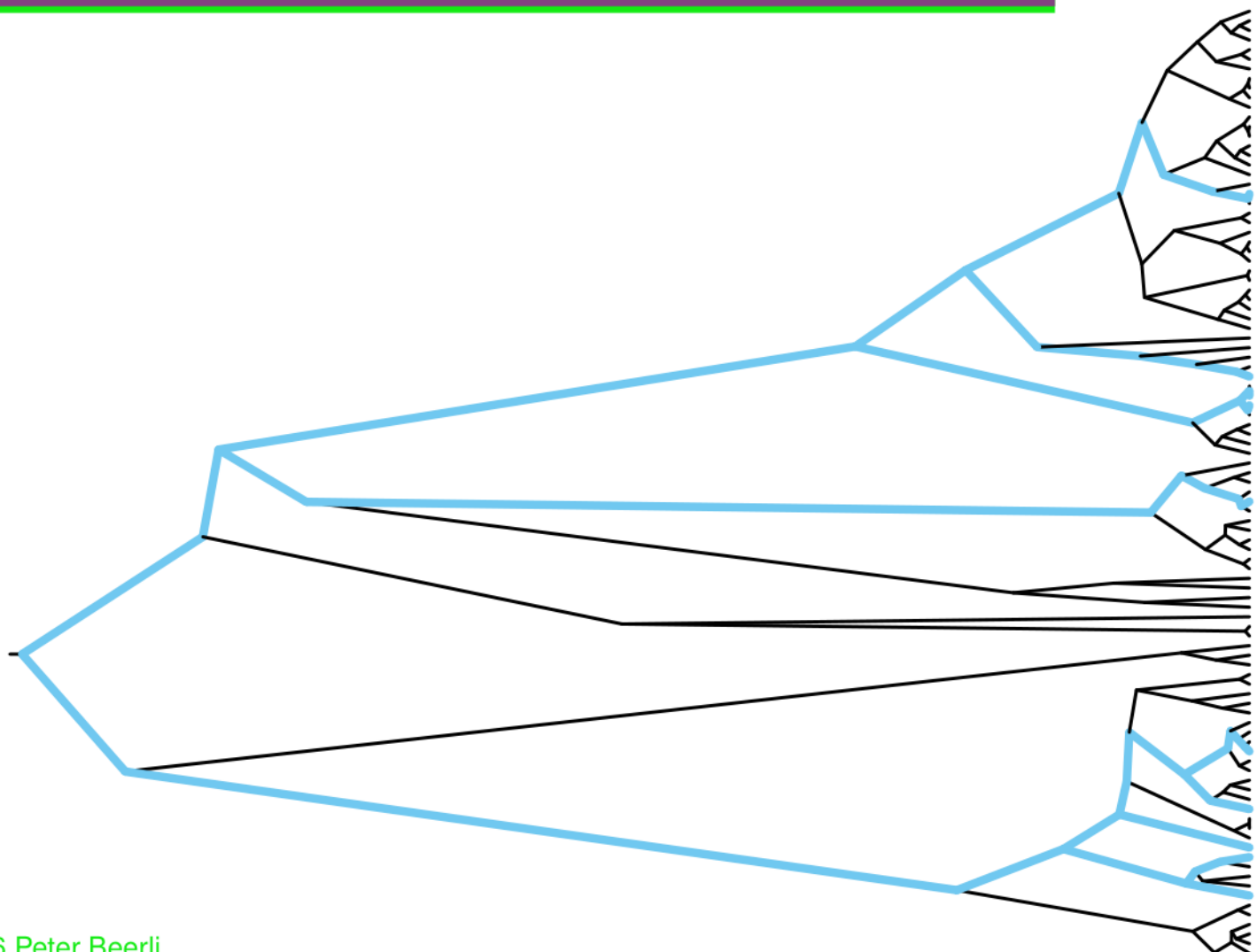
Required samples is small



Required samples is small

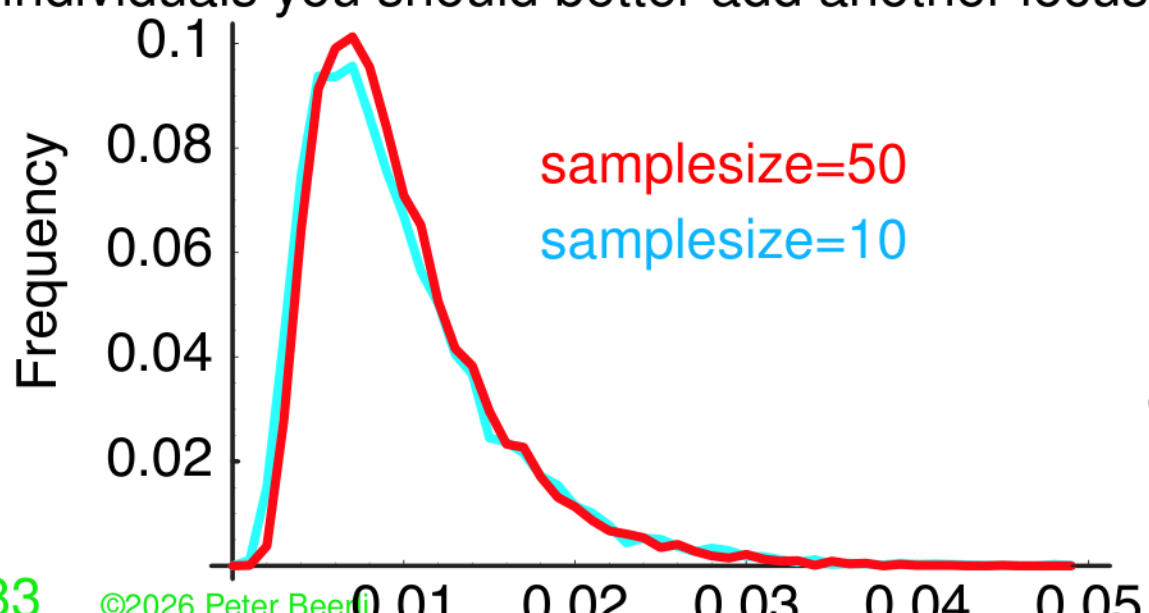


Required samples is small



Required samples is small

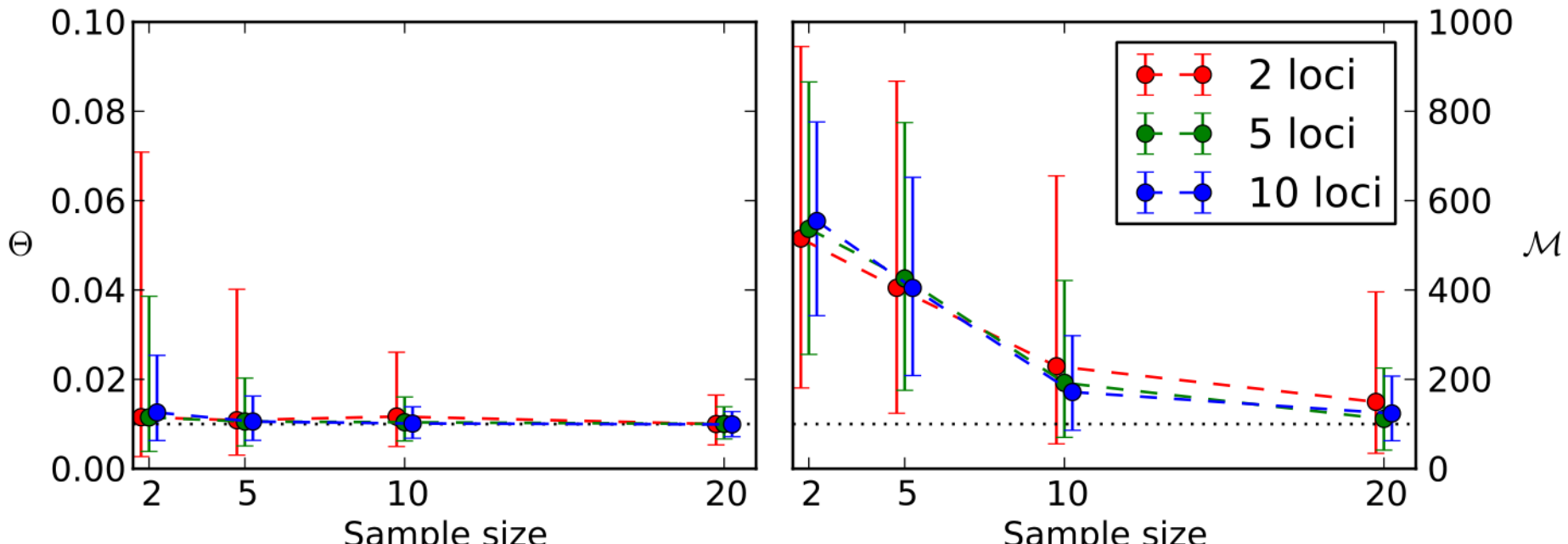
- The time to the most recent common ancestor is robust to different sample sizes.
- Simulated sequence data from a single population have shown that after 8 individuals you should better add another locus than more individuals.



Felsenstein (2005)

Pluzhnikov and Donnelly (1996)

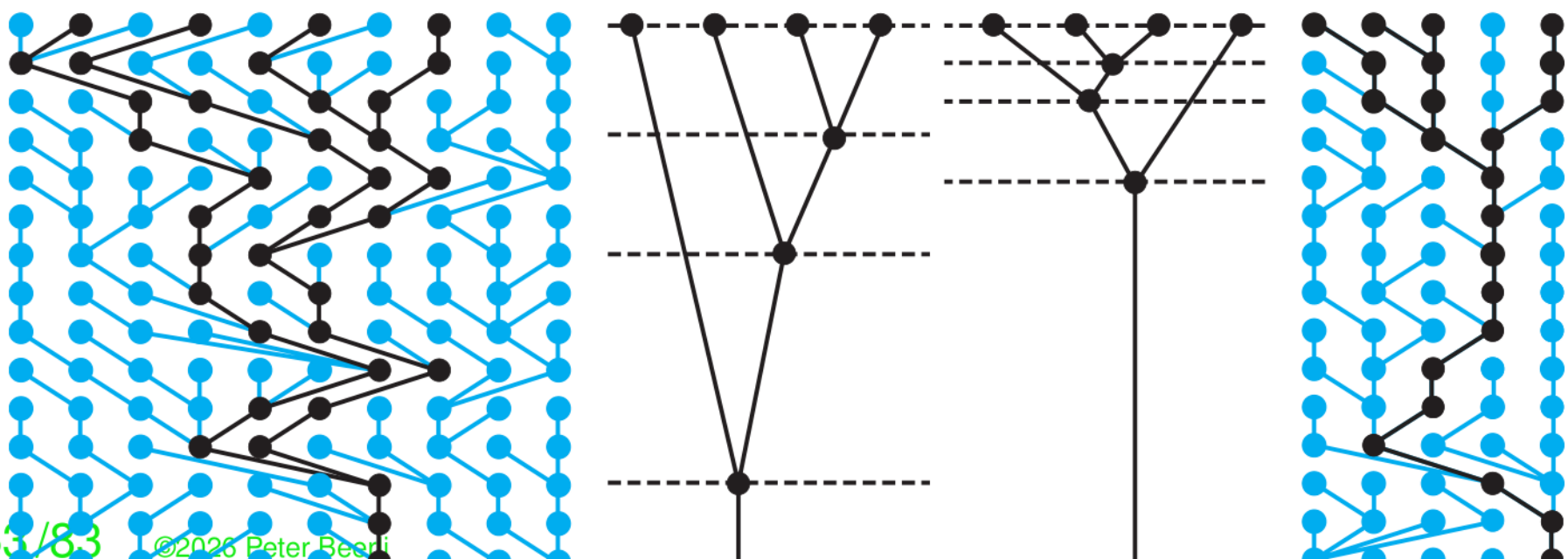
Required number of samples is small



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; $M = 100$

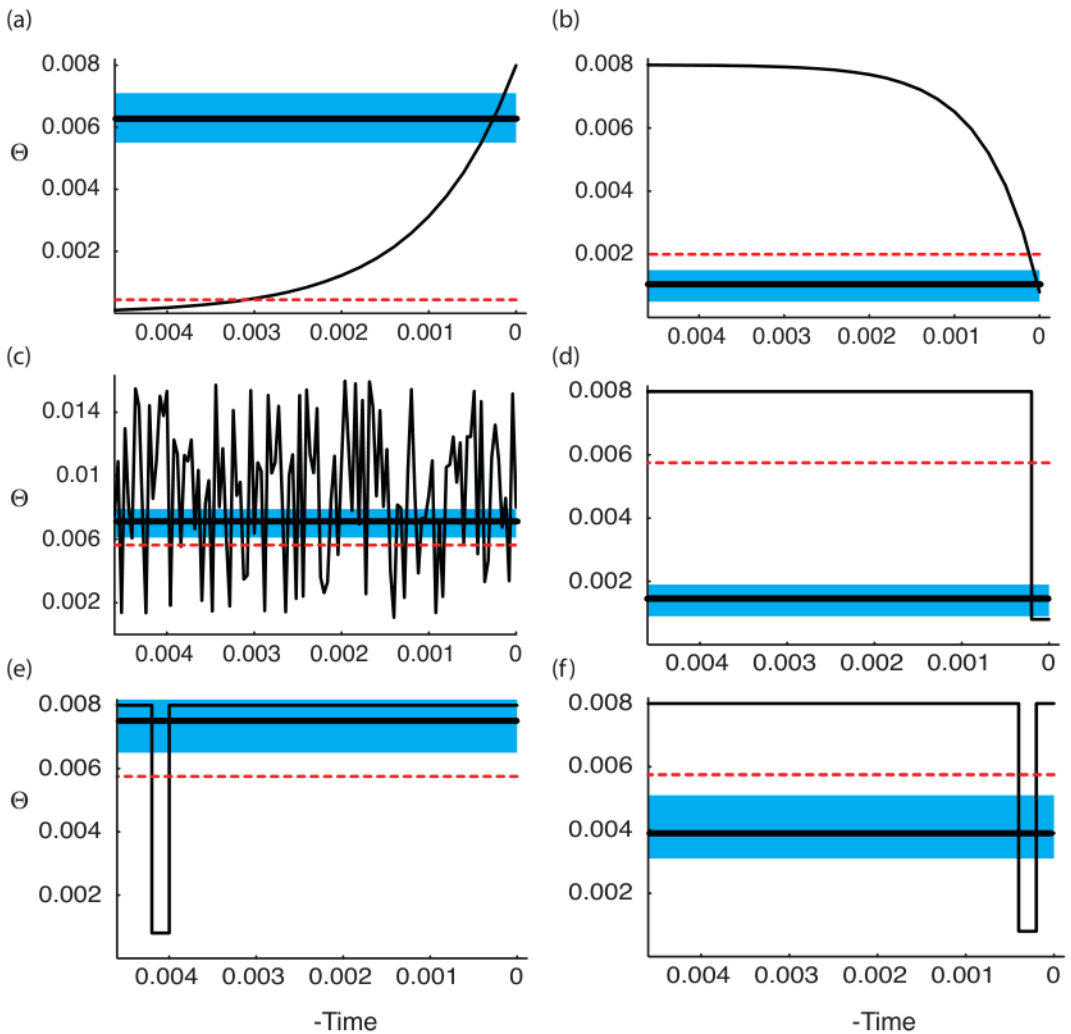
Average of parameters over long time

Researchers from the frequency-based camp (those that use F_{ST}) 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 ...



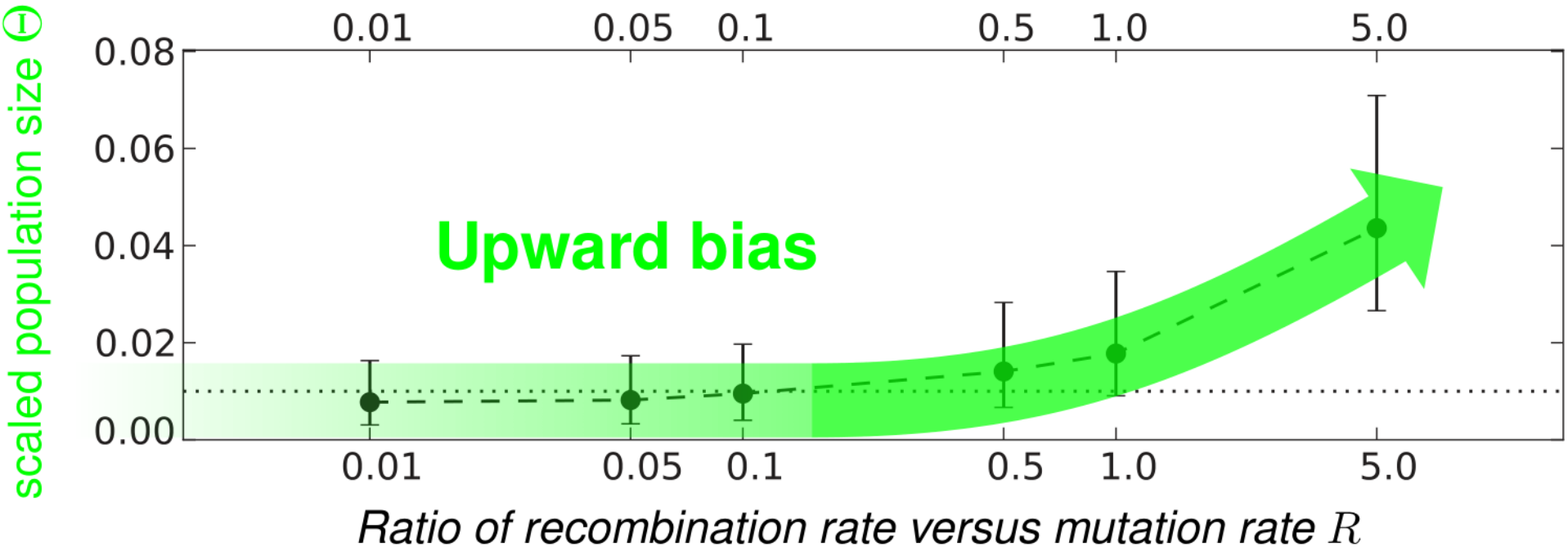
Average of parameters over long time

True value
MIGRATE estimate
Support interval
Harmonic mean



Ignoring recombination

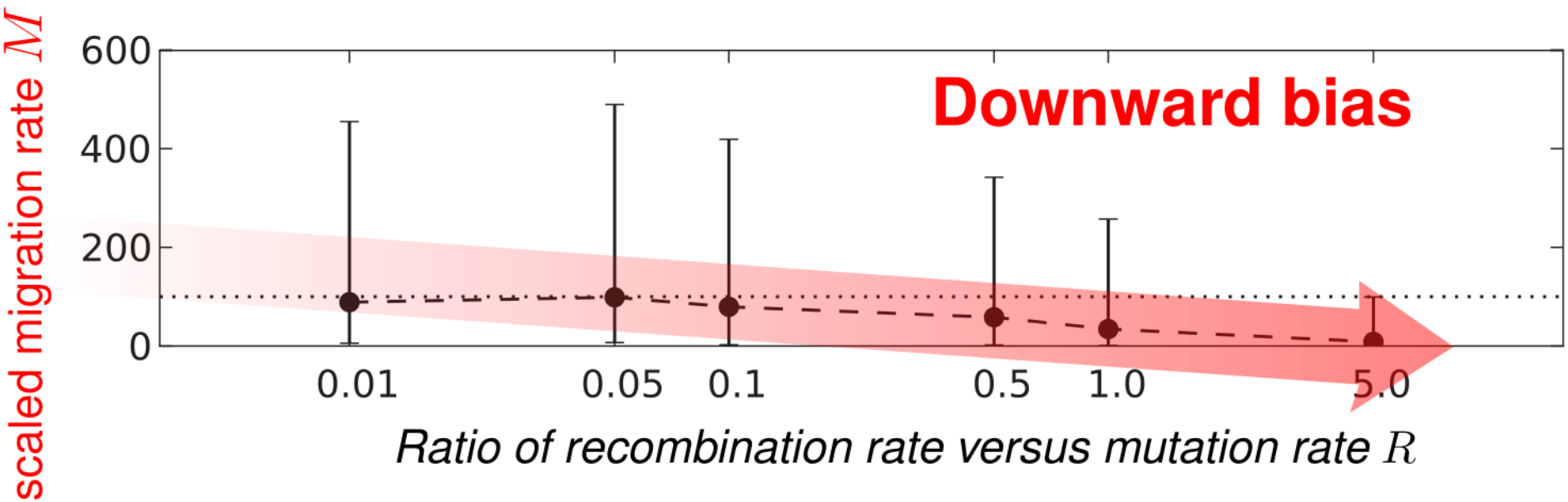
~500 simulated datasets



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

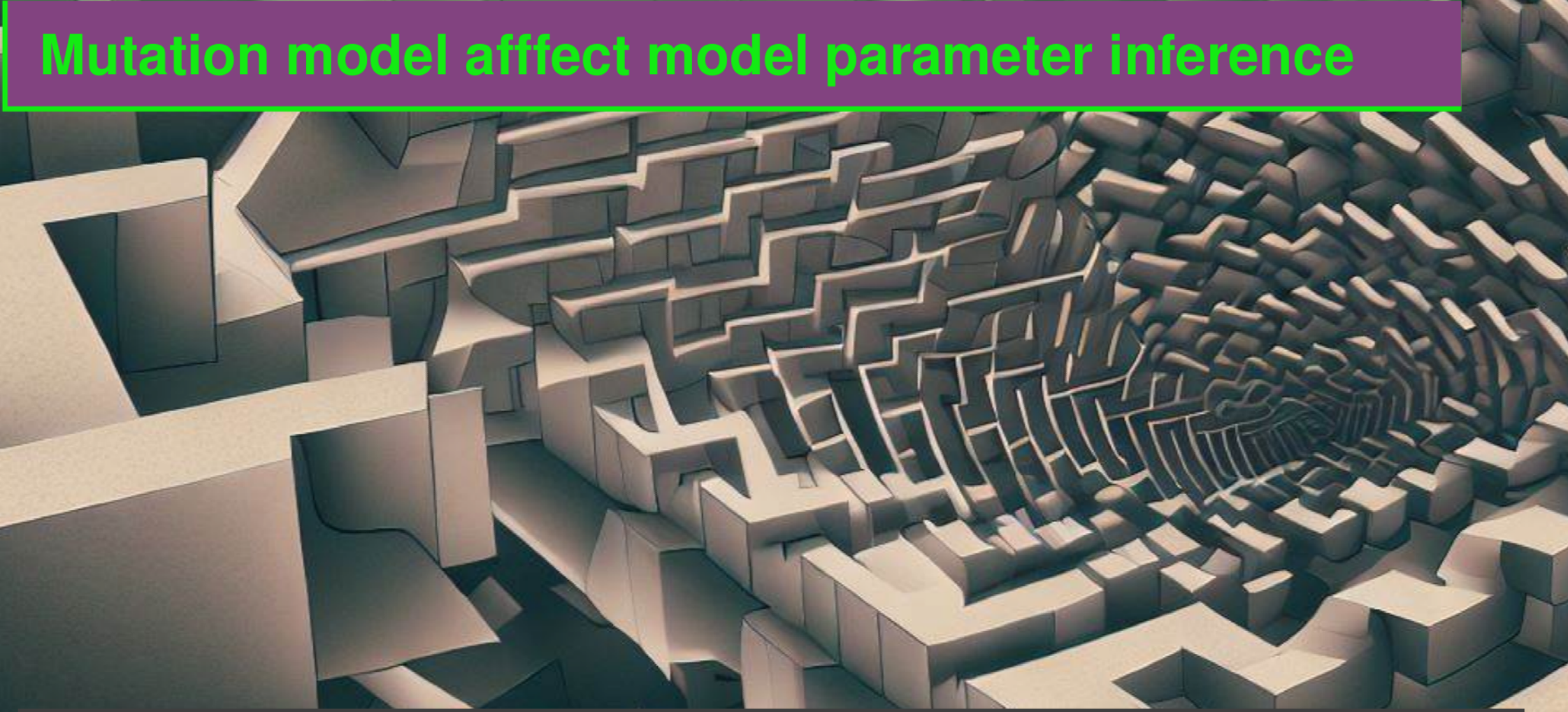
Ignoring recombination

~500 simulated datasets



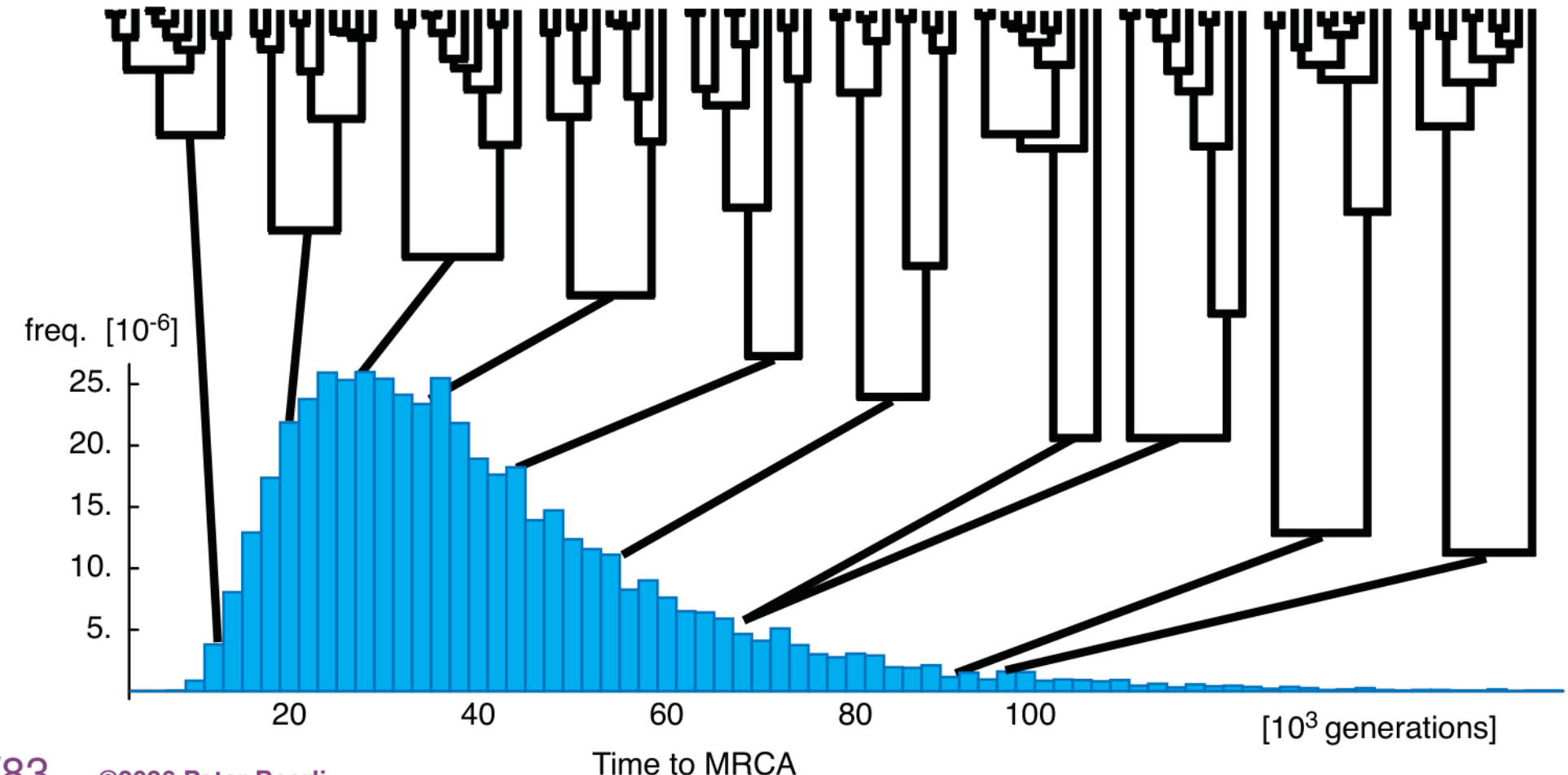
Averages with 95% credibility intervals of runs with different mutation-scaled recombination rates $R = C/\mu$. The dotted lines mark the 'true' values.

Mutation model afffect model parameter inference



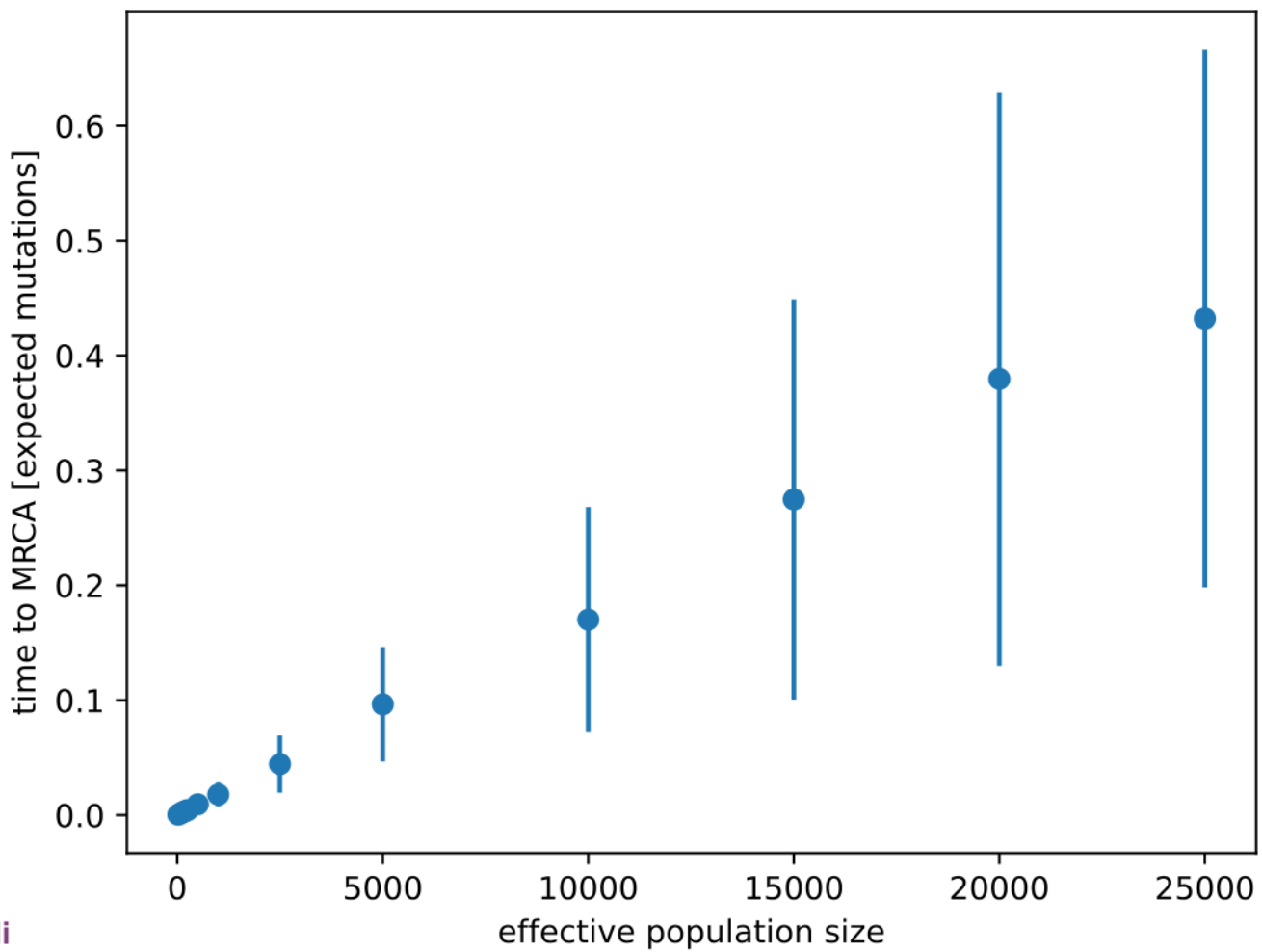
alternative titles:
“How to not manipulate your data” or “Down the rabbit hole?”

the Coalescent and Tree depth

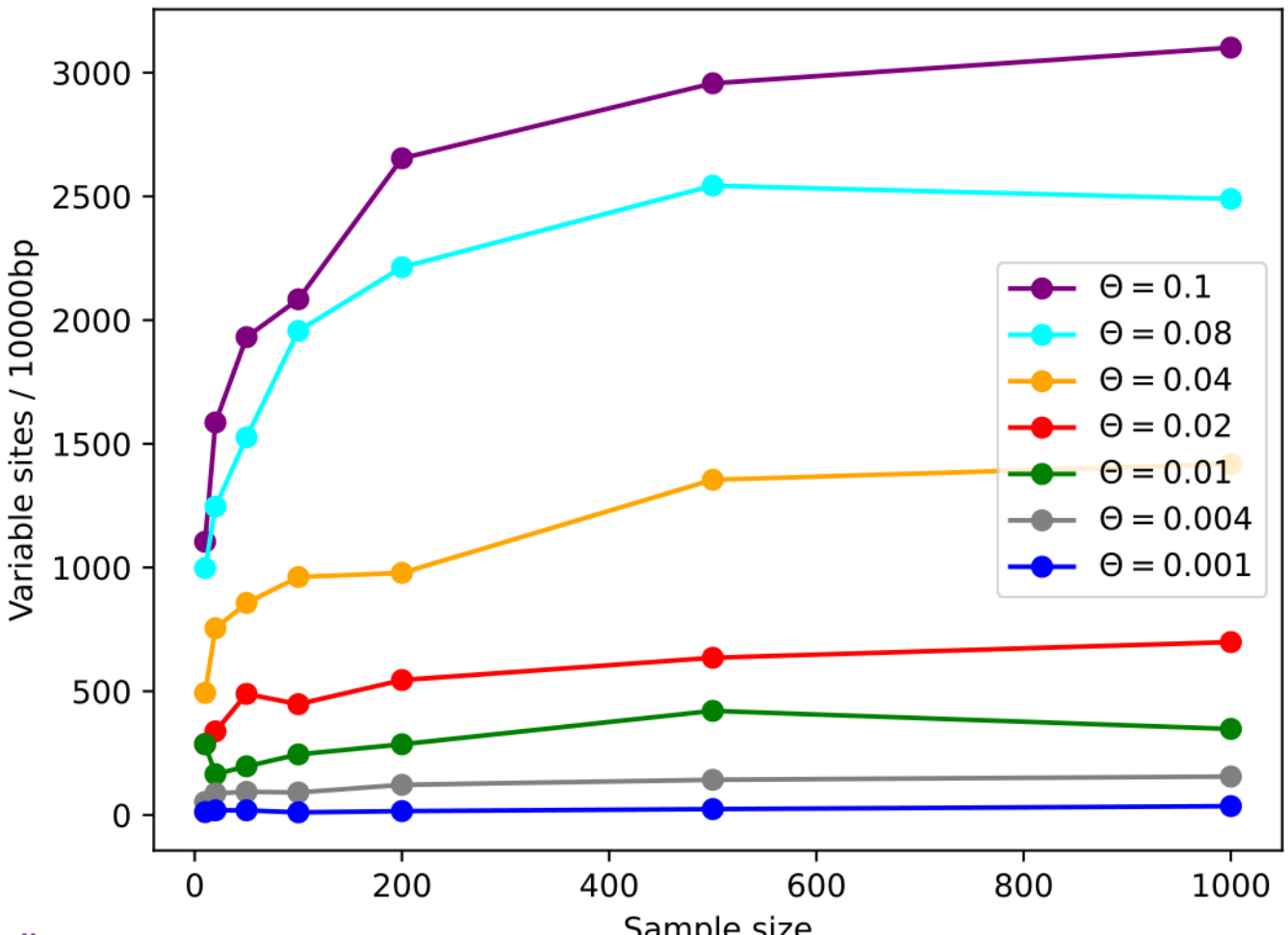


the Coalescent and Tree depth

$$\mathbb{E}(\Theta) \sim \tau_{\text{MRCA}}$$



the Coalescent and detected mutations



the Coalescent to estimate population size

$$\text{Watterson's } \Theta = \frac{S_n}{\sum_{i=i}^{n-1} \frac{1}{i}}$$

Ewens (2004) points out that the number of variable sites S_n is not a sufficient statistic for Θ and that the data may hold more information. And true! Several researchers developed MCMC-based approaches to estimate Θ, for example using Bayesian inference:

$$p(\Theta|D,\mu) = \frac{p(\Theta)\int_G p(G|\Theta)p(D|G,\mu)dG}{p(D|\mu)}$$

How to deliver the mutations to the inference method

Two main methods:

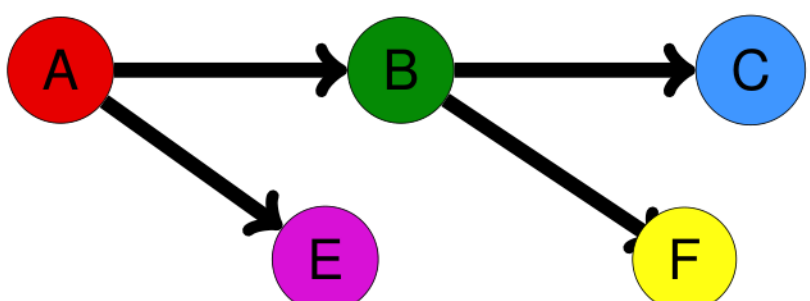
Aligned DNA sequences for many loci: We need a finite mutation model: many to pick from: Jukes-Cantor, Kimura, Felsenstein, Tamura-Nei, GTR
These models usually assume a Markov property

Single nucleotide polymorphism data: Usually these are treated as diallelic markers, to fit the infinite sites model used in the **Site Frequency spectra**

Mutation model history

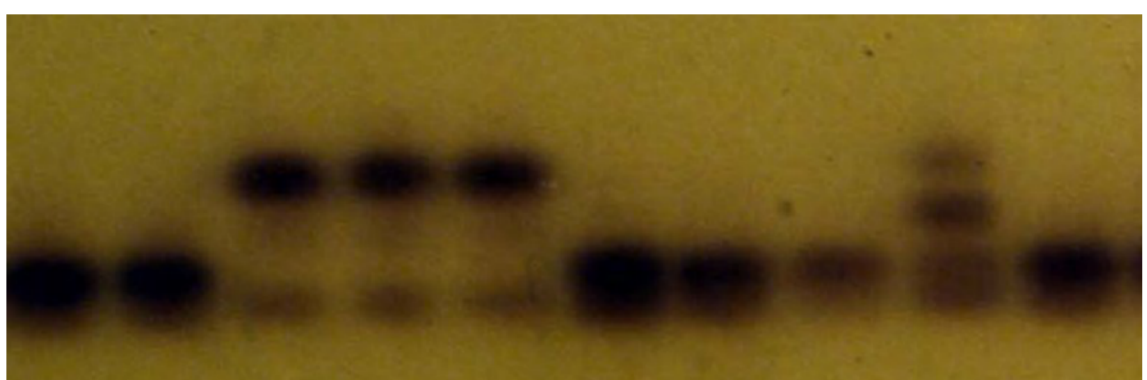
Early $\rightarrow$ Mid $\rightarrow$ Late

Early: Mutation models became a ”thing” when electrophoretic allozyme markers were used to look at variability in natural populations, starting with papers such by Lewontin and Hubby ($\sim$1964). The **infinite allele model** became famous



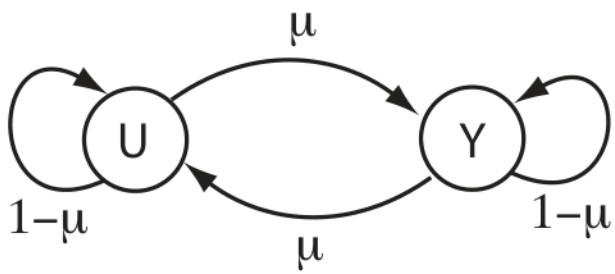
Mutation model history

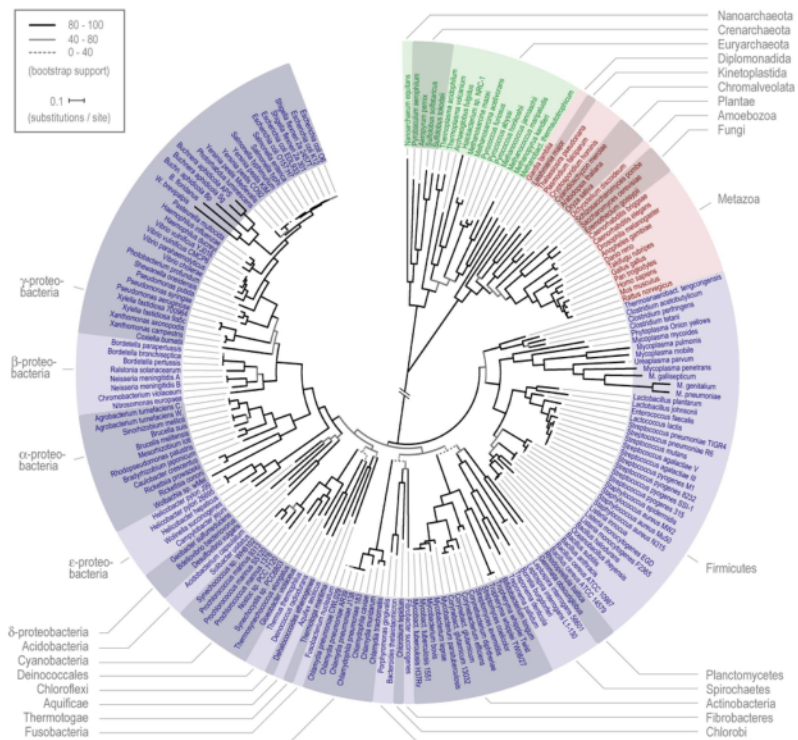
Mid (popgen): These were the haydays of enzyme electrophoresis and many population genetic studies based on allozymes were generated, using the infinite allele model (Kimura and Crow, 1964) or a variant of the ladder model (Ohta and Kimura, 1973) that then became the standard for microsatellite data.



Mutation model history

Mid (phylogeny): Researchers started. to sequence DNA, such as 5S rRNA and mtDNA, and were able to work on phylogenetic trees of species; Likelihood analyses of phylogenetic problems using **finite mutation models** became feasible. Models that explicitly model the transition between nucleotides over the course of time, many variants created a considerable alphabet soup of models: JC69, K2P, F81, F84, HKY, TN93, GTR, ...





Mutation model history

Mid - Late: Some population geneticists started to tinker with sequence data and used the **infinite sites model** (Kimura 1969). For example, Strobeck (1984) evaluated the population size of two Drosophila species assuming an infinite sites model.

Let $N \gg 1$ be the number of diploid individuals in the population each generation (thus there are $2N$ copies of a gene in the population). The $2N$ genes in one generation are drawn randomly with replacement from the $2N$ genes in the previous generation. A gene is assumed to consist of an infinite number of sites at which mutation can occur. Since the rate of mutation at each site is small, the probability of two mutations occurring at the same site is zero. Let $\mu \simeq O(1/N)$ be the mutation rate of neutral alleles per gene per generation. It is also assumed that there is no recombination between the sites.

Drosophila virilis: n=10, a_1=4, a_2=6

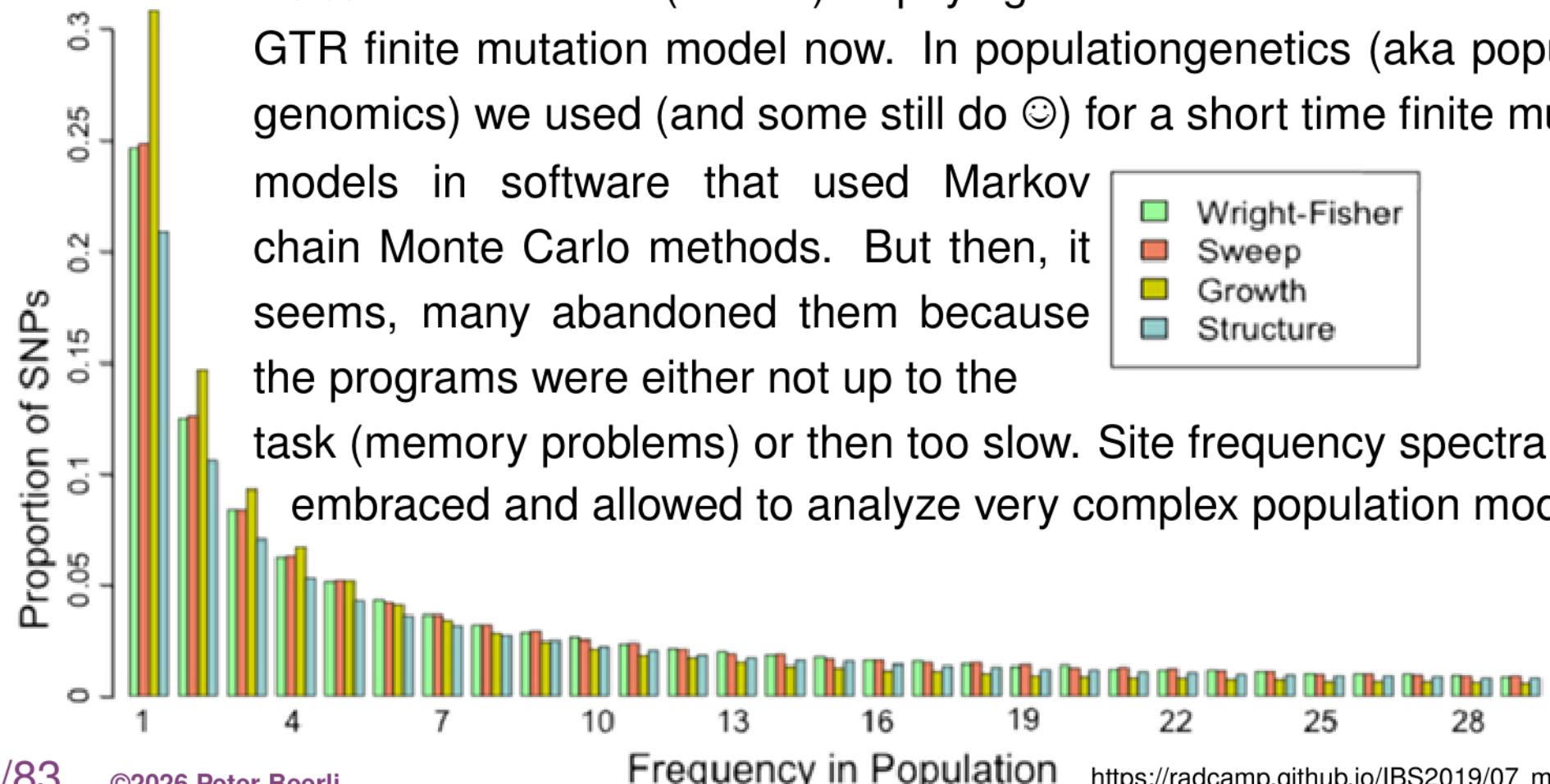
Infinite allele: $\theta_{\text{Ewens}} = 1.97$

Variable site: $\theta_{\text{Watterson}} = 0.35$

Infinite site: $\theta_{\text{Strobeck}} = 0.34$

Mutation model history: “Today’s state of the Art”

Late: It seems that (almost) all phylogeneticists use the time reversible GTR finite mutation model now. In populationgenetics (aka population genomics) we used (and some still do ☺) for a short time finite mutation models in software that used Markov chain Monte Carlo methods. But then, it seems, many abandoned them because the programs were either not up to the task (memory problems) or then too slow. Site frequency spectra were embraced and allowed to analyze very complex population models.



https://radcamp.github.io/IBS2019/07_momi2_API.html

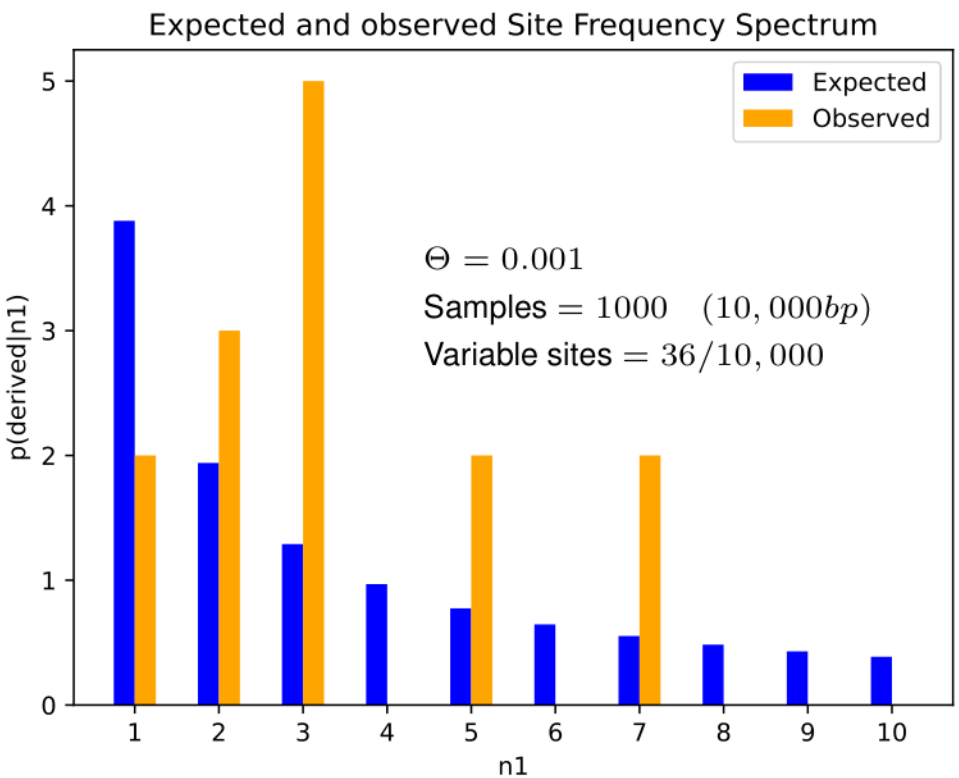
How to ‘accommodate’ data to the infinite sites model

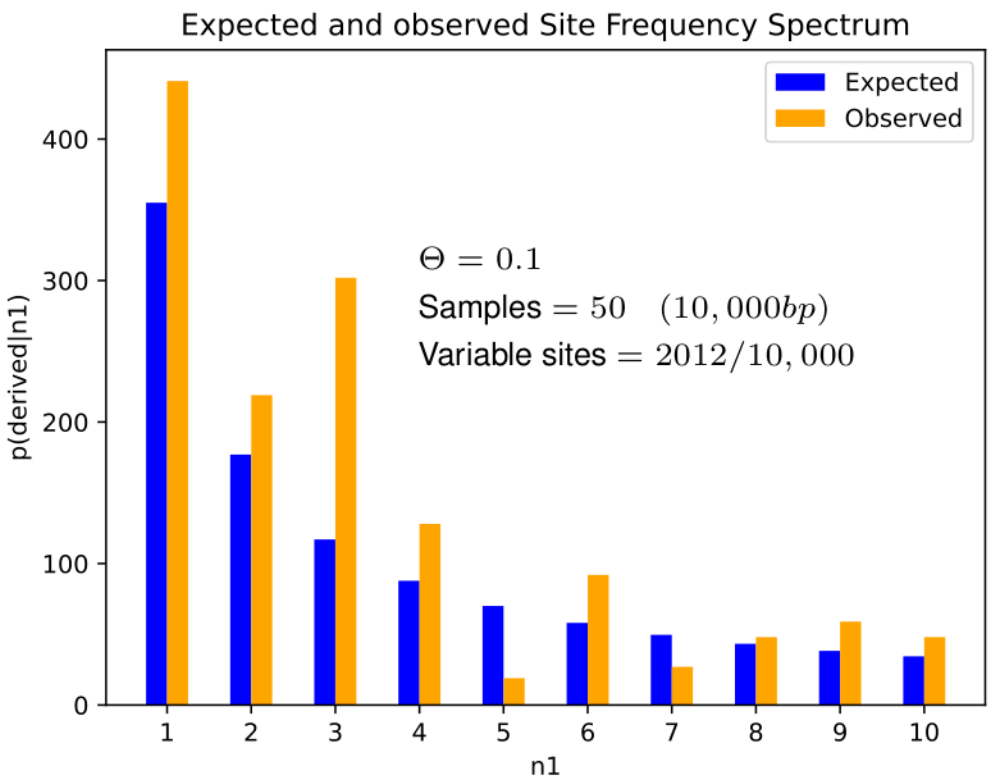
- A major and minor allele has to be picked (usually, we use an outgroup to define the ancestral allele; or pick the major allele as the ancestral and call it the ‘folded’ SFS)
  With low variability and no sequencing error, this works great

- With high variability, one may discard the tri-allelic states or pick the two most common as the bi-allelic marker

- If we assume there are sequencing errors, then many will not recognize derived alleles that occur only once as a real allele

Site frequency spectra based on SNPs





Expected allele frequencies: $\frac{n!}{i(n-1)!)} \frac{|S1_{k-1}^{n-i}|}{|S1_k^n|}$

SNP ascertainment issues

SNPs and population parameters: Single nucleotide polymorphisms are usually reported as an ancestral allele and the alternative allele (2-state). This works fine under the assumption that populations are small and a mutation rate is small.

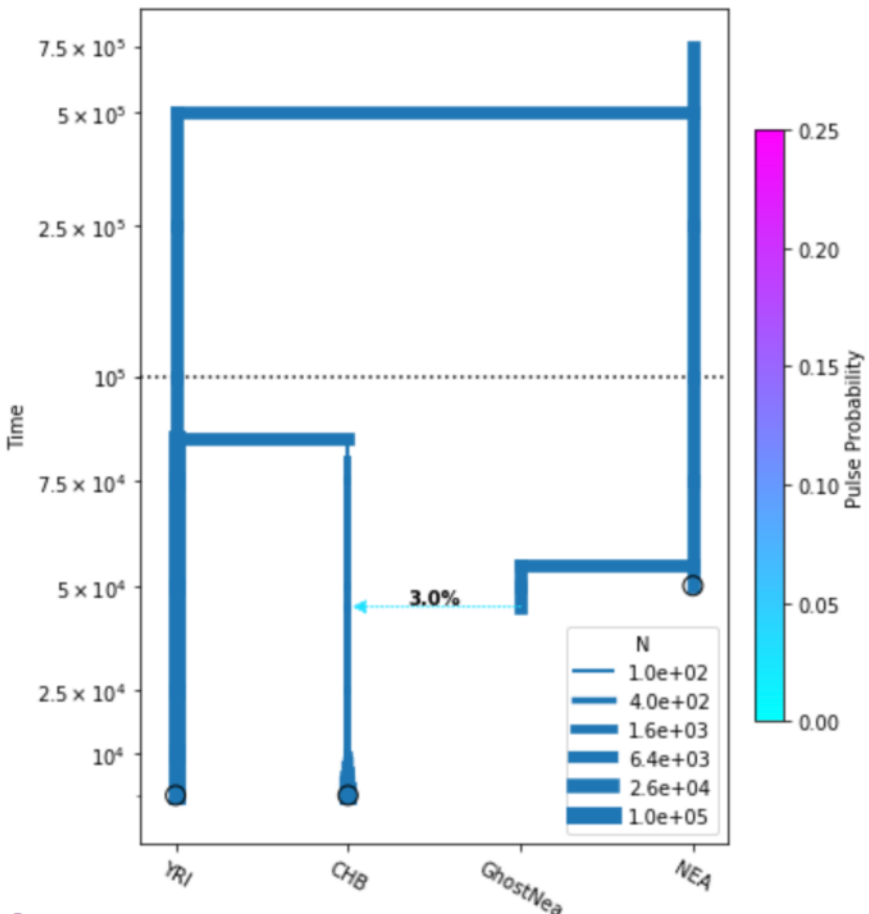
Modern parallel sequencing allows us to retrieve SNPs without bias; with enough coverage, we can find them all, and if we do not remove lower frequency alleles, then we may get good estimates of the site frequency spectrum.

Removal of low-frequency SNPs without correction will lead to errors.

Analyses in population genomics

Population model parameter estimation: The coalescent with many samples becomes rather intractable when we assume complicating forces such as gene flow, recombination, population size changes, admixture, population splitting. This was one of the reasons for the development of methods that depend on the SFS, but one may wonder whether we have traded one problem with another untractable one.

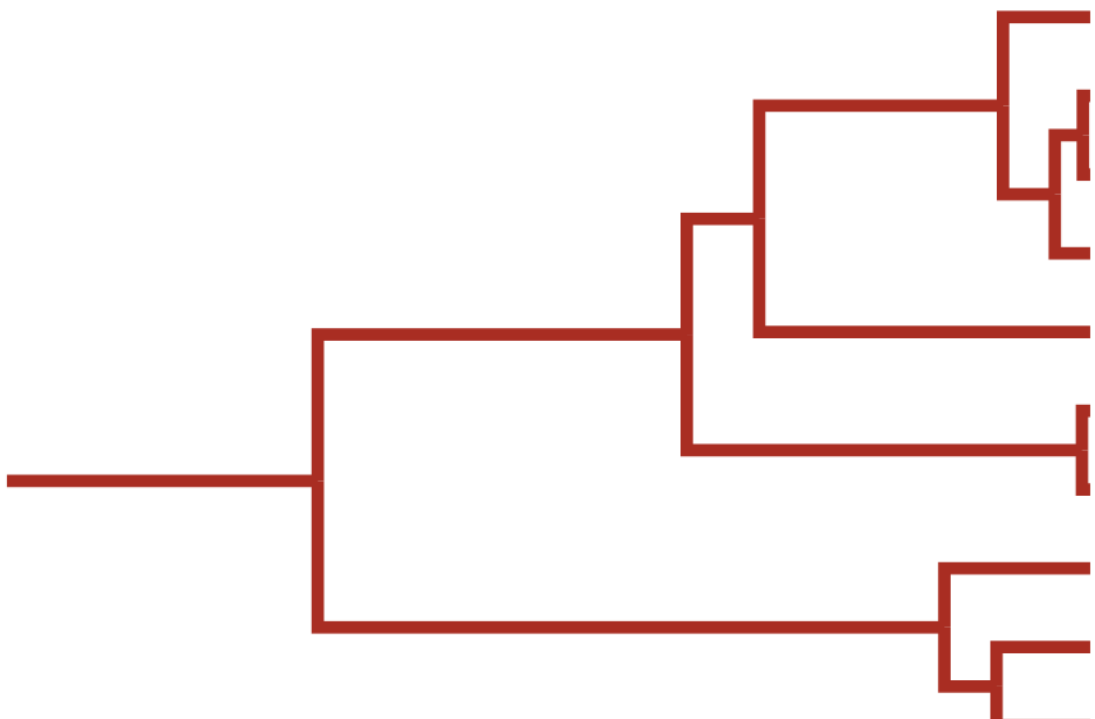
Example of a site frequency spectrum method



Momi2: momi (MOran Models for Inference) is a Python package that computes the expected sample frequency spectrum (SFS) and uses it to fit demographic history. In short: we take SNP dataset with n population, create a multipopulation SFS, create a specific population model and find parameter setting of this particular model so that we can generate an expected SFS that is close the estimated SFS. We assume an infinite sites mutation model.

Analyses in population genomics

A sample of a single population:

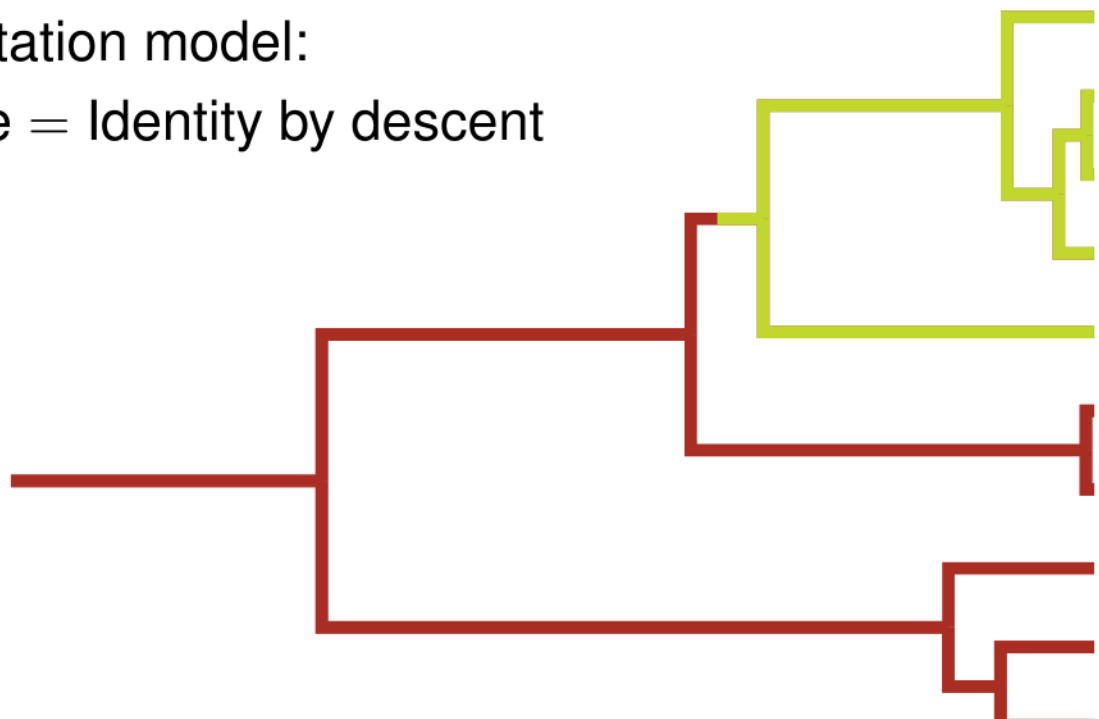


Analyses in population genomics

A sample of a single population:

Infinite mutation model:

allelic state = Identity by descent

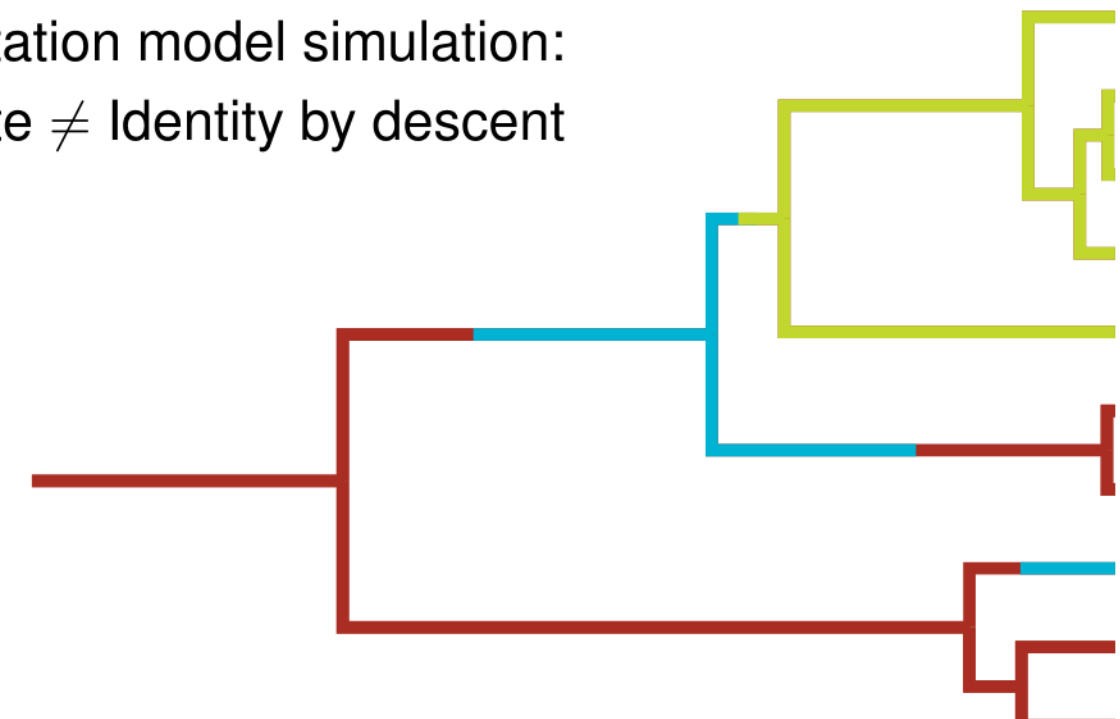


Analyses in population genomics

A sample of a single population:

Finite mutation model simulation:

allelic state $\neq$ Identity by descent



Analyses in population genomics

The last example shows three different alleles: Even if we would only see 2 alleles, then we may guess that the mutation rate with finite mutation models may be higher than with an infinite mutation model.

Variability is the measure for almost everything in population genetics! Low variability suggests low population size; with little variability, we also assume that two populations are more similar, leading to estimate high gene flow, or recent divergence times, ...

Scenarios that we should explore, but I do not see lots of work on that.

Infinite sites (aka 2-allele SNPs) vs finite sites models

An example

A recent report on genomic sequences:

The Anopheles gambiae 1000 Genomes Consortium: Genetic diversity of the African malaria vector *Anopheles gambiae*. Nature, 552(7683):96–100, Dec 2017.

They identified 52,525,957 high-quality SNPs, of which 21% had three or more alleles, an average of one variant allele every 2.2 bases of the accessible genome

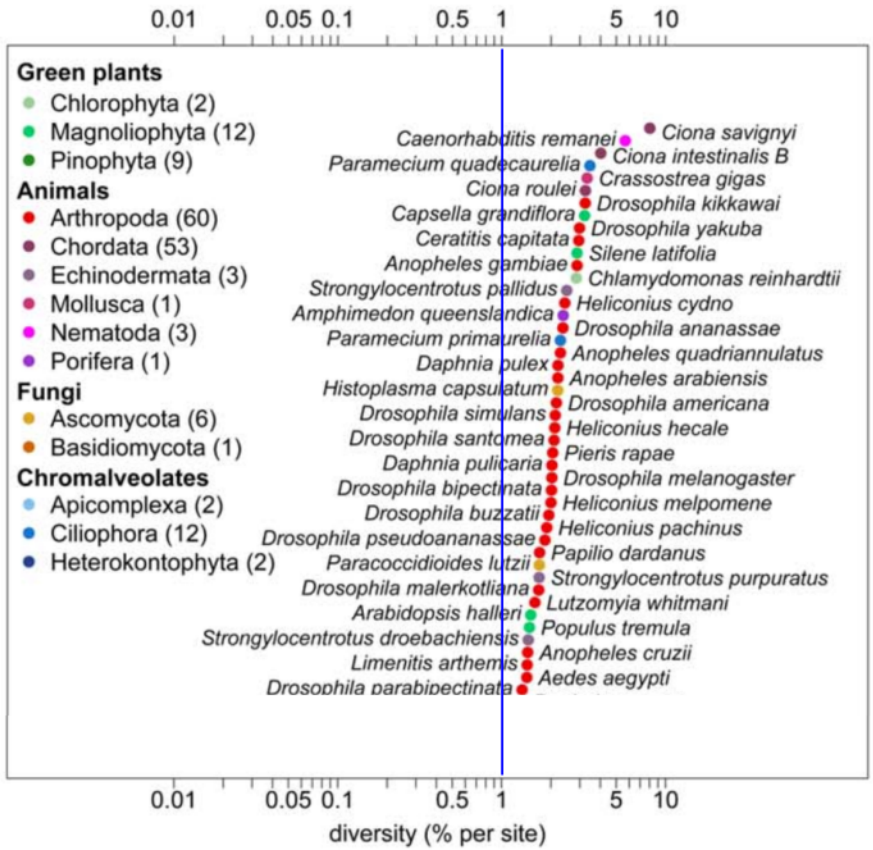
These types of data are usually
NOT analyzed with finite mutation models

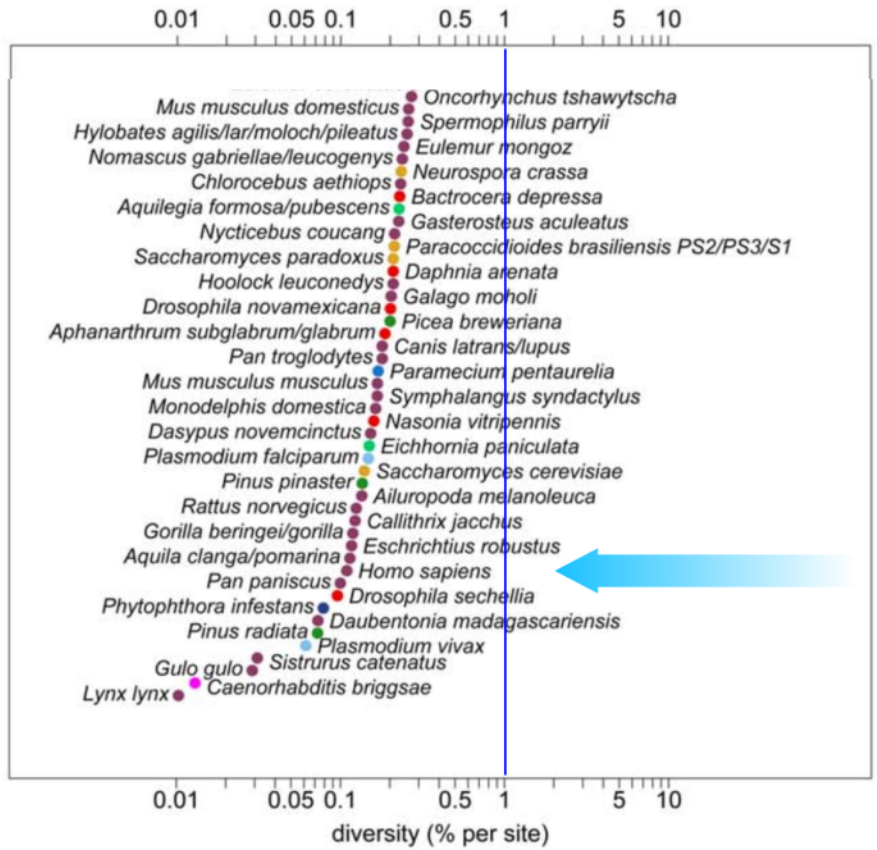
Diversity measures and the coalescent

Table 1: Frequency and number of variable sites are used in the analysis (this example is only for the simulations of two populations with a total of 20 individuals with a divergence time of 1.0 coalescent units and different population sizes from overall $\Theta{=}0.002$ to $\Theta{=}0.2$ ($\Theta{=}4N_e\mu$, mutation rate μ is per site). The table shows the total number of variable sites over 100,000 bp.

Population Size	Freq. Variable Sites	SNPs	Tri-Allelic	Tetra-Allelic
0.002	0.00888	888	0	0
0.005	0.02124	2124	4	0
$\star$ 0.010	0.03915	3915	36	0
0.020	0.07900	7900	124	1
0.050	0.19273	19273	847	15
$\star$ 0.100	0.36277	36277	3477	116
0.150	0.46862	46862	5872	291
0.200	0.60330	60330	10596	797

Genetic Diversity within species





Leffler et al. (2012) Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species?

We have so much data, does it matter?

- If your species or species group has very small population sizes or had small population sizes for a long time in the past, it may be just fine to discard the rare tri-allelic sites.

- If you do not know the effective population sizes or your species have large differences in their population size, I suggest not trusting the site frequency-based methods and comparing their results with alternatives.

Extension to the SFS

- Wakeley et al. 2023: Modeling rare alleles using recurrent mutations from a common allele
- Jenkins et al. 2014: Allowing for 0, 1, or 2 mutations that are distinguishable (non-nested)
- Hobolth and Wiuf 2009: extension to nested mutations.

Their work is progress, but need to be expanded beyond single populations. We may need an SFS allowing for calculations based on identity by state instead of numbered mutations. Or improve speed on finite-mutation models.

Outlook

We need to develop methods that work for all levels of diversity, I am worried, in particular, about results using any species with large population sizes (pathogens?)

Supported by NSF DBI2019989
and Florida State University

Outlook

- We will have a lab on Friday where you will learn about Bayesian model selection with MIGRATE using 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.