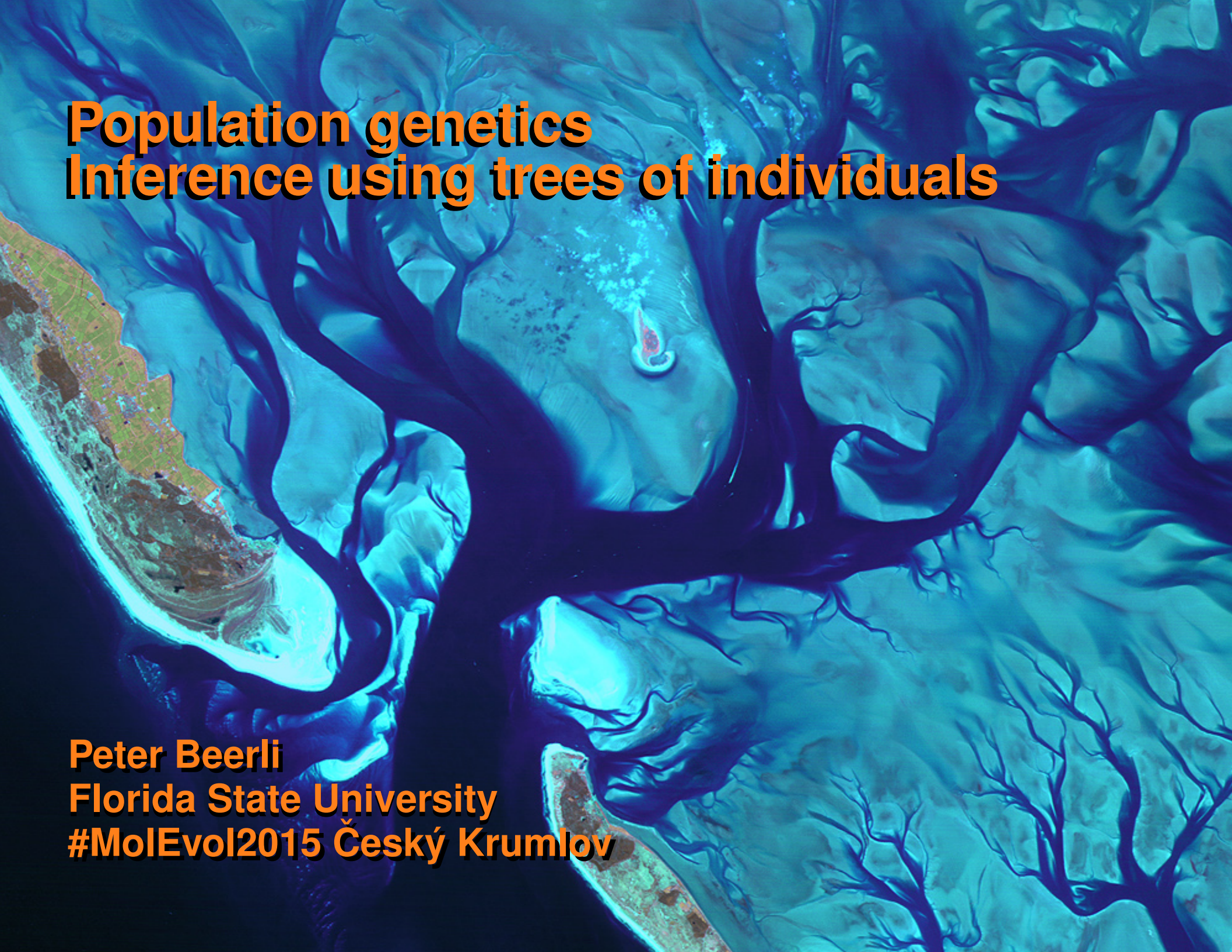
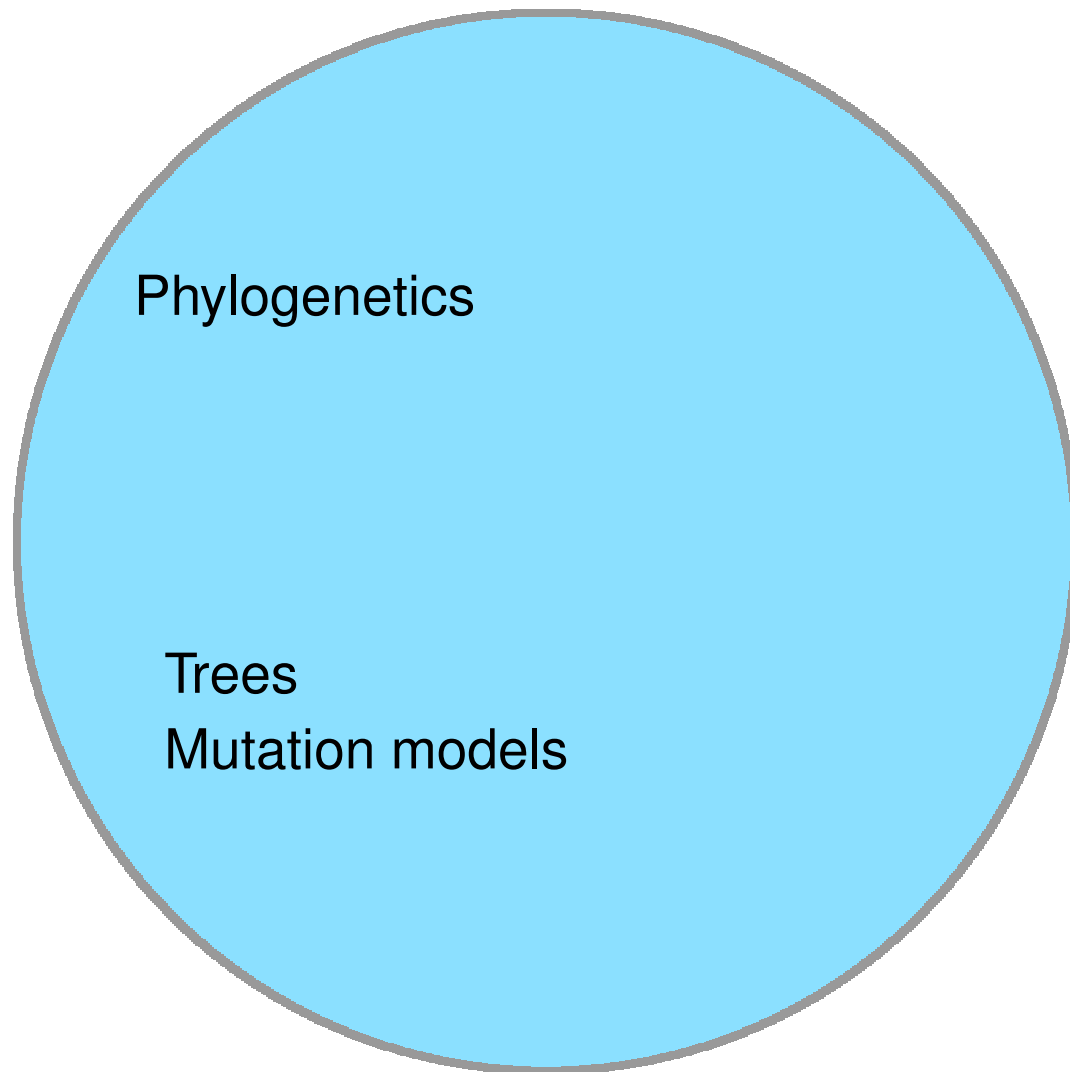




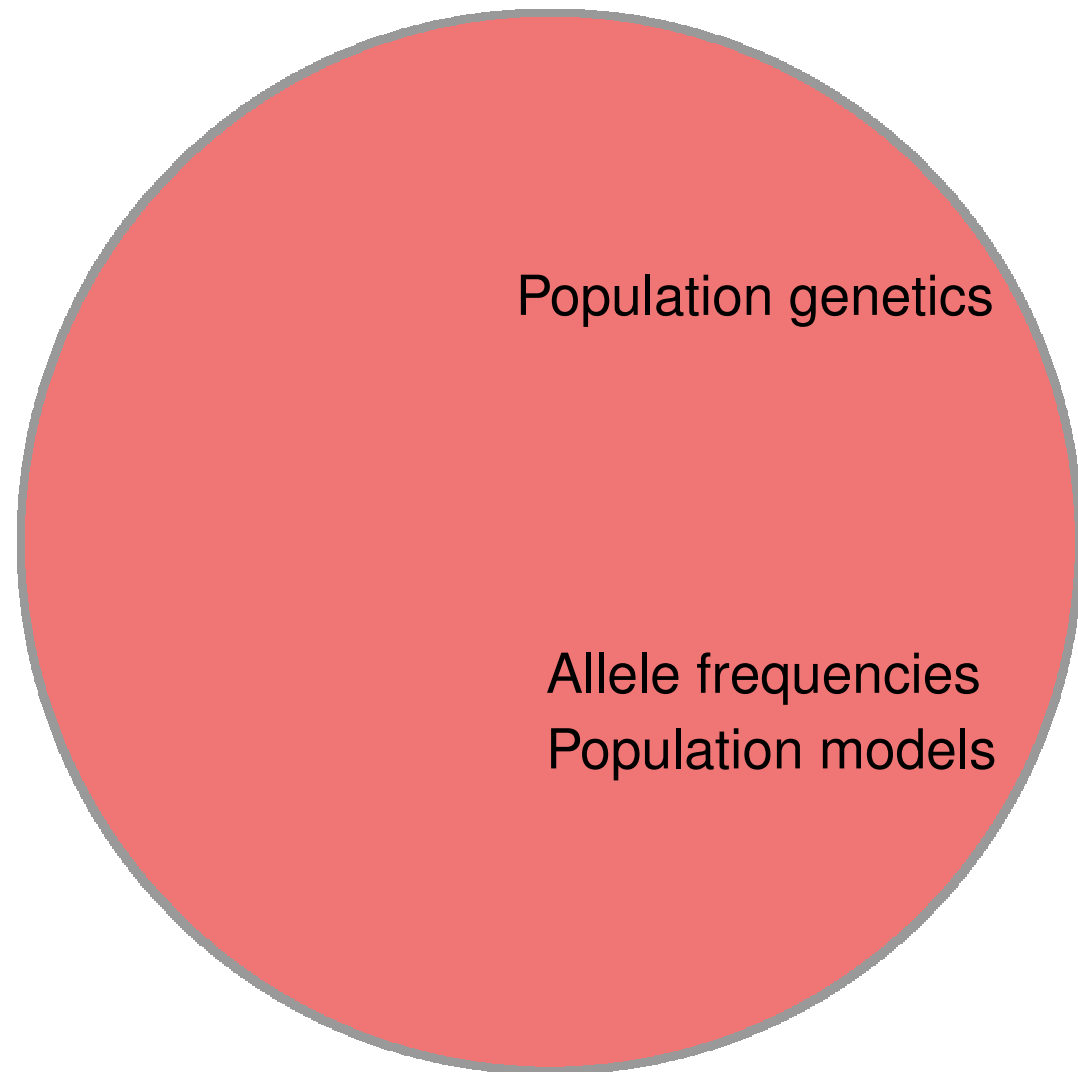
Population genetics Inference using trees of individuals

Peter Beerli
Florida State University
#MolEvol2015 Český Krumlov

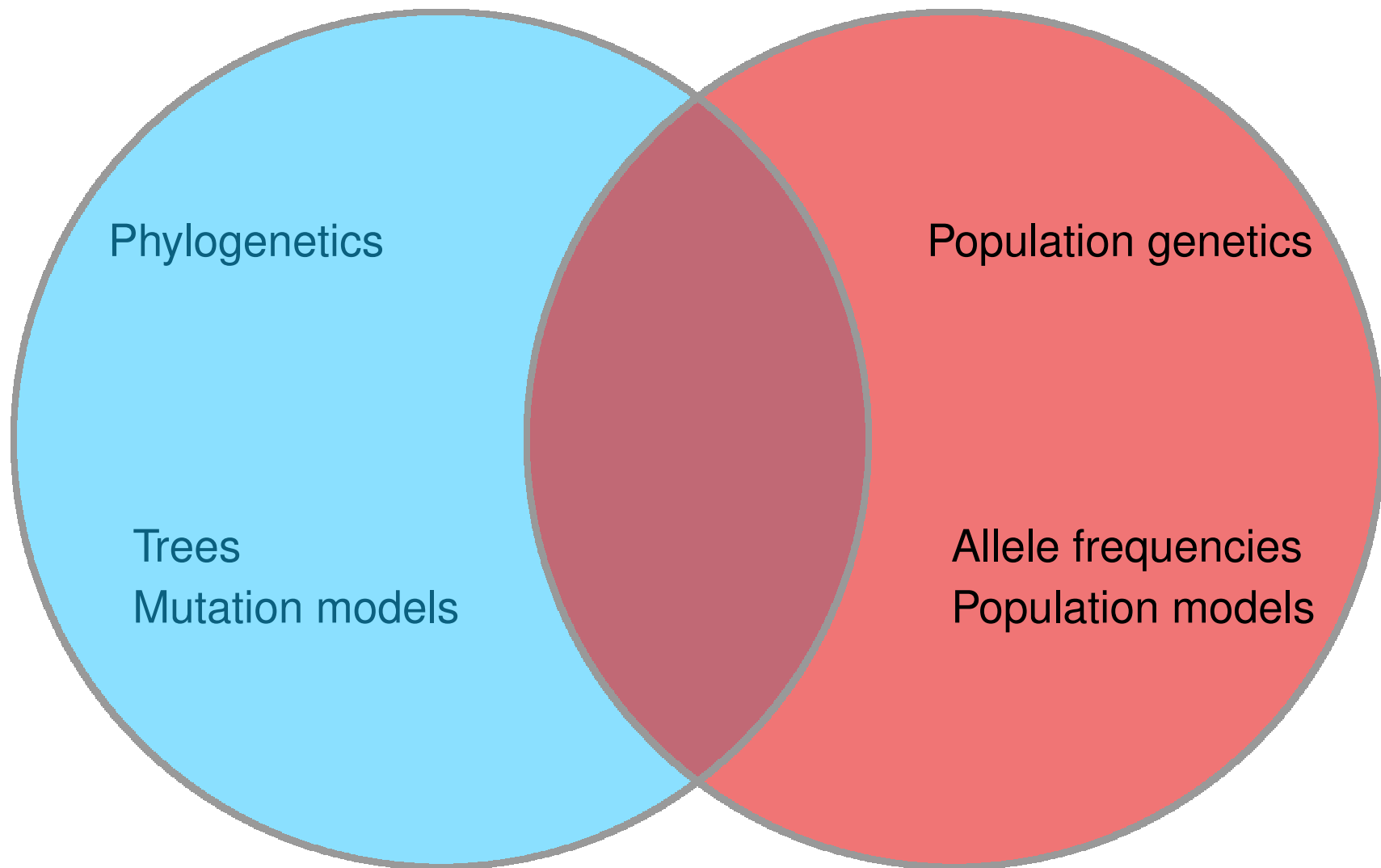
The big overview



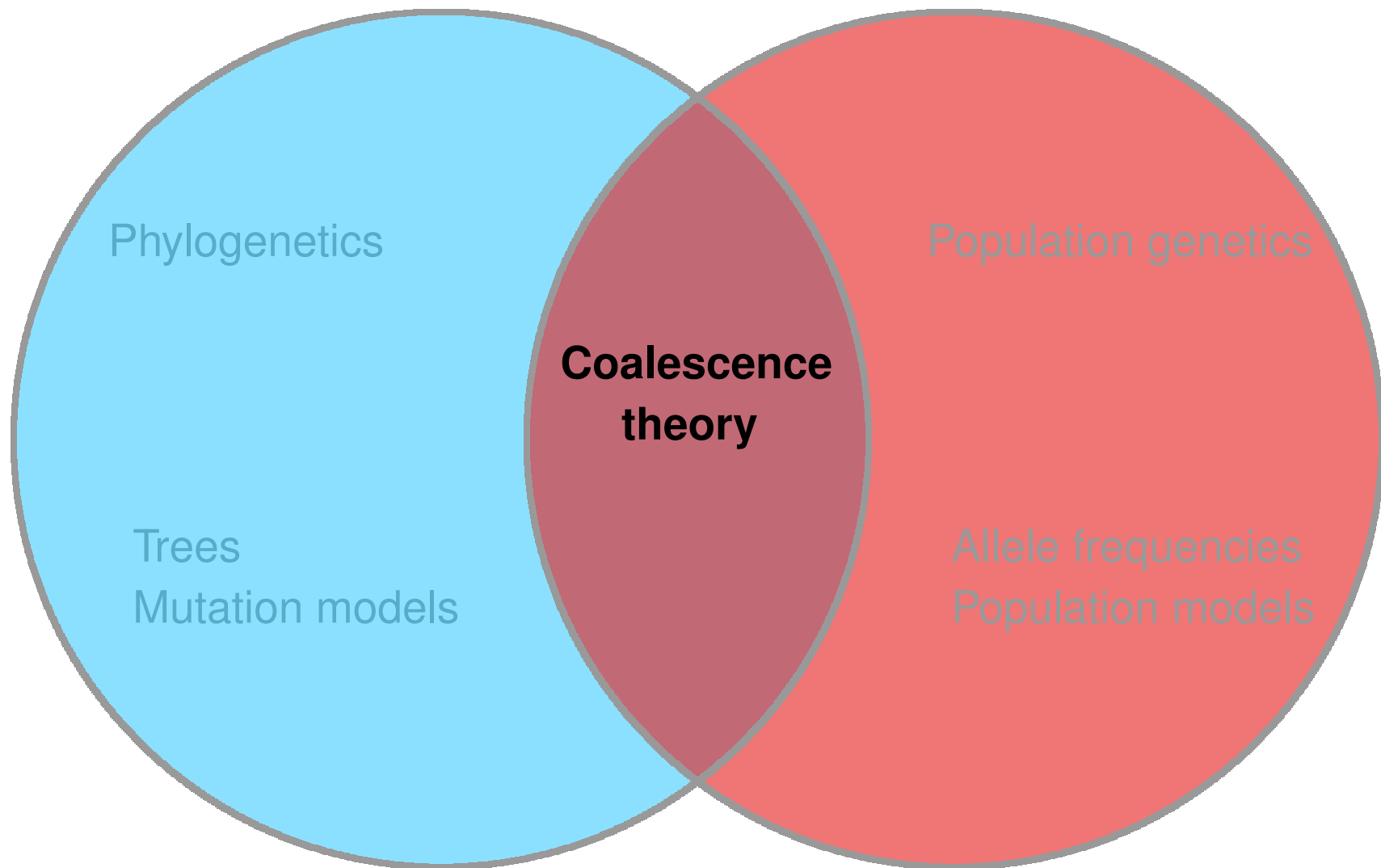
The big overview



The big overview



The big overview



Coalescence theory as a tool for population genetics

co•a•lesce |ˌkōəˈles|

verb [intrans.]

come together and form one mass or whole : *the puddles had **coalesced into** shallow streams | the separate details coalesce to form a single body of scientific thought.*

- [trans.] combine (elements) in a mass or whole : *to help coalesce the community, they established an office.*

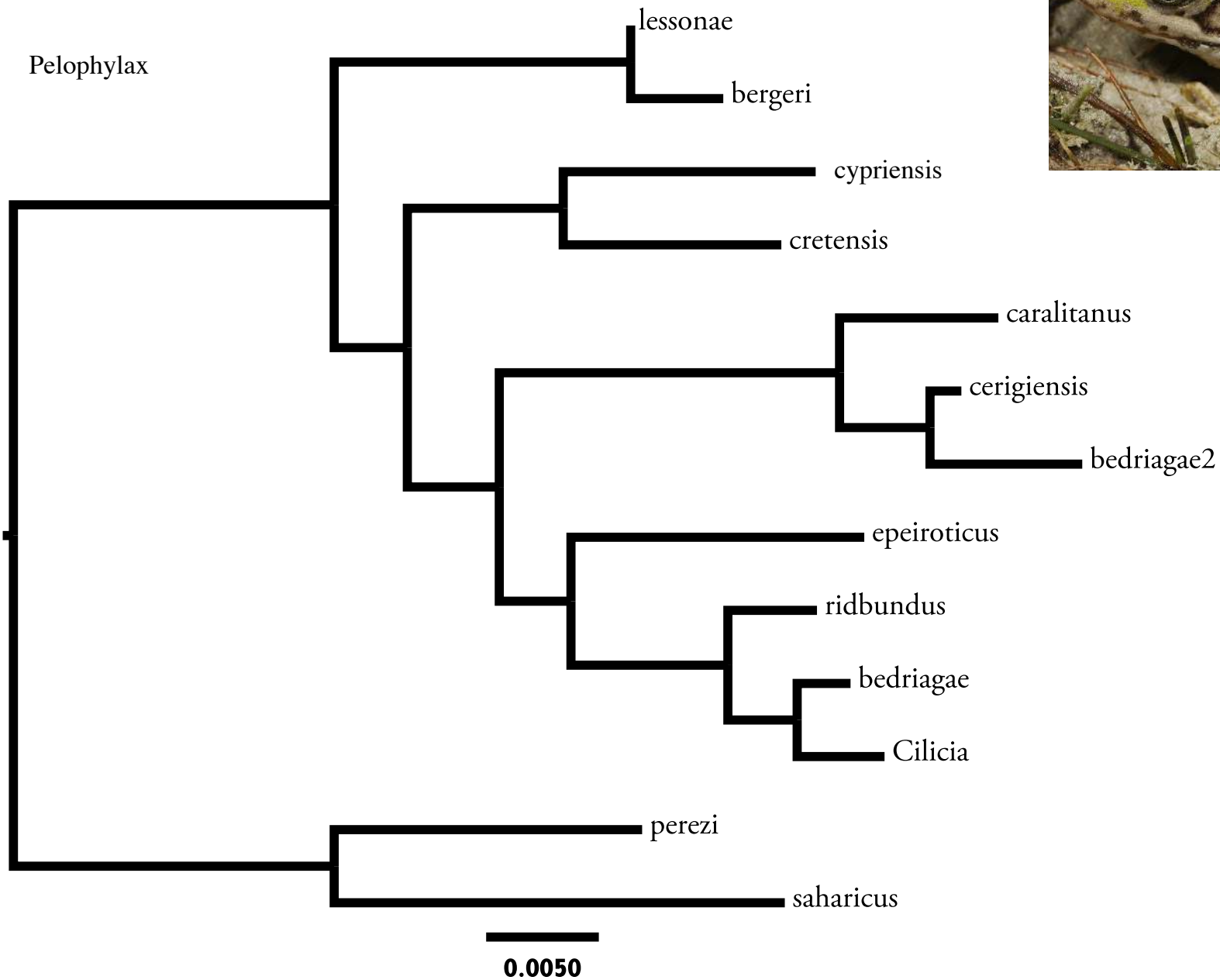
DERIVATIVES

co•a•les•cence |-ˈlesəns| noun

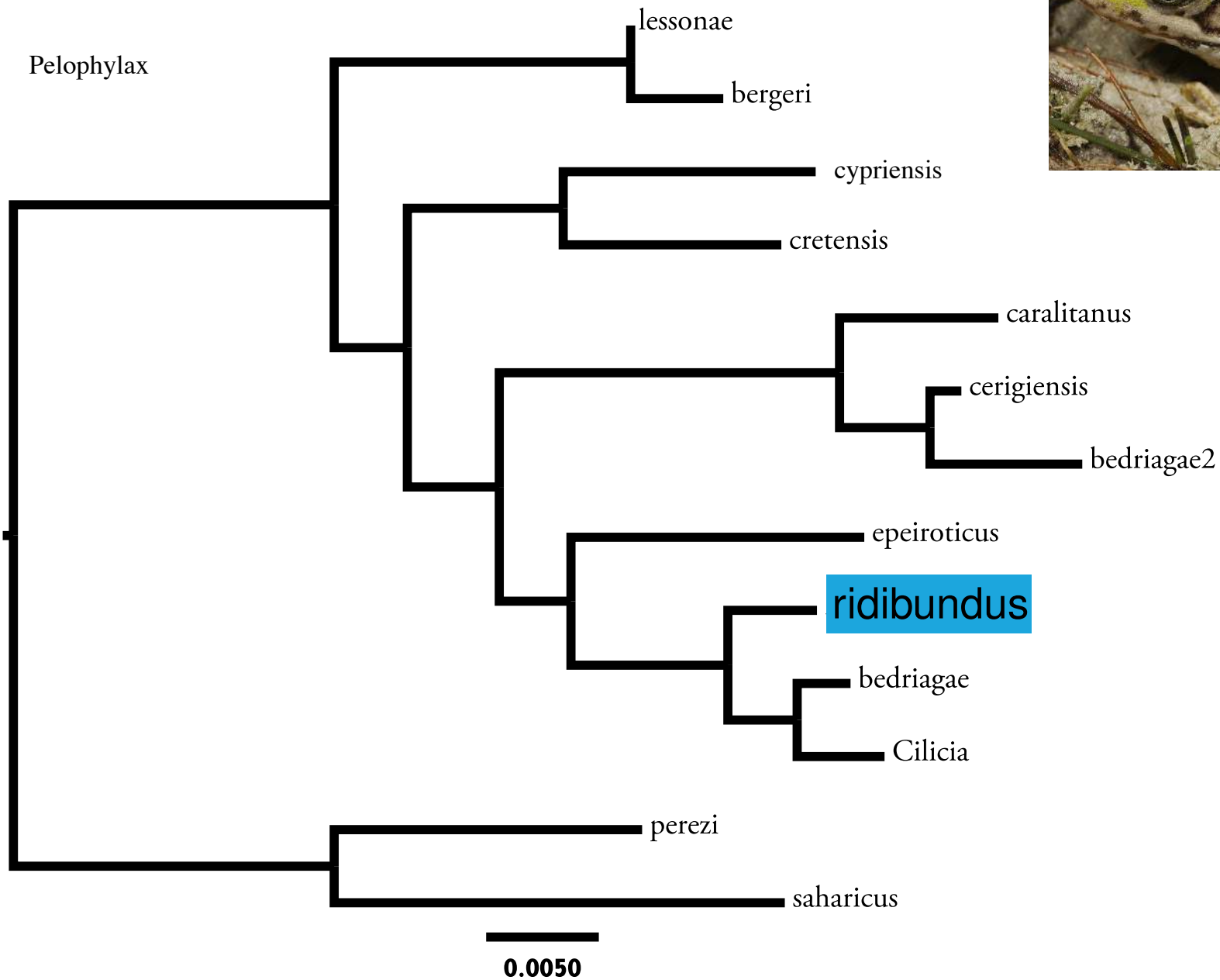
co•a•les•cent |-ˈlesənt| adjective

ORIGIN mid 16th cent. (in the sense [bring together, unite]): from Latin *coalescere*, from *co-* (from *cum* ‘with’) + *alescere* ‘grow up’ (from *alere* ‘nourish’).

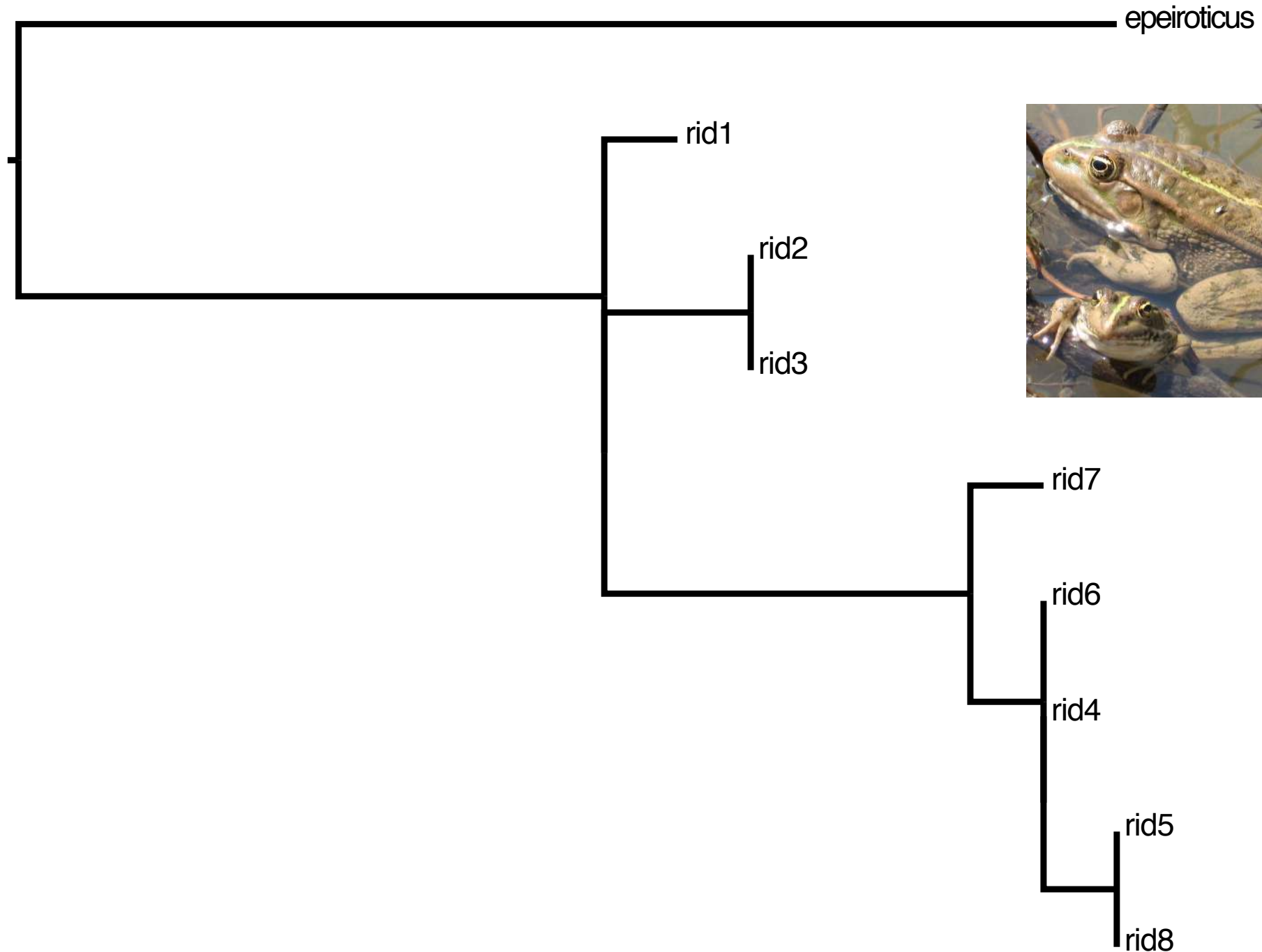
Species trees



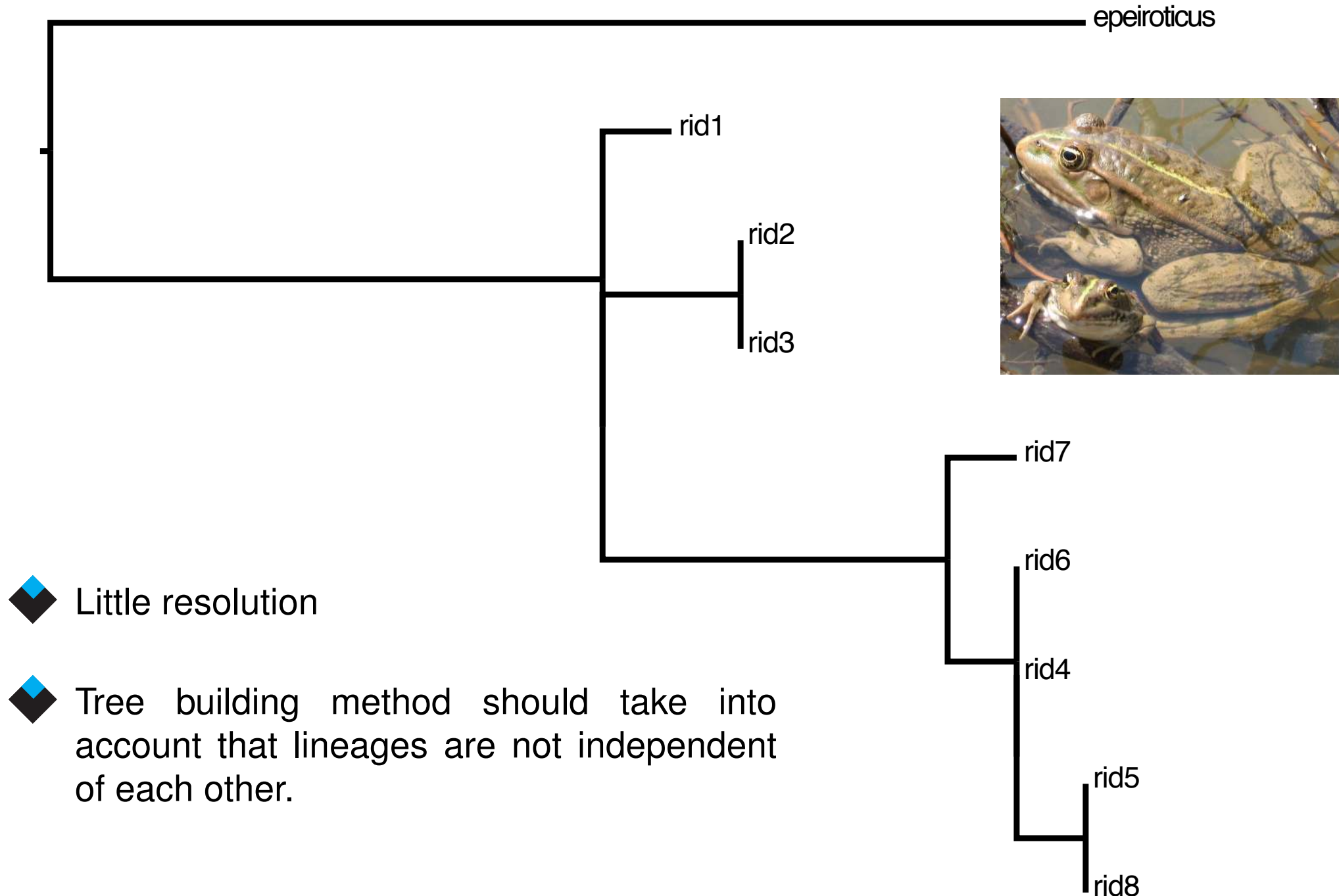
Species trees

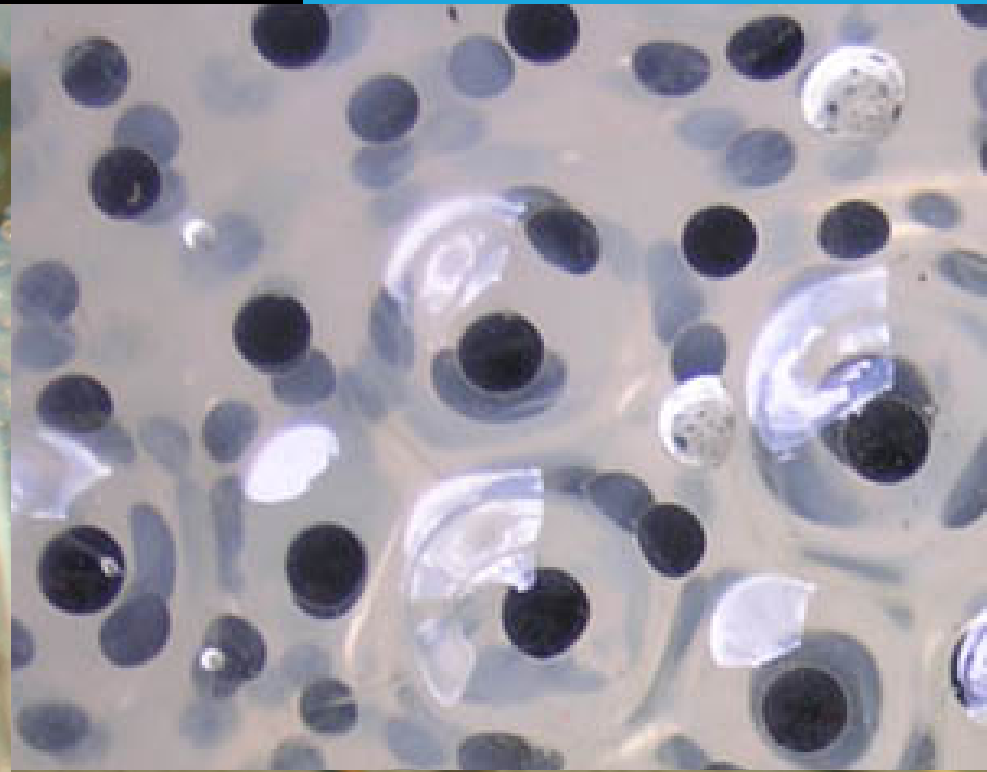


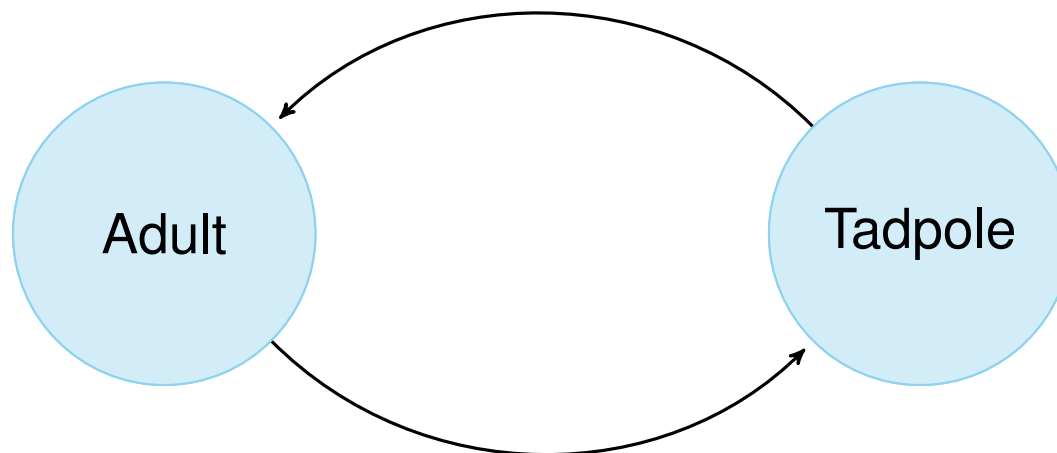
Tree of individuals of same species



Tree of individuals of same species







Wright-Fisher population model

- ◆ All individuals live one generation and get replaced by their offspring
- ◆ All have same chance to reproduce, all are equally fit
- ◆ The number of individuals in the population is constant



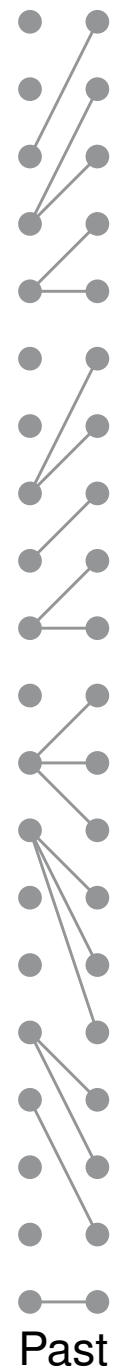
Past

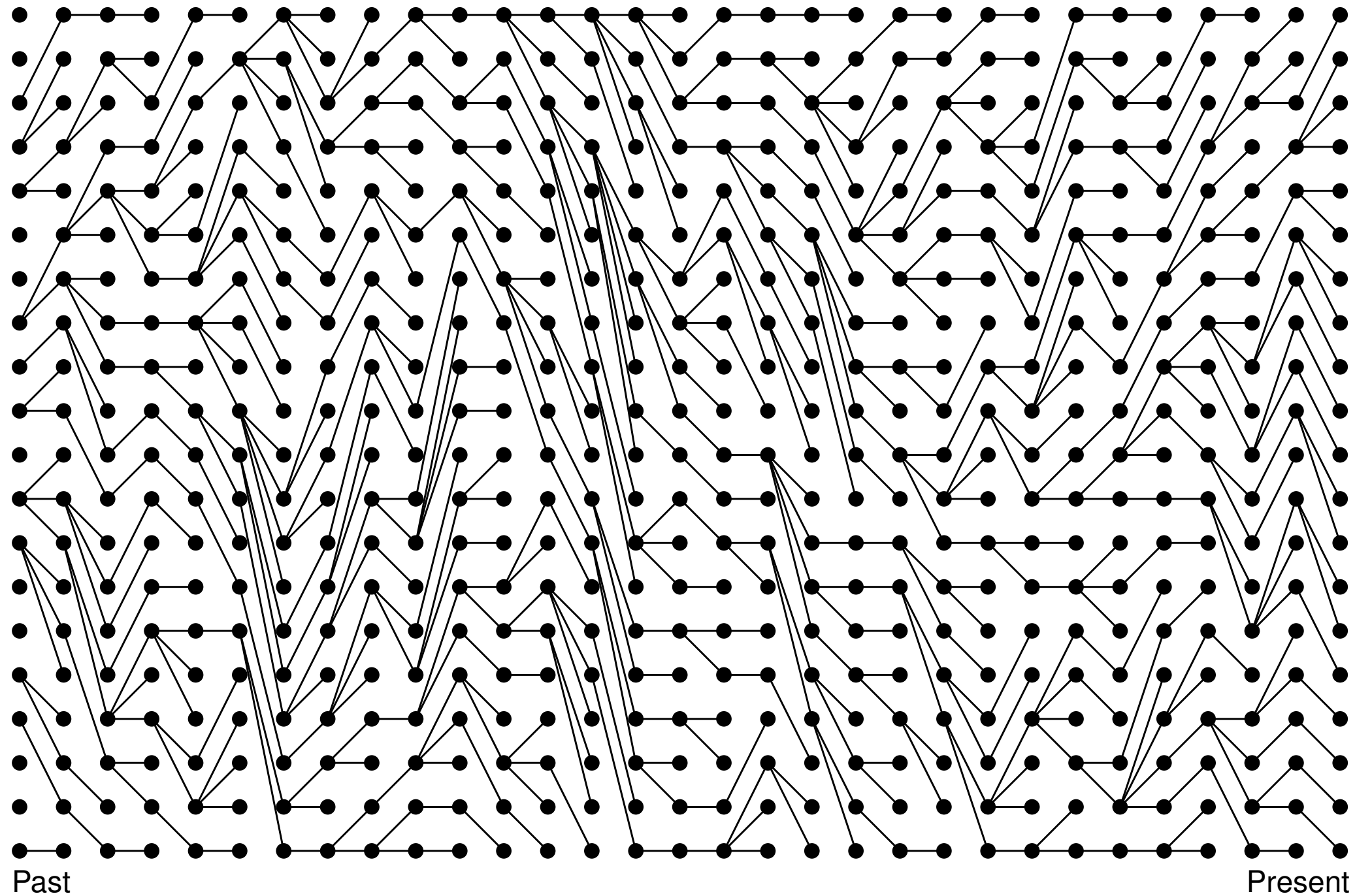
Present



Past

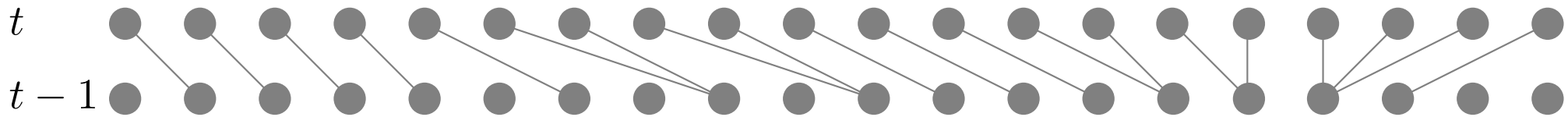
Present





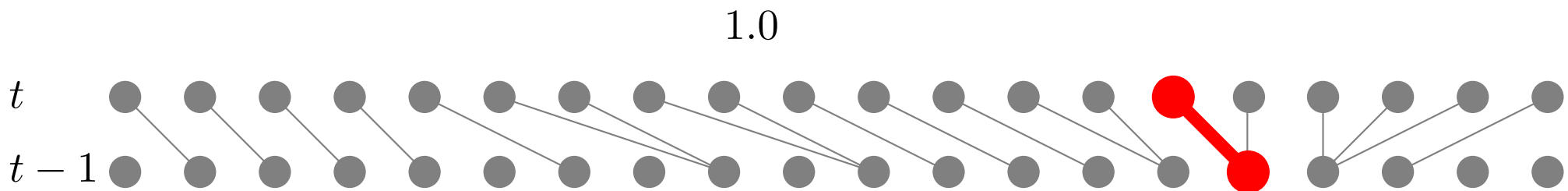


Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t - 1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is



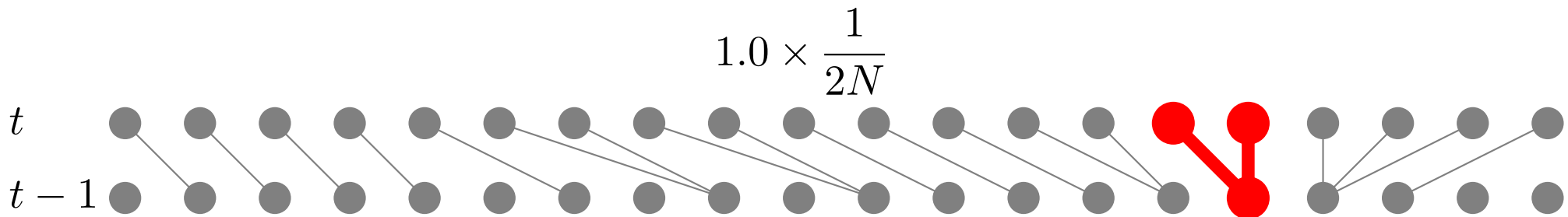


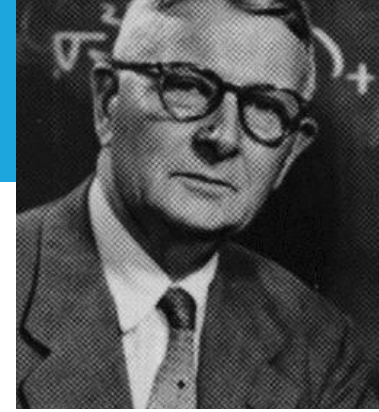
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t - 1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is





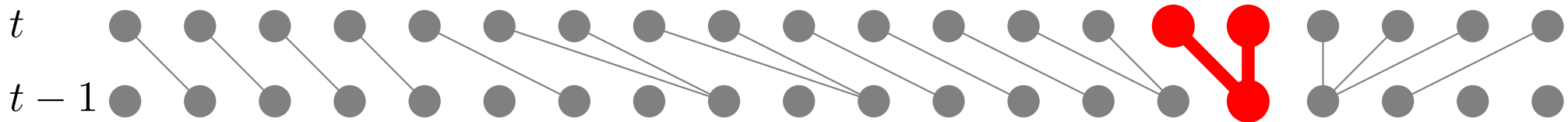
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t - 1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is





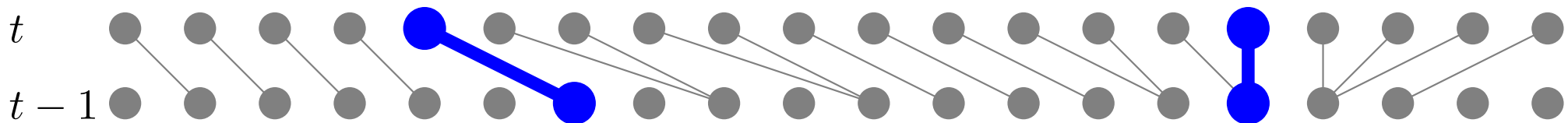
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t - 1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in last generation is

$$\frac{1}{2N}$$



The probability that two randomly picked chromosome do not have a common ancestor is

$$1 - \frac{1}{2N}$$



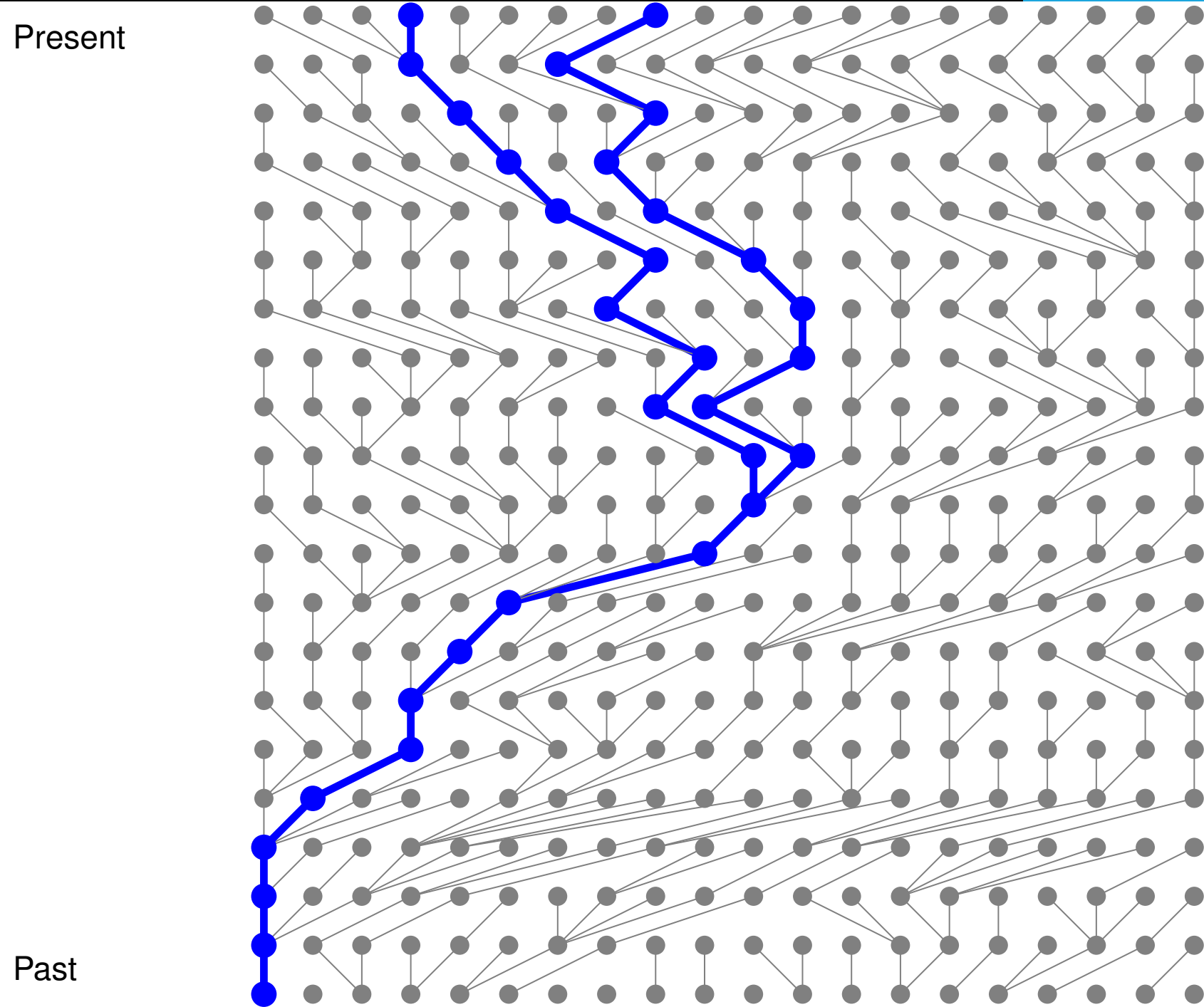


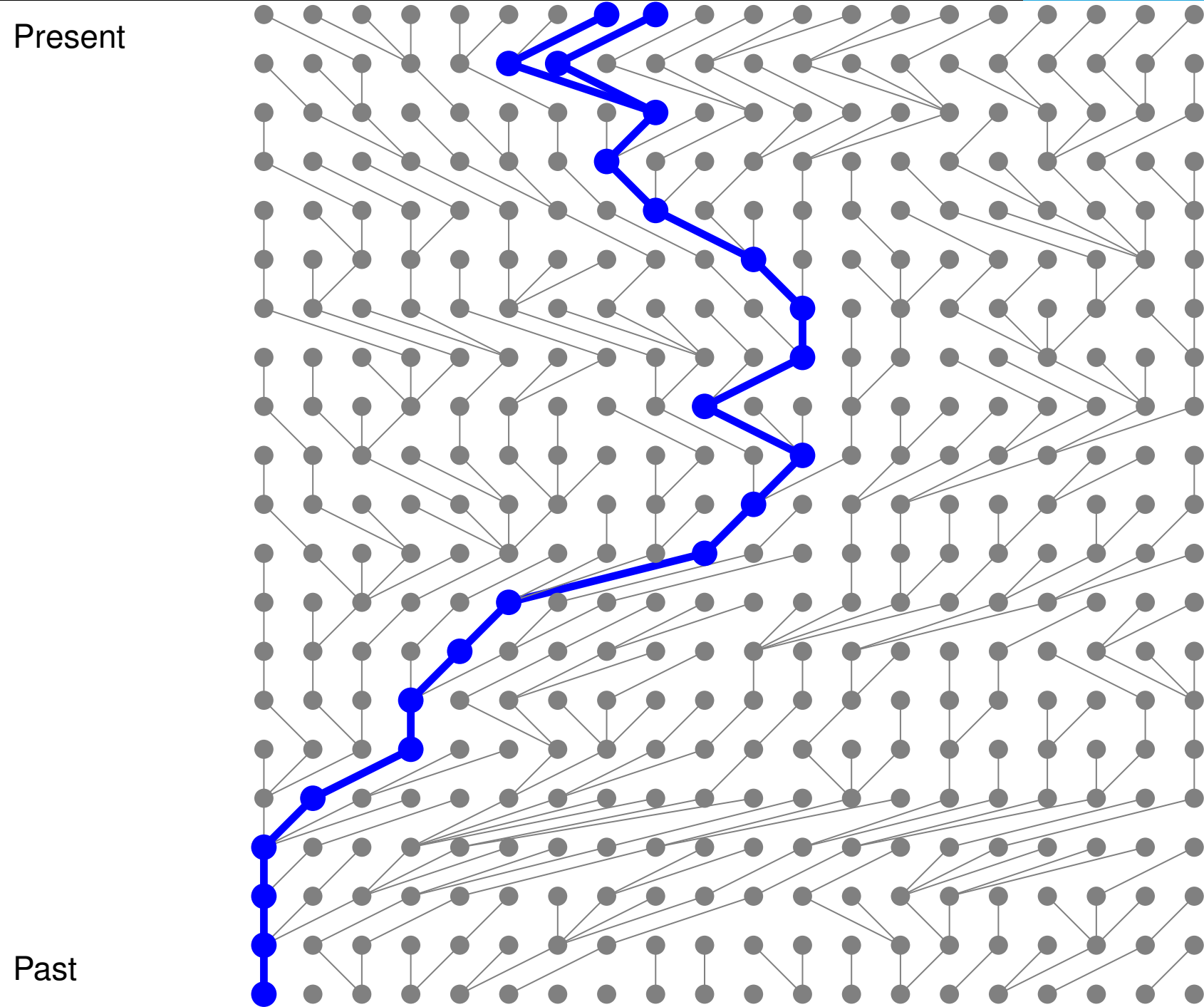
If we know the genealogy of the two individuals then we can calculate the probability as

$$P(\tau|N) = \left(1 - \frac{1}{2N}\right)^\tau \left(\frac{1}{2N}\right)$$

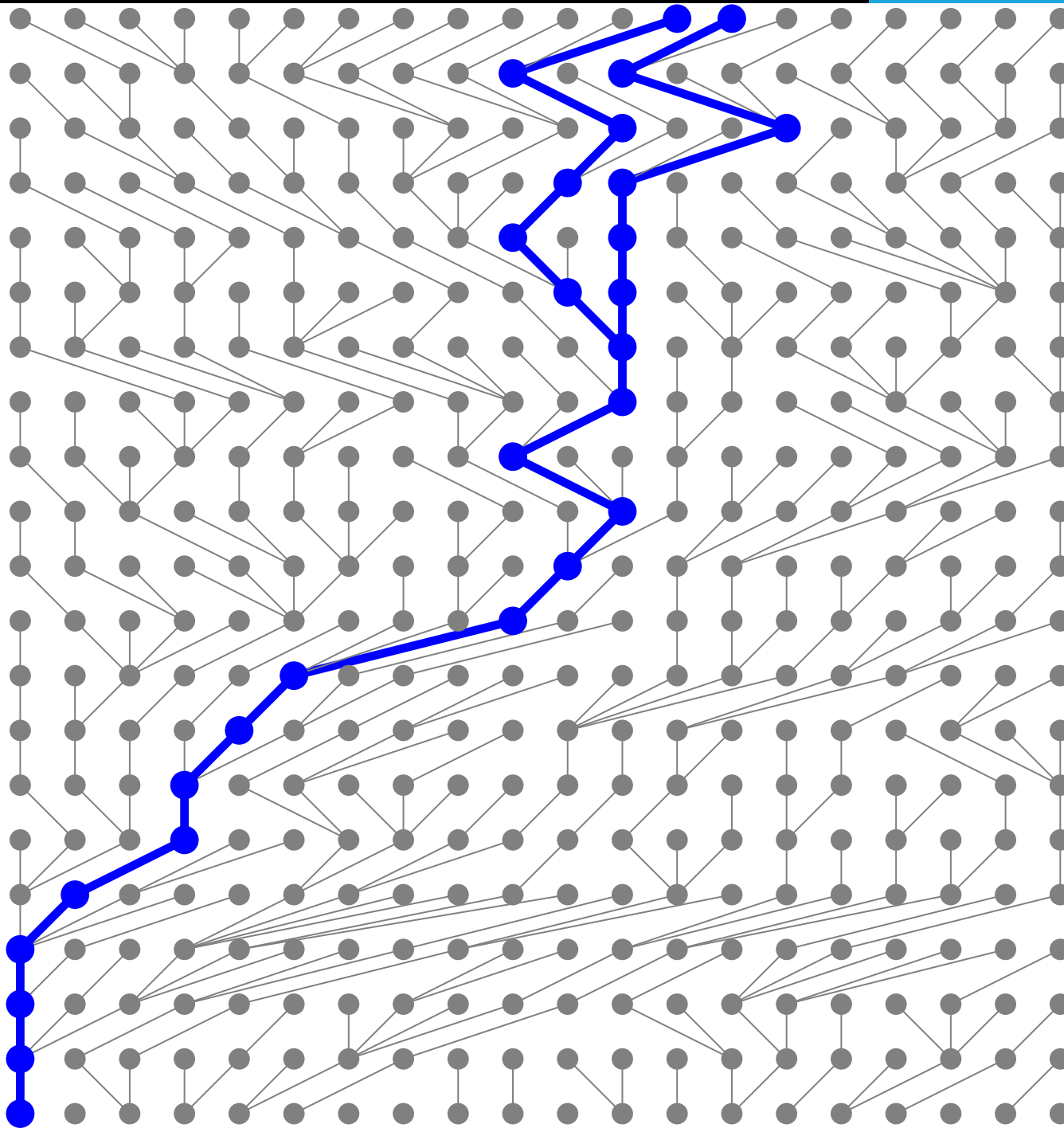
where τ is the number of generations with no coalescence. This formula is the Geometric Distribution and we can calculate the expectation of the waiting time until two random individuals coalesce:

$$\mathbb{E}(\tau) = 2N$$

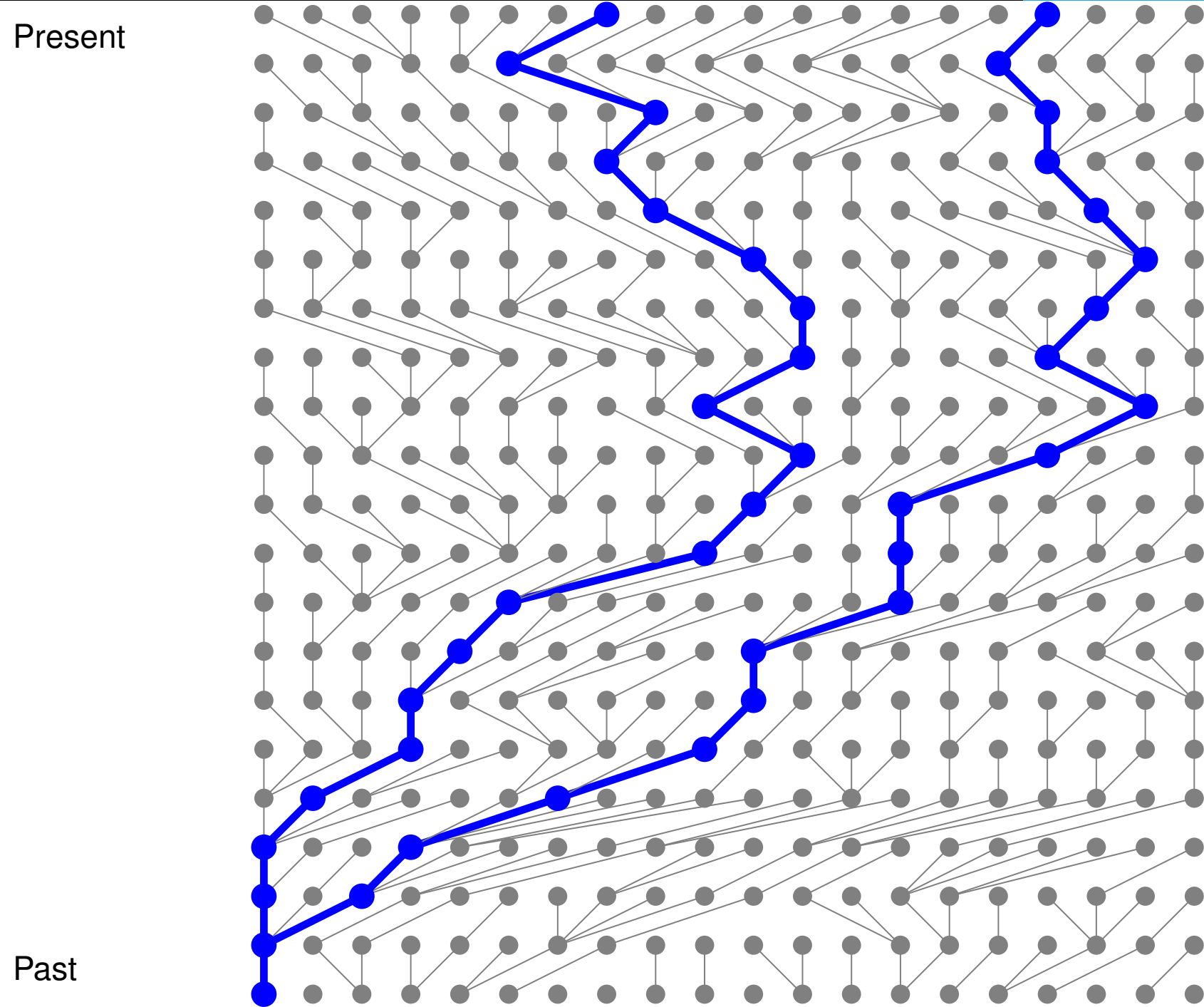




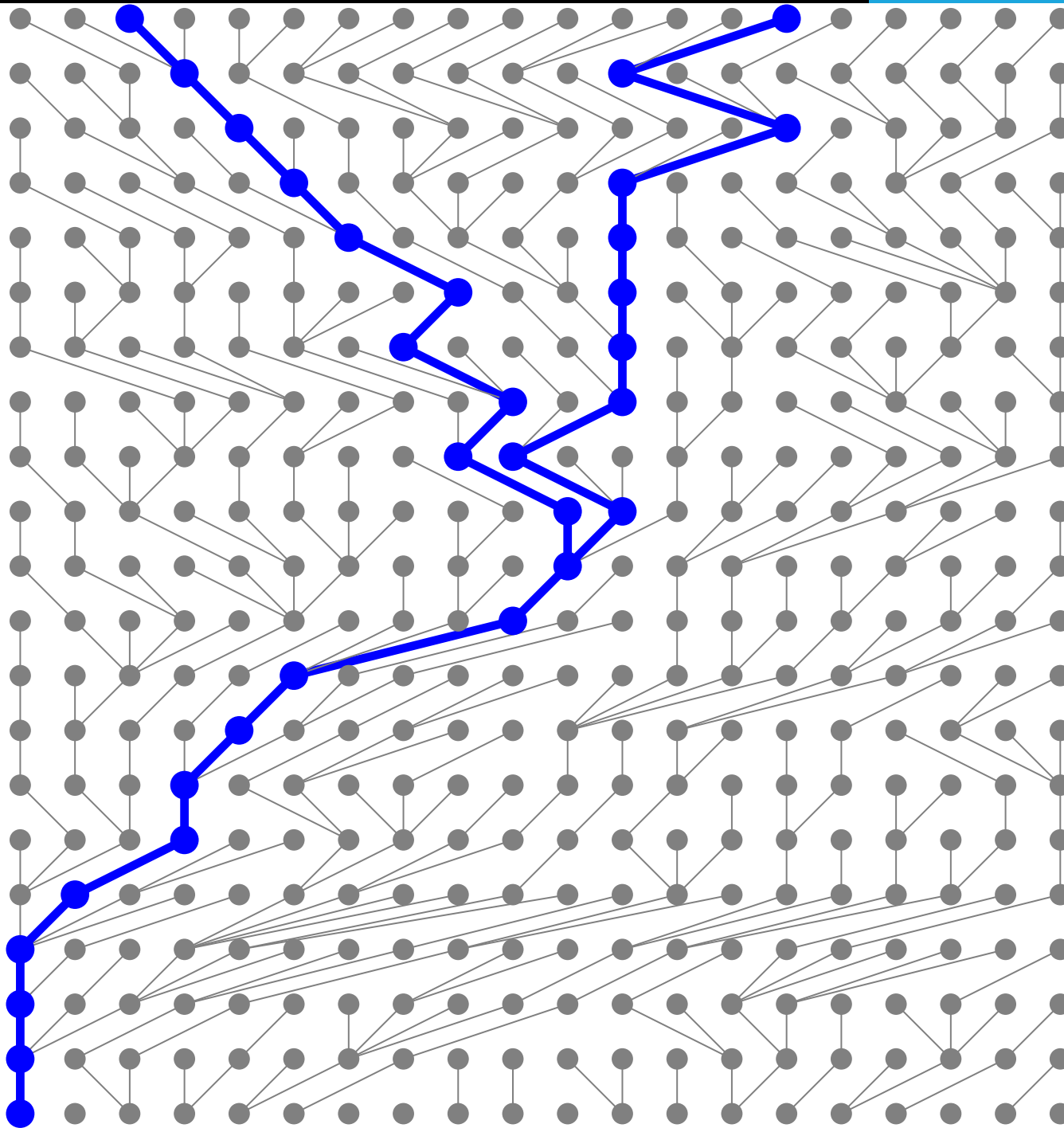
Present



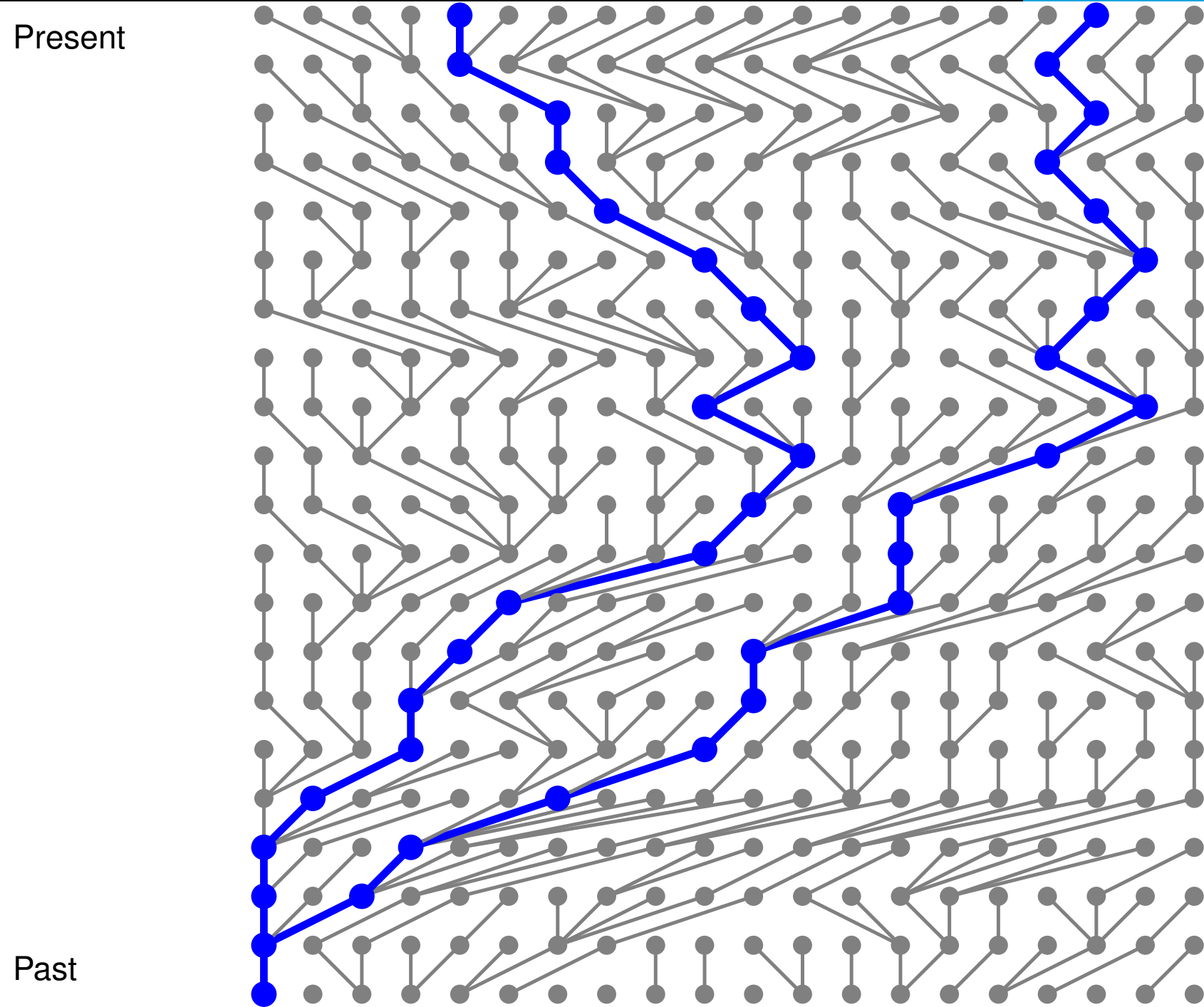
Past

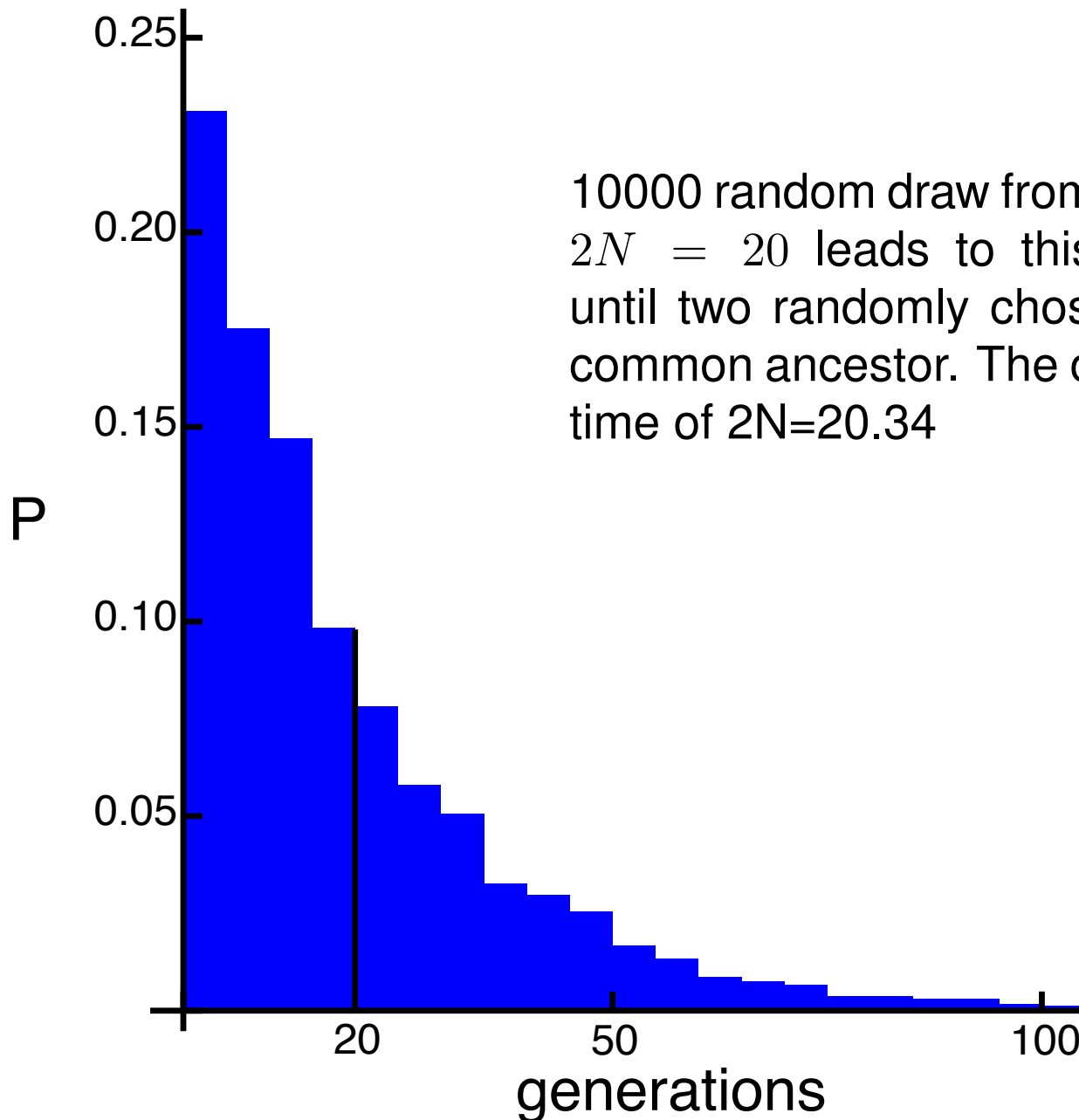


Present



Past



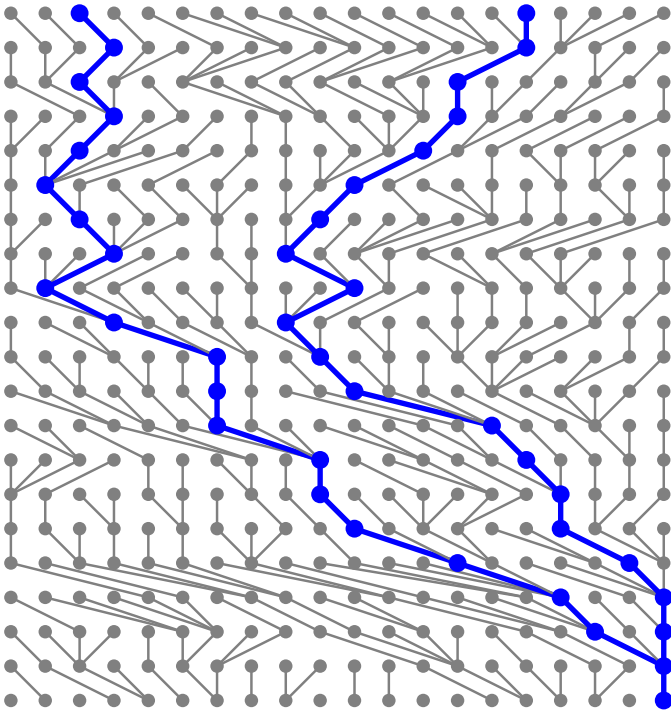


10000 random draw from a population with size $2N = 20$ leads to this distribution of times until two randomly chosen individuals have a common ancestor. The observed mean waiting time of $2N=20.34$

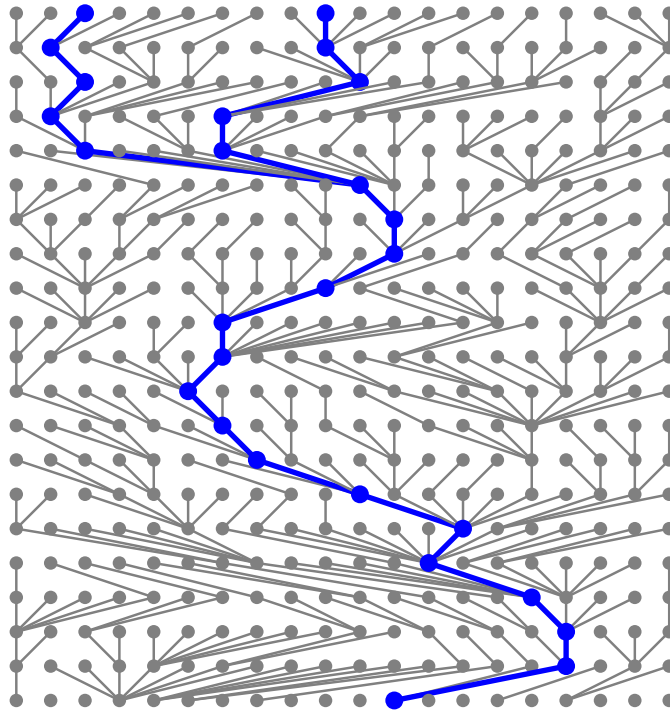
- ◆ For the time of coalescence in a sample of two, we will wait on average $2N$ generations assuming it is a Wright-Fisher population
- ◆ The model assumes that the generations are discrete and non-overlapping
- ◆ Real populations do not necessarily behave like a Wright-Fisher (the '*ideal*' population)
- ◆ *We assume that calculation using Wright-Fisher populations can be extrapolated to real populations.*

Other population models

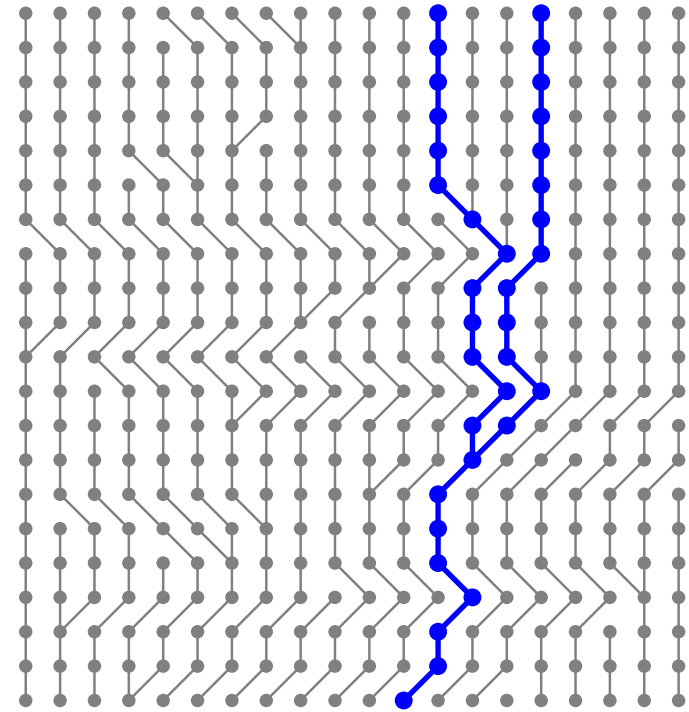
Wright-Fisher



Canning

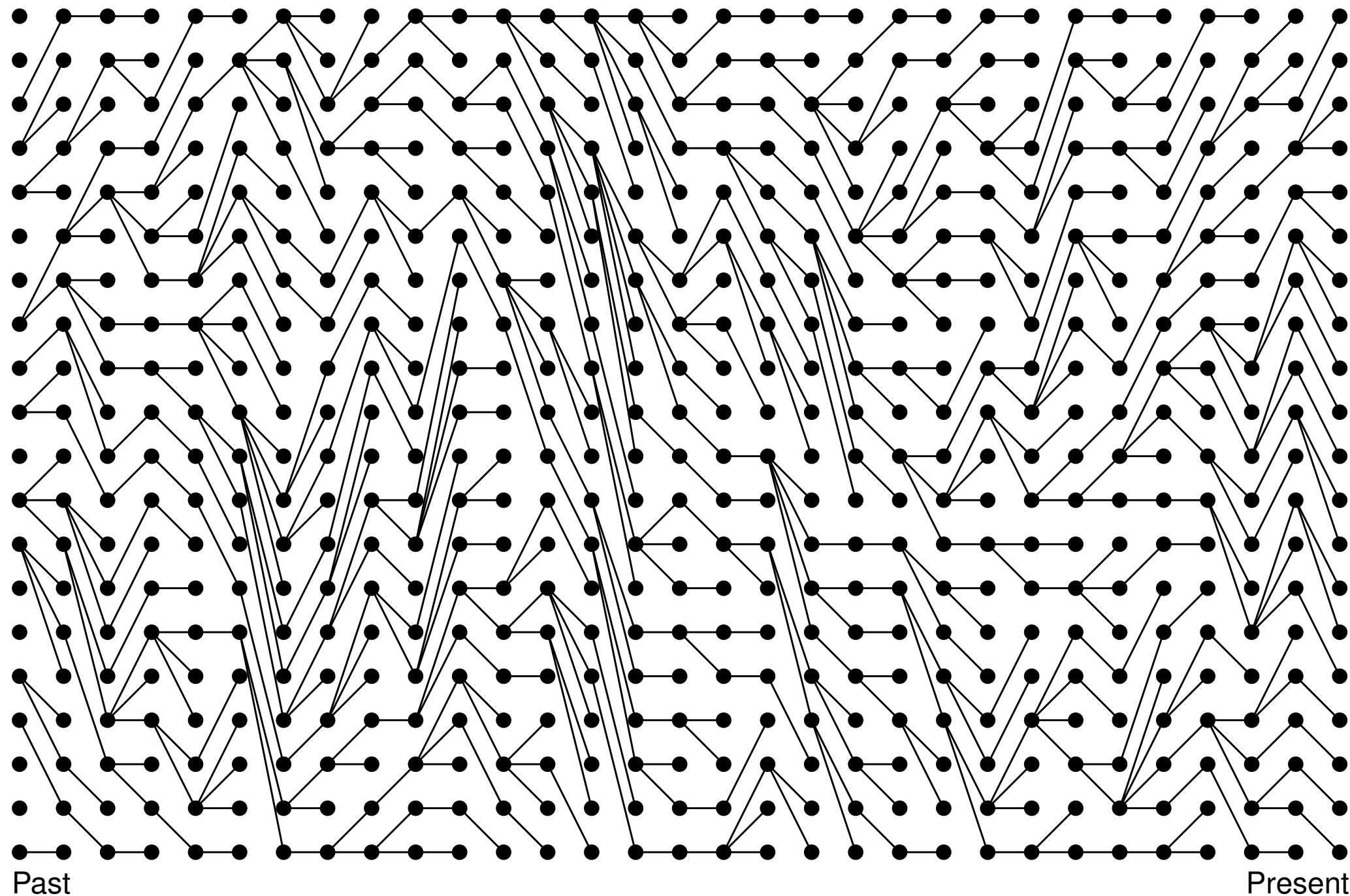


Moran



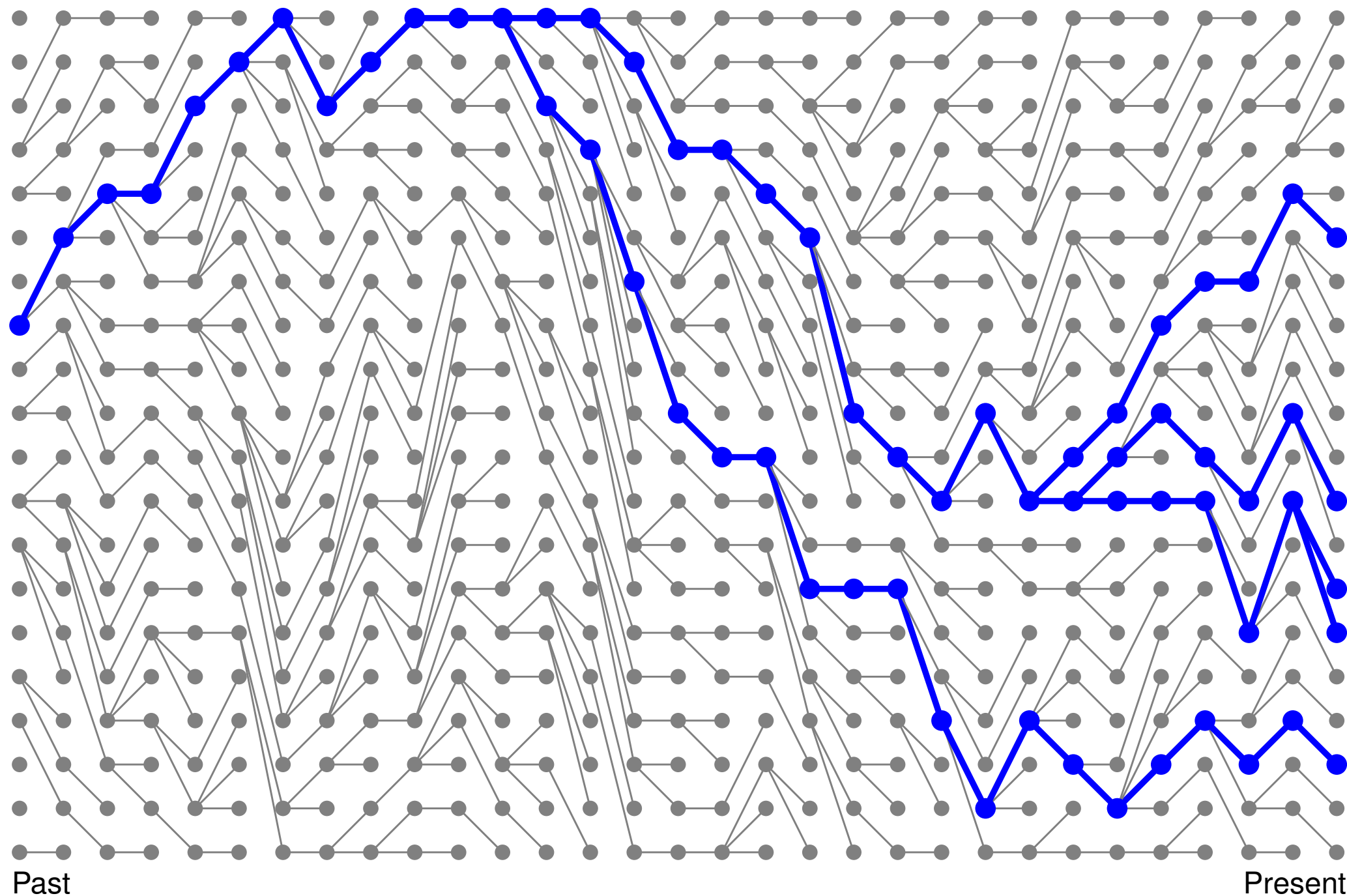
Sample larger than TWO

Wright-Fisher



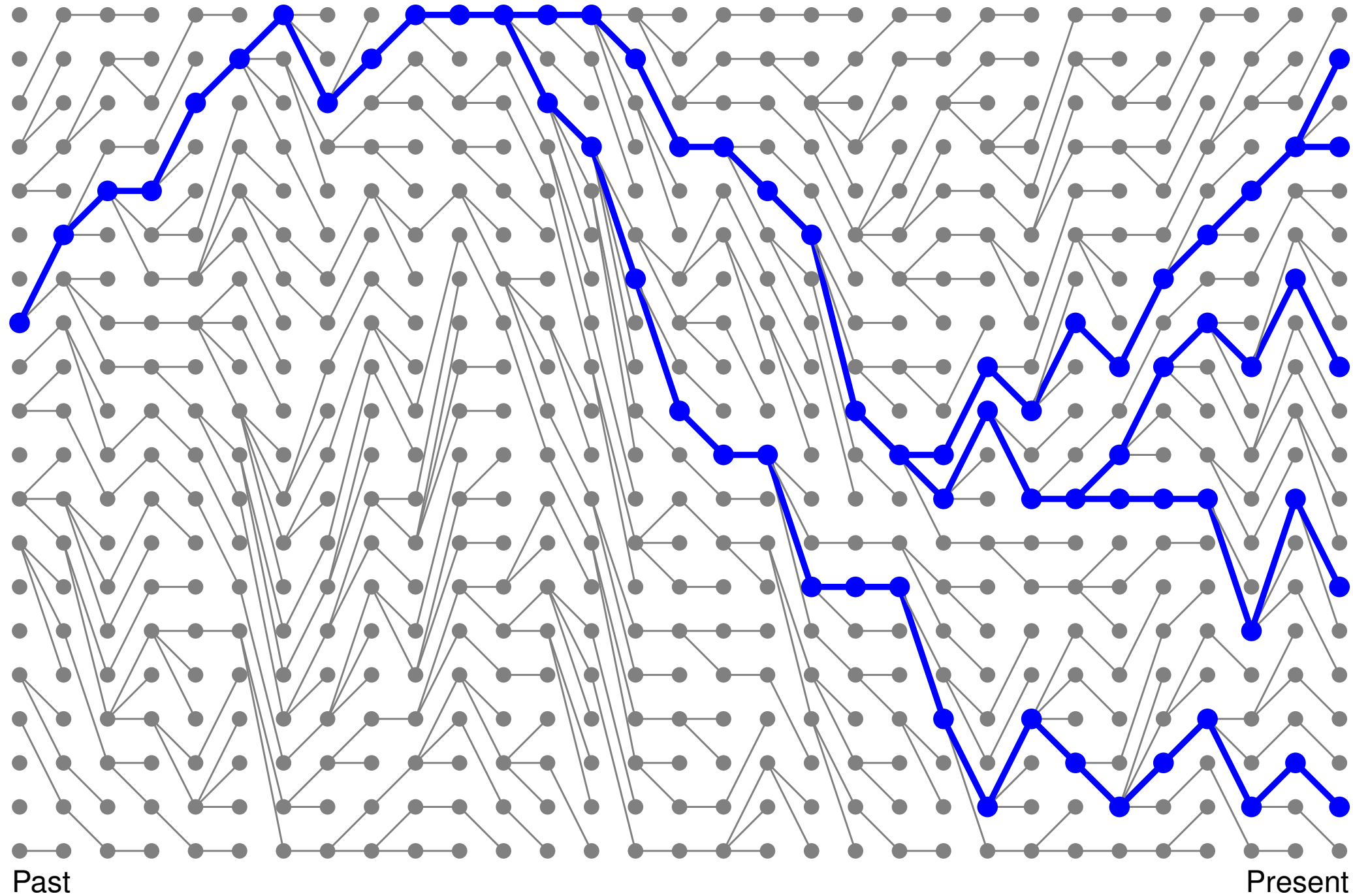
Sample larger than TWO

Wright-Fisher



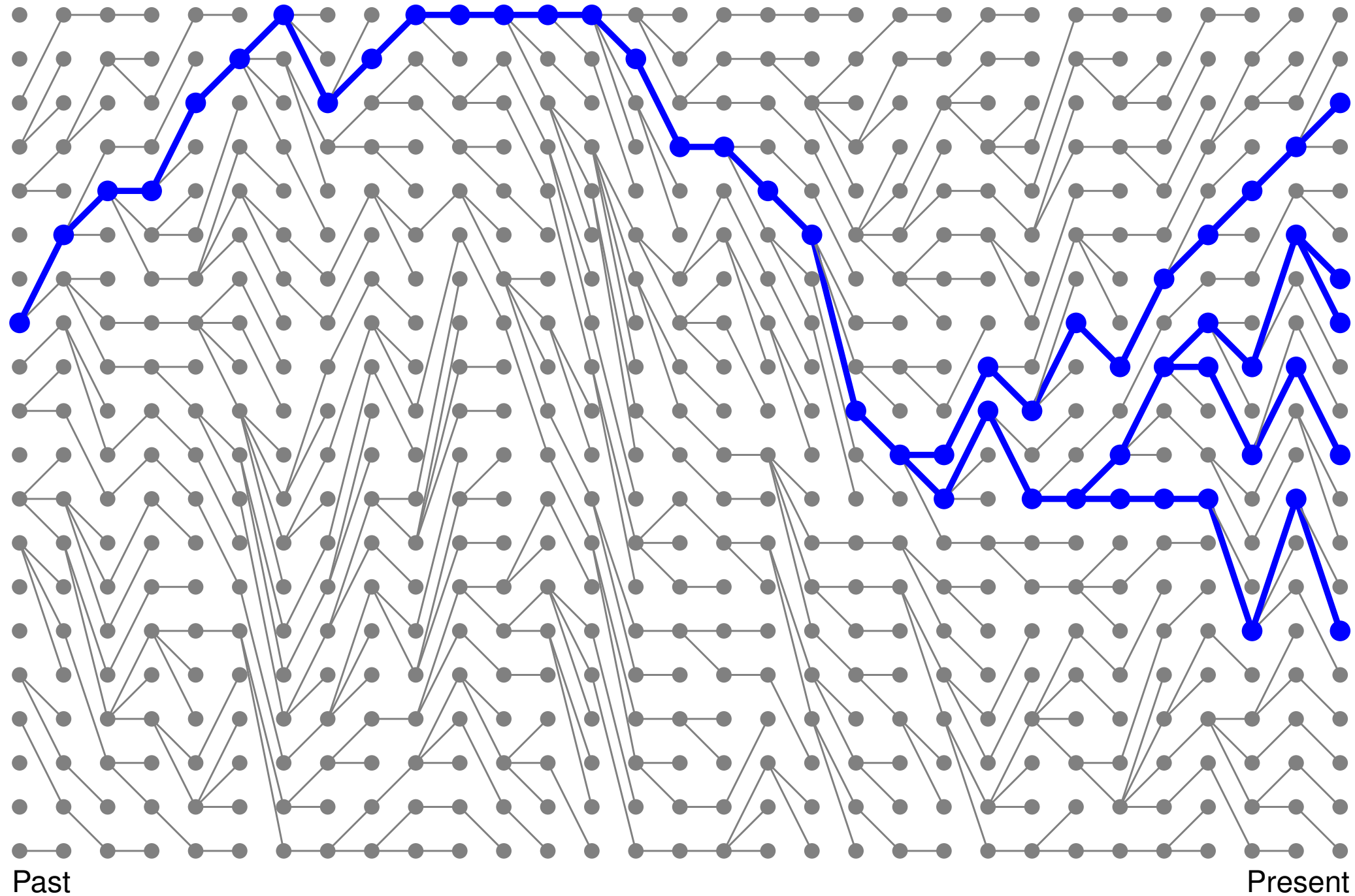
Sample larger than TWO

Wright-Fisher



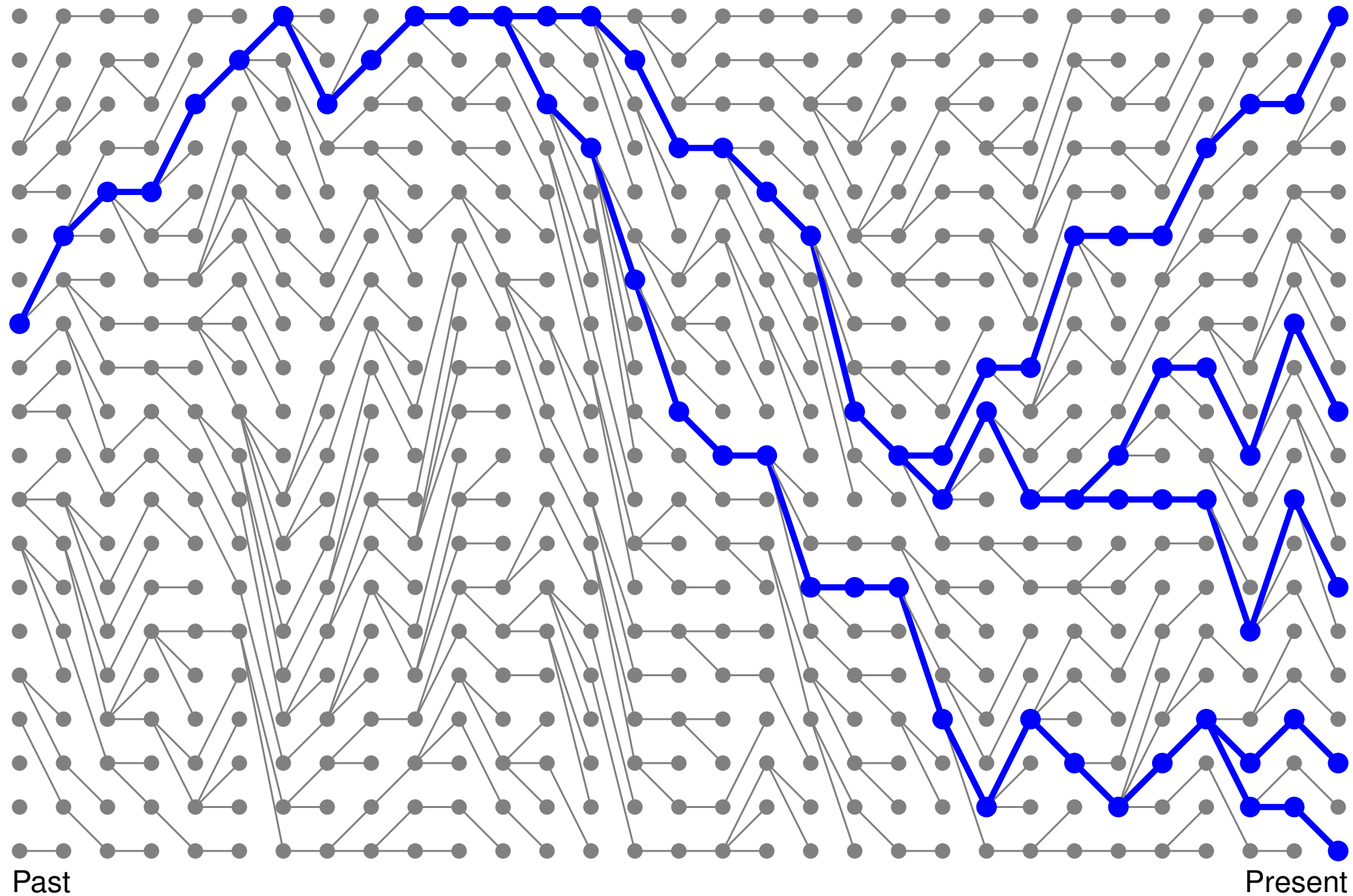
Sample larger than TWO

Wright-Fisher



Sample larger than TWO

Wright-Fisher



Samples larger than two

Sir J. F. C. Kingman described in 1982 the n -coalecent. He showed the behavior of a sample of size n , and its probability structure looking backwards in time.

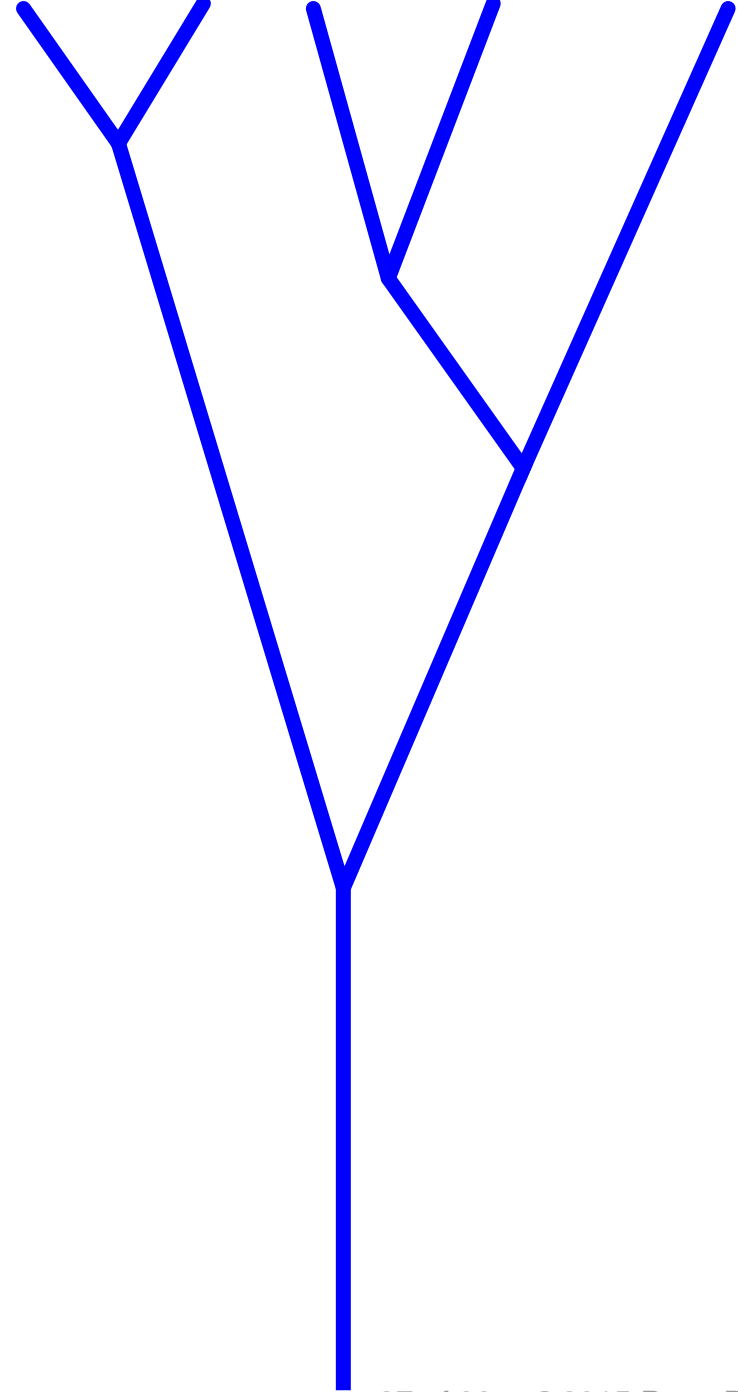
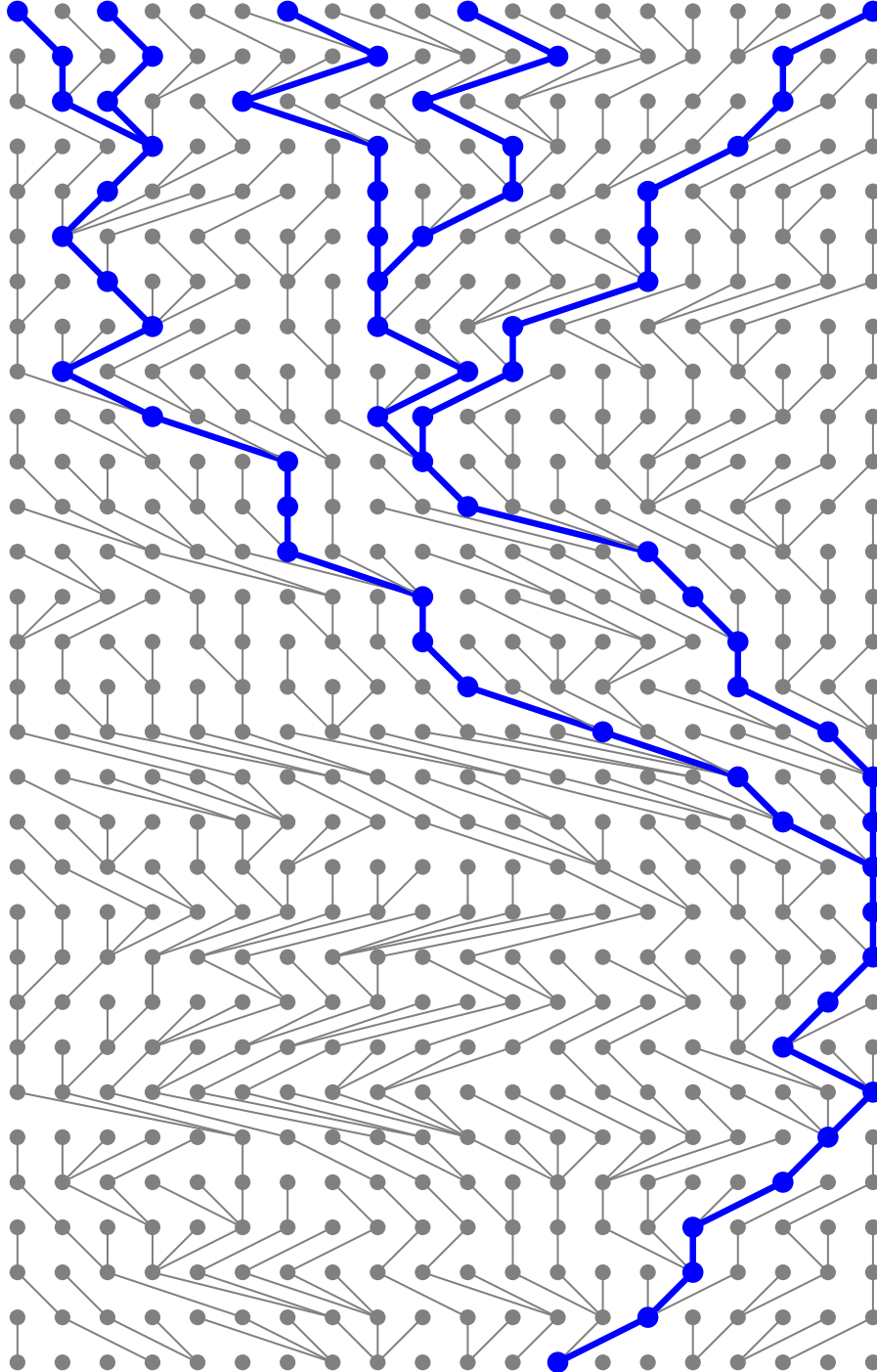
General findings:

$$\text{coalescence rate} = \binom{n}{2} = \frac{n(n-1)}{2}$$

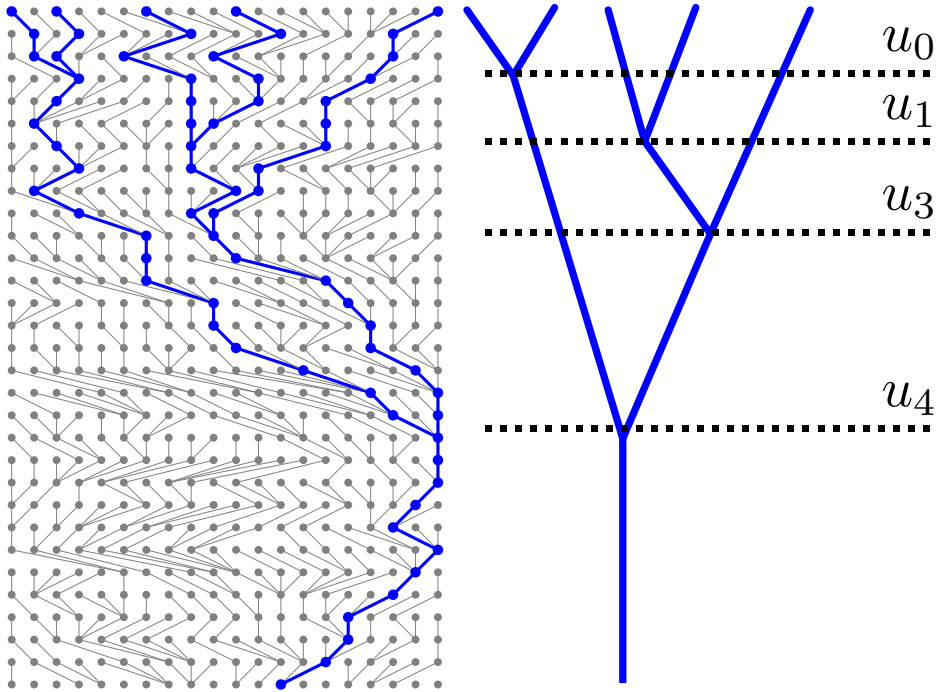
Once a coalescence happened n is reduced to $n - 1$ because two lineages merged into one. He then imposed a continuous approximation of the Canning's exchangeable model to get results.



Samples larger than two

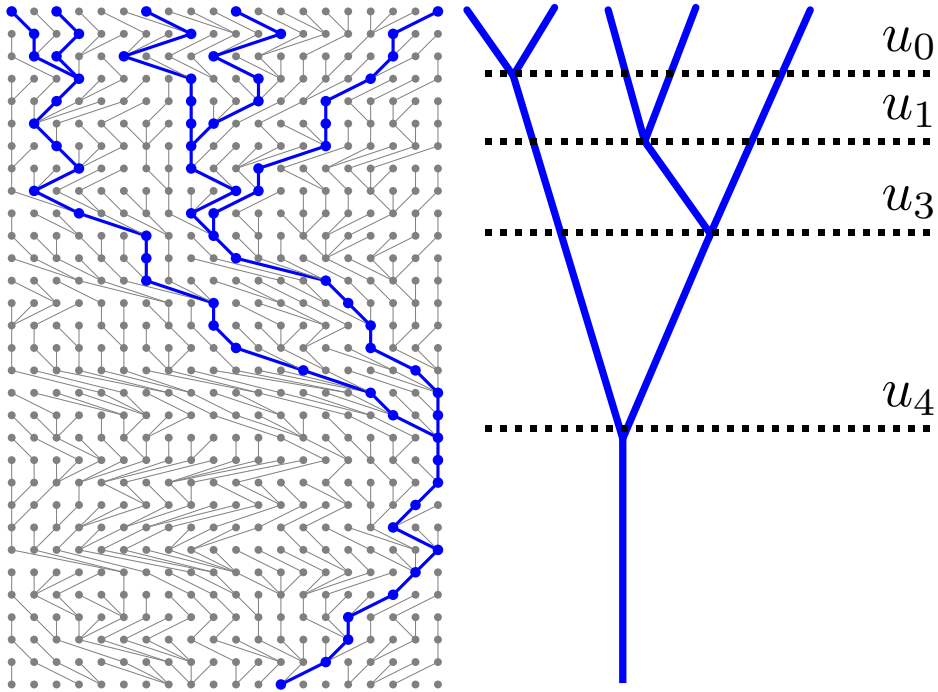


Samples larger than two



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample n and the total population size N .

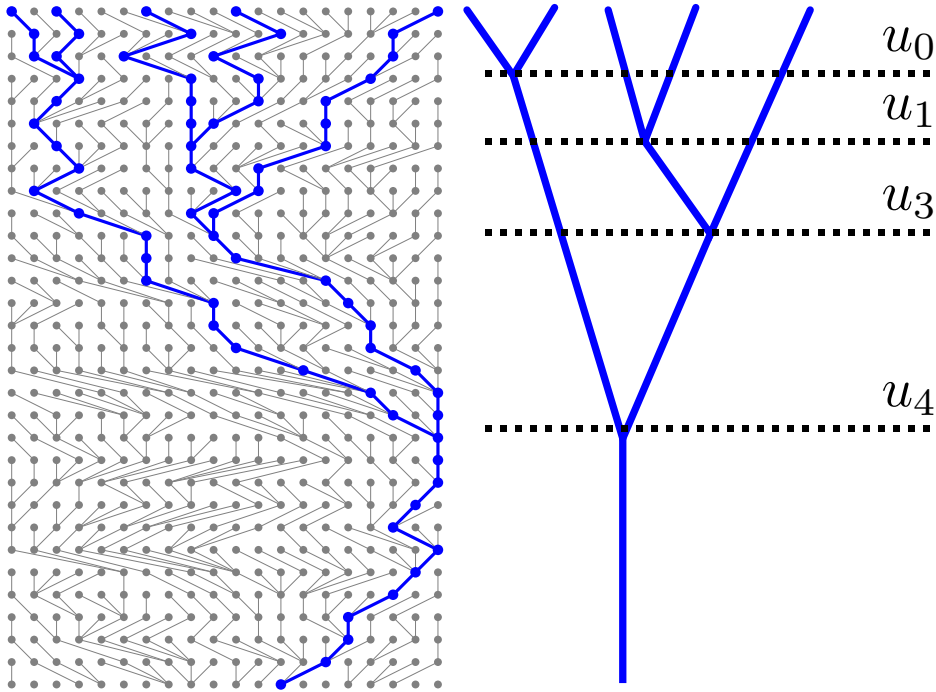
Samples larger than two



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample n and the total population size N .

Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:

Samples larger than two



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample n and the total population size N .

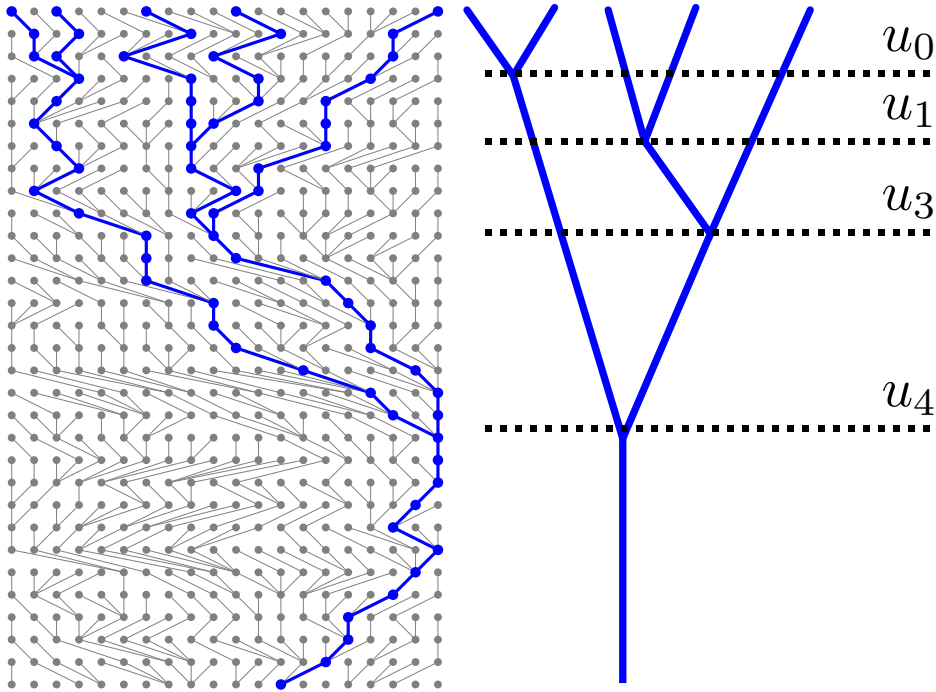
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with an exponential distribution:

$$P(u_j|N) = e^{-u_j\lambda}\lambda$$

with the scaled coalescence rate

$$\lambda = \binom{k}{2} \frac{1}{2N} \times \text{Prob}(\text{others do not coalesce})$$

Samples larger than two



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample n and the total population size N .

Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:

$$P(u_j|N) = e^{-u_j\lambda}\lambda$$

with the scaled coalescence rate

$$\lambda = \binom{k}{2} \frac{1}{2N} = \frac{k(k-1)}{2(2N)} = \frac{k(k-1)}{4N}$$

Chance of coalescence in a particular generation

The standard coalescence uses the assumption that only one coalescence happen within a particular generation. This is a questionable assumption, but does it matter?

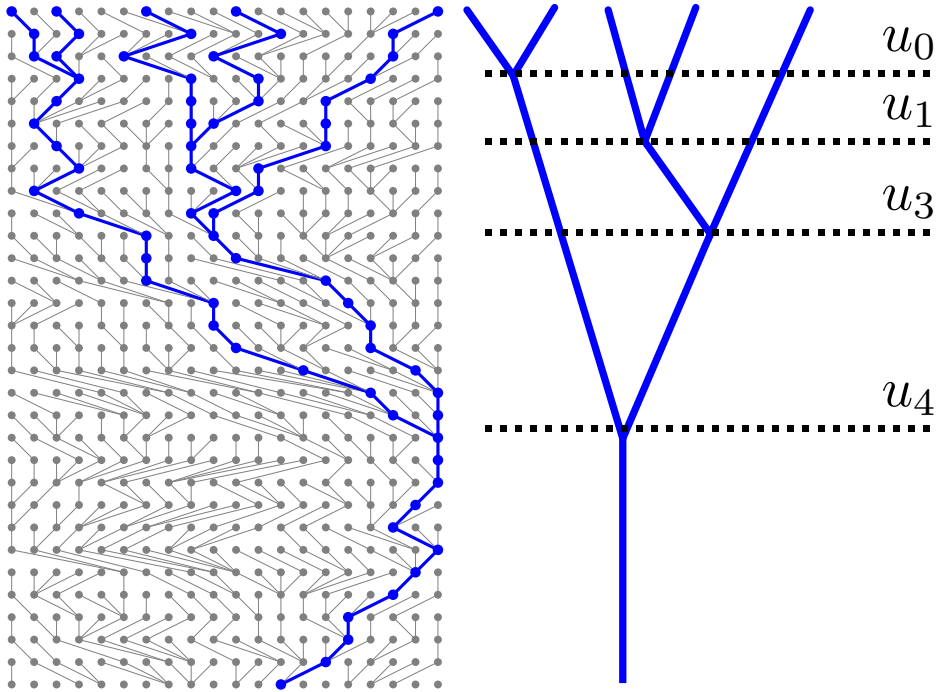
Here are the exact probabilities of 0, 1, or more coalescences with 10 lineages in populations of different sizes:

N	0	1	>1
100	0.796	0.187	0.017
1000	0.978	0.022	0.000
10000	0.998	0.002	0.000

Note that increasing the population size by a factor of 10 reduces the coalescent rate for pairs by about 10-fold, but reduces the rate for triples (or more) by about 100-fold.

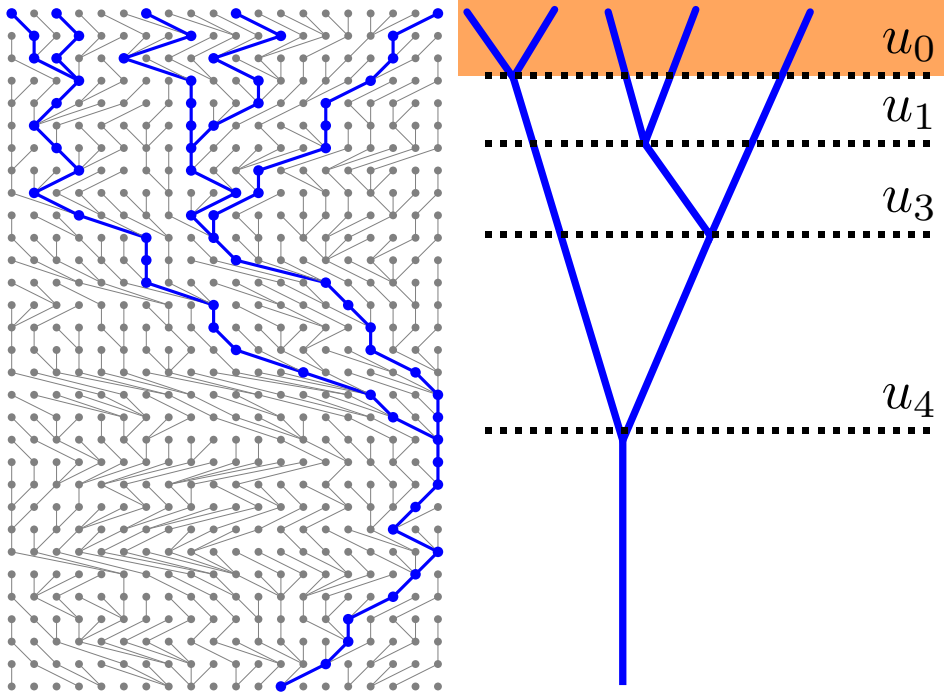
Samples larger than two

the coalescent



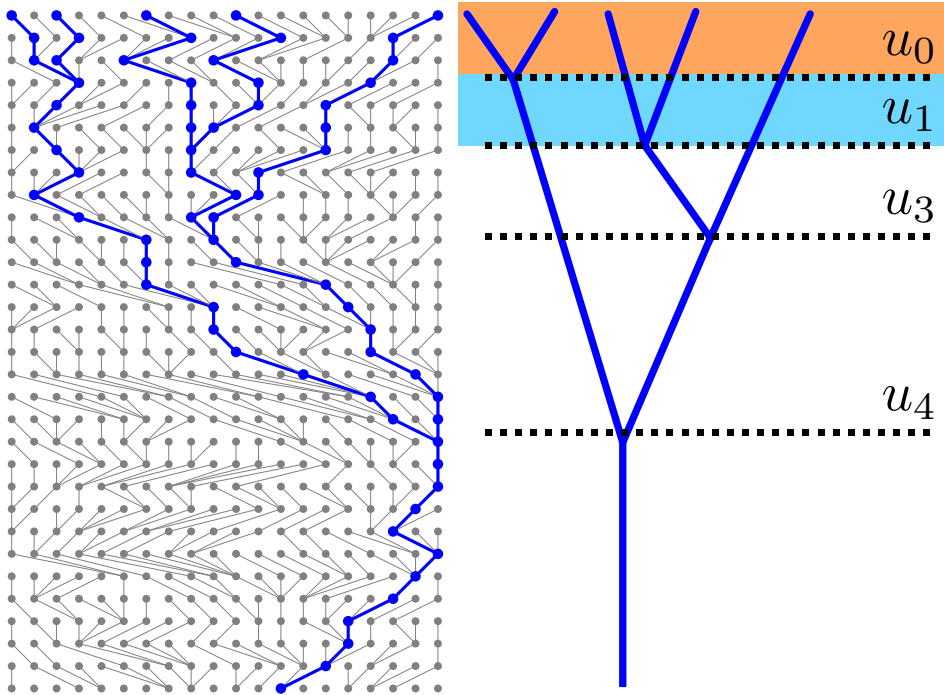
We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$P(G|N)$$



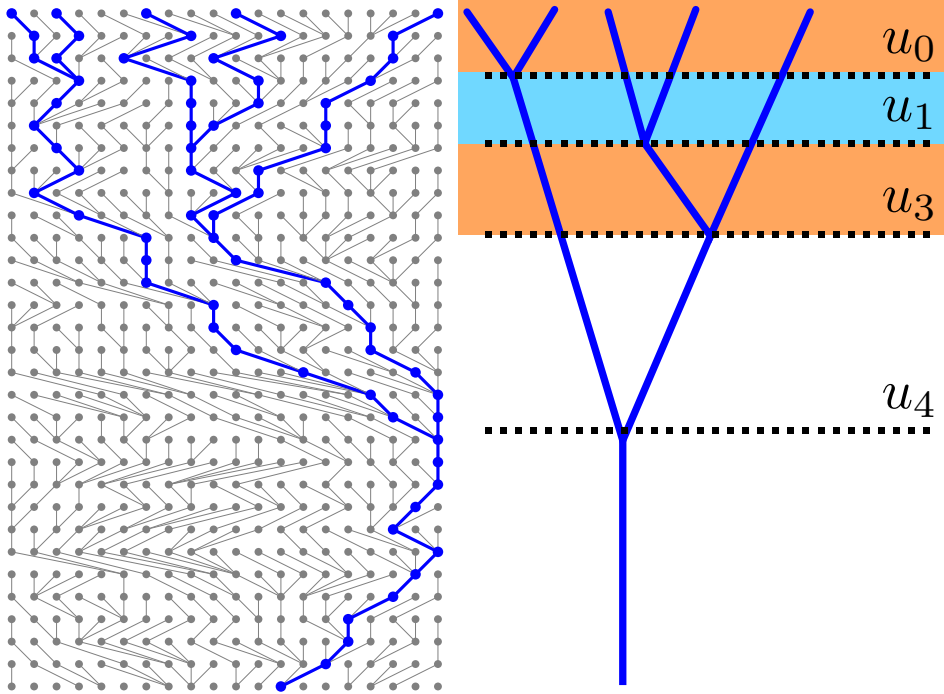
We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$P(G|N) = P(u_0|N, i_1, i_2) \times$$



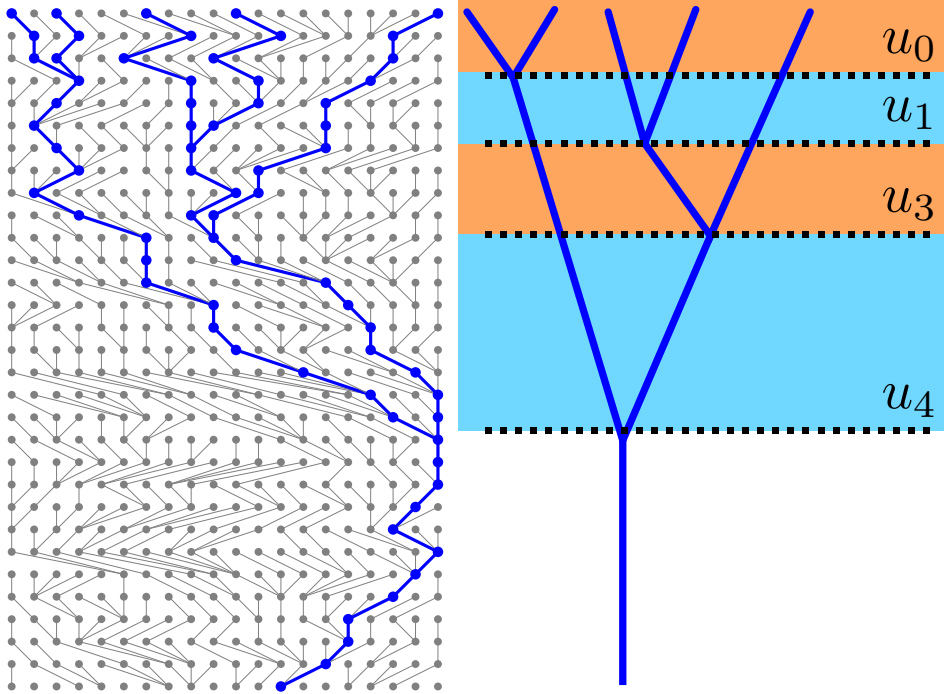
We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$P(G|N) = P(u_0|N, i_1, i_2) \times P(u_1|N, i_3, i_4)$$



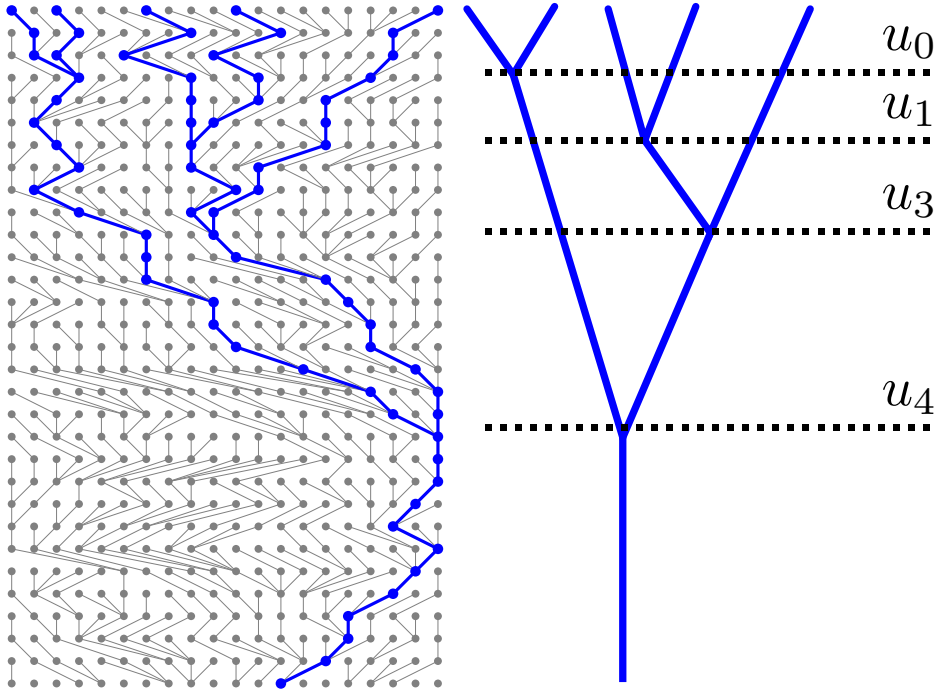
We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$\begin{aligned} P(G|N) = & P(u_0|N, i_1, i_2) \\ & \times P(u_1|N, i_3, i_4) \\ & \times P(u_3|N, i_{3,4}, i_5) \end{aligned}$$



We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

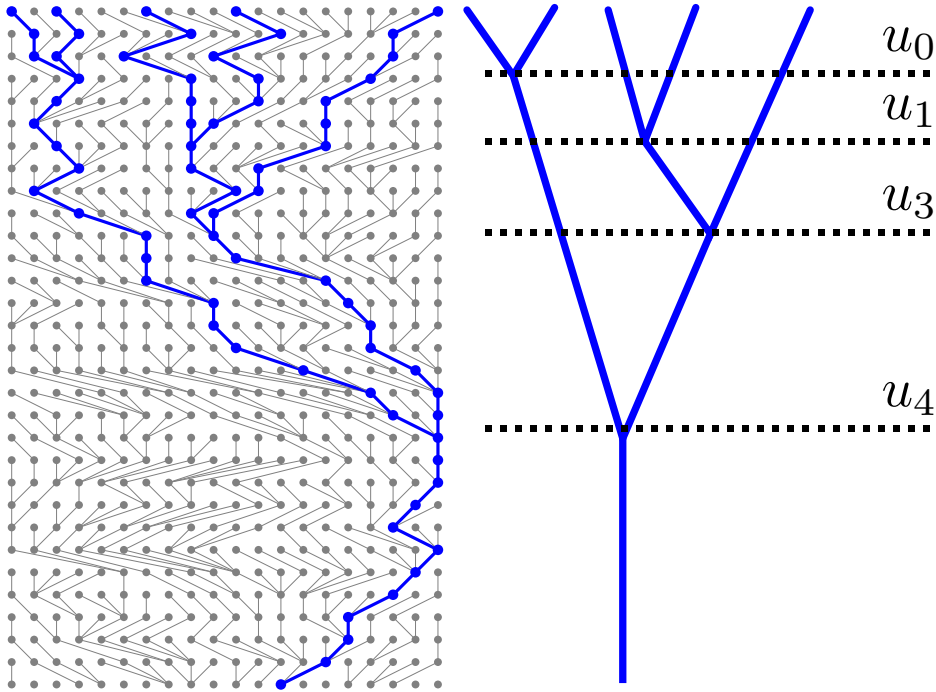
$$\begin{aligned} P(G|N) = & P(u_0|N, i_1, i_2) \\ & \times P(u_1|N, i_3, i_4) \\ & \times P(u_3|N, i_{3,4}, i_5) \\ & \times P(u_4|N, i_{1,2}, i_{3,4,5}) \end{aligned}$$



We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$\begin{aligned} P(G|N) = & P(u_0|N, i_1, i_2) \\ & \times P(u_1|N, i_3, i_4) \\ & \times P(u_3|N, i_{3,4}, i_5) \\ & \times P(u_4|N, i_{1,2}, i_{3,4,5}) \end{aligned}$$

$$P(G|N) = \prod_{j=0}^T e^{-u_j \frac{k_j(k_j-1)}{4N}} \frac{2}{4N}$$



Each interval u_j is independent of the others, the expected length of the interval is the inverse of the coalescent rate. Thus we can sum these expectations to get to expectation of the depth of the genealogy.

$$\mathbb{E}(\tau_{\text{MRCA}}) = \text{Sum of the expectation of each time interval} = \sum_{j=0}^J \frac{4N}{k_j(k_j - 1)}$$

$$\lim_{k \rightarrow \infty} \mathbb{E}(\tau_{\text{MRCA}}) = 2N + \frac{2}{3}N + \frac{1}{3}N + \frac{1}{5}N + \frac{2}{15}N + \dots = 4N$$

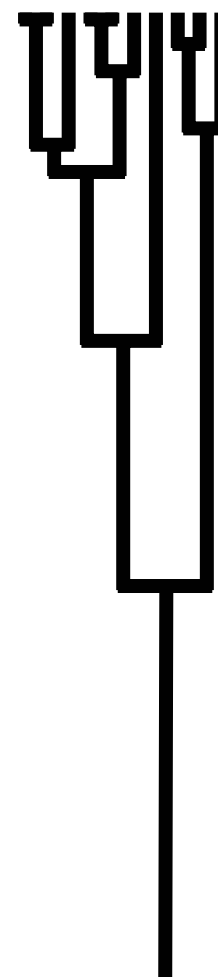
$$\lim_{k \rightarrow \infty} \sigma(\tau_{\text{MRCA}}) = 4N$$

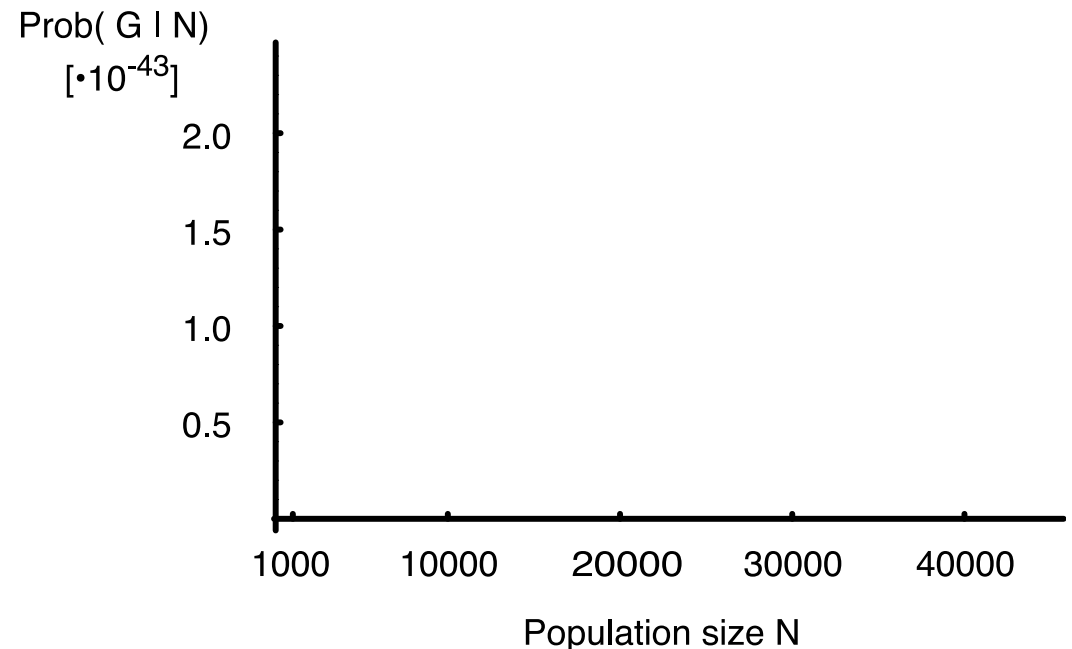
If we know the genealogy G with certainty then we can calculate the population size N . Finding the maximum probability $P(G|N, k)$ is simple, we evaluate all possible values for N and pick the value with the highest probability.

If we know the genealogy G with certainty then we can calculate the population size N . Finding the maximum probability $P(G|N, k)$ is simple, we evaluate all possible values for N and pick the value with the highest probability.



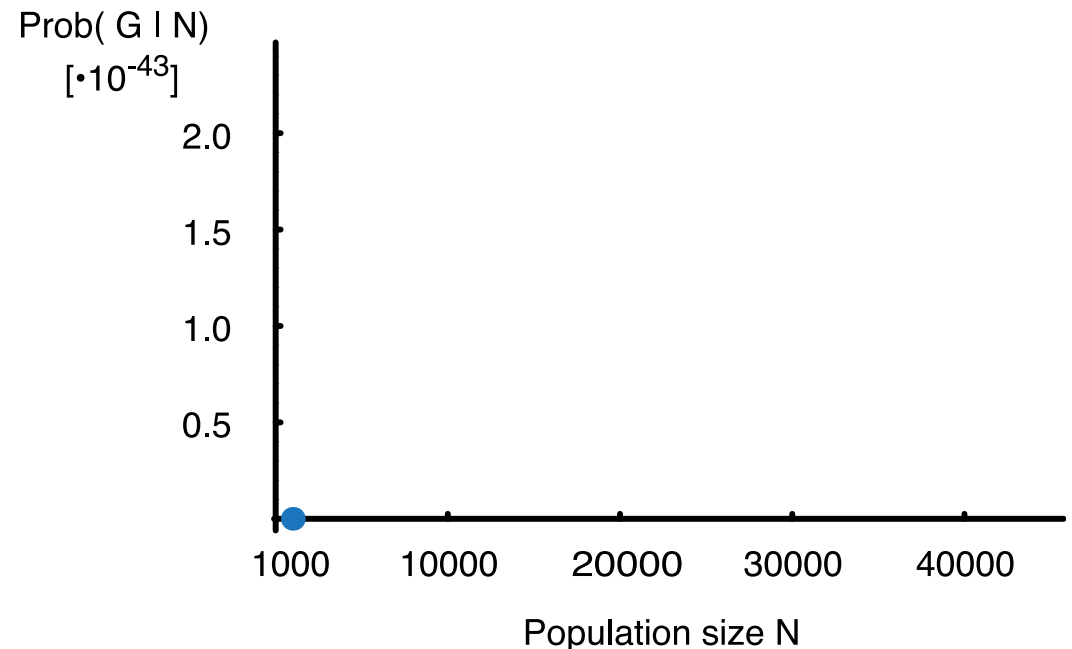
If we know the genealogy G with certainty then we can calculate the population size N . Finding the maximum probability $P(G|N, k)$ is simple, we evaluate all possible values for N and pick the value with the highest probability.





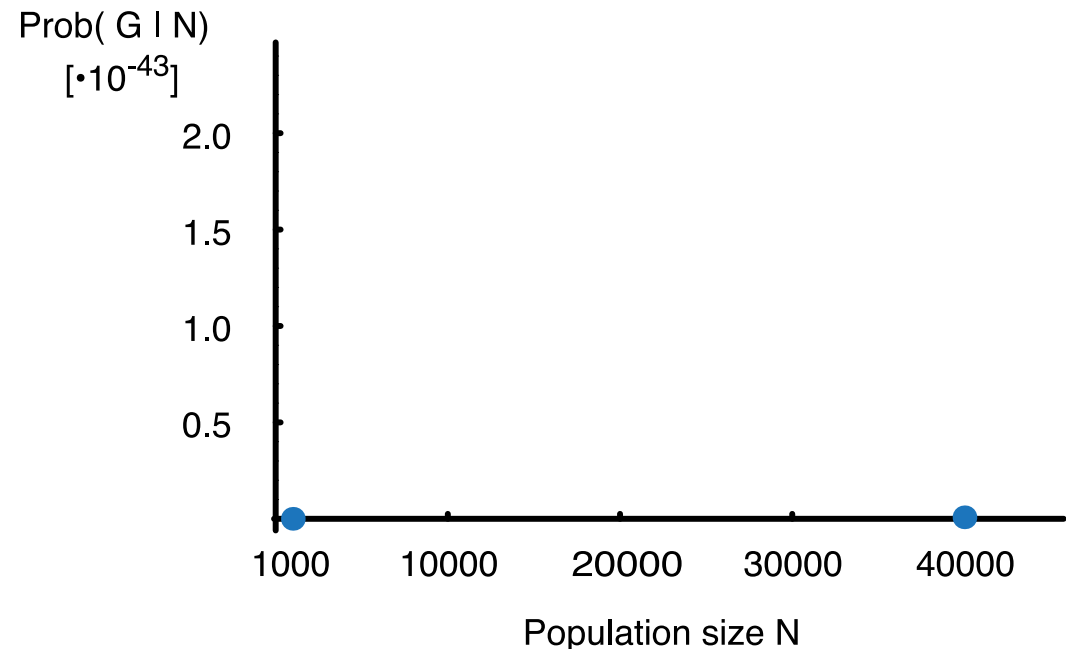
If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



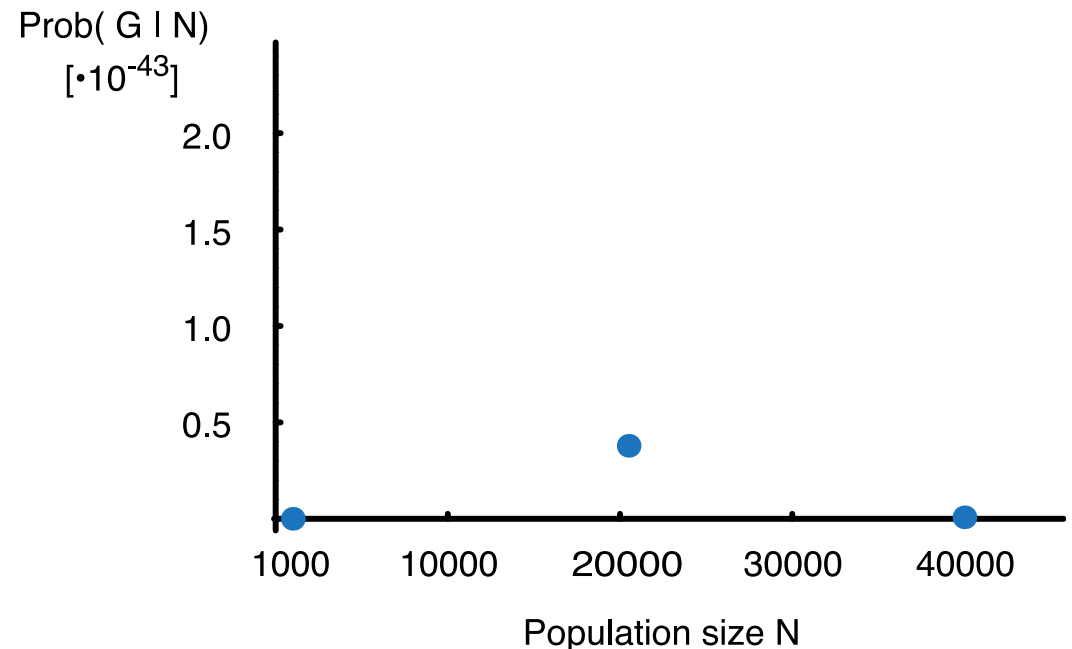
If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



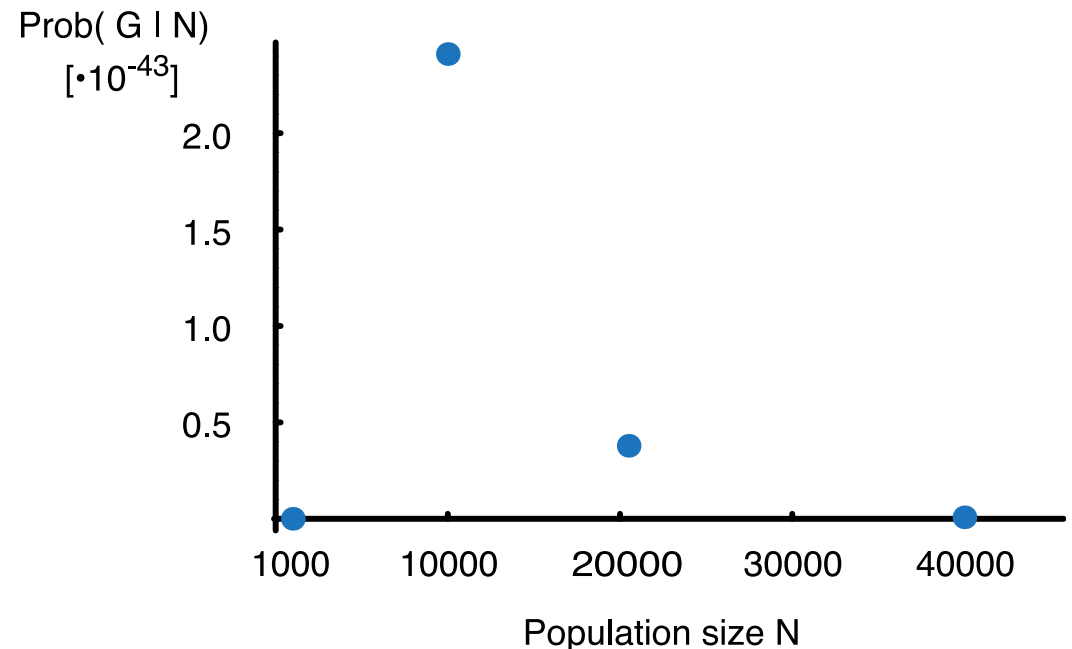
If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



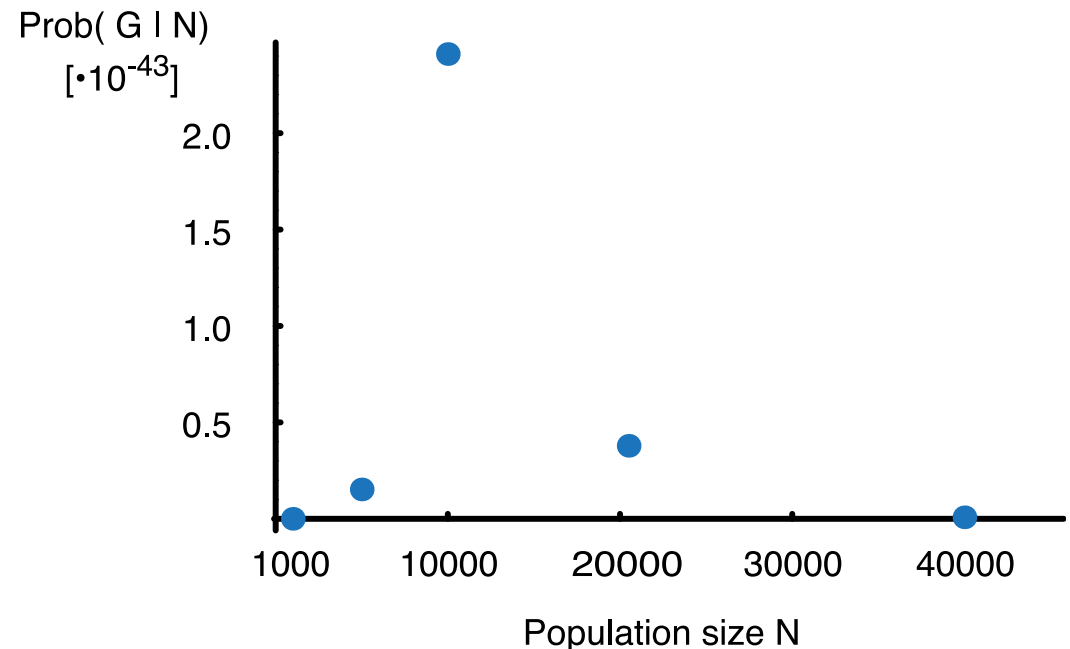
If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



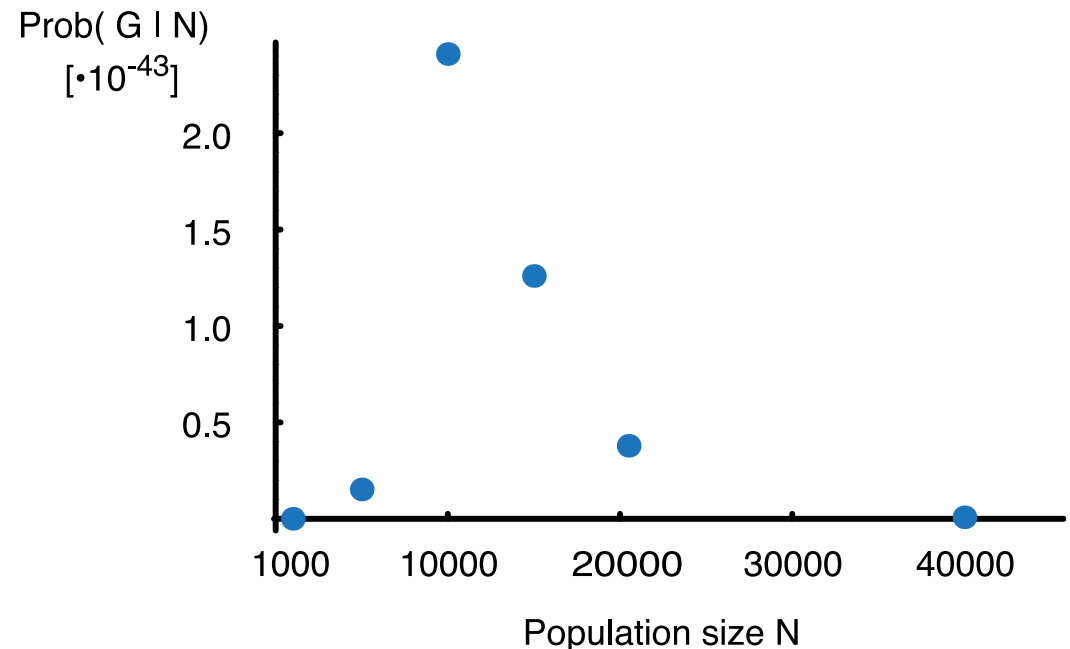
If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



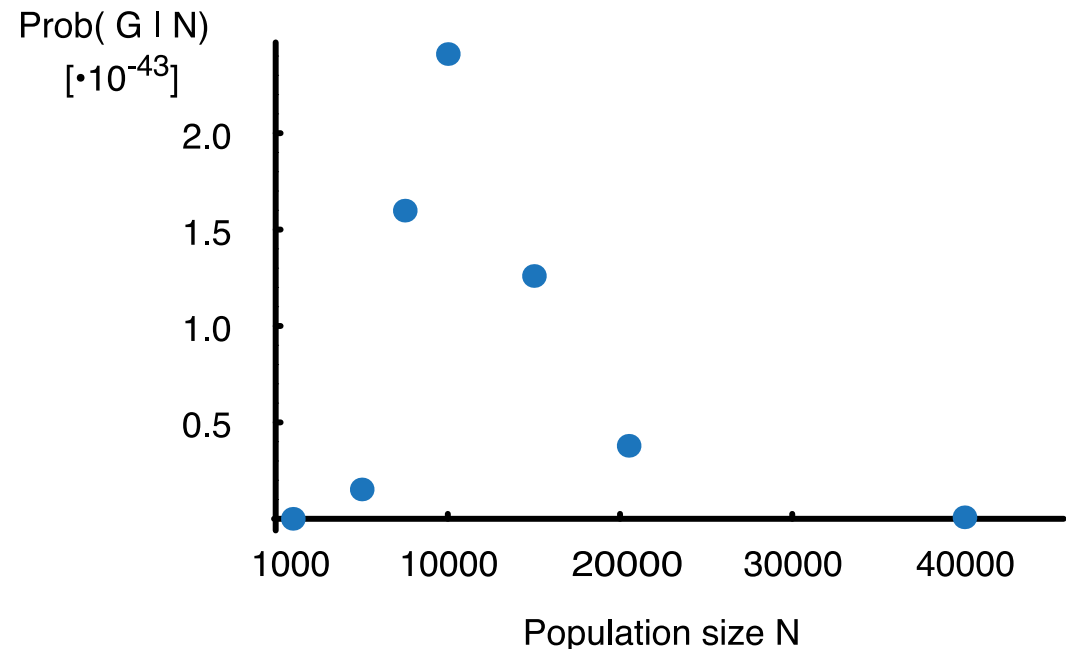
If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



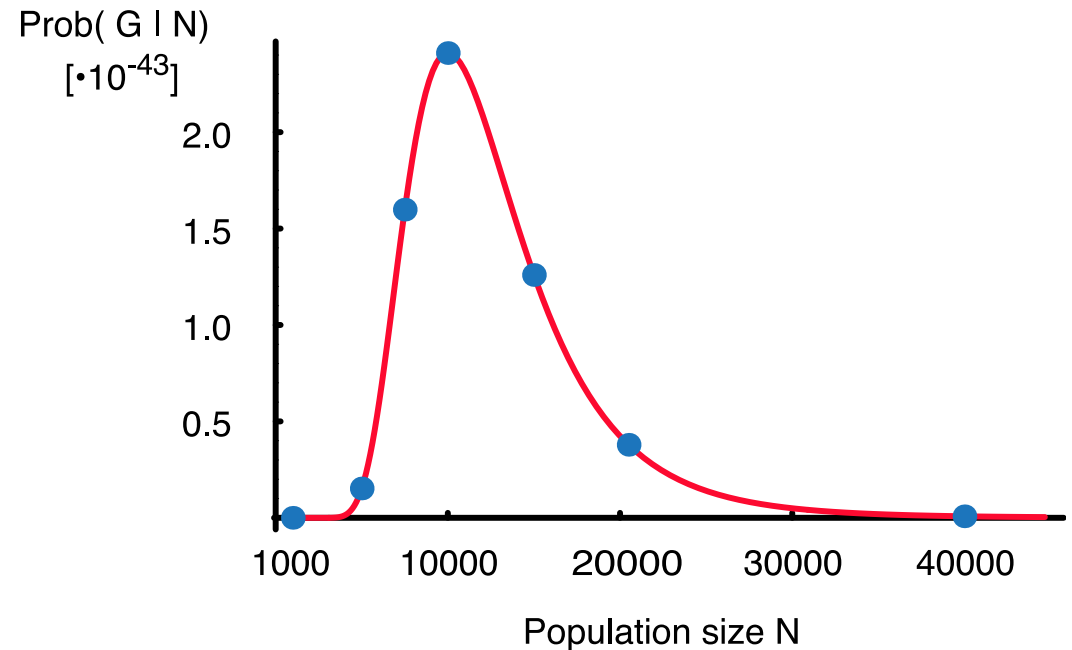
If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



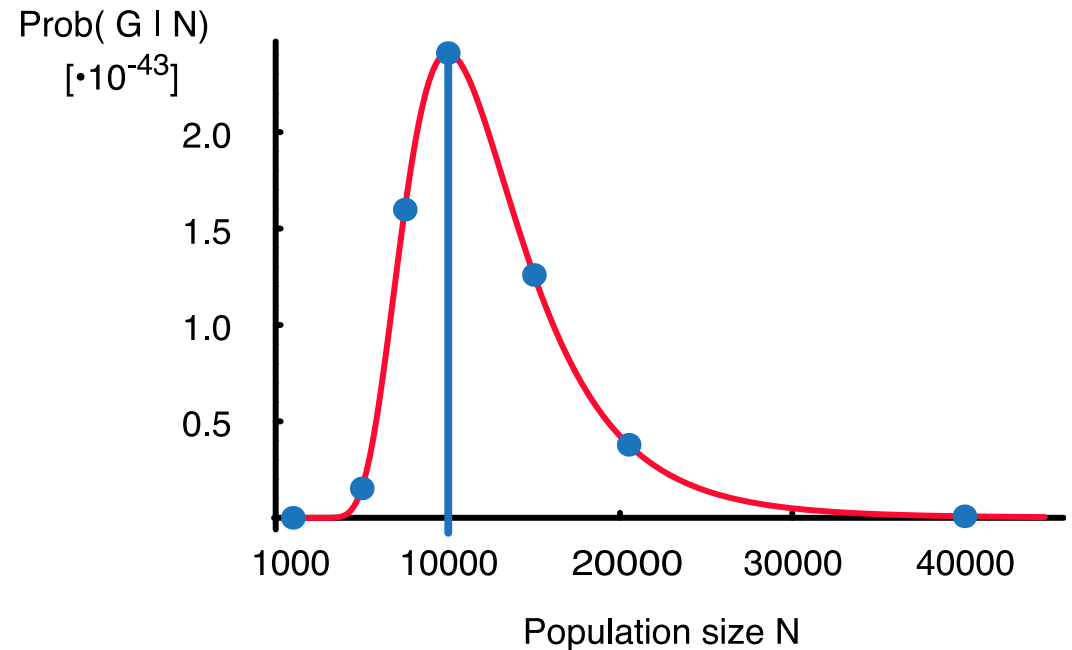
If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



If an oracle gives us the true relationship tree G then we can calculate the population size N .

$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$



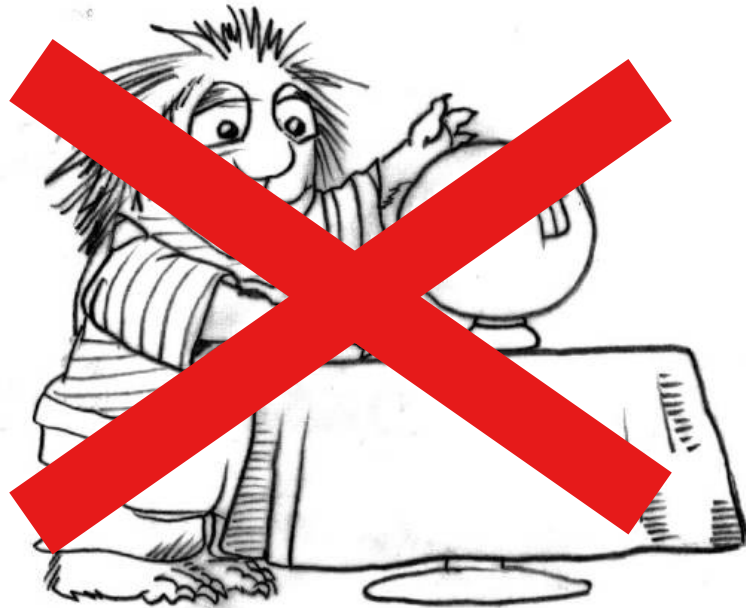
If an oracle gives us the true relationship tree G then we can calculate the population size N .

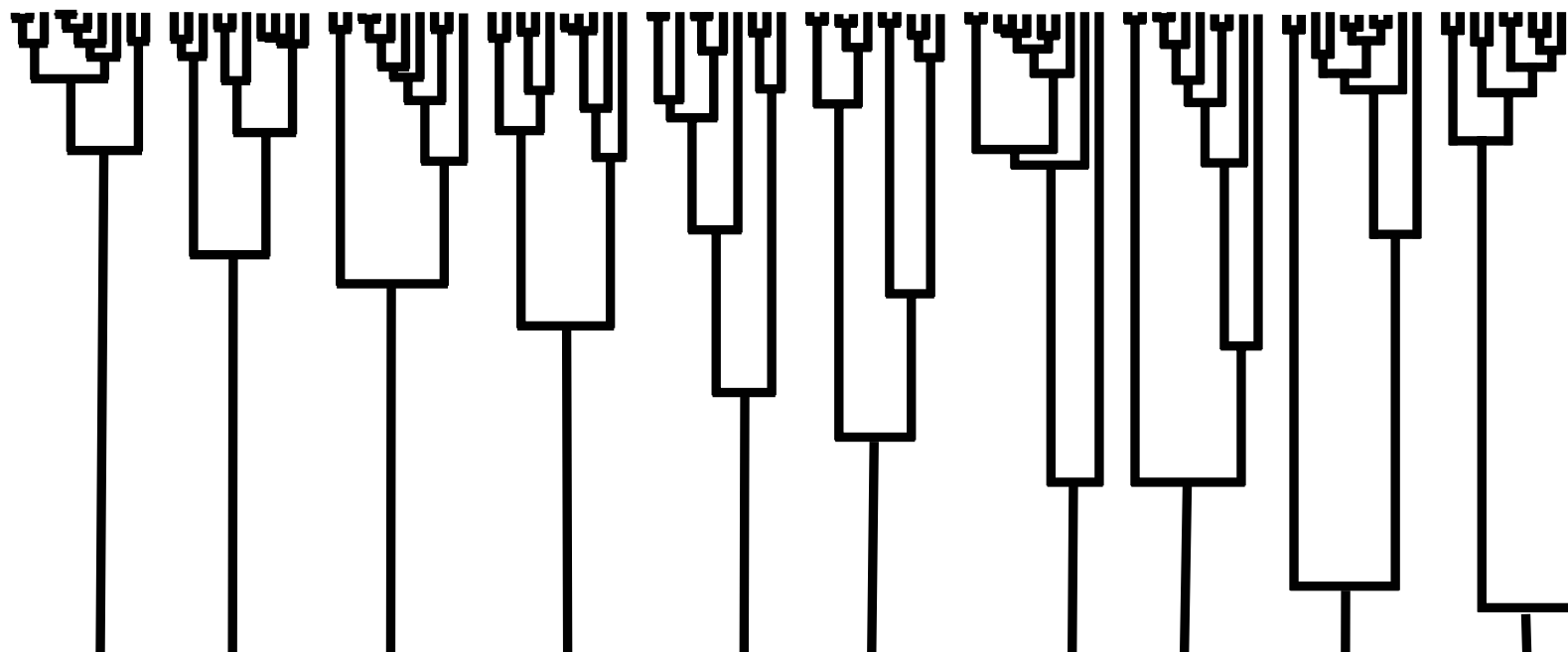
$$p(G|N, n) = \prod_{k=2}^n \exp \left(-u_k \frac{k(k-1)}{4N} \right) \frac{2}{4N}$$

Population size estimation

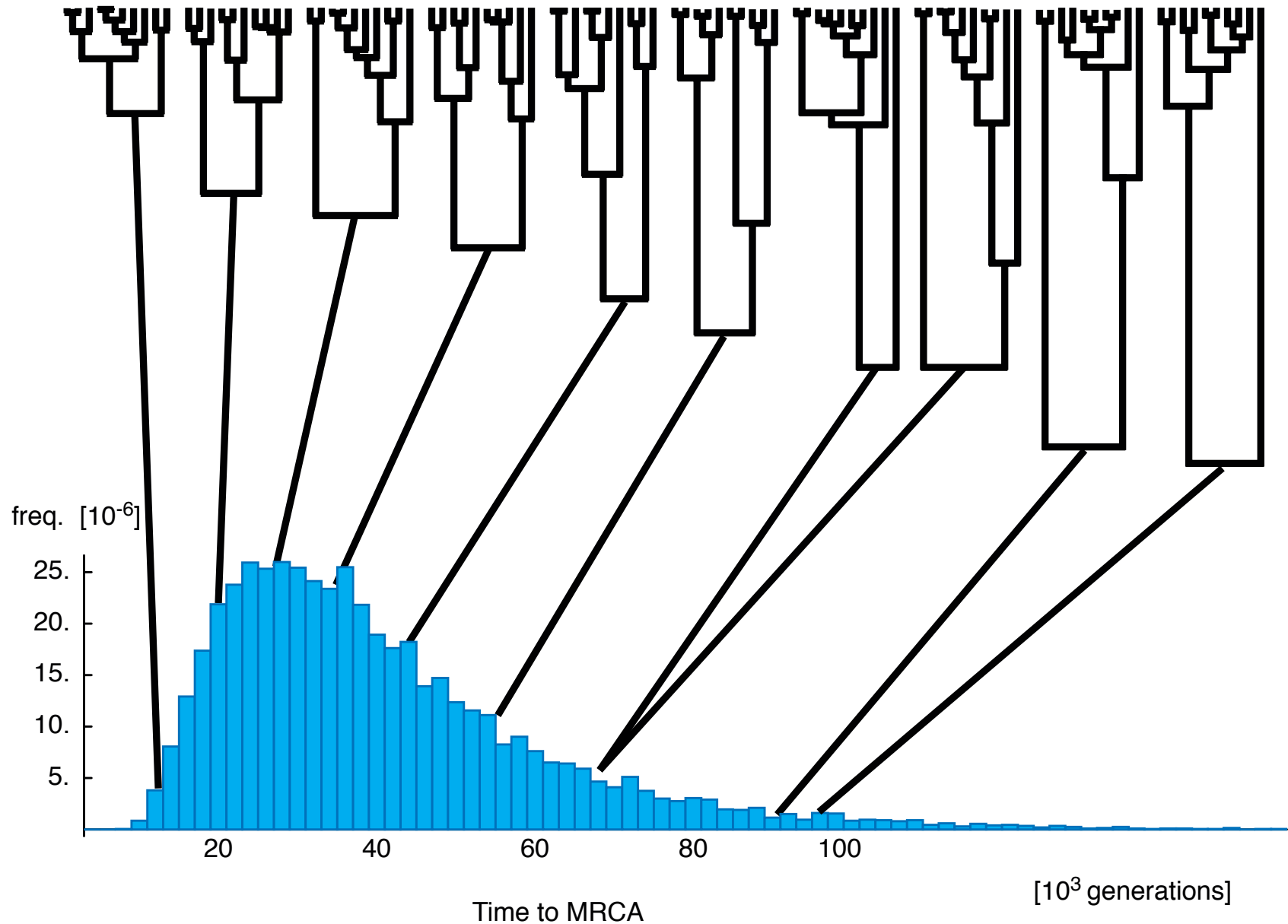
There are at least two problems with the oracle-approach:

- ❖ There is no oracle to gives us clear information!
- ❖ We do not record genealogies, our data are sequences, microsatellite loci!
- ❖ What about the variability of the coalescence process?





All genealogies were simulated with the same population size $N_e = 10,000$



MRCA = most recent common ancestor (last node in the genealogy)

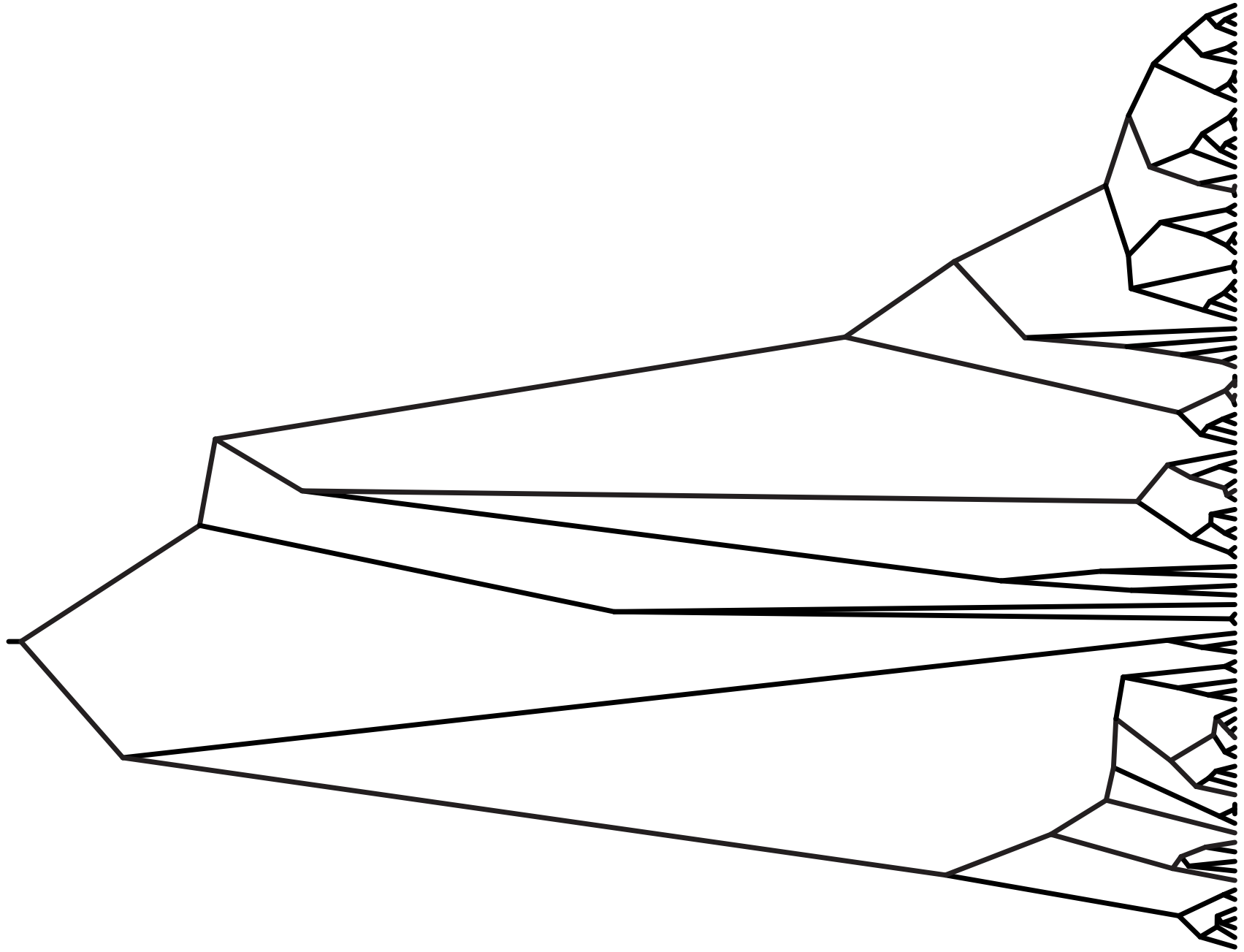
- ◆ All individuals have the same fitness (no selection).
- ◆ All individuals have the same chance to be in the sample (random sampling).
- ◆ The coalescent allows only merging two lineages per generation. This restricts us to have a much smaller sample size than the population size.

$$n \ll N$$

- ◆ Yun-Xin Fu (2005) described the exact coalescent for the Wright-Fisher model and derived a maximal sample size $n < \sqrt{4N}$ for a diploid population. Although this may look like a severe restriction for the use of the coalescence in small populations, it turned out that the coalescence is rather robust and that even sample sizes close to the effective population size are not biasing immensely.

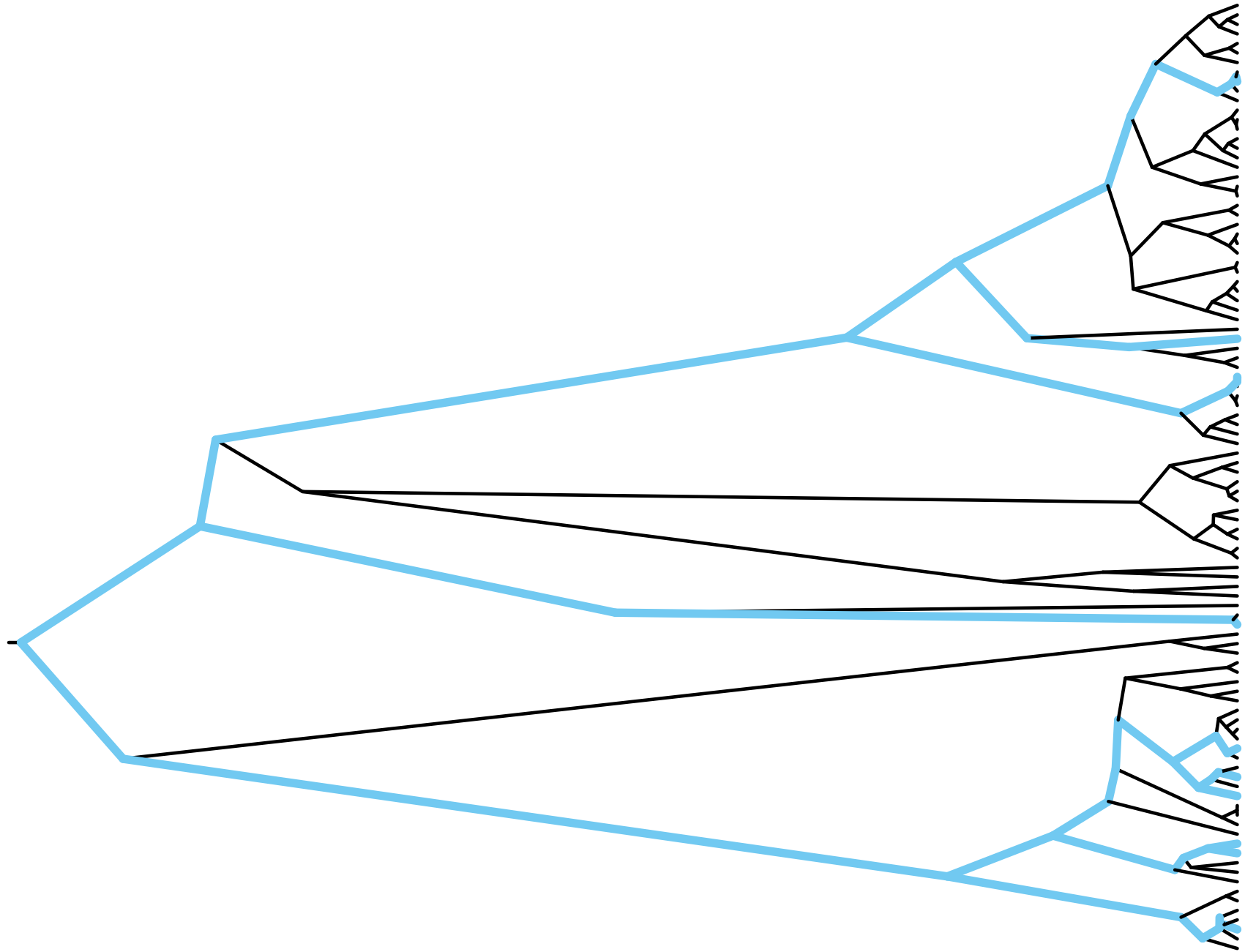
Kingman's n -coalescent is an approximation

Sample size



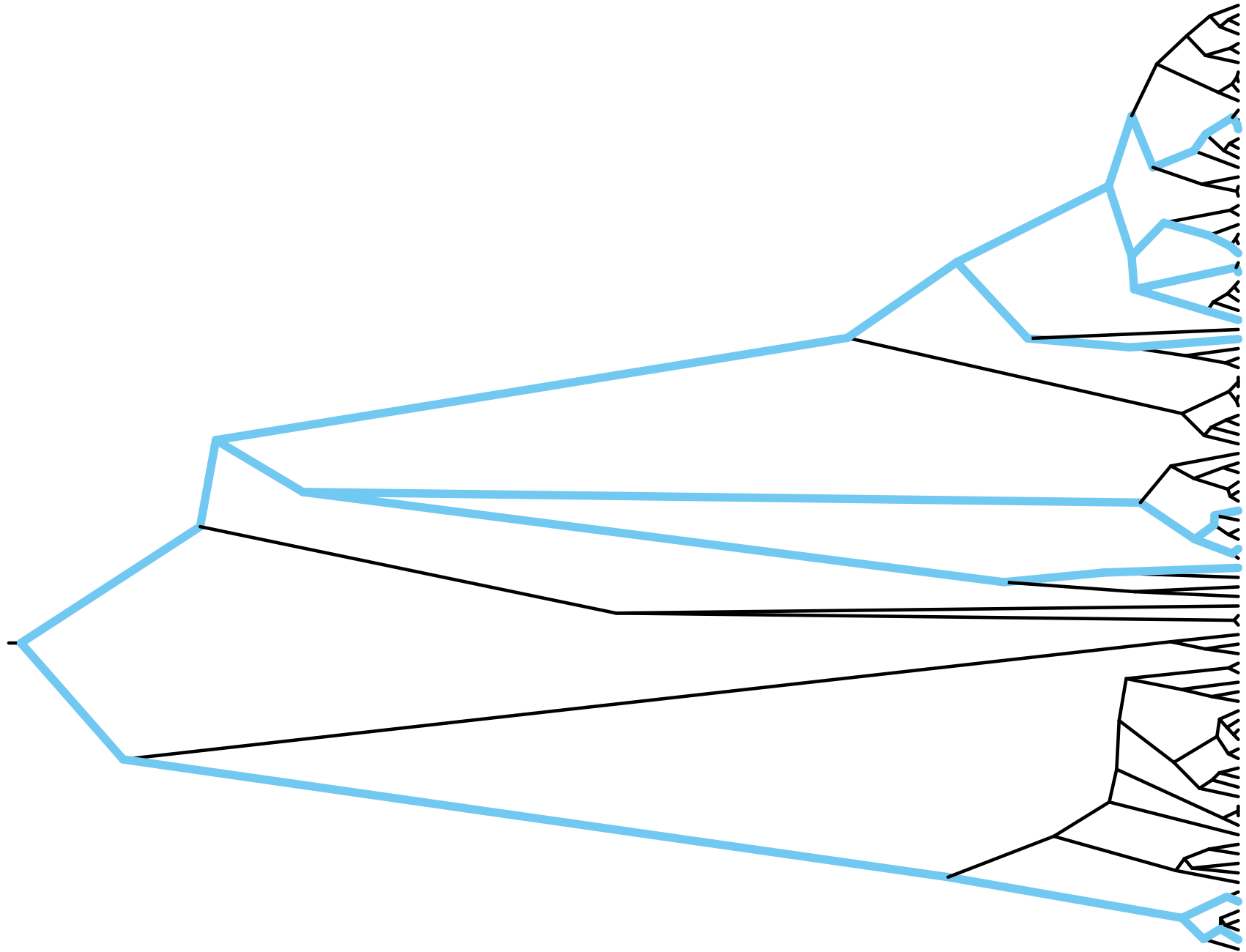
Kingman's n -coalescent is an approximation

Sample size



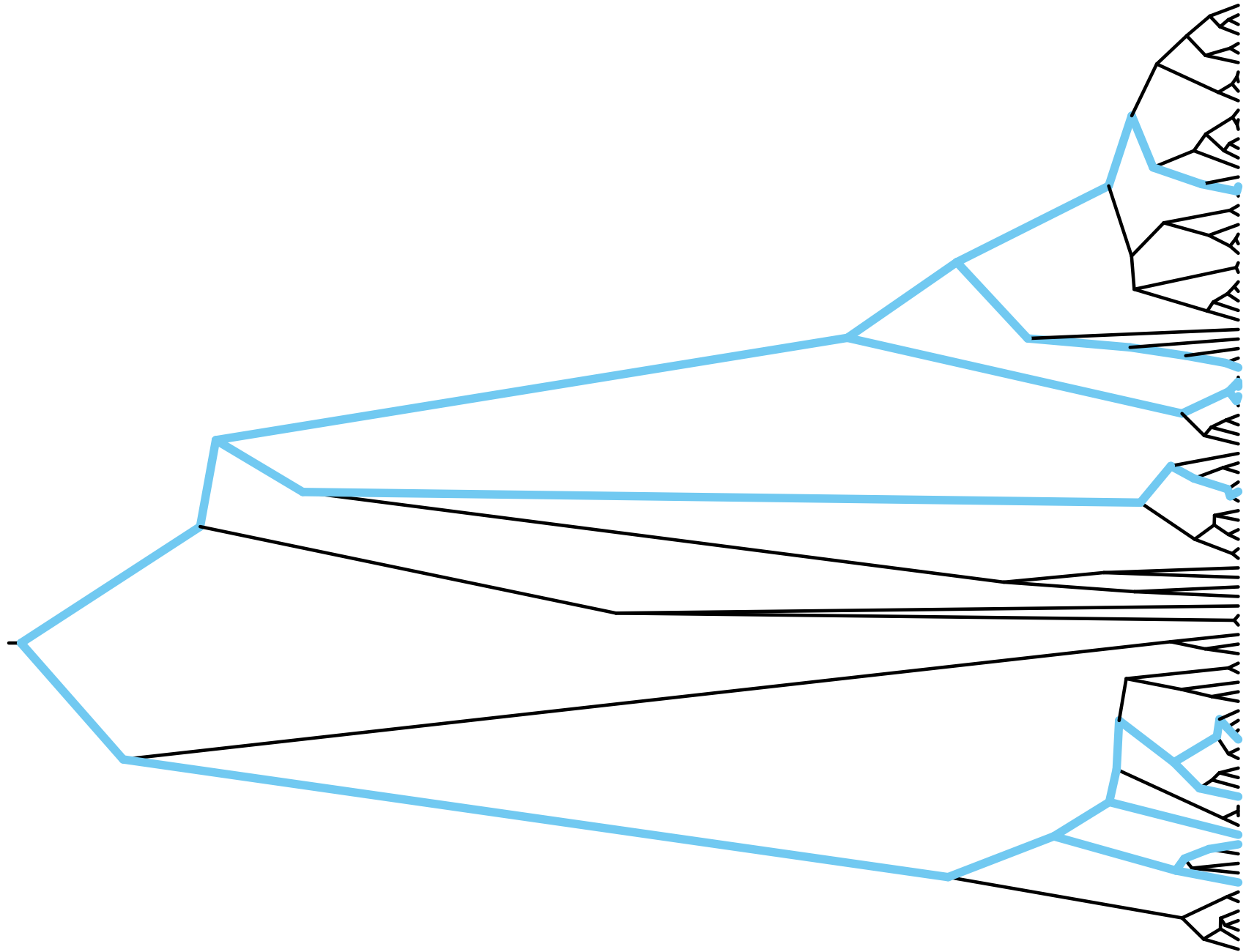
Kingman's n -coalescent is an approximation

Sample size



Kingman's n -coalescent is an approximation

Sample size



- ❖ Large samples coalesce on average in $4N$ generations.
- ❖ The time to the most recent common ancestor (TMRCA) has a large variance
- ❖ Even a sample with few individuals can most often recover the same TMRCA as a large sample.
- ❖ The sample size should be much smaller than the population size, although severe problems appear only with sample sizes of the same magnitude as the population size, or with non-random samples because Kingman's coalescence process assumes that maximally two sample lineages coalesce in any generation.
- ❖ With a known genealogy we can estimate the population size. Unfortunately, the true genealogy of a sample is rarely known.

Genealogy and data

our data looks like this:

[illegible]

Genealogy and data

our data looks like this:

rid1
 rid2
 rid3
 rid4
 rid5
 rid6
 rid7
 rid8
 rid9ty1b
 rid10ty2b
 bed1
 cyp1
 cyp2
 cyp3
 cyp4
 cilwest1
 cilwest2
 cileast1
 cileast2
 cf.car1
 cf.cara2
 cf.cara3
 cf.car4
 cf.cer1
 cf.cer2
 cf.cer3
 cf.cer4
 cf.bed1
 cf.bed2
 cf.bed3
 cf.bed4
 cf.bed5
 cf.bed6
 cf.bed7
 cf.bed8
 epe6-GR
 epe7-GR
 cre04a-GR
 cre5-GR

Finding the best genealogy from such data is difficult

Genetic data and the coalescent

- ◆ Finite populations lose alleles due to genetic drift
- ◆ Mutation introduces new alleles into a population at rate μ
- ◆ With $2N$ chromosomes we can expect to see every generation $2N\mu$ new mutations. The population size N is positively correlated with the mutation rate μ .
- ◆ With genetic data sampled from several individuals we can use the mutational variability to estimate the population size.

Population size

The observed genetic variability

$$\mathcal{S} = f(N, \mu, n).$$

Different N and appropriate μ can give the same number of mutations. For example, for 100 loci sampled from 20 individuals with 1000bp each, we get :

N	μ	$4N\mu$	$\hat{S}$	σ_S^2
1250	10^{-5}	0.05	153.95	16.25
12500	10^{-6}	0.05	152.89	16.05

Using genetic variability alone therefore **does not allow** to disentangle N and μ .

With **multiple dated samples** and known generation time we **can** estimate N and μ independently.

Mutation-scaled population size

By convention we express most results as the compound $N\mu$ and an inheritance scalar x , for simplicity we call this the **mutation-scaled population size**

$$\Theta = xN\mu,$$

where μ is the mutation rate per generation and per site. With a mutation rate per locus we use θ .

◆ for diploids: $\Theta = 4N\mu$.

◆ for haploids: $\Theta = 2N\mu$.

◆ For mtDNA in diploids with strictly maternal inheritance this leads to $\Theta = 2N_f\mu$, and if the sex ratio is 1 : 1 then $\Theta = N\mu$

Most real populations do not behave exactly like Wright-Fisher populations, therefore we subscript N and call it the **effective** population size N_e , and consider Θ the **mutation-scaled EFFECTIVE population size**.

Mutation-scaled population size

By convention we express most results as the compound $N\mu$ and an inheritance scalar x , for simplicity we call this the **mutation-scaled population size**

$$\Theta = xN\mu,$$

where μ is the mutation rate per generation and per site. With a mutation rate per locus we use θ .

◆ for diploids: $\Theta = 4N\mu$.

◆ for haploids: $\Theta = 2N\mu$.

◆ For mtDNA in diploids with strictly maternal inheritance this leads to $\Theta = 2N_f\mu$, and if the sex ratio is 1 : 1 then $\Theta = N\mu$



Gag Grouper starts out as a female and later in life becomes male.

Most real populations do not behave exactly like Wright-Fisher populations, therefore we subscript N and call it the **effective** population size N_e , and consider Θ the **mutation-scaled EFFECTIVE population size**.

Historical humpback whale population size

Humpback whales in the North Atlantic: Census population size around 12,000.



Historical humpback whale population size

using the data by Joe Roman and Stephen R. Palumbi (Science 2003 301: 508-510)

$\Theta = 2N_{\text{♀}}\mu$ 0.01529 Population size of the North Atlantic population, estimated using migrate

$N_{\text{♀}} = \frac{\Theta}{2\mu}$ 31,854 with $\mu = 2.0 \times 10^{-8} \text{bp}^{-1} \text{year}^{-1}$ and a generation time of 12 years

$N_e = N_{\text{♀}} + N_{\text{♂}}$ 63,708 Sex ratio is 1:1

$N_B = 2N_e$ 127,417 ratio N_B/N_e assumed, using other data

$N_T = N_B \frac{N_{\text{juveniles}} + N_{\text{adults}}}{N_{\text{adults}}}$ 203,867 from catch and survey data (used a ratio of 1.6)

Using the infinite sites model we use the number of variable sites S per locus to calculate the mutation-scaled population size:

$$\theta_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

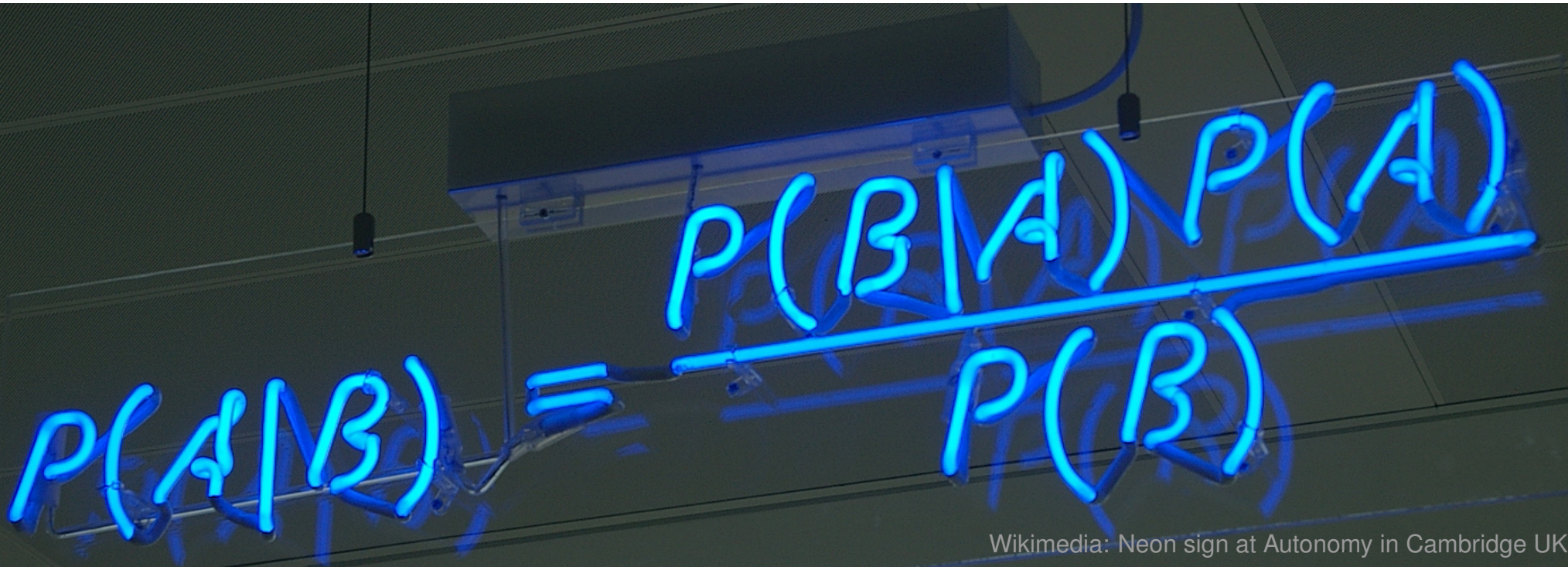
from a sample of n individuals. For a single population the Watterson's estimator works marvelously well, but it is vulnerable to population structure.

Watterson's θ_W uses a mutation rate per locus! To compare with other work use mutation rate per site.

For Bayesian inference we want to calculate the probability of the model parameters given the data $p(\text{model}|\mathcal{D})$.

Coalescent to describe the population genetic processes.

Mutation model to describe the change of genetic material over time.



We calculate the **Posterior distribution** $p(\Theta|D)$ using Bayes' rule

$$p(\Theta|D) = \frac{p(\Theta)p(D|\Theta)}{p(D)}$$

where $p(D|\Theta)$ is the **likelihood** of the parameters.



$$p(D|\Theta, G) = p(G|\Theta)p(D|G)$$

$$p(G|\Theta)$$



The probability of a genealogy given parameters.

$$p(D|G)$$



The probability of the data for a given genealogy.
Phylogeneticists know this as the tree-likelihood.

$$p(D|\Theta) = \int_G p(G|\Theta)p(D|G)dG$$

$p(G|\Theta)$



The probability of a genealogy given parameters.

$p(D|G)$



The probability of the data for a given genealogy.
Phylogeneticists know this as the tree-likelihood.

$$p(D|\Theta) = \sum_G p(G|\Theta)p(D|G)$$

$p(G|\Theta)$



The probability of a genealogy given parameters.

$p(D|G)$

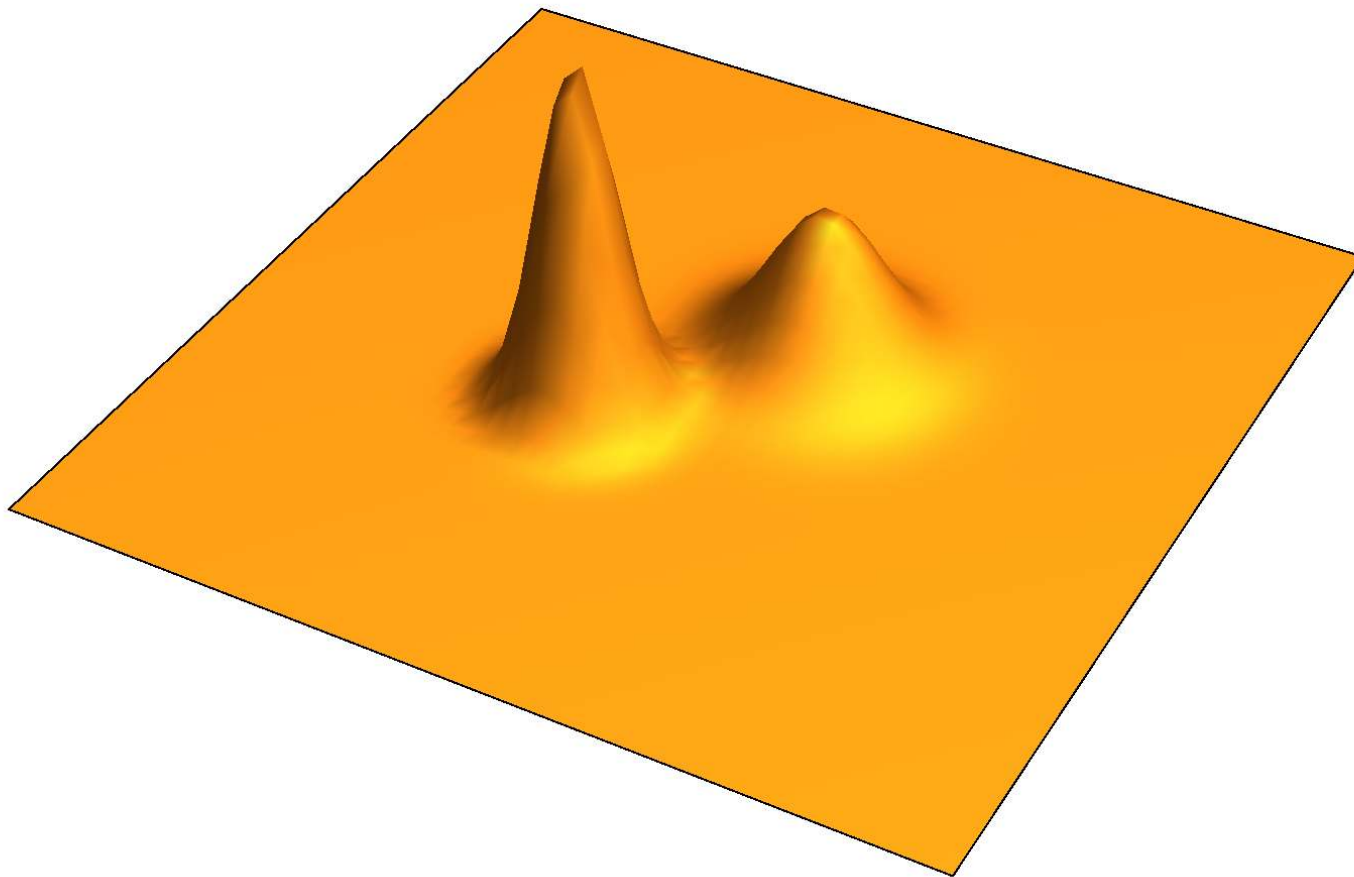


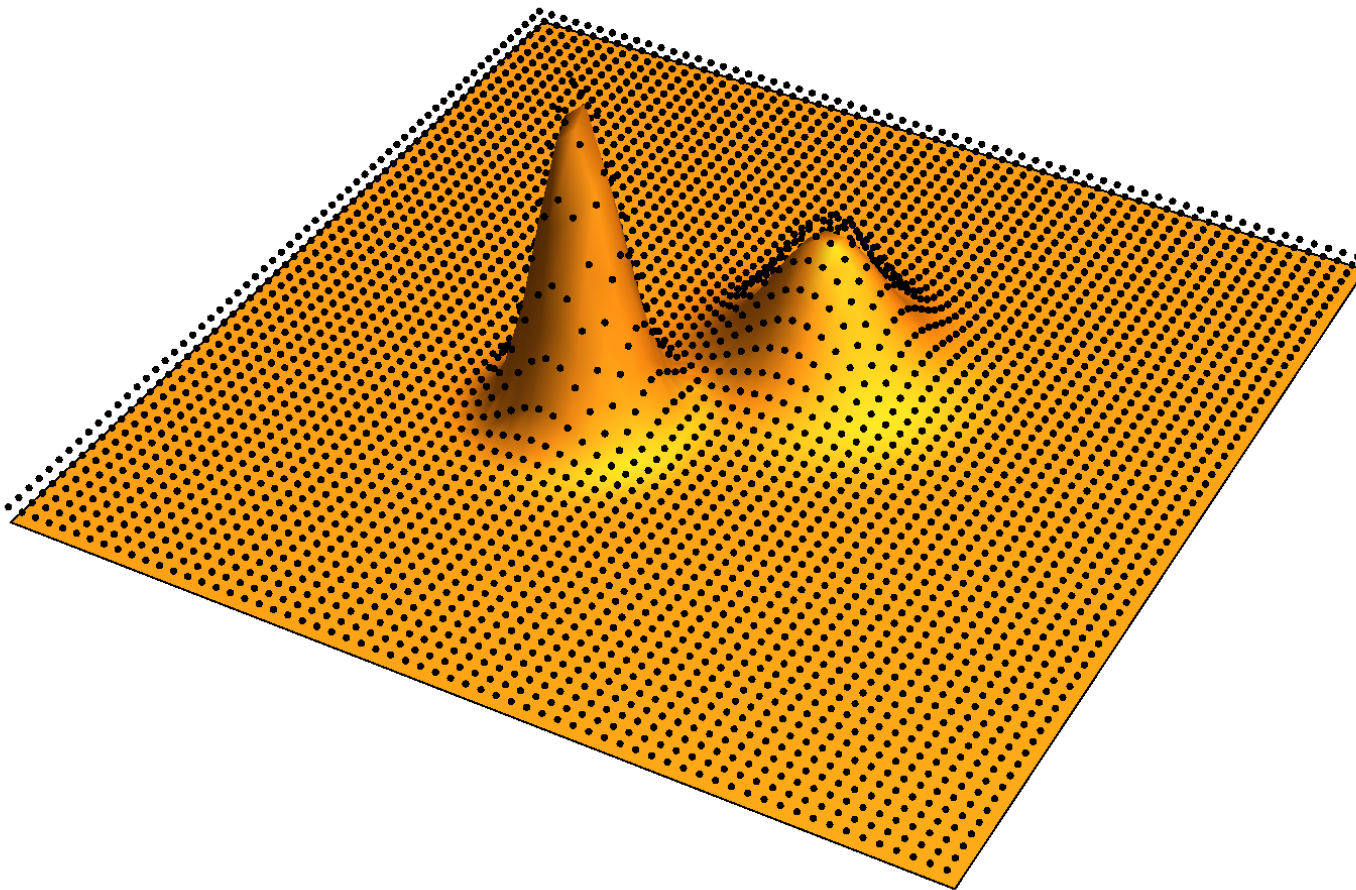
The probability of the data for a given genealogy.
Phylogeneticists know this as the tree-likelihood.

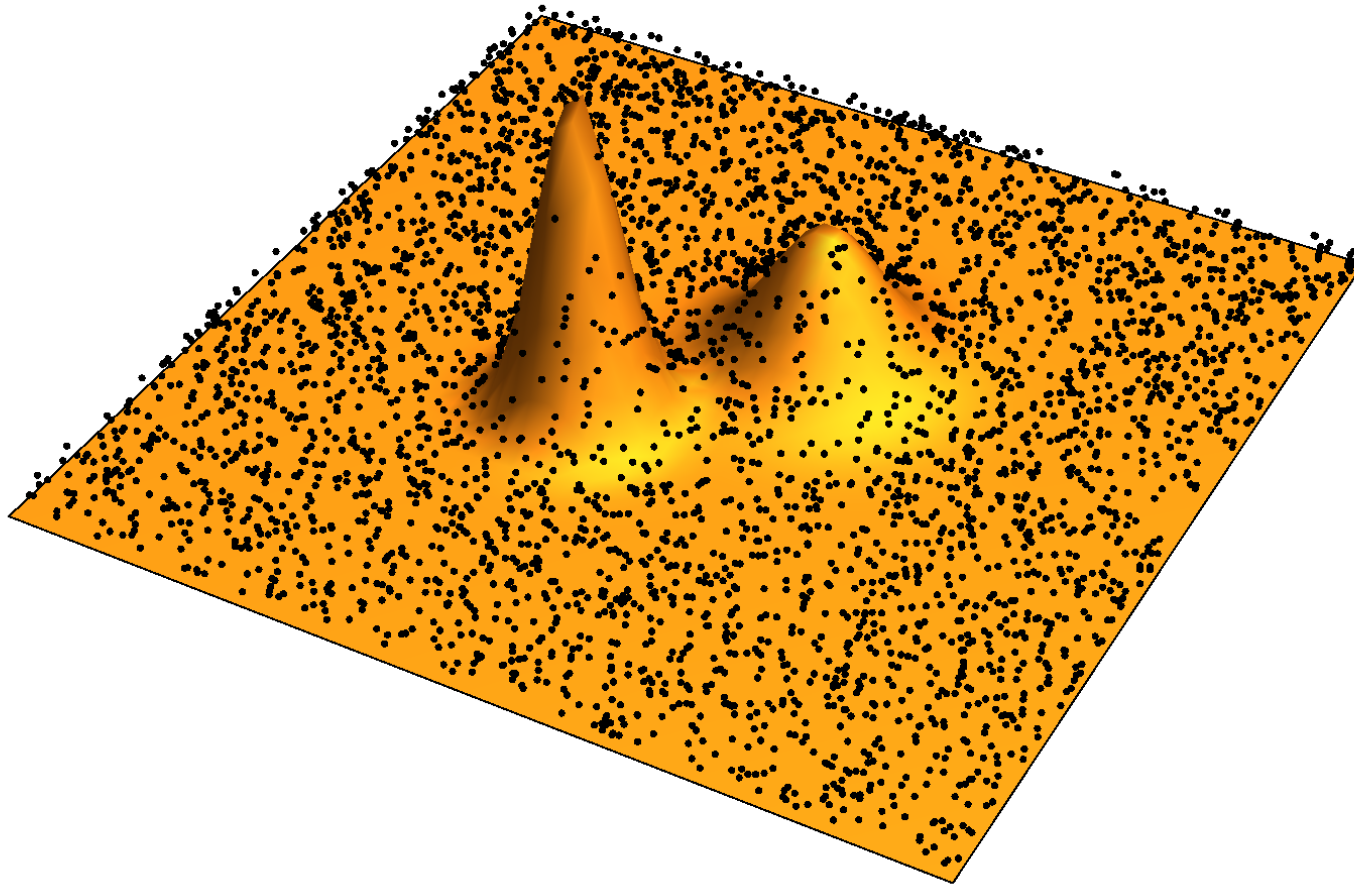
Problem with integration formula

$$p(D|\Theta) = \int_G p(G|\Theta)p(D|G)dG$$

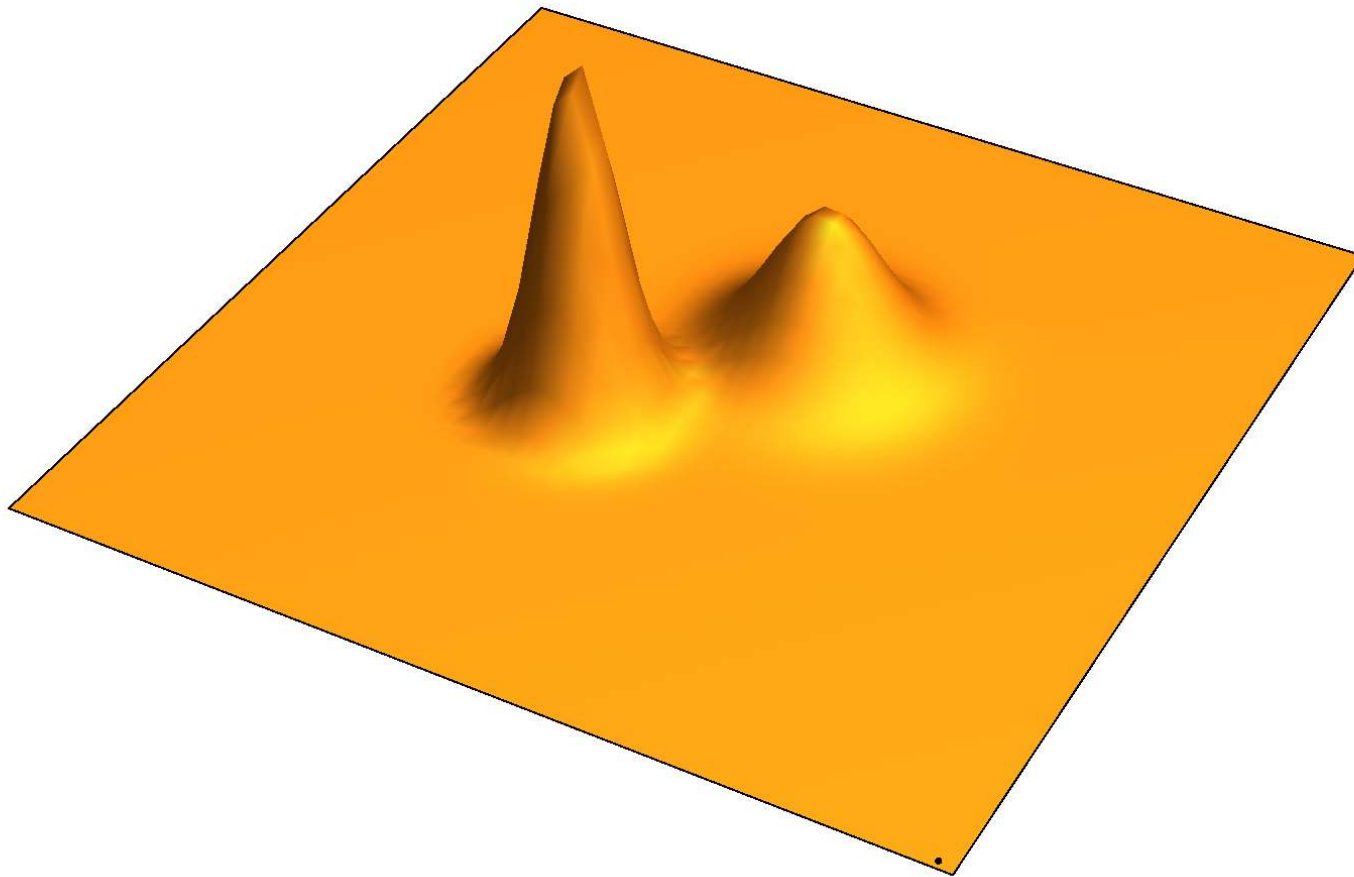
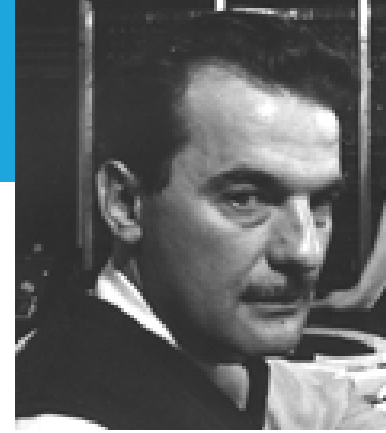
The number of possible genealogies is very large and for realistic data sets, programs need to use Markov chain Monte Carlo methods.



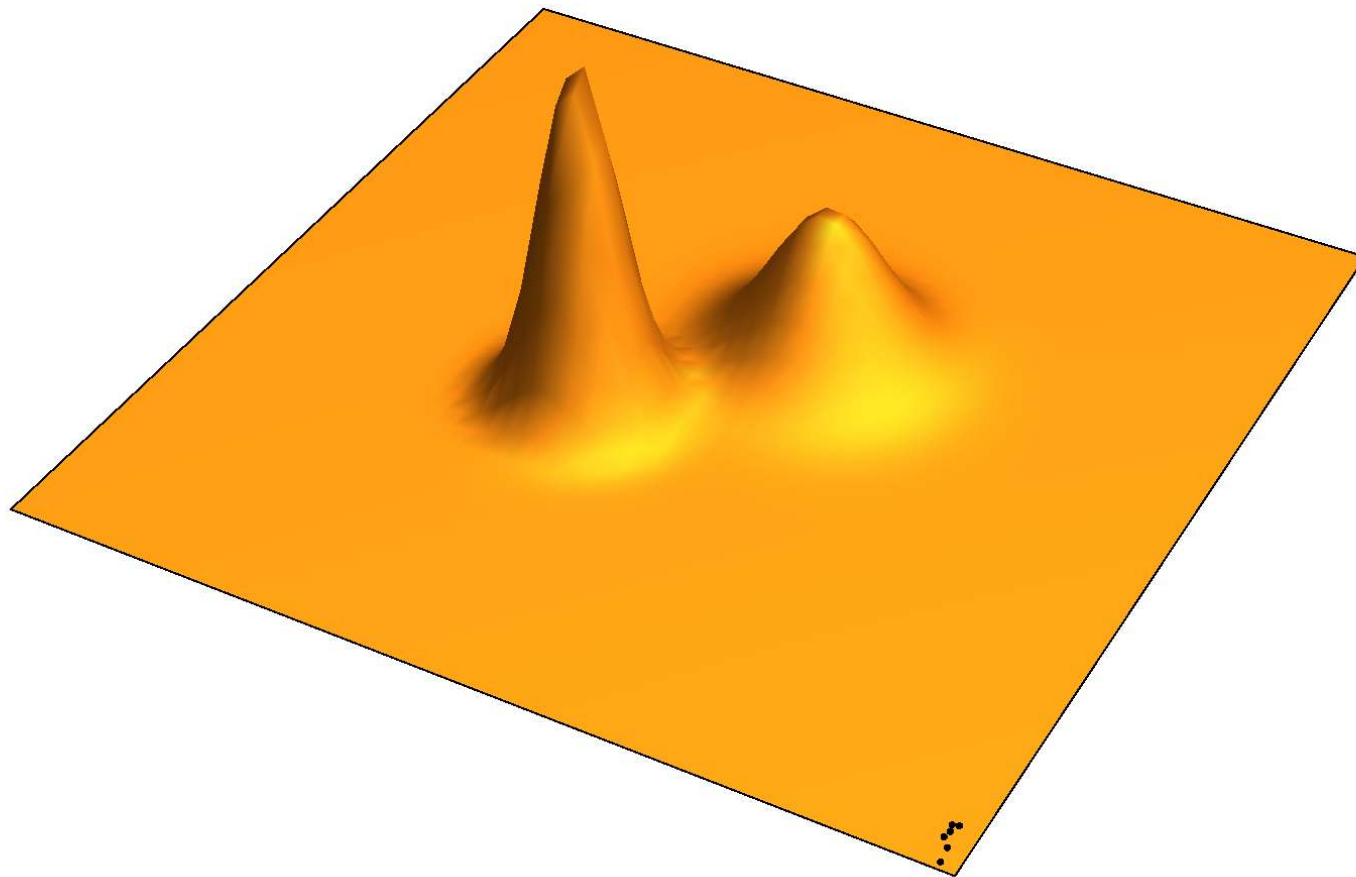
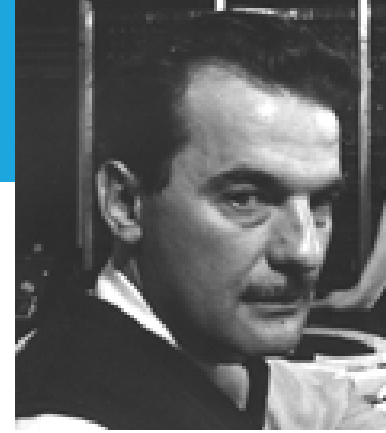




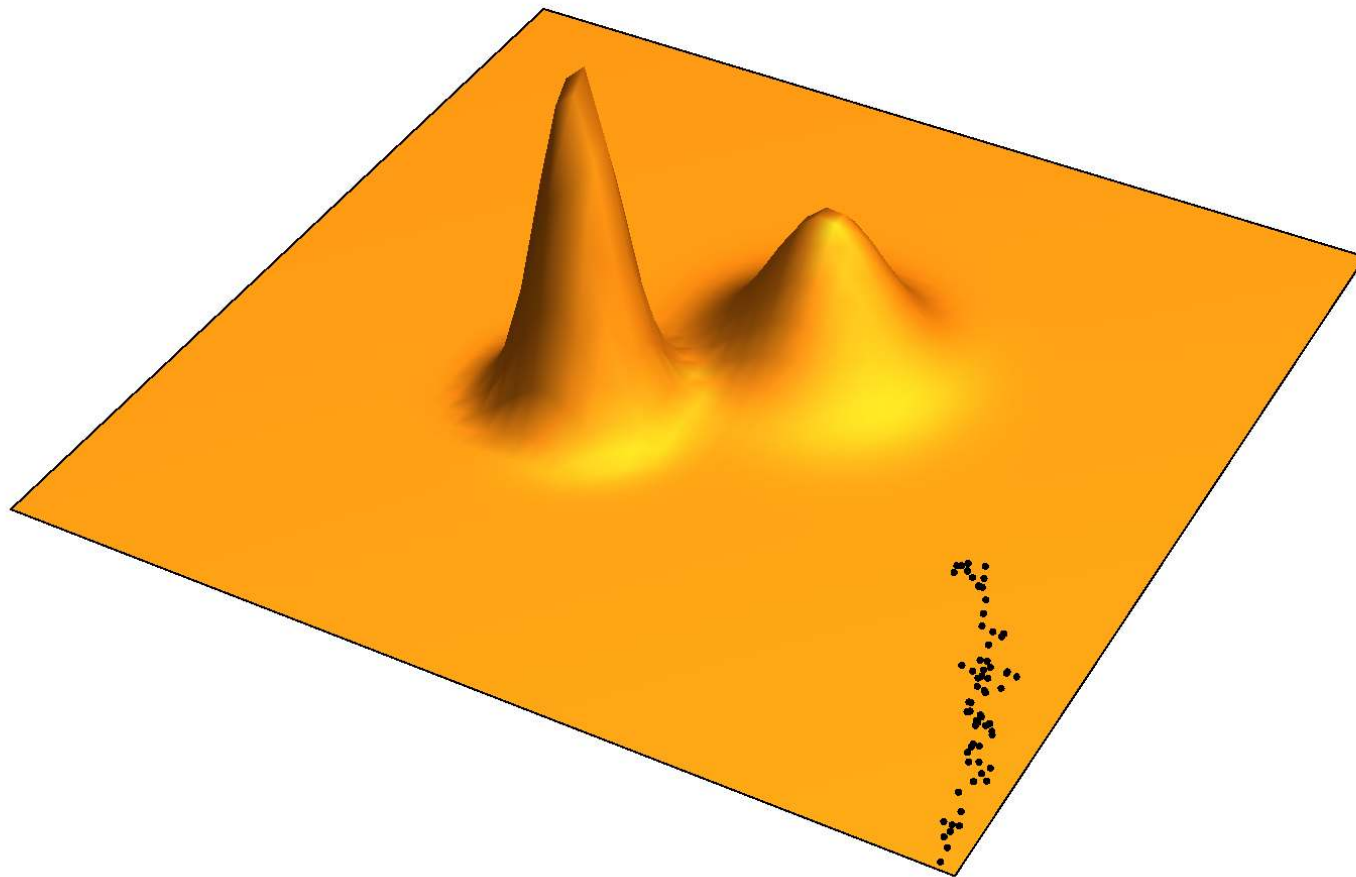
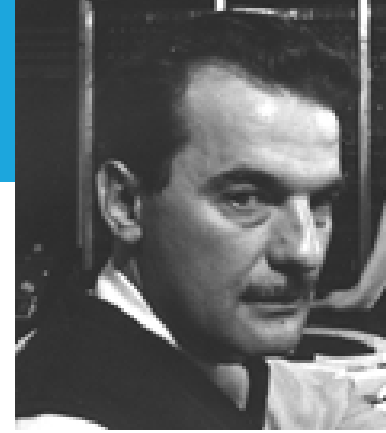
Metropolis-Hastings algorithm



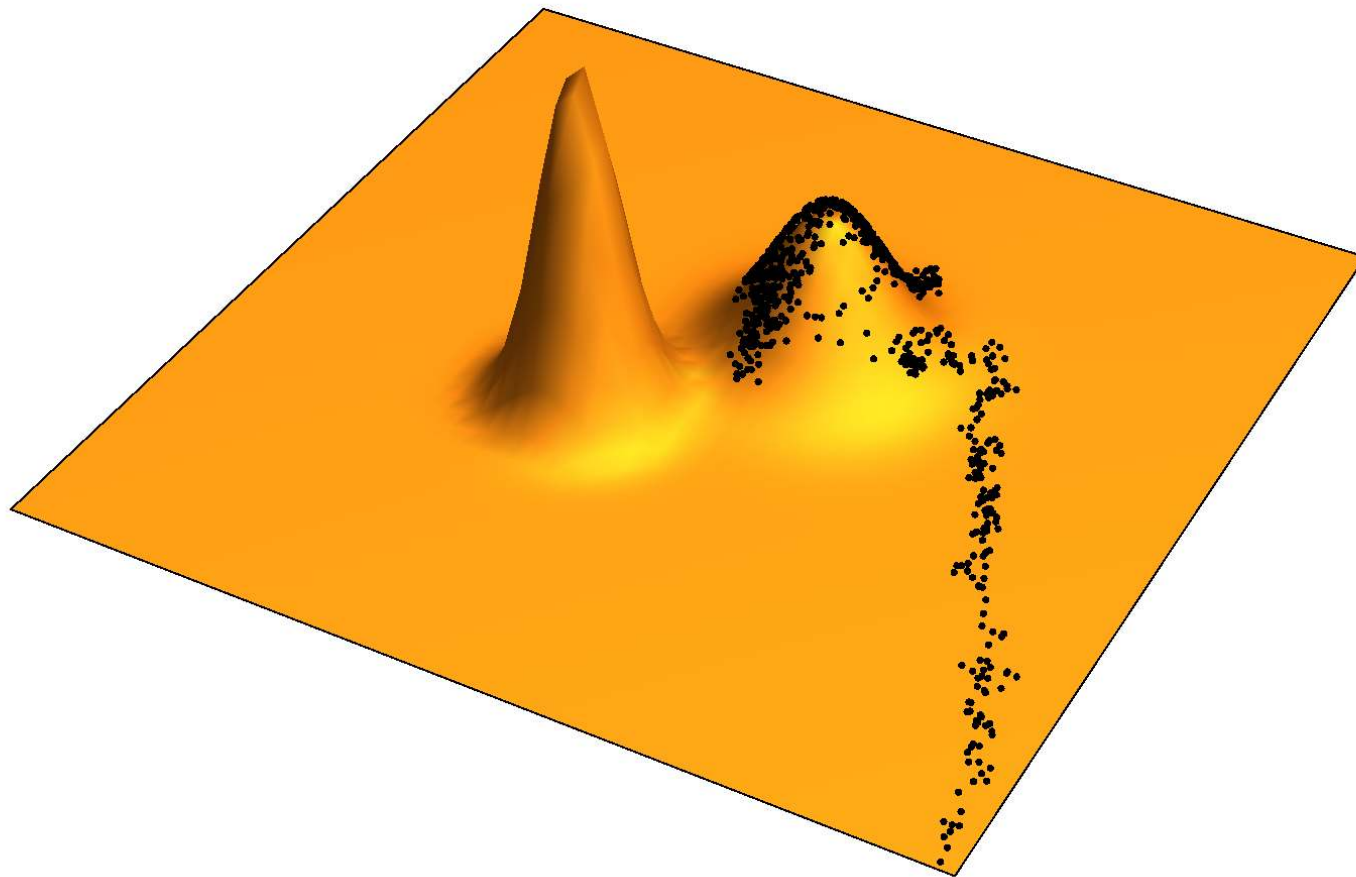
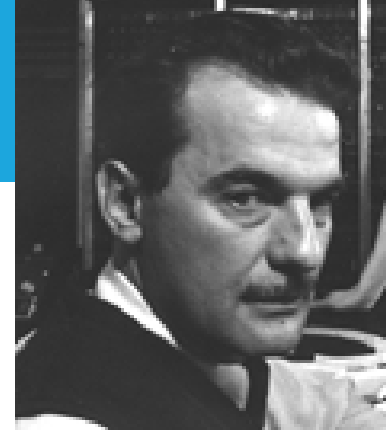
Metropolis-Hastings algorithm



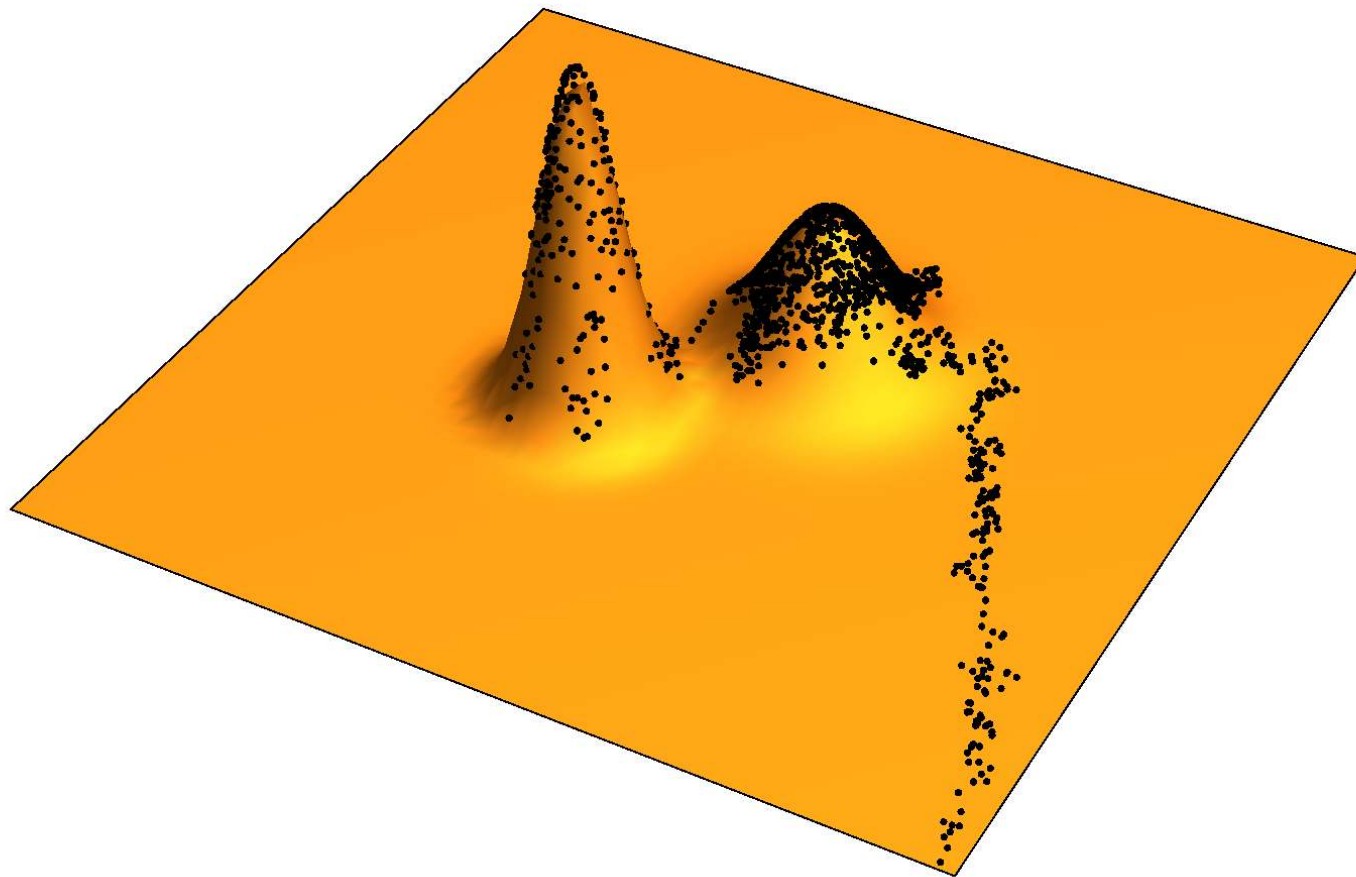
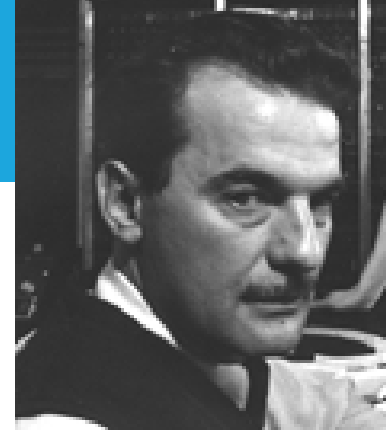
Metropolis-Hastings algorithm



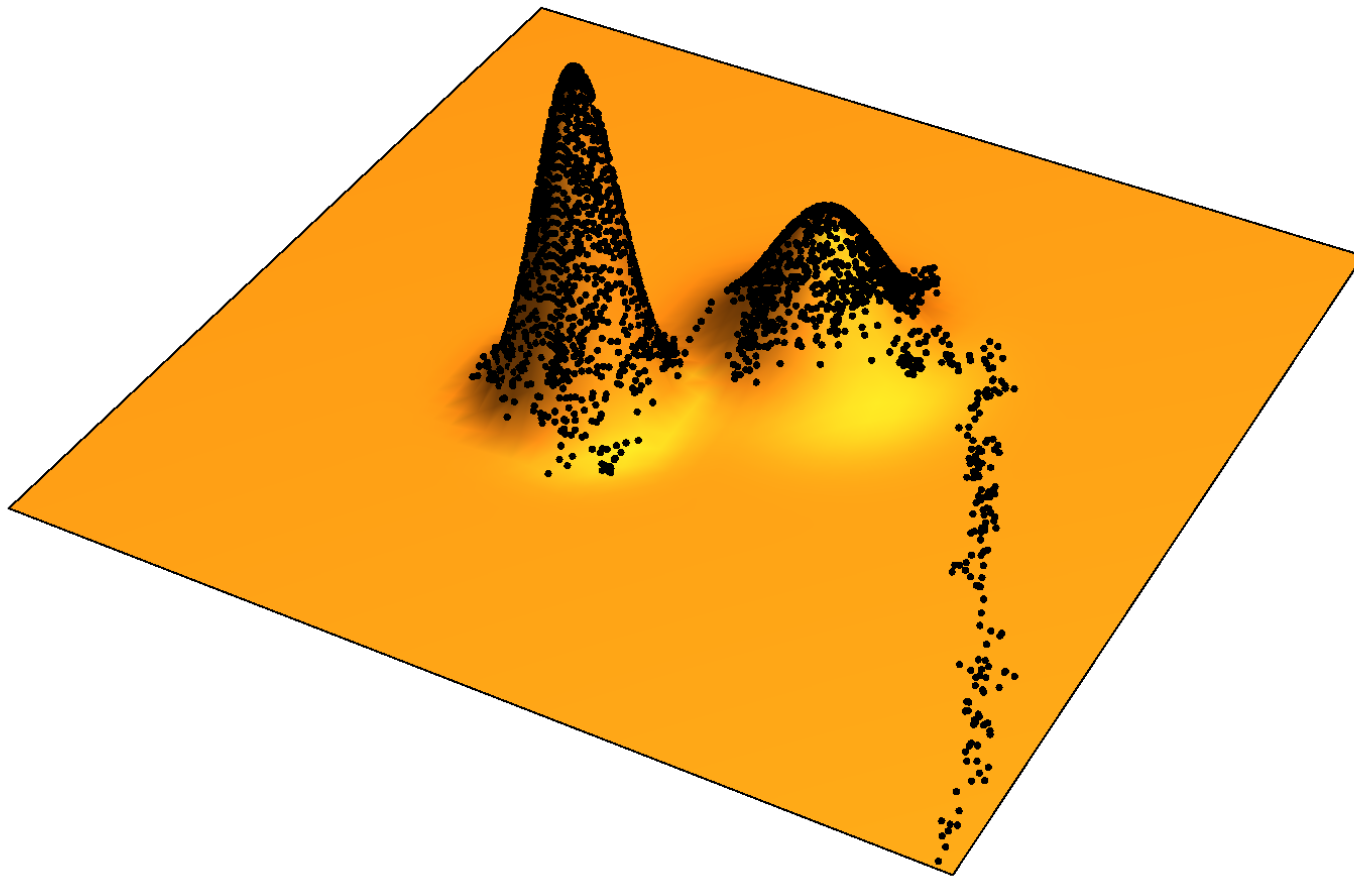
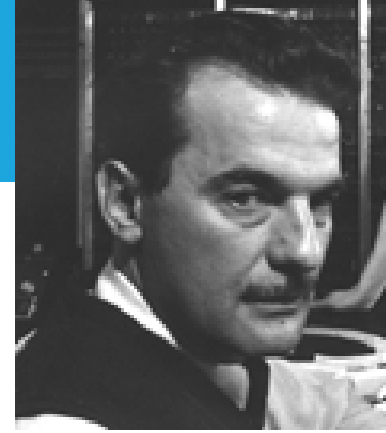
Metropolis-Hastings algorithm



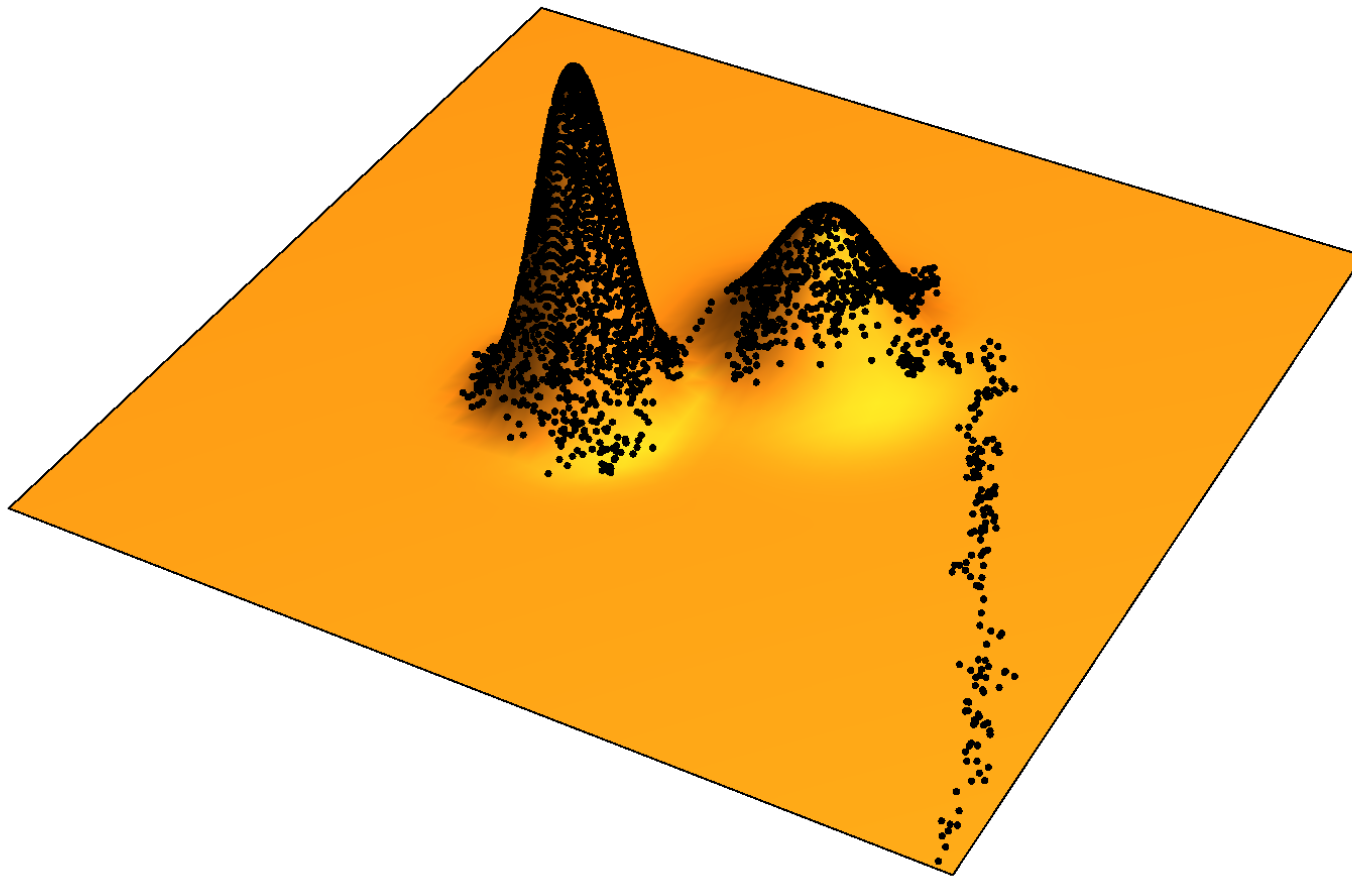
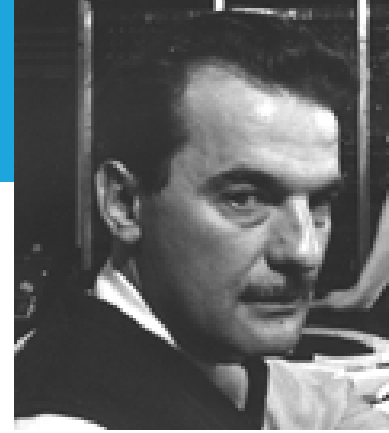
Metropolis-Hastings algorithm



Metropolis-Hastings algorithm



Metropolis-Hastings algorithm



Inference of population size

Nuu-Chah-Nulth



Around 1930 – Friendly Cove, Vancouver Island



Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 8720–8724, October 1991
Evolution

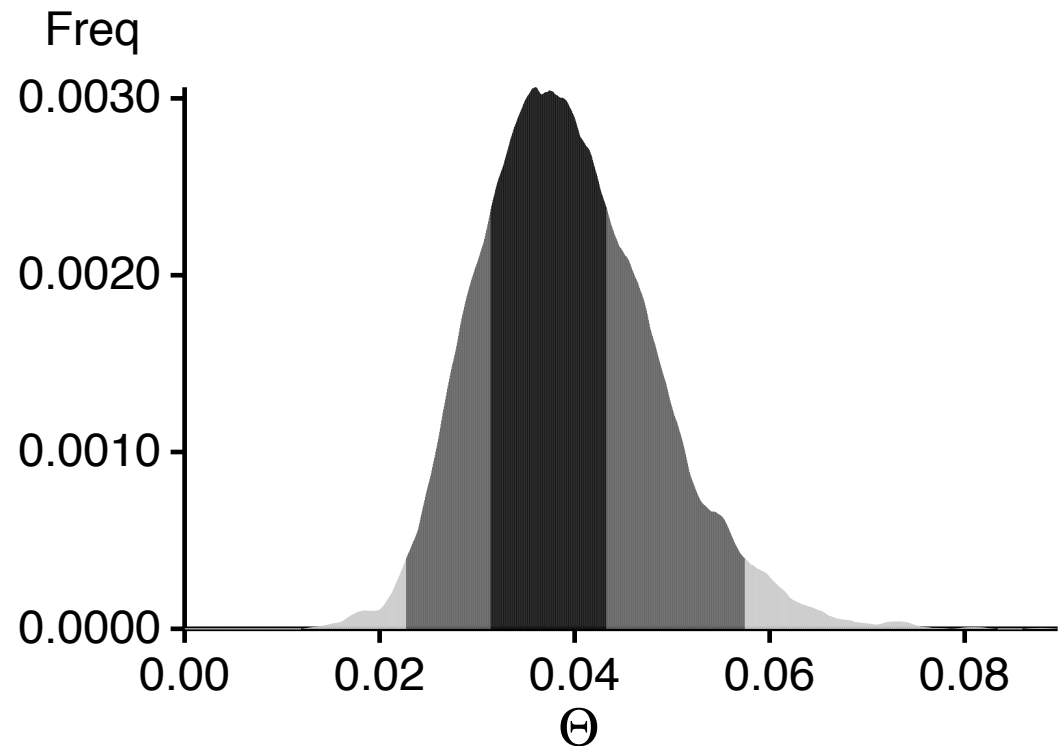
Extensive mitochondrial diversity within a single Amerindian tribe

(population genetics/molecular anthropology/Pacific Northwest/human evolution)

R. H. WARD*, BARBARA L. FRAZIER*, KERRY DEW-JAGER*, AND SVANTE PÄÄBO†

*Department of Human Genetics, School of Medicine, University of Utah, Salt Lake City, UT 84132; and †Department of Zoology, University of Munich, Luisenstrasse 14, D-8000 Munich 2, Federal Republic of Germany

[The Nuu-Cha-Nulth are organized in 14 nations totaling 8147 (Nuuchahnulth tribal council Indian registry from February 2006)]



Bayesian inference: $\Theta = 0.036$

Ward *et al* calculated $\Theta_{Ewens} = 0.043$

With a mutation rate of 0.32/site/million year and a generation time of 27 years we get $N_{\text{females}} = 2082$. Assuming same numbers of men and women and on average 2 children we get $N = 8328$.

References

Coalescent:

Nuu-Cha-Nulth population size: J. Felsenstein. 1971. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* 68:581-597; R. H. Ward, B. L. Frazier, Kerry Dew-Jager, and S. Pääbo. 1991. Extensive mitochondrial diversity within a single Amerindian tribe. *PNAS* 88:8780-8724; Sigurðardóttir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P. 2000. The mutation rate in the human mtDNA control region. *Am J Hum Genet.* 66:1599-609; S. Matsumura and P. Forster. 2008. Generation time and effective population size in Polar Eskimos. *Proc. R. Soc. B* 275:1501-1508.

Sample size: Felsenstein, J.2005. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *MBE* 23: 691-700. Pluzhnikov A, Donnelly P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144: 1247-1262.