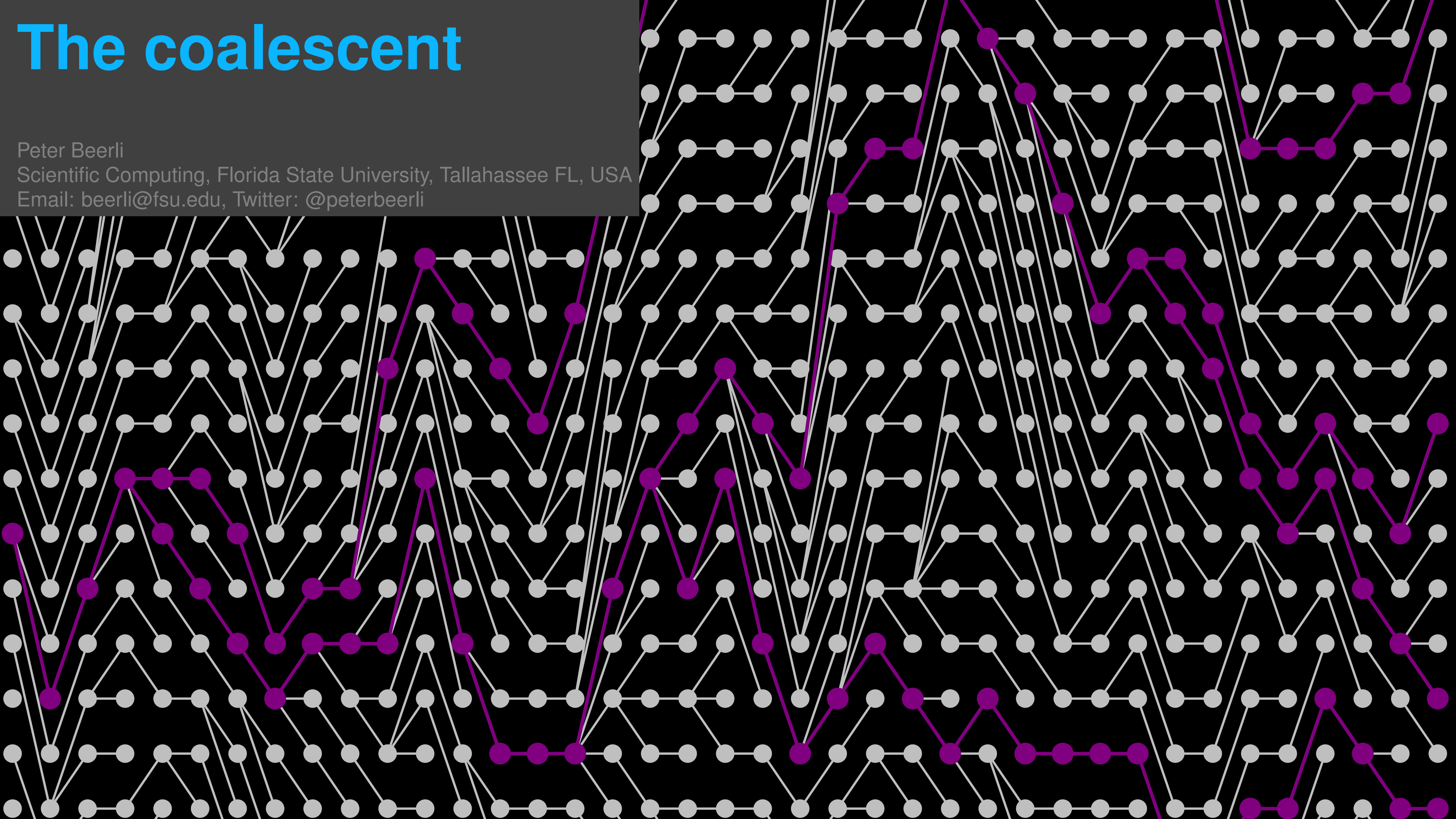
The coalescent

Peter Beerli
Scientific Computing, Florida State University, Tallahassee FL, USA
Email: beerli@fsu.edu, Twitter: @peterbeerli



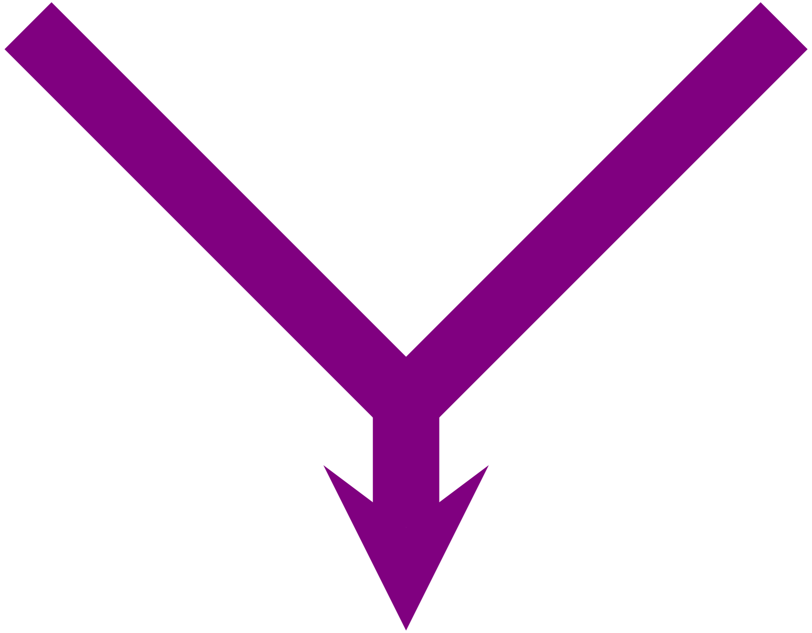
The Coalescent ???????

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.*



Unraveling the mystery: Coalescence theory

Summary:

- To understand biological processes we need models
- Population genetics is the discipline that links natural processes with mathematical understanding
- Coalescence theory is a probabilistic model that explains population genetic processes

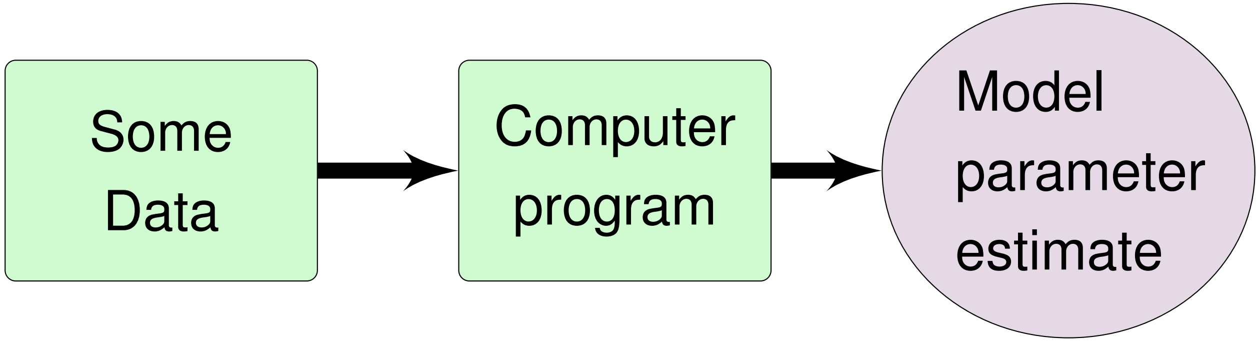
Unraveling the mystery: Coalescence theory

Summary:

- To understand biological processes we need models
  - Data and models
- Population genetics is the discipline that links natural processes with mathematical understanding
  - Population model [Wright-Fisher population]
- Coalescence theory is a probabilistic model that explains population genetic processes
  - The coalescent in detail
  - An example how we would use the coalescent for inference

Data and models

Practical Biologists:



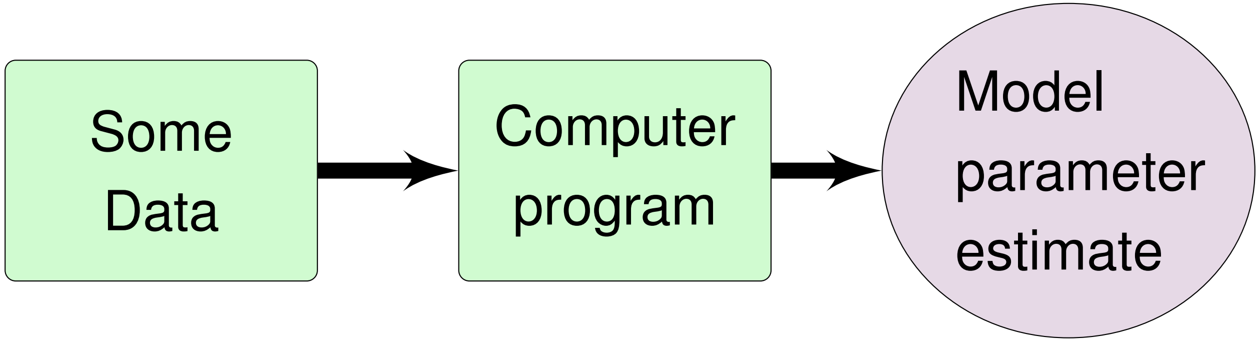
Some Data

Computer program

Model parameter estimate

Data and models

Practical Biologists:

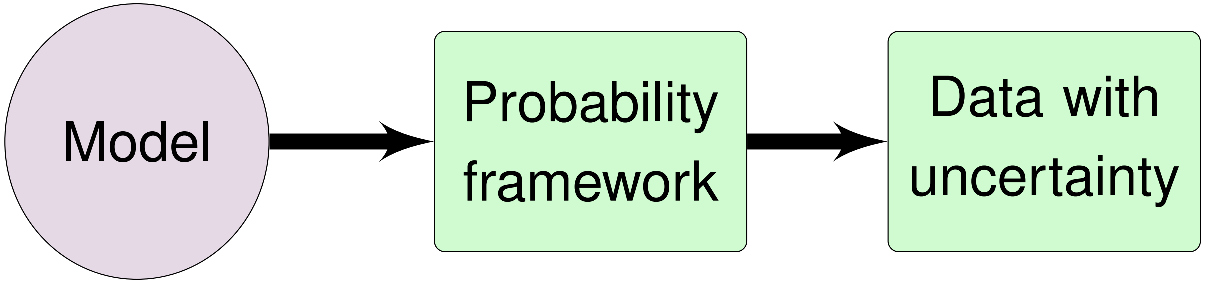


Some Data

Computer program

Model parameter estimate

Theorist:



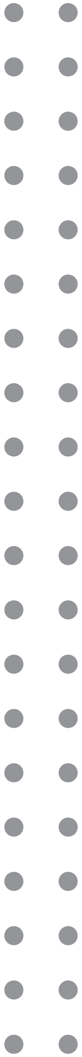
Model

Probability framework

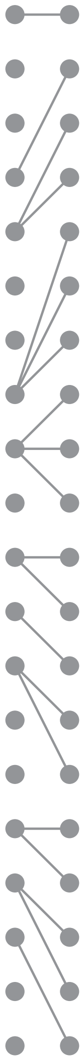
Data with uncertainty

Population model

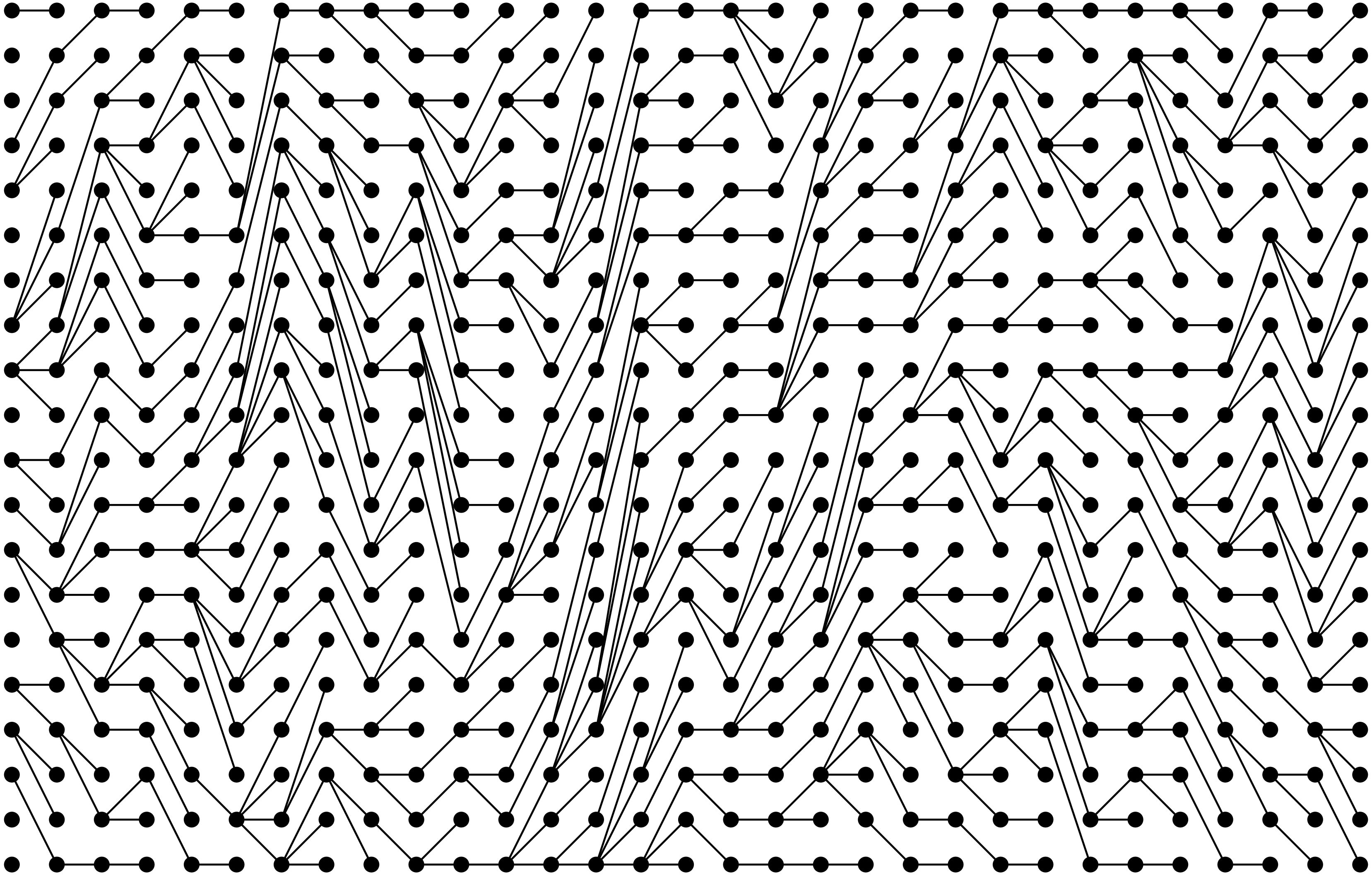
Population model



Population model

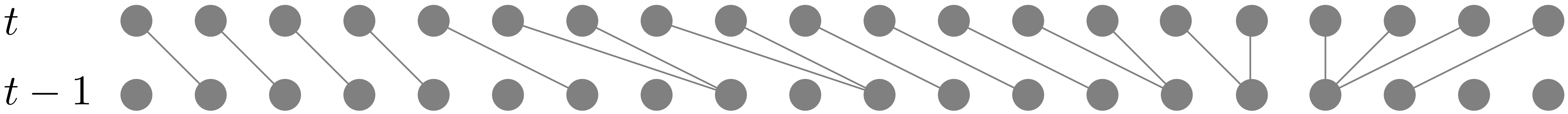


Population model



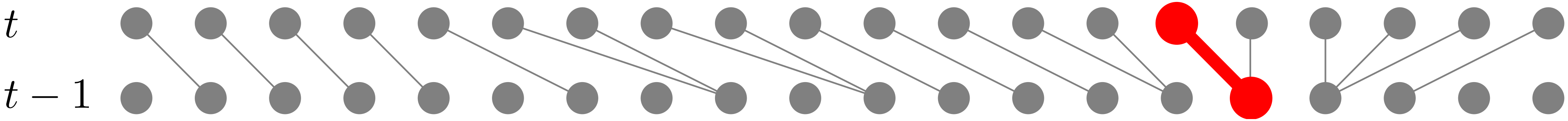
Population model: probability of having an ancestor

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 having a common ancestor in the last generation is ?



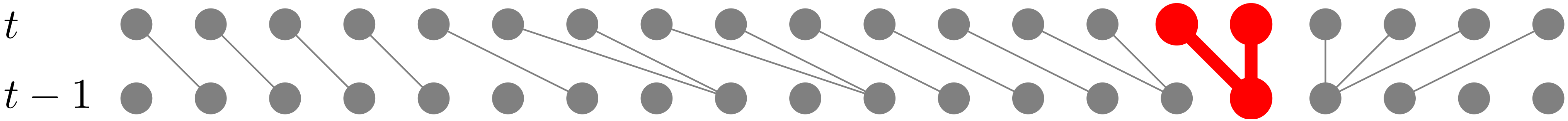
Population model: probability of having an ancestor

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 having a common ancestor in the last generation is 1.0



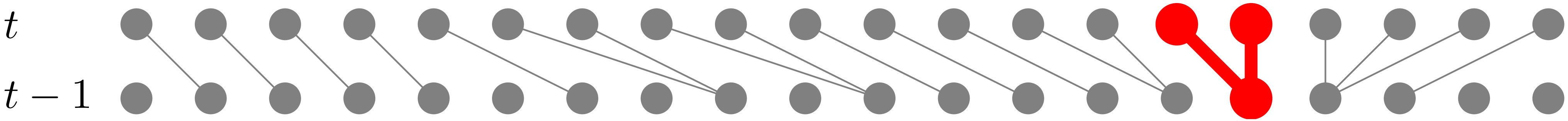
Population model: probability of having the same ancestor

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 $1.0 \times \frac{1}{2N}$

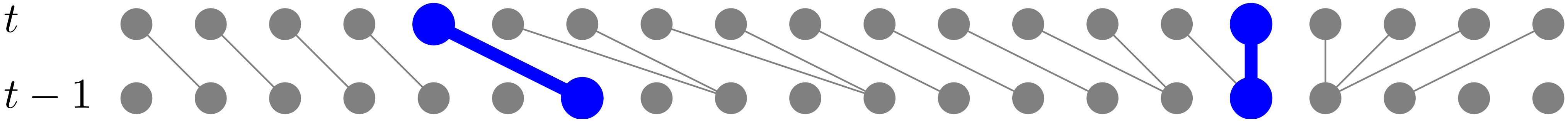


Population model: probability of having NOT the same ancestor

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 $\frac{1}{2N}$



The probability that two randomly picked chromosome do not have a common ancestor is $1-\frac{1}{2N}$



Time to coalescence of two lineages

The probability that two individuals share a common parent after t generations is

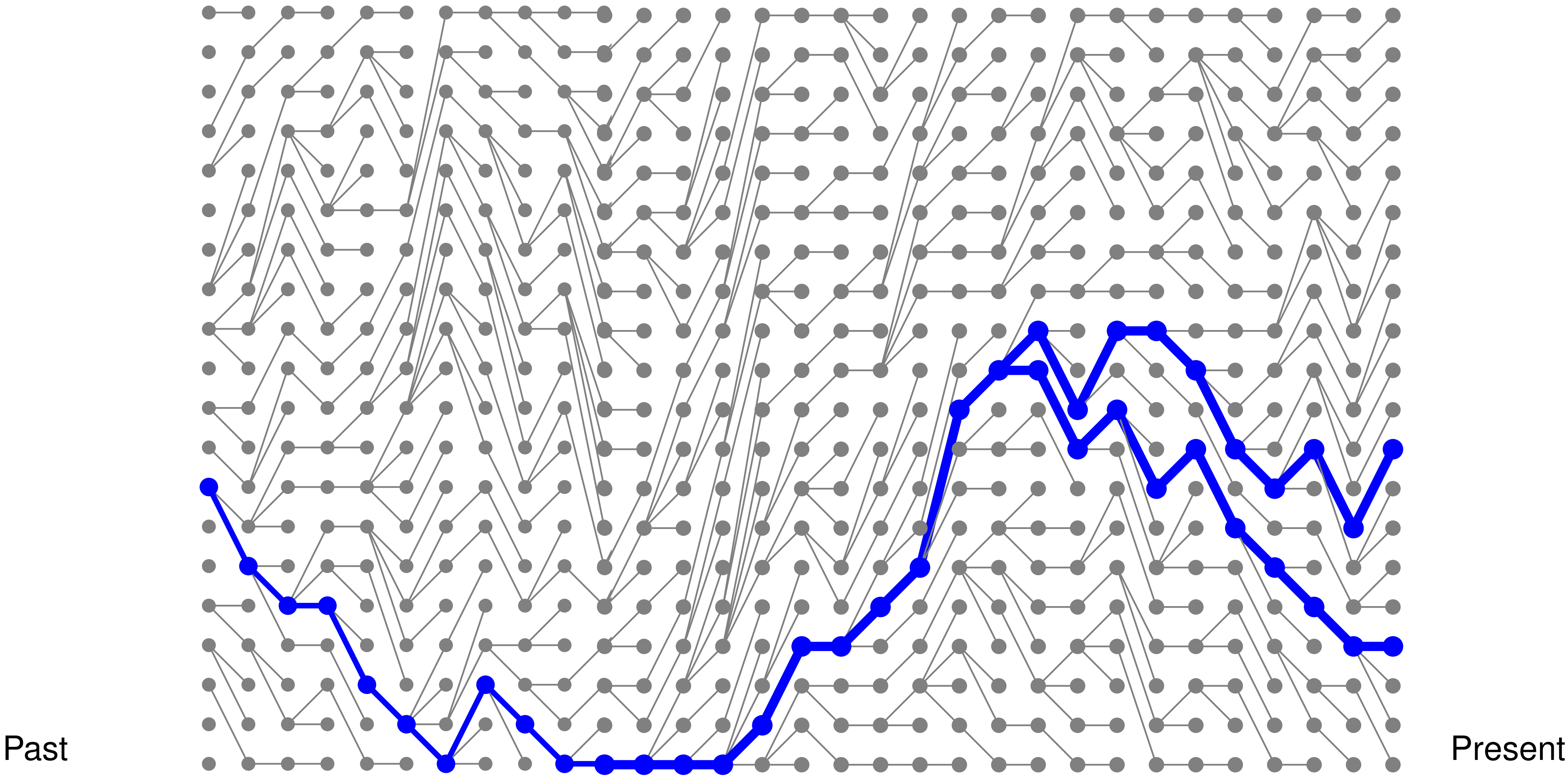
$$\mathrm{P}(t|N) = \underbrace{\left(1-\frac{1}{2N}\right)\times\left(1-\frac{1}{2N}\right)\ldots\times\left(1-\frac{1}{2N}\right)}_{t\text{ times}}\left(\frac{1}{2N}\right)$$

$$= \left(1-\frac{1}{2N}\right)^{t}\left(\frac{1}{2N}\right)$$

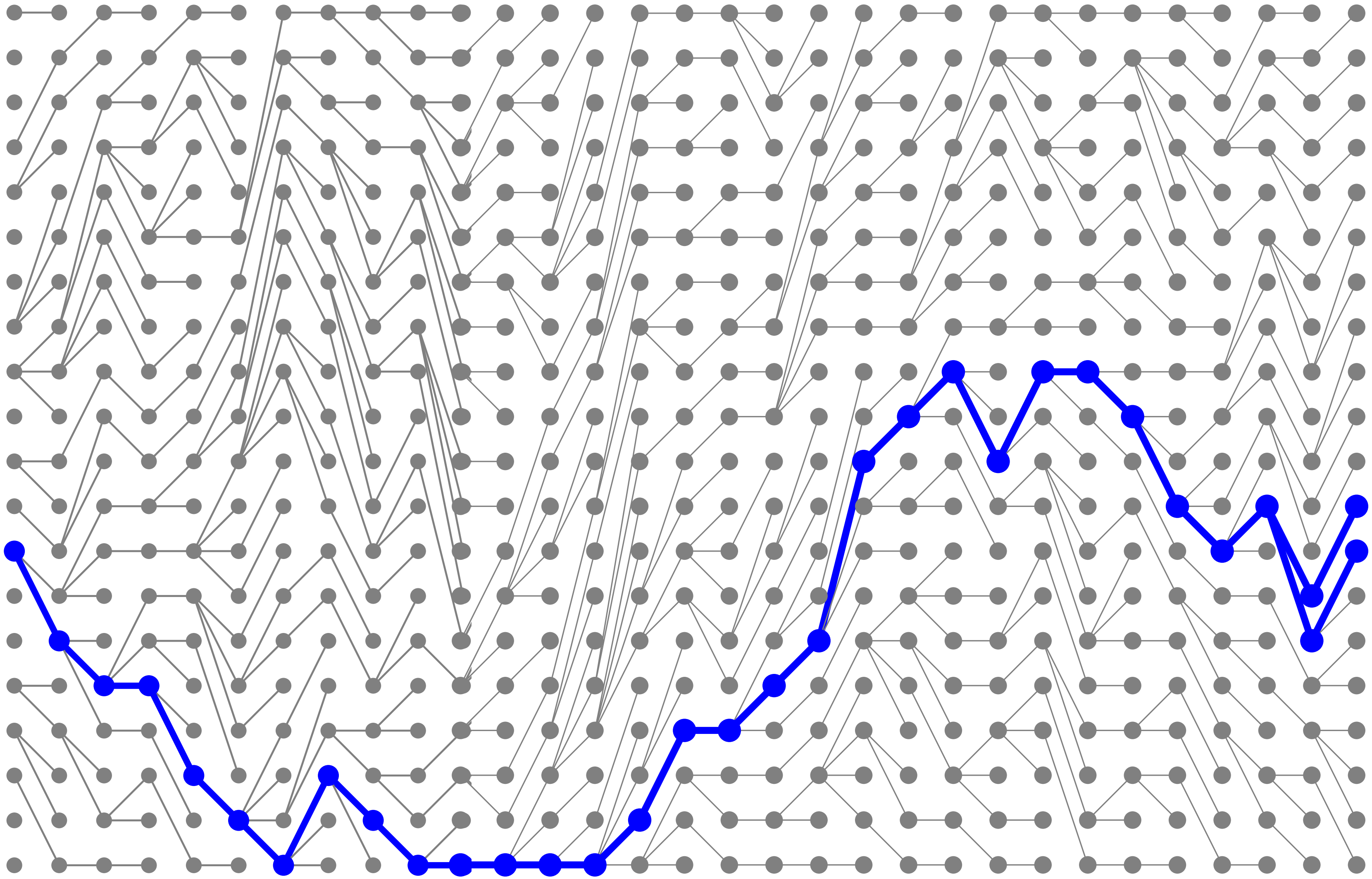
where t is the number of generations with no coalescence. This formula is known as the Geometric Distribution and we can calculate the expectation of the waiting time until two random individuals coalesce as

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

Sample of two individuals



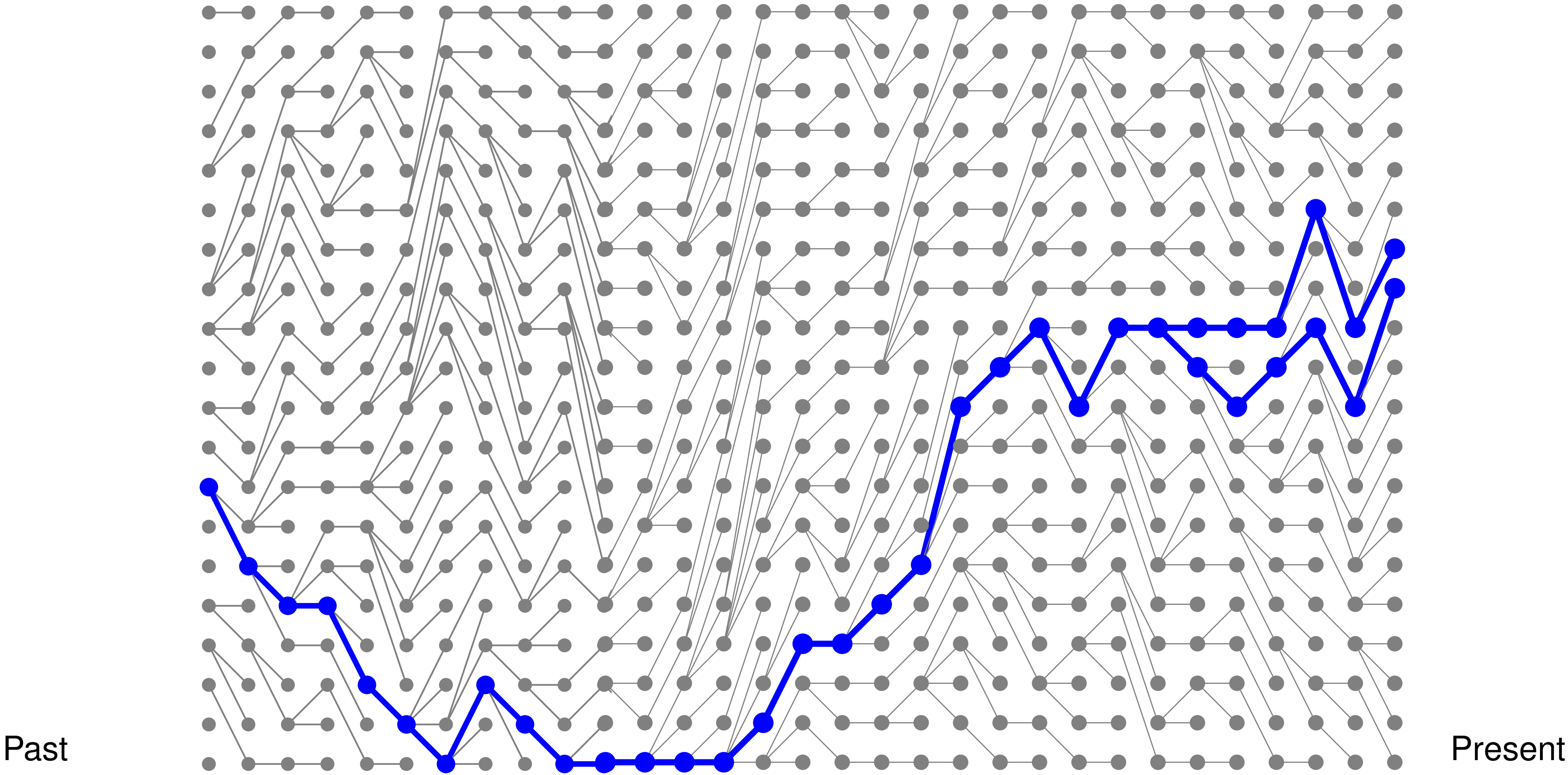
Sample of two individuals



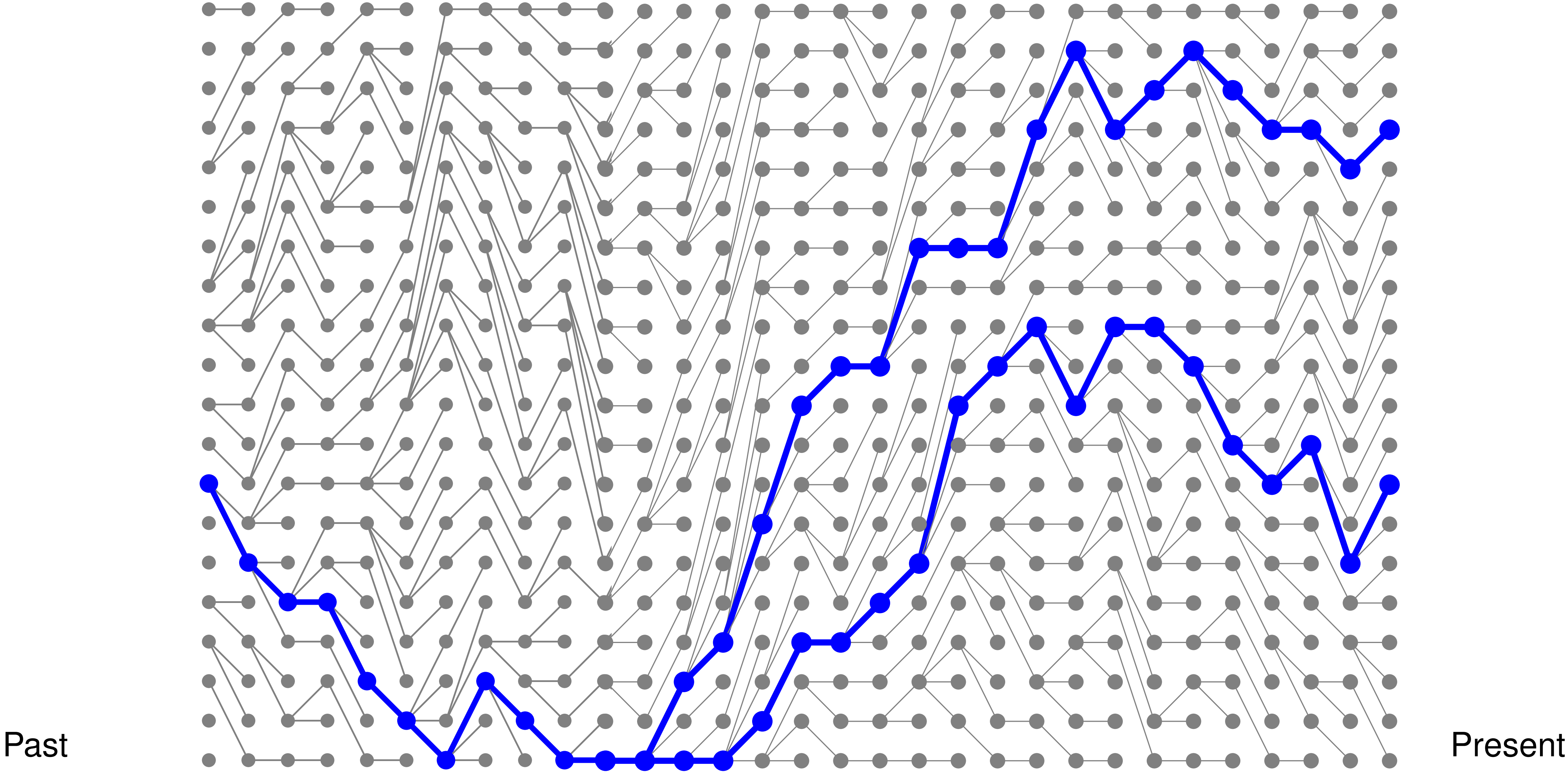
Past

Present

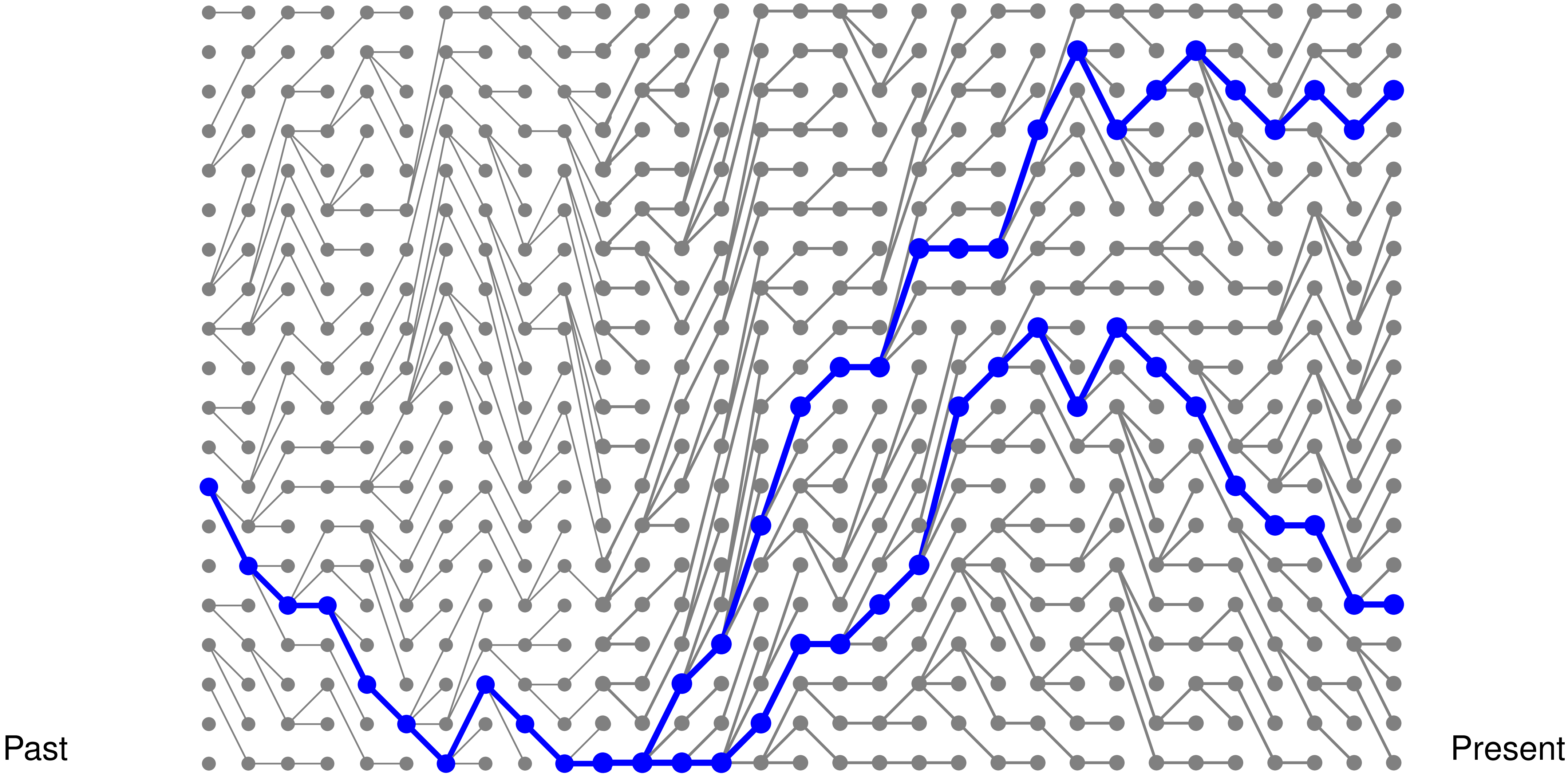
Sample of two individuals



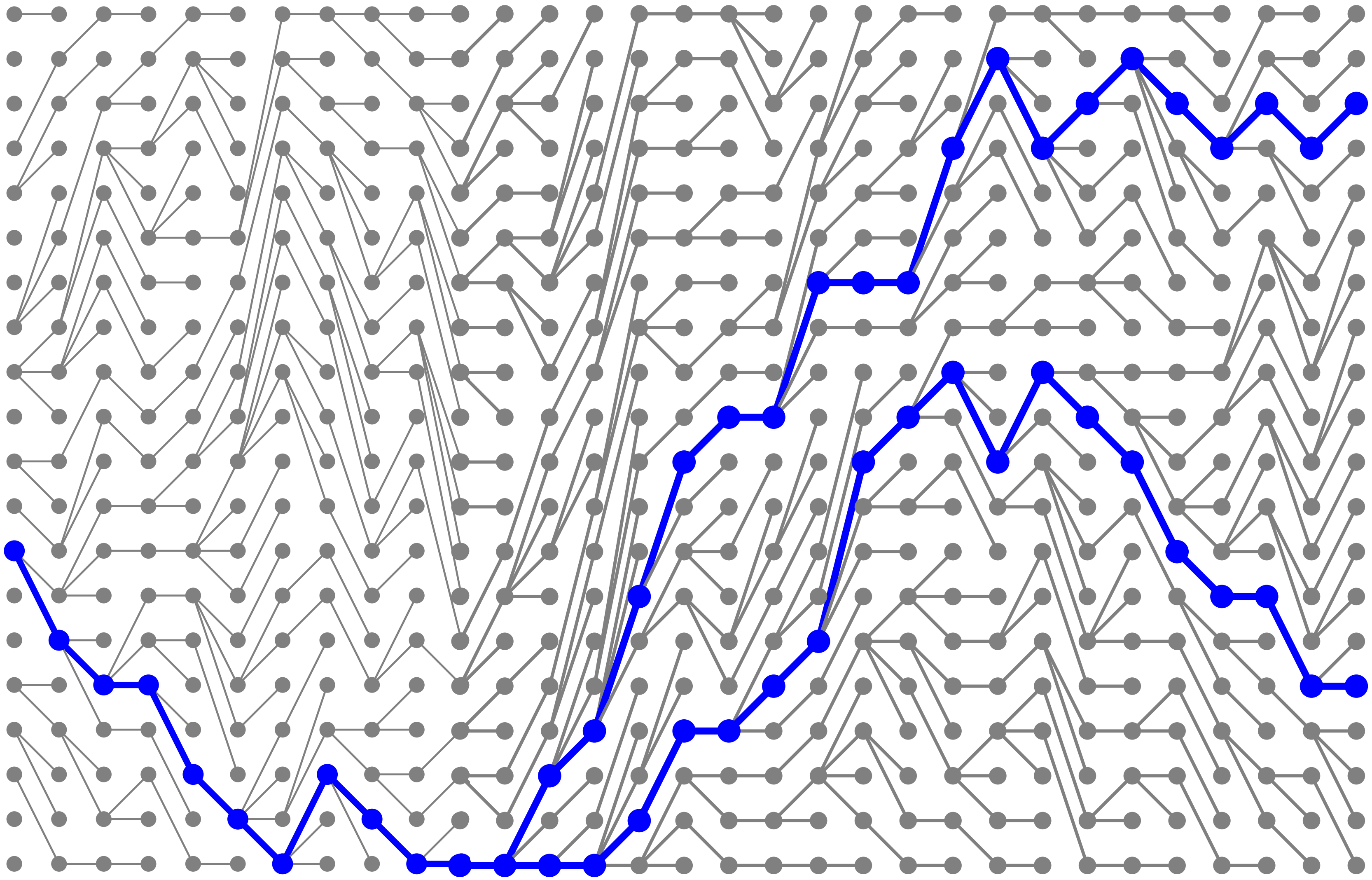
Sample of two individuals



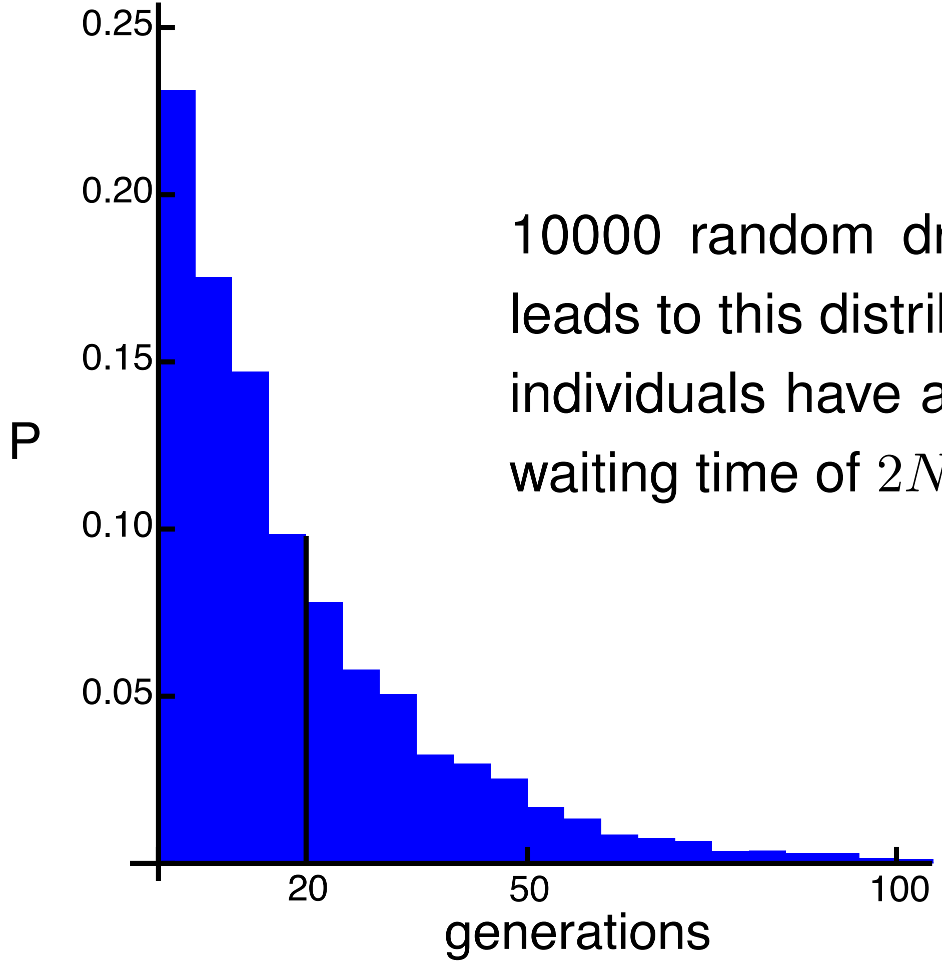
Sample of two individuals



Sample of two individuals



Time of coalescence of two lineages: Probability Distribution

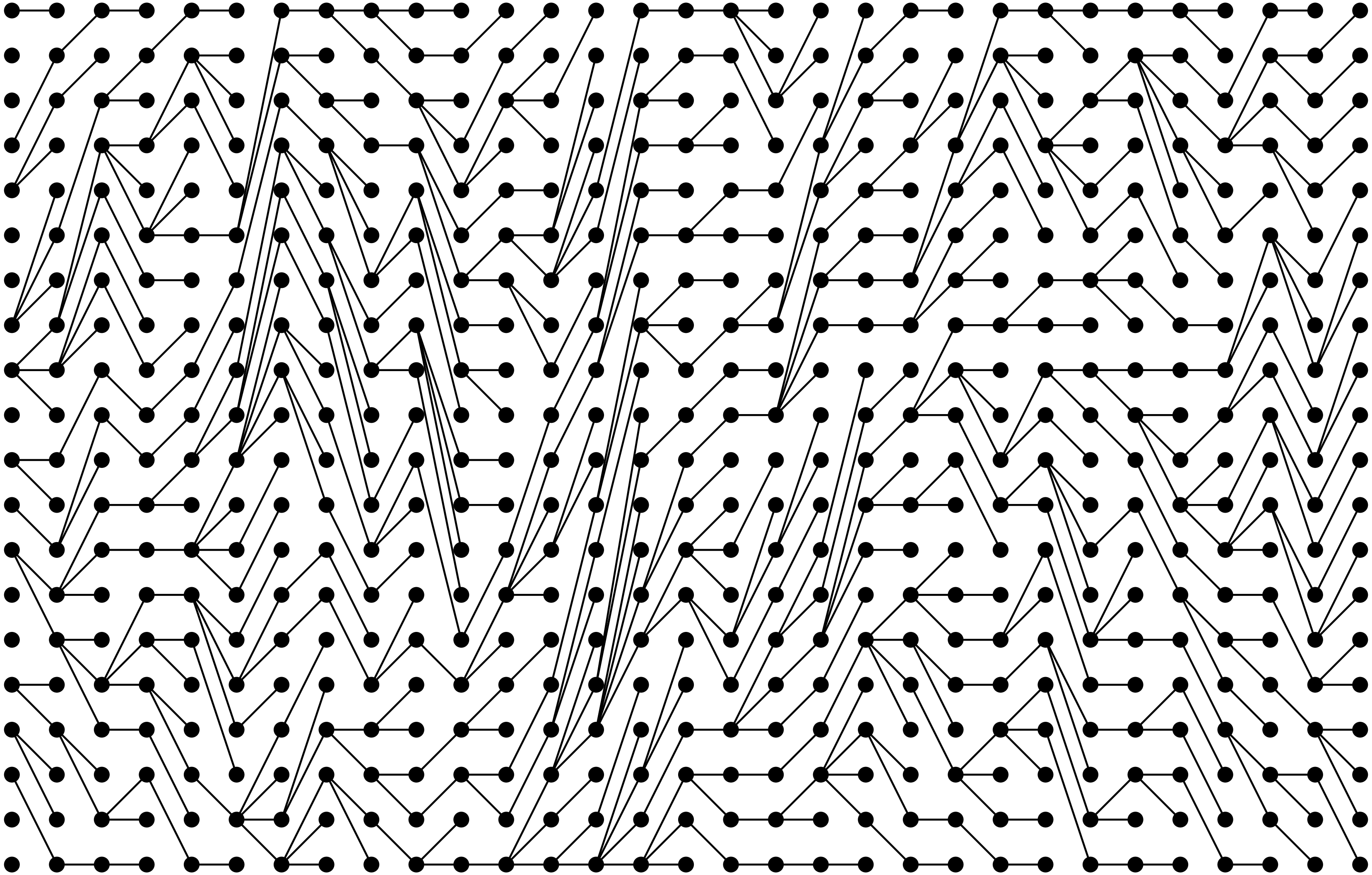


10000 random draw from a population with size $2N = 20$ leads to this distribution of times until two randomly chosen individuals have a common ancestor. The observed mean waiting time of $2N = 20.34$

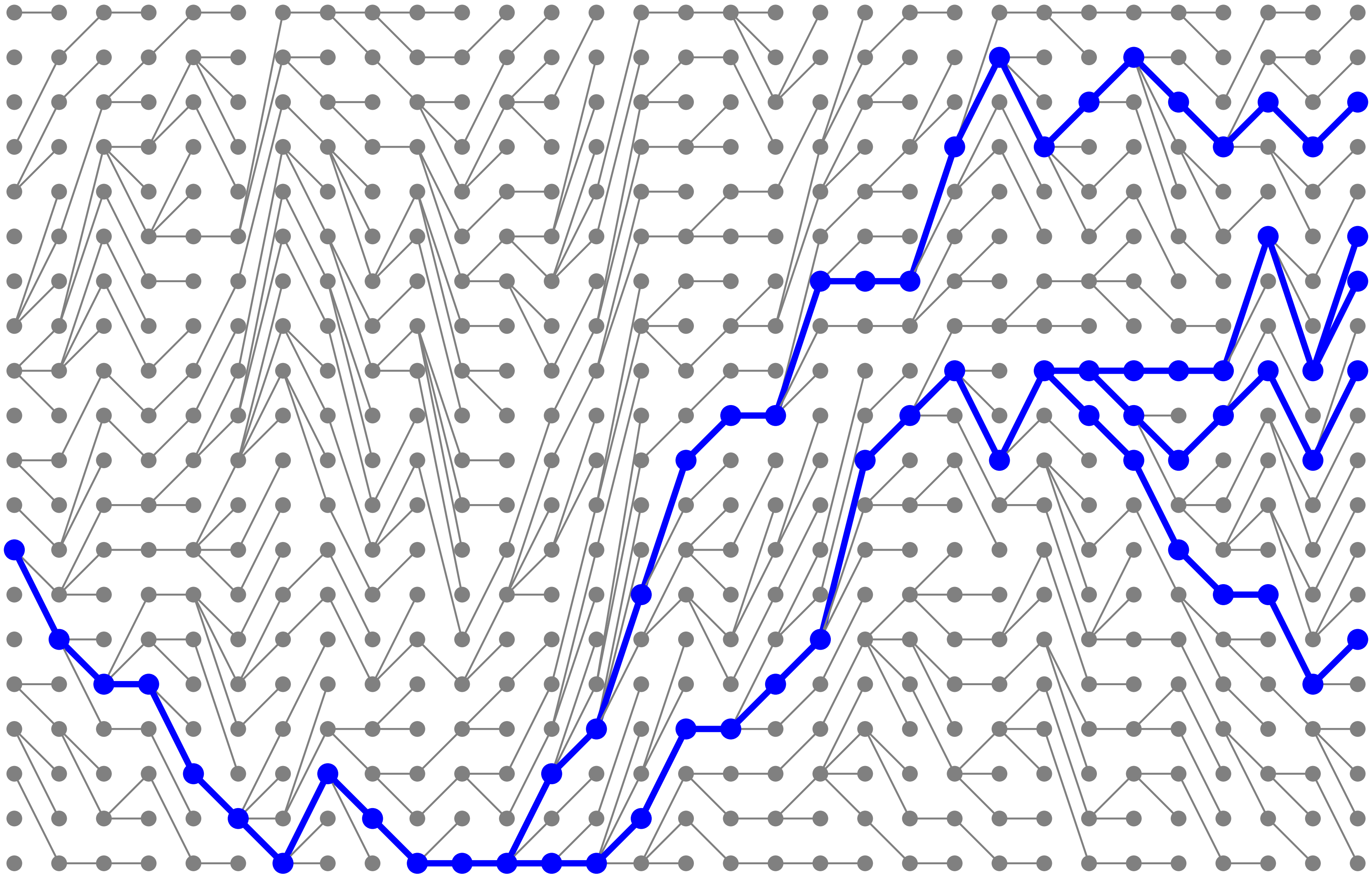
Observations: Coalescence of two lineages

- 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.*

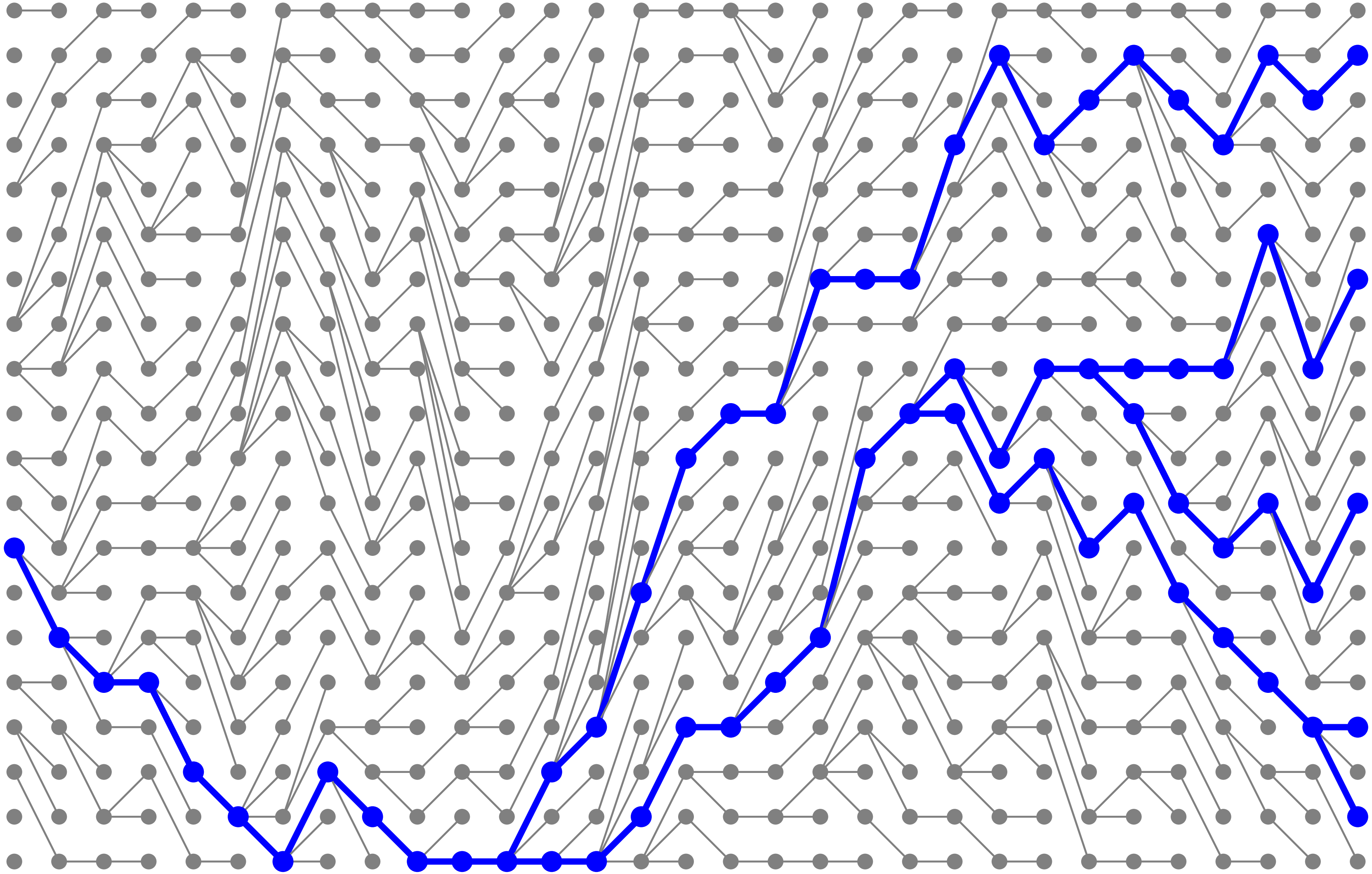
Sample larger than TWO



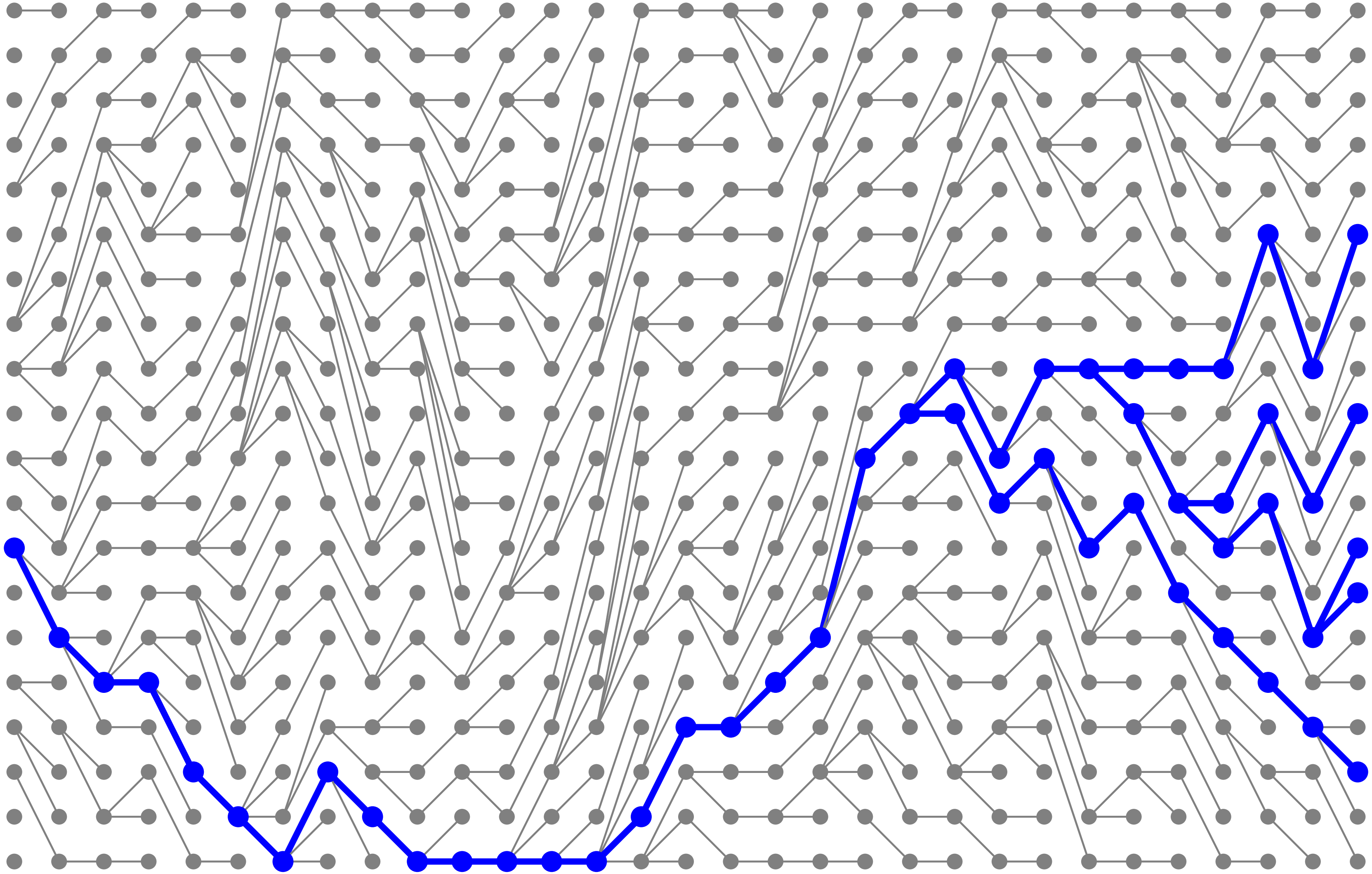
Sample larger than TWO



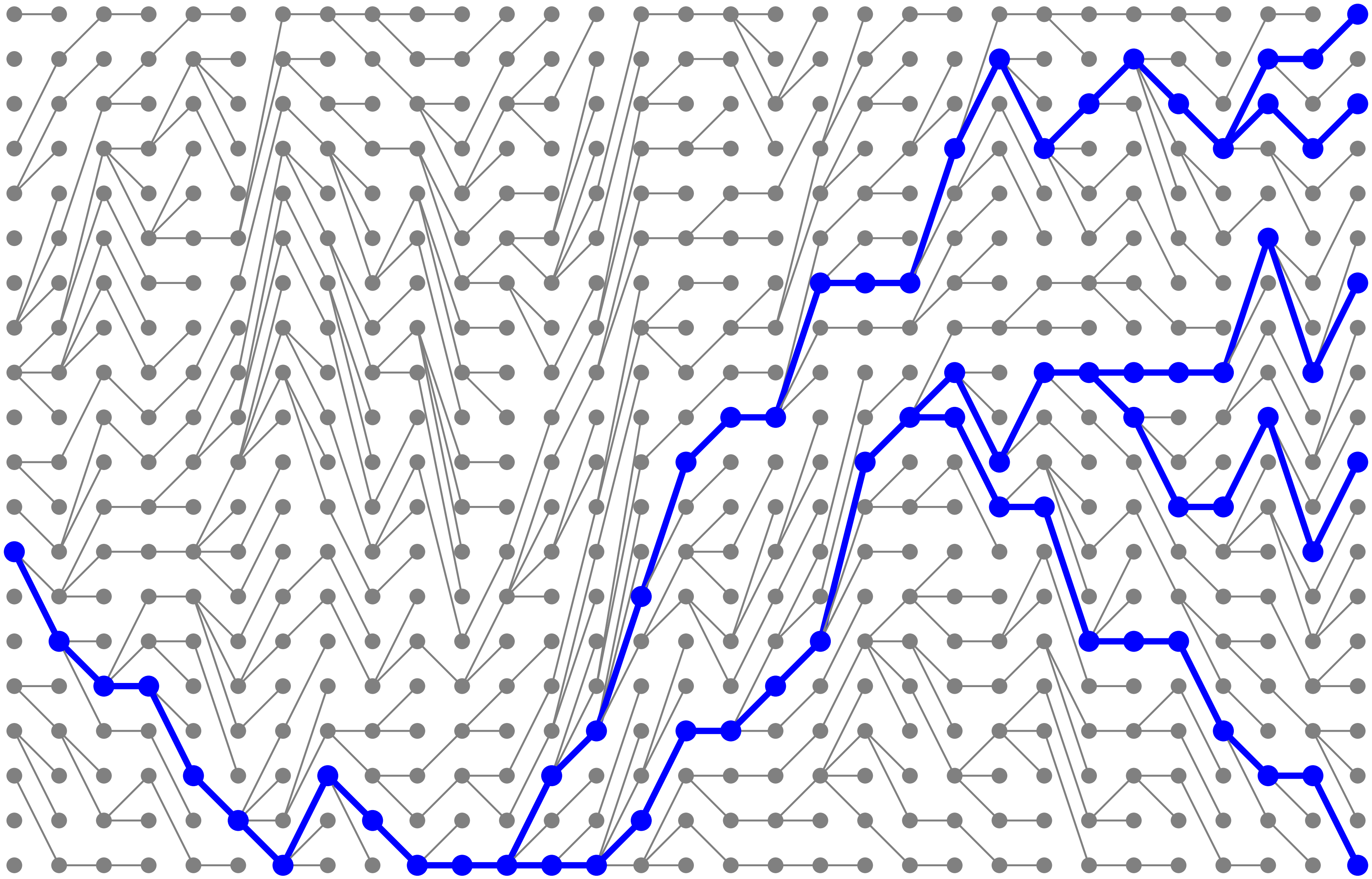
Sample larger than TWO



Sample larger than TWO

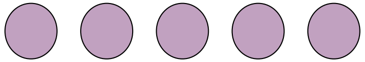


Sample larger than TWO



Coalescence theory

John Kingman described in 1982 the n-coalecent. He showed the behavior of a sample of size n (instead of n I will use k in the following slides), and its probability structure looking backwards in time.



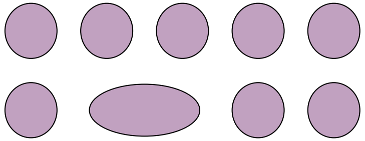
General findings:

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

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

Coalescence theory

John Kingman described in 1982 the n-coalecent. He showed the behavior of a sample of size n (instead of n I will use k in the following slides), and its probability structure looking backwards in time.



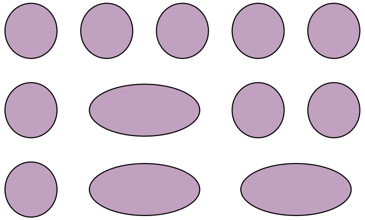
General findings:

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

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

Coalescence theory

John Kingman described in 1982 the n-coalecent. He showed the behavior of a sample of size n (instead of n I will use k in the following slides), and its probability structure looking backwards in time.



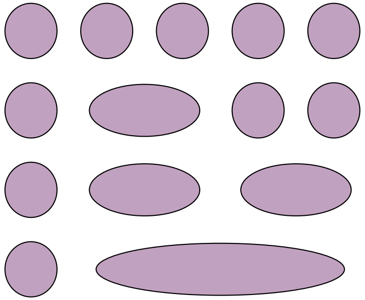
General findings:

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

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

Coalescence theory

John Kingman described in 1982 the n-coalecent. He showed the behavior of a sample of size n (instead of n I will use k in the following slides), and its probability structure looking backwards in time.



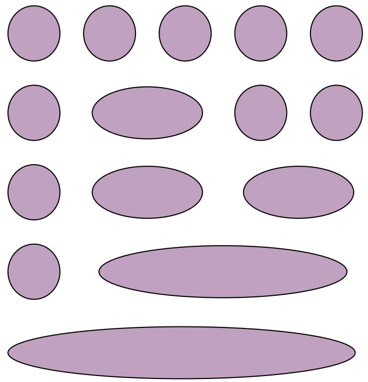
General findings:

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

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

Coalescence theory

John Kingman described in 1982 the n-coalecent. He showed the behavior of a sample of size n (instead of n I will use k in the following slides), and its probability structure looking backwards in time.



General findings:

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

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

Timescale

Sewall Wrights result on two lineages can be approximated:

In the discrete Wright-Fisher model we calculate the probability of non-coalescent during t generation; By using a suitable timescale τ such that one unit of scaled time corresponds to $2N$ generations, we can simplify to an continuous process

$$\left(1 - \frac{1}{2N}\right)^{t} = \left(1 - \frac{1}{2N}\right)^{(2N)\tau} \rightarrow e^{-\tau},$$

as N goes to infinity.

Timescale

Sewall Wrights result on two lineages can be approximated:

In the discrete Wright-Fisher model we calculate the probability of non-coalescent during t generation; By using a suitable timescale τ such that one unit of scaled time corresponds to $2N$ generations, we can simplify to an continuous process

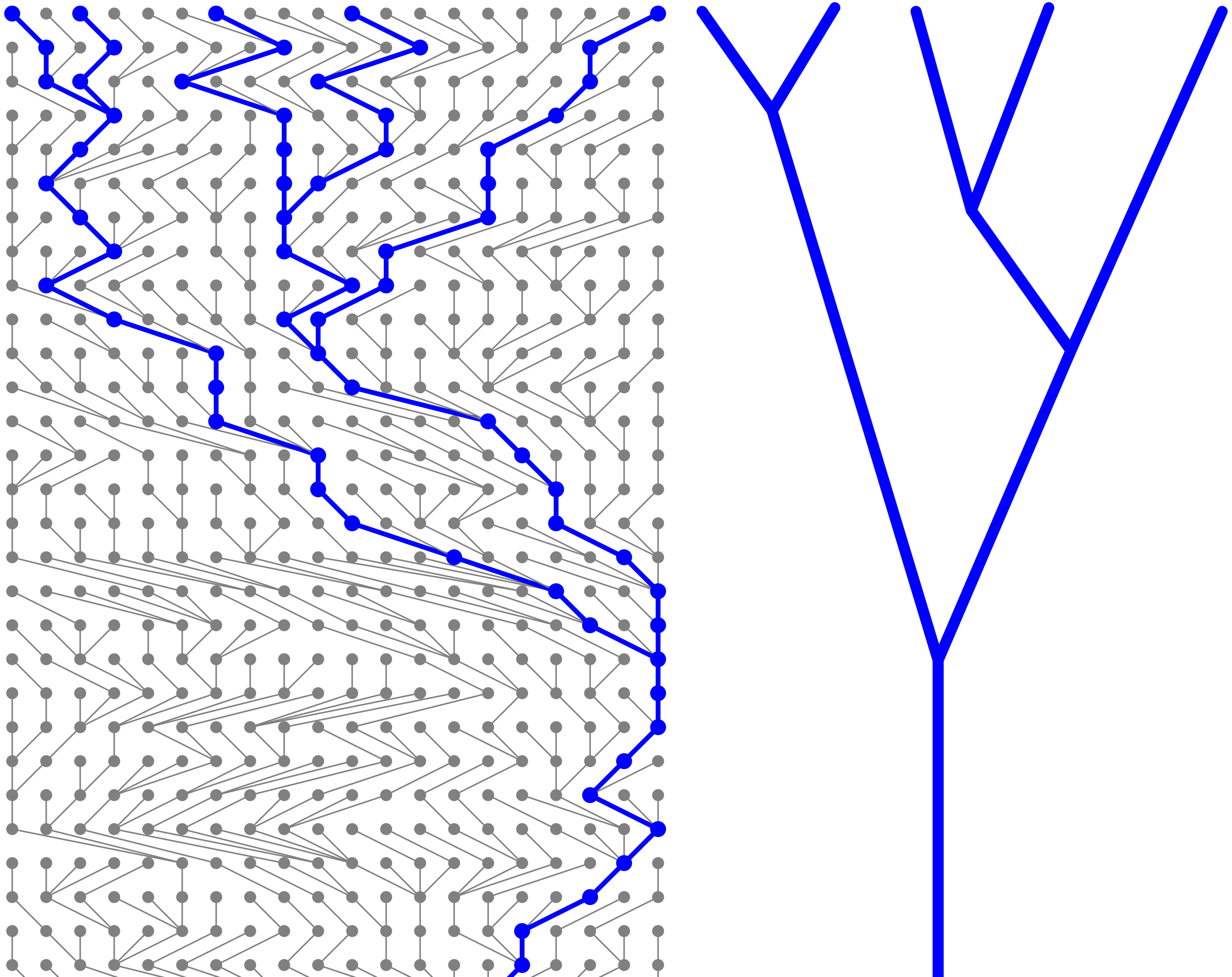
$$\left(1-\frac{1}{2N}\right)^{t}=\left(1-\frac{1}{2N}\right)^{(2N)\tau}\rightarrow e^{-\tau},$$

as N goes to infinity. For more than two lineages we use Kingman's result and use

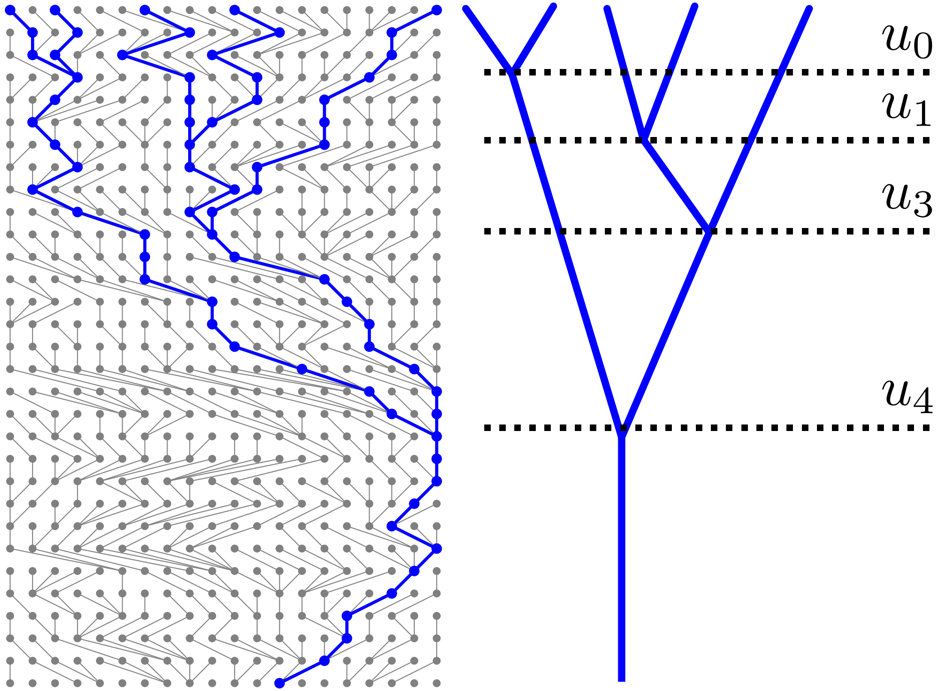
$$e^{-\tau\binom{k}{2}}$$

for the probability of non-coalescence of k lineages during the time interval τ; we will elaborate on τ soon.

Probability calculations on a genealogy

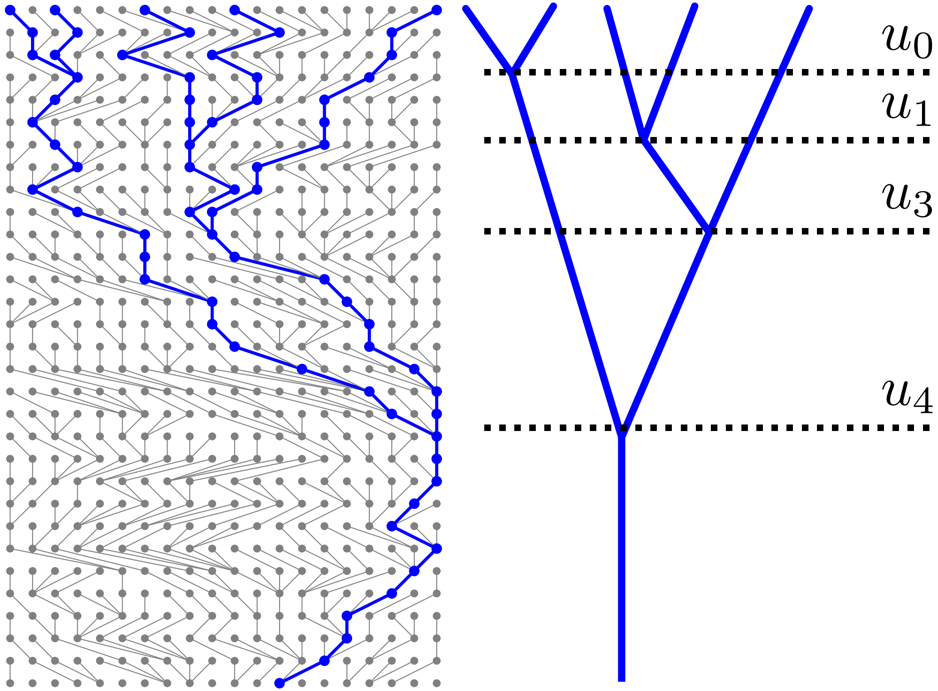


Probability calculation on genealogy



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

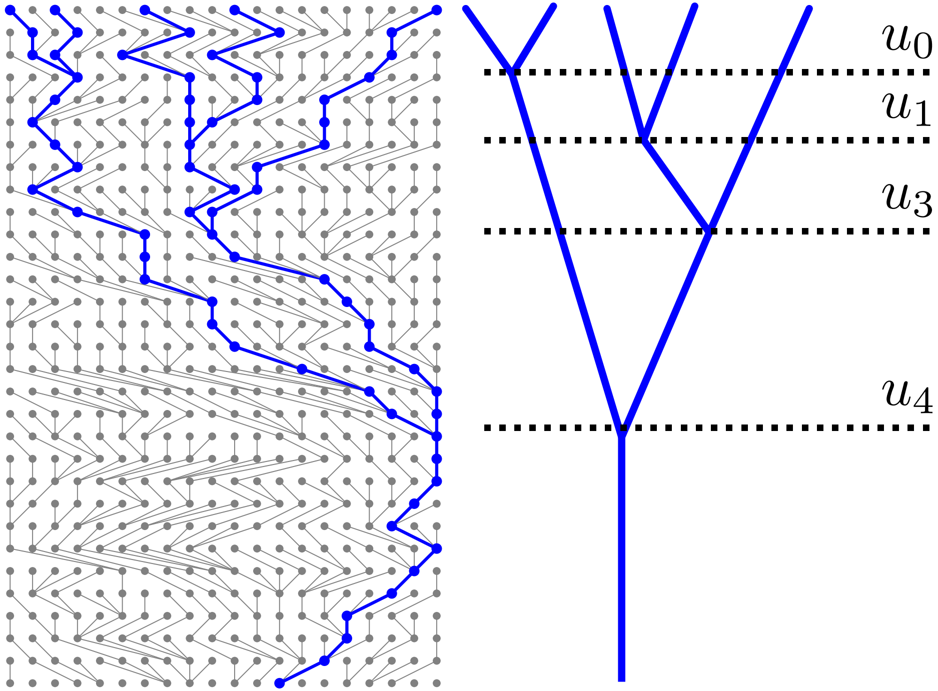
Probability calculation on genealogy



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample size k 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:

Probability calculation on genealogy: coalescent rate



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

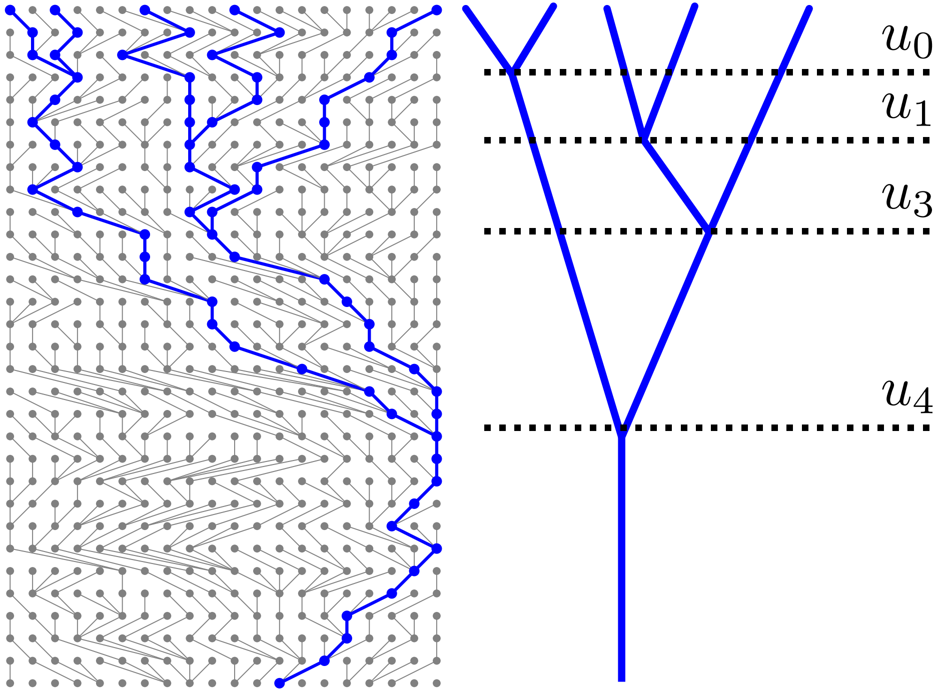
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with an exponential distribution:

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

with the scaled coalescence rate

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

Probability calculation on genealogy: coalescent rate



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

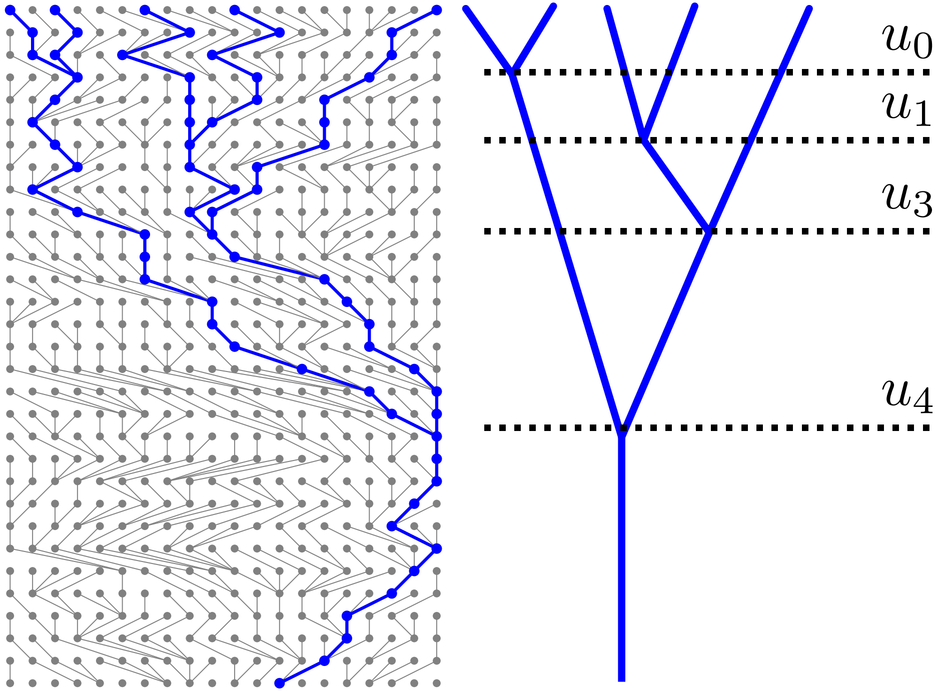
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with an exponential distribution:

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

with the scaled coalescence rate

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

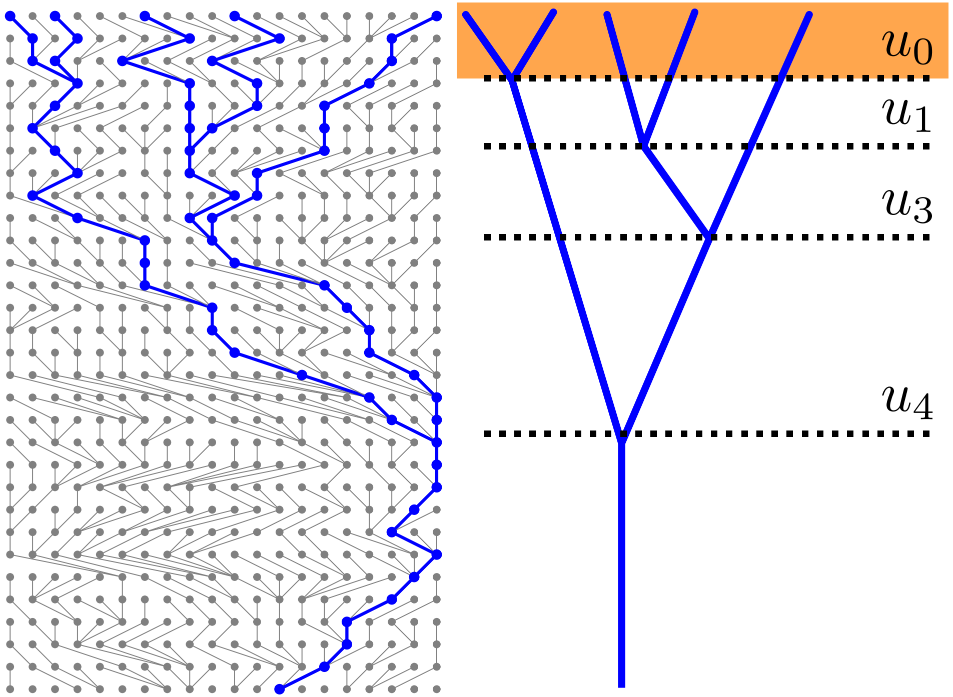
Probability calculation on genealogy



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:

$$\mathrm{P}(G|N)$$

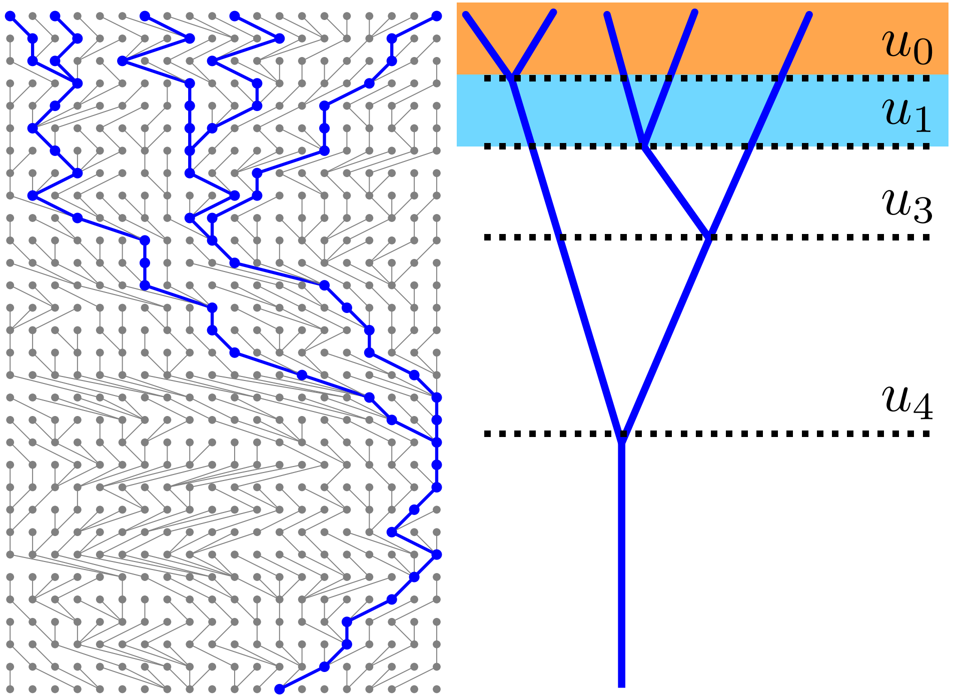
Probability calculation on genealogy



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:

$$\mathrm{P}(G|N) = \quad \mathrm{P}(u_0|N, i_1, i_2)$$

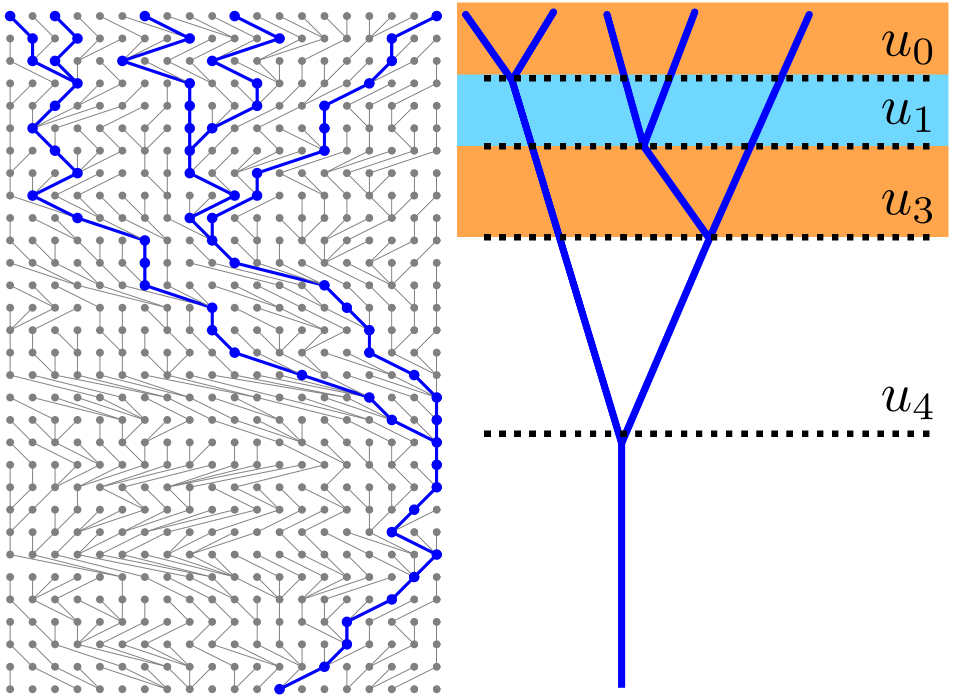
Probability calculation on genealogy



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:

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

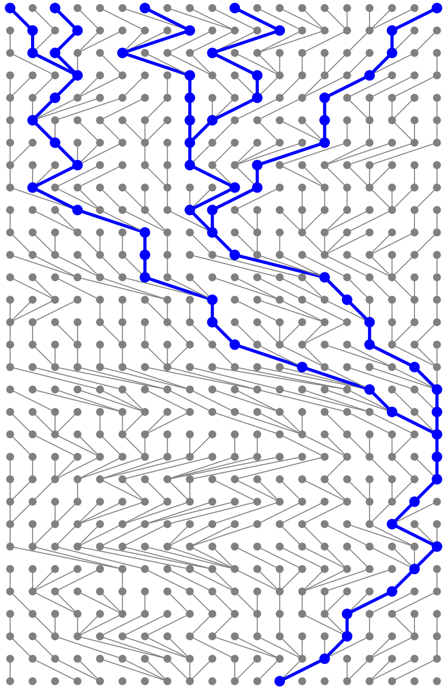
Probability calculation on genealogy

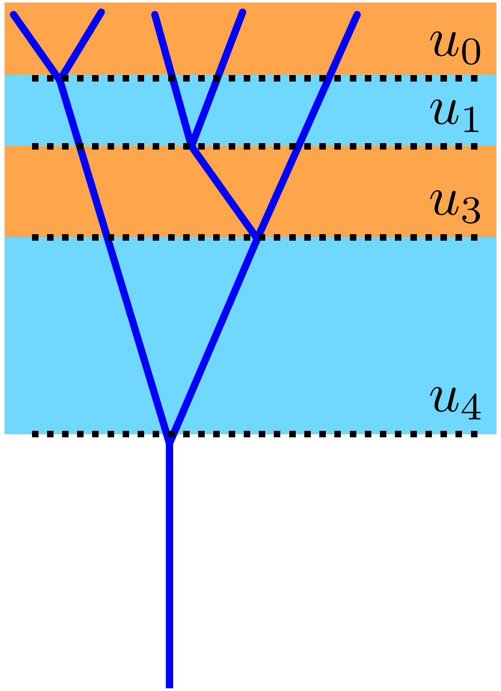


We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

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

Probability calculation on genealogy



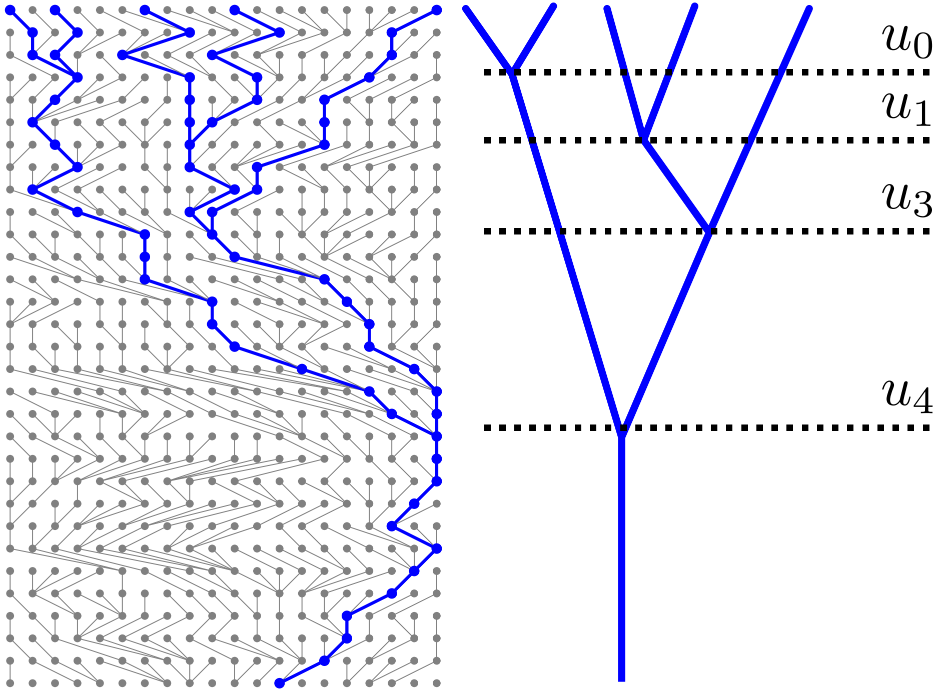


We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

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

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

Expected time of coalescence: large number of samples



The expectations of the total time to coalescence is the sum of the expectations for each interval. Each interval has expectation

$$\mathbb{E}(u) = \frac{4N}{k(k-1)}$$

this leads to the expectation for the time of the most recent common ancestor

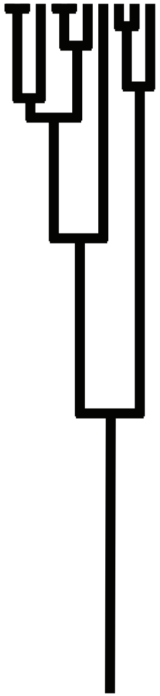
$$\mathbb{E}(\tau_{\text{MRCA}}) = \text{Sum of the expectation of each time interval} = \sum_{j=0}^{J} \frac{4N}{k_j(k_j-1)}$$

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

What is it good for?

If we know the genealogy G with certainty then we can calculate the population size N.

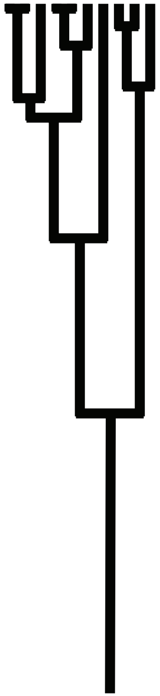
Finding the maximum probability $\mathrm{P}(G|N,k)$ is simple: we evaluate all possible values for N and pick the value with the highest probability.

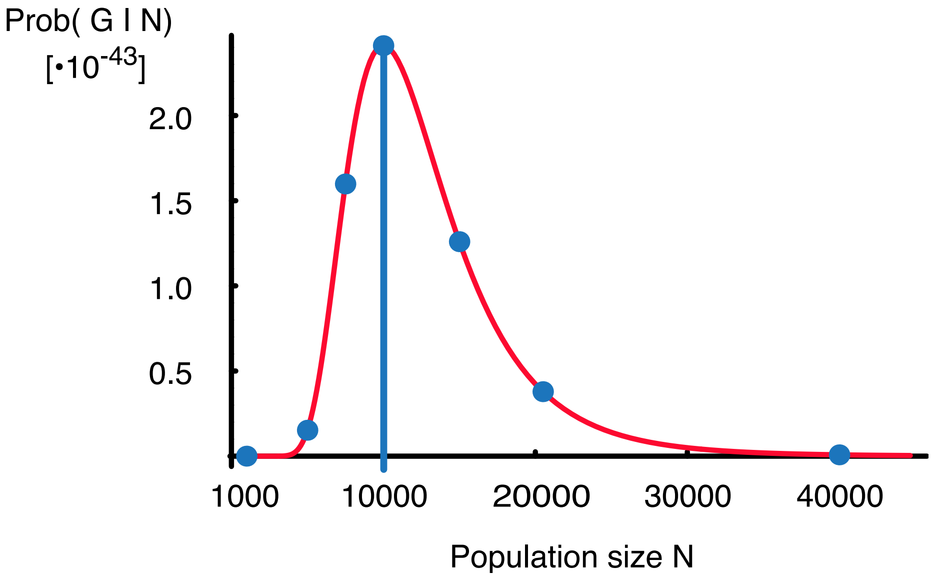


What is it good for? naive estimation of population size

If we know the genealogy G with certainty then we can calculate the population size N.

Finding the maximum probability $\mathrm{P}(G|N,k)$ is simple: we evaluate all possible values for N and pick the value with the highest probability.

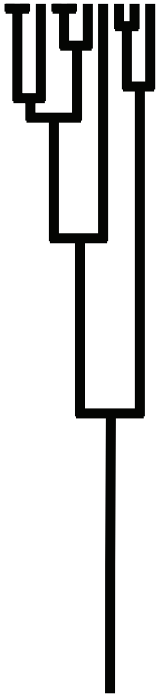


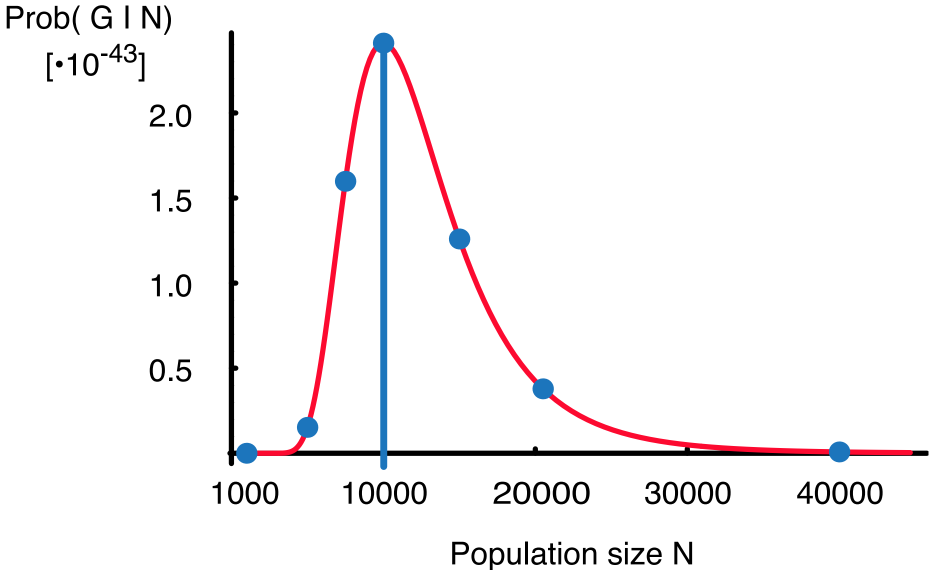


What is it good for?

If we know the genealogy G with certainty then we can calculate the population size N.

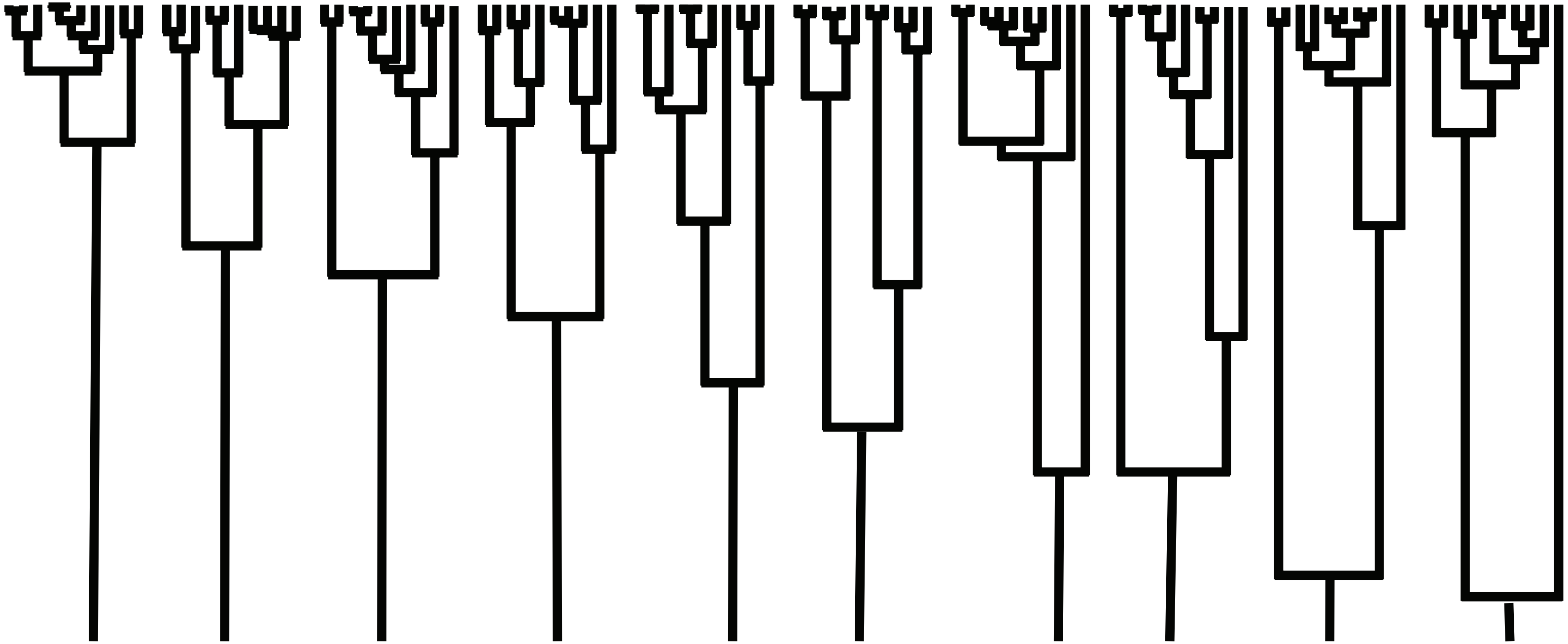
Finding the maximum probability $\mathrm{P}(G|N,k)$ is simple: we evaluate all possible values for N and pick the value with the highest probability.





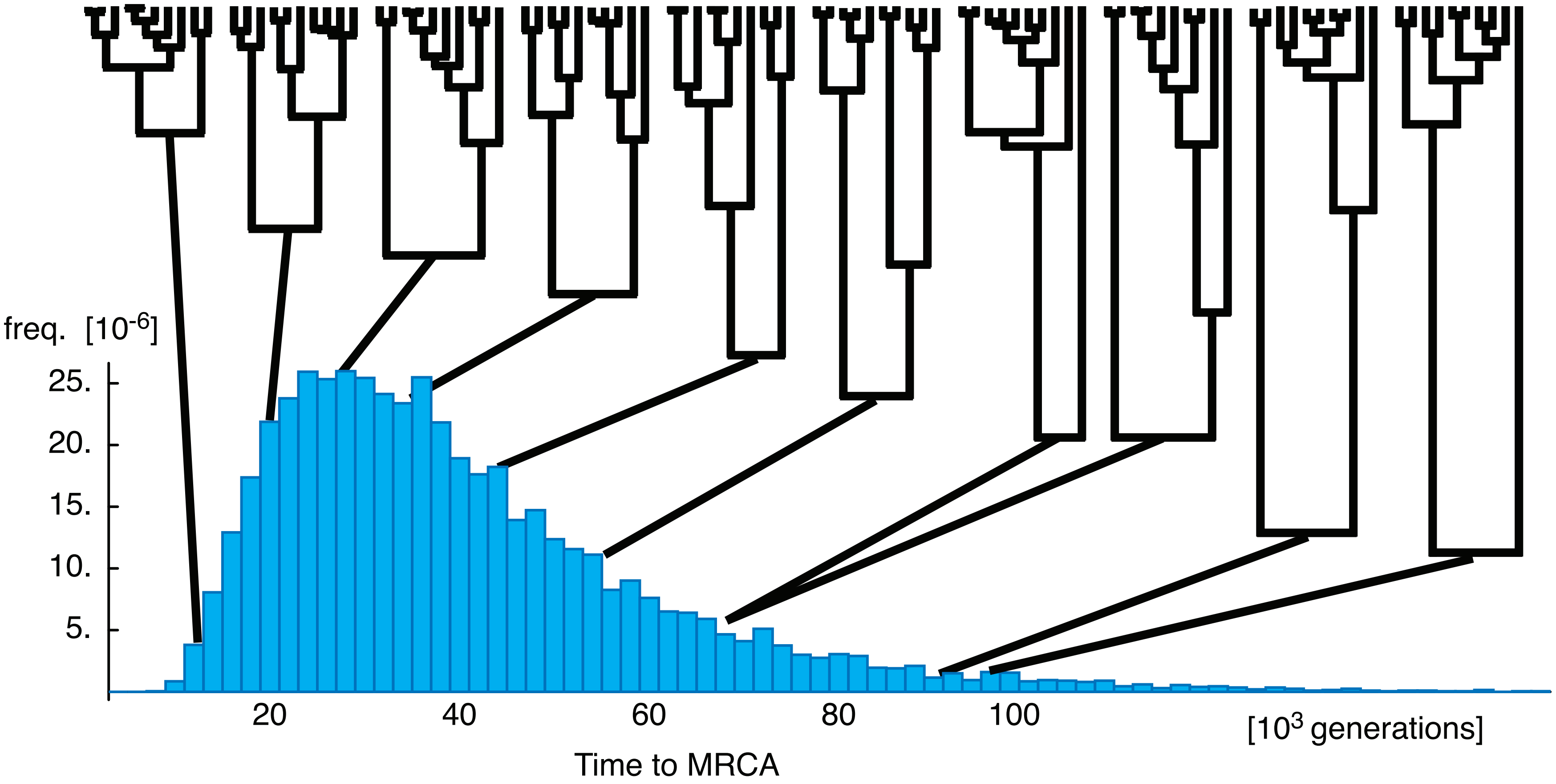
Can we really know the genealogy in all detail? NO

Variability of the coalescent process



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

Distribution of the Time to the Most Recent Common Ancestor



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

Kingman's n-coalescent is an approximation

- 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 to have a much smaller sample size than the population size:

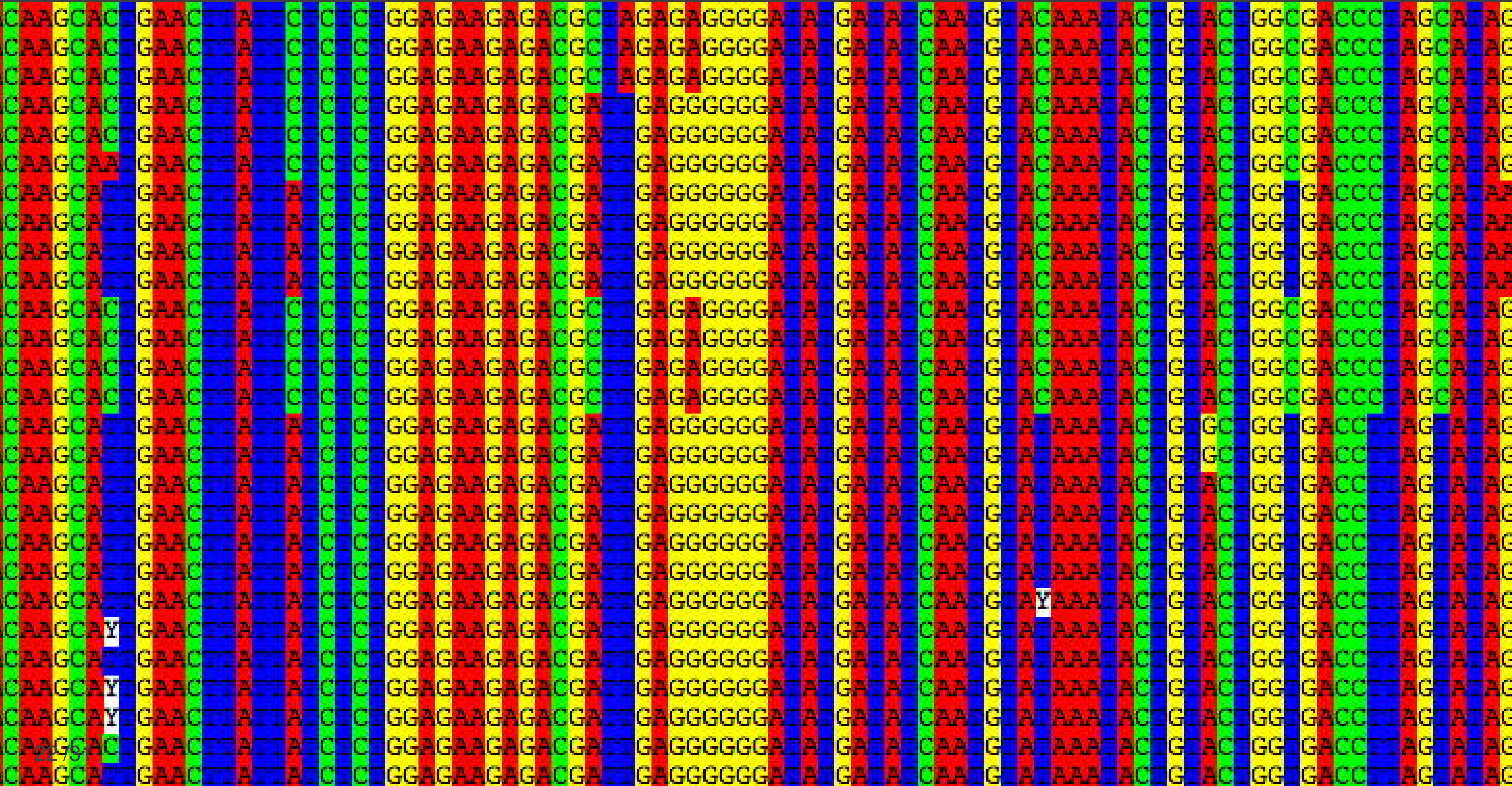
$$n \ll N$$

- 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.

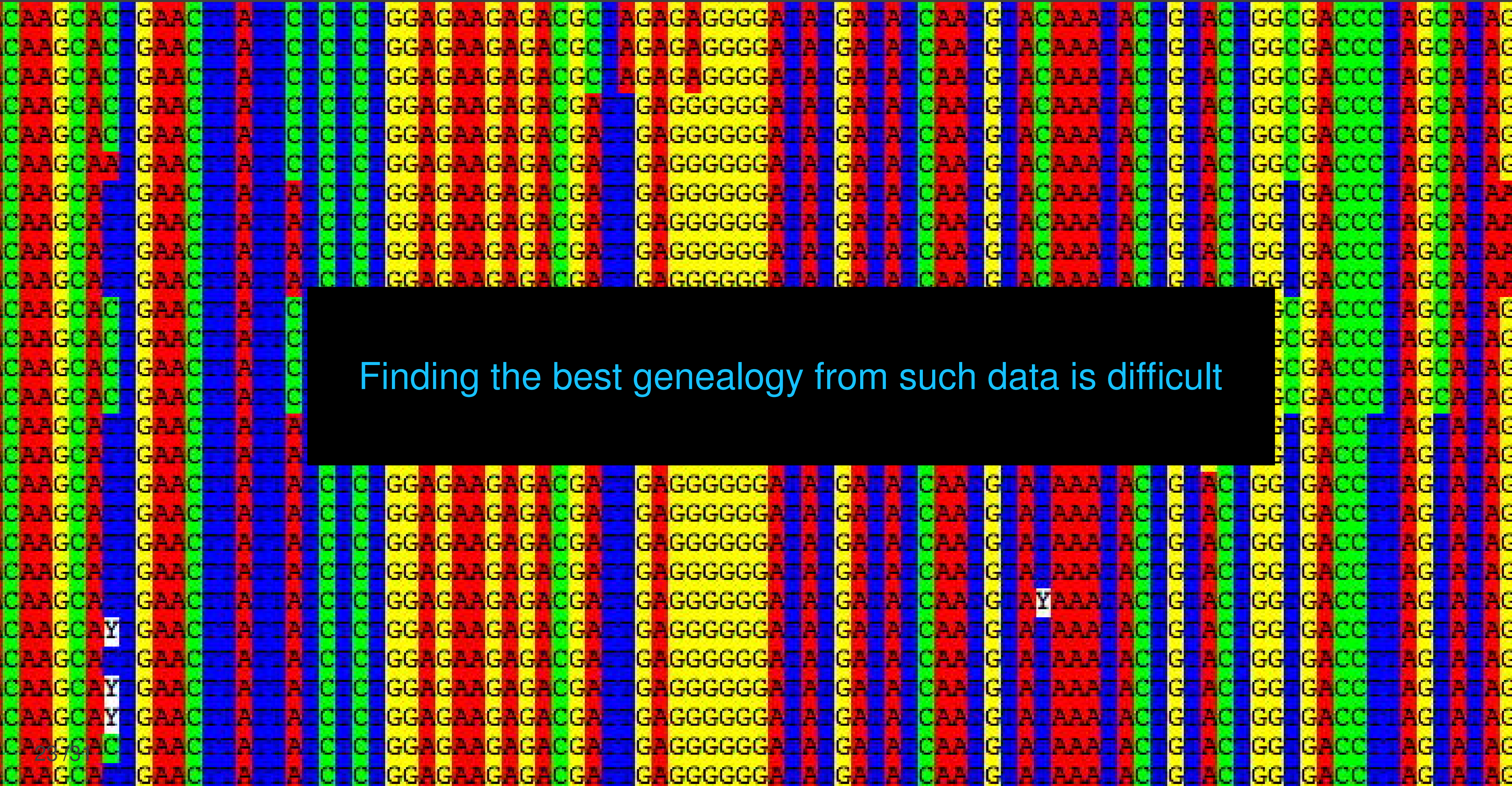
- Large samples coalesce on average in $4N$ generations.

- The time to the most recent common ancestor (TMRCA) has a large variance!

Genealogy and data



Genealogy and data



Finding the best genealogy from such data is difficult

Infer population size from data

- We can estimate a single best genealogy by assuming a model how DNA changes and use maximum likelihood
- We know that the coalescent has high variance

We could integrate over all possible genealogies using the fit between data and tree as a weight (likelihood).

Infer population size from data

- We can estimate a single best genealogy by assuming a model how DNA changes over time and use maximum likelihood
- We know that the coalescent has high variance

We could integrate over all possible genealogies using the fit between data and tree as a weight (likelihood).

Mutation model deals with mutations per generation,
our timescale becomes scaled by mutation rate which we do not know!

Mutation-scaled population size

The observed genetic variability S is

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

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.

We express the compound $N\mu$ and an inheritance scalar x

as the mutation-scaled population size $\Theta = xN\mu$

where μ is the mutation rate per generation and per site.

Mutation-scaled population size

The inheritance scalar x is different for different data types, for DNA sequences from

- diploids: $\Theta = 4N\mu$.
- haploids: $\Theta = 2N\mu$.
- 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.

Estimating Population size using Bayesian inference

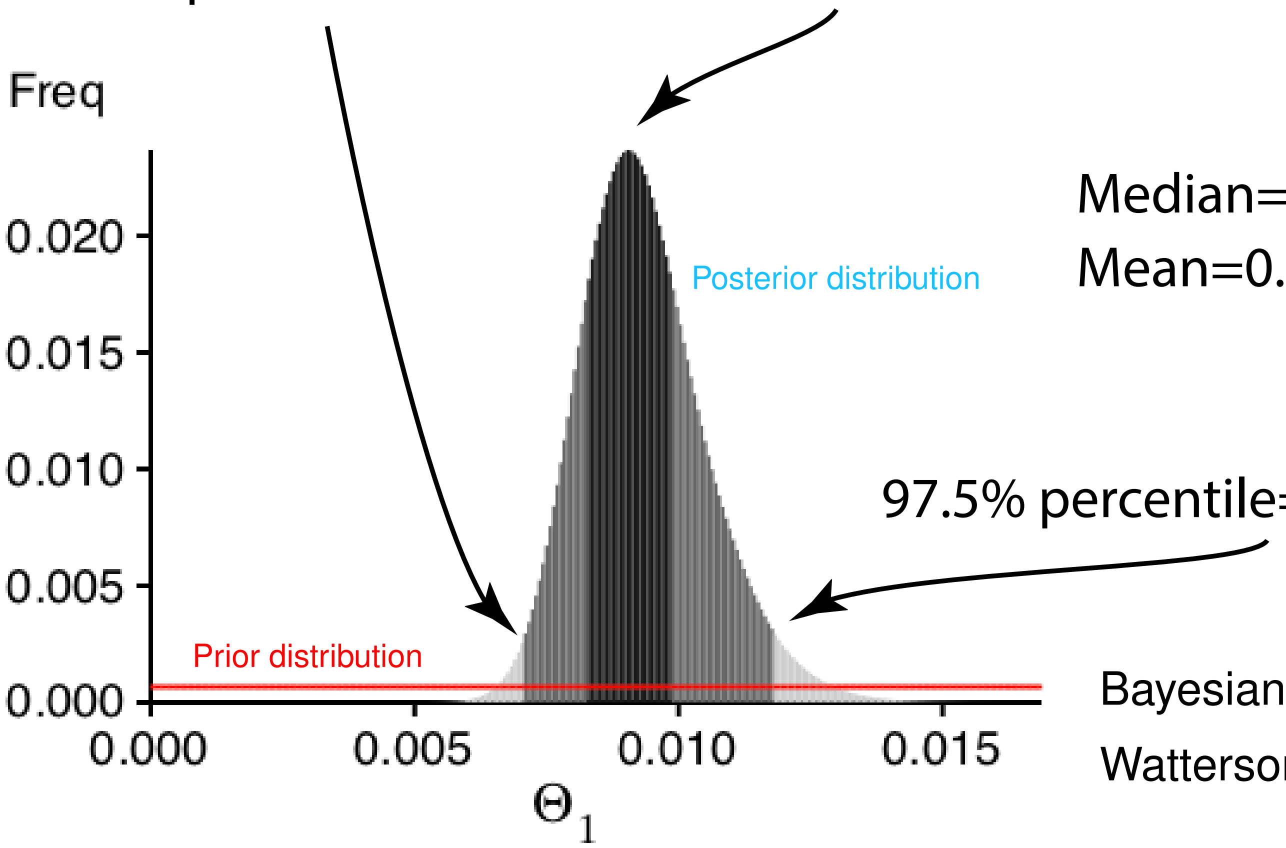
$$\begin{aligned}
\mathrm{p}(\Theta|\mathrm{D}) &= \frac{\mathrm{p}(\Theta)\mathrm{p}(\mathrm{D}|\boldsymbol{\Theta})}{\mathrm{p}(\mathrm{D})} \\
&= \frac{\mathrm{p}(\Theta)\int_{\mathrm{G}}\mathrm{p}(\mathrm{G}|\boldsymbol{\Theta})\mathrm{p}(\mathrm{D}|\mathrm{G})\mathrm{d}\mathrm{G}}{\mathrm{p}(\mathrm{D})}
\end{aligned}$$

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

Example: Inference of population size

2.5% percentile=0.007

Mode=0.00903



Median=0.00934

Mean=0.00934

97.5% percentile=0.0118

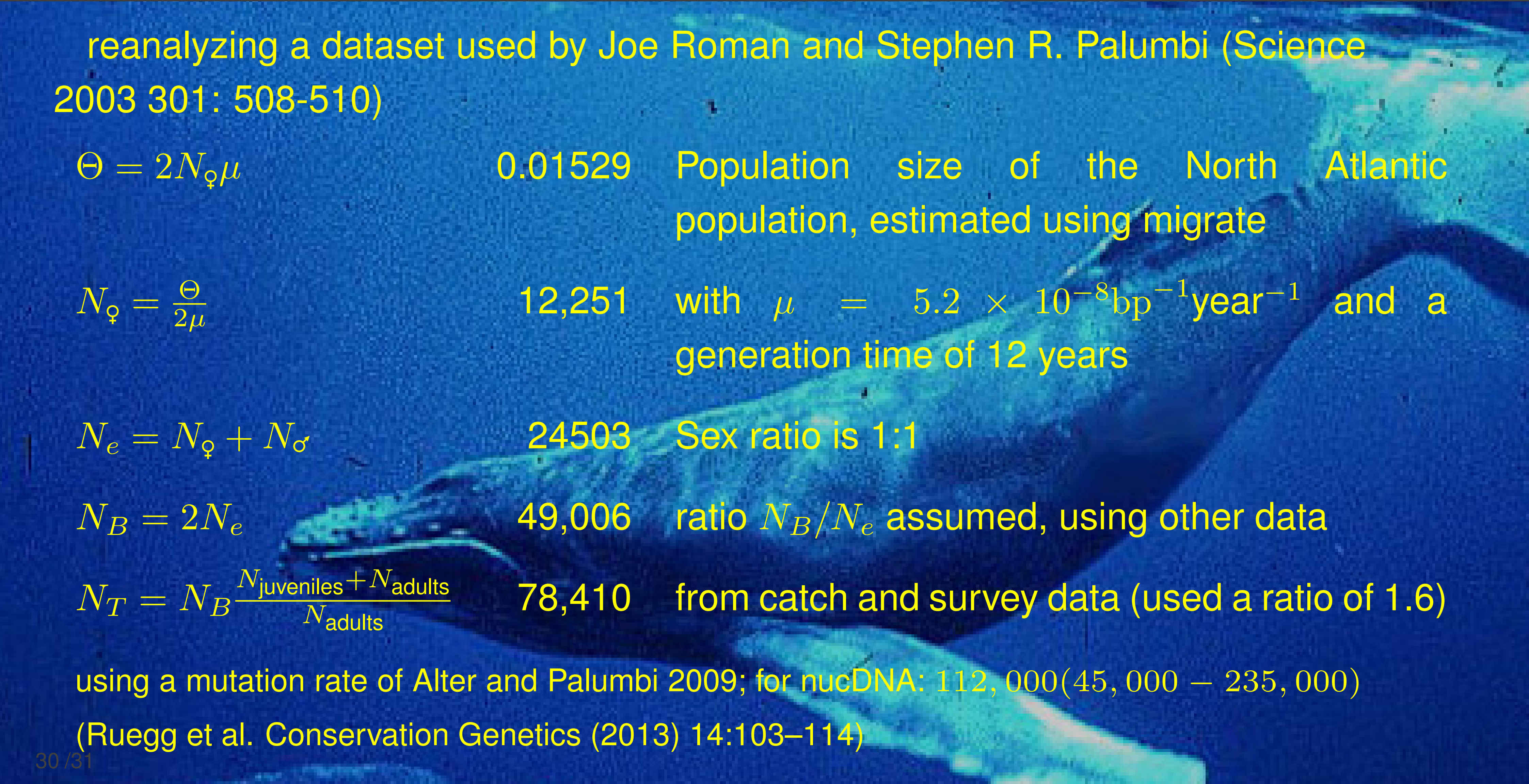
Bayesian inference: $\Theta = 0.00903$

Watterson Estimator $\Theta_W = 0.01003$

Historical humpback whale population size

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

Historical humpback whale population size



reanalyzing a dataset used by Joe Roman and Stephen R. Palumbi (Science 2003 301: 508-510)

$\Theta = 2N_{♀}\mu$	0.01529	Population size of the North Atlantic population, estimated using migrate
$N_{♀} = \frac{\Theta}{2\mu}$	12,251	with $\mu = 5.2 \times 10^{-8}\mathrm{bp}^{-1}$year^{-1} and a generation time of 12 years
$N_e = N_{♀} + N_{♂}$	24503	Sex ratio is 1:1
$N_B = 2N_e$	49,006	ratio N_B/N_e assumed, using other data
$N_T = N_B\frac{N_{\text{juveniles}}+N_{\text{adults}}}{N_{\text{adults}}}$	78,410	from catch and survey data (used a ratio of 1.6)

using a mutation rate of Alter and Palumbi 2009; for nucDNA: $112,000(45,000 - 235,000)$ (Ruegg et al. Conservation Genetics (2013) 14:103–114)

Summary and Outlook

- Genetic data and the coalescent allow estimating long term (historical) population sizes.
- There are many different programs to infer population sizes that use the coalescent: some are summary statistics, others are probabilistic Bayesian approaches.
- The basic coalescent is the foundation of many extension that include other population genetics forces, such as population size changes through time, gene flow among multiple population, recombination, admixture, ...

This talk was created with partial support by USA National Science Foundation grants DBI 1564822 and DBI 2019989.