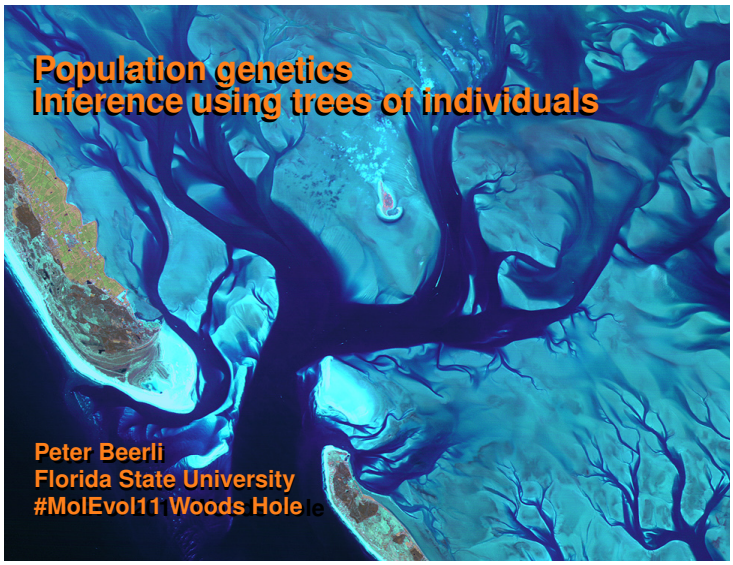


Outline and logistics

- 9:00 – 12:00 **Coalescence theory as a tool for population genetics**
  Inferences based on the coalescent
  Extensions of the basic coalescent
- 10:30 – 10:55 Break

Problems that need to be solved

HEALTH (old and new)

Problems that need to be solved

HEALTH (old and new)

Problems that need to be solved

Questions

- What is the rate of emergence of new diseases?
  How many strains of influenza could there be?
  Why are some influenza strains deadly and others not?
  How fast do new strains adapt to humans (other species)?
- How do diseases spread?
  Are there recurrent patterns of emergence (old strains maintenance) ?
  What are the most common routes of distributions of diseases?

Problems that need to be solved

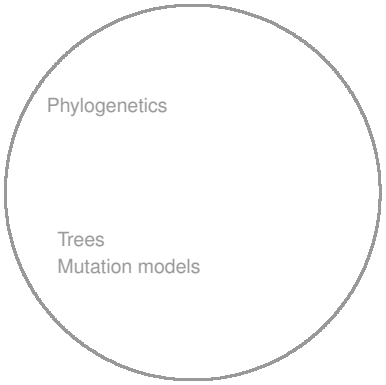
Conservation

Problems that need to be solved

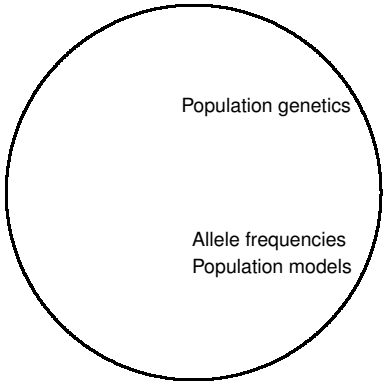
- How small can sustainable population of endangered species be?
  How can we maintain the genetic variability within a population?
  How do diseases affect rare species?
- How are populations connected?
  What are the dynamics in a landscape? How many individuals need to exchange among populations to keep the genetic variability high?
  What was the connectivity among populations in the past? In the future?

How do we approach problems like these?

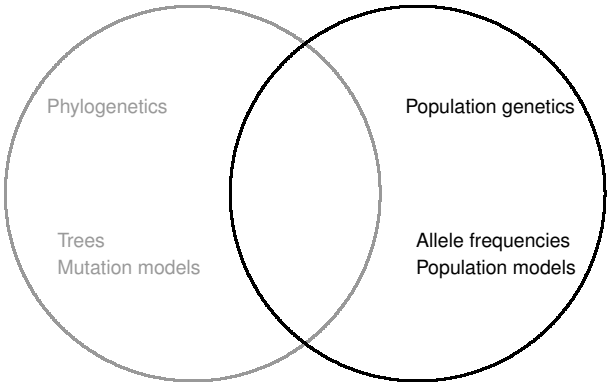
How do we approach problems like these?



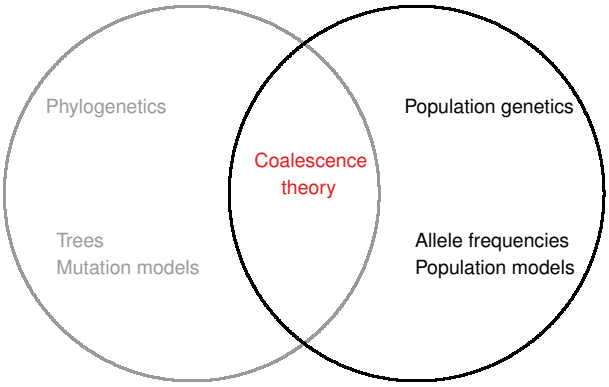
How do we approach problems like these?



How do we approach problems like these?



How do we approach problems like these?



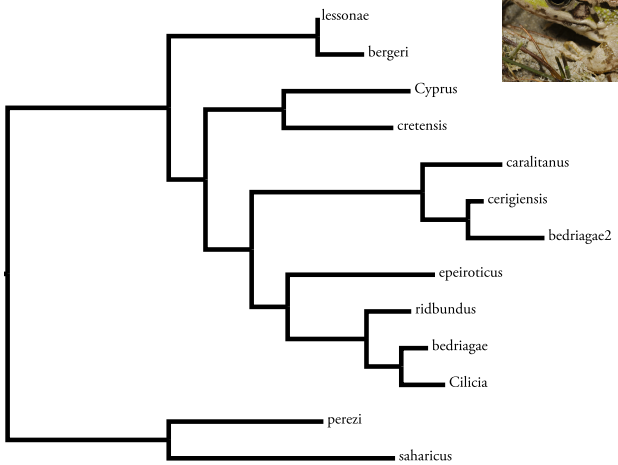
Coalescence theory as a tool for population genetics

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.*

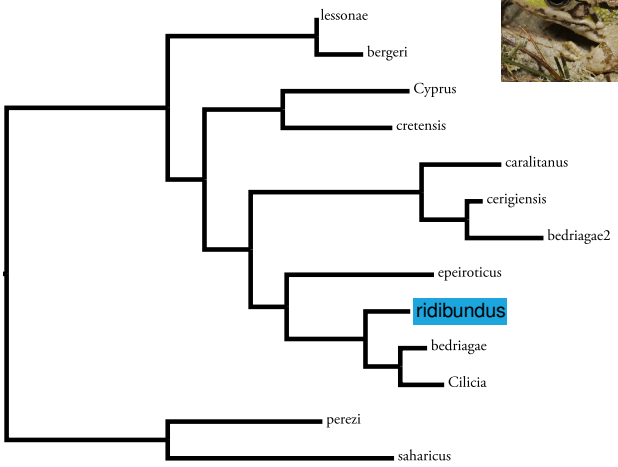
DERIVATIVES
co•a•les•cence |-ˈlesəns| noun
co•a•les•cent |-ˈlesənt| adjective

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

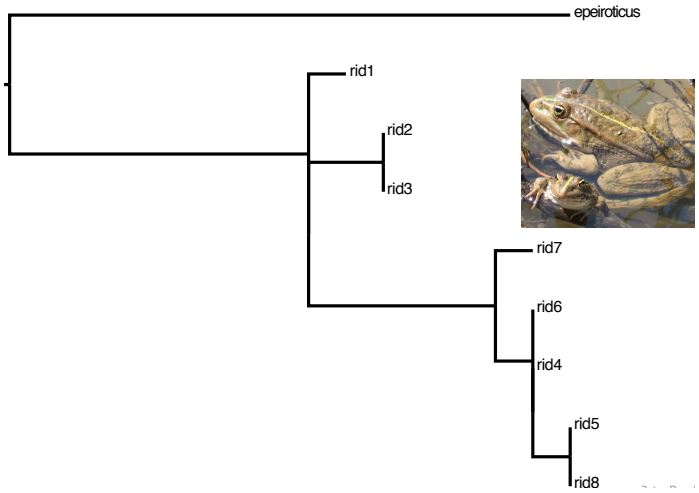
Species trees



Species trees



Tree of individuals of same species



Tree of individuals of same species

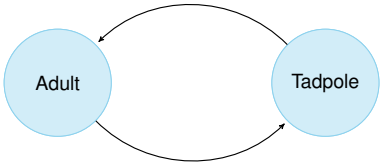
- Little resolution
- Tree building method should take into account that lineages are not independent of each other.
- Time-scale expected mutations

Interaction among individuals

Life cycle

Interaction among individuals

Life cyle



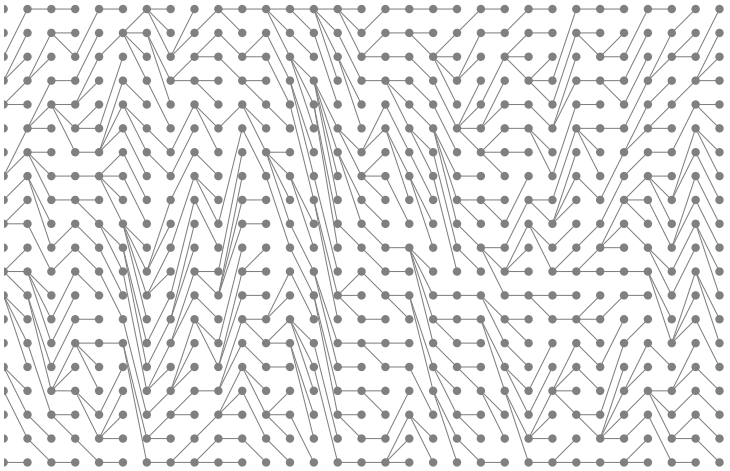
Wright-Fisher population model

- 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

As a result the individuals in generation n are a random draw from the previous generation $n-1$.

Population model

Wright-Fisher

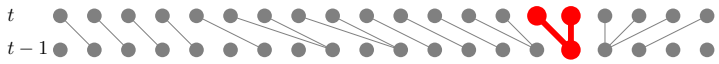


Population model

Wright

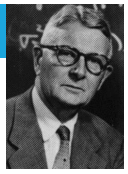
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in last generation is

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



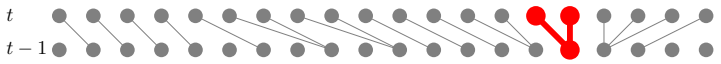
Population model

Wright



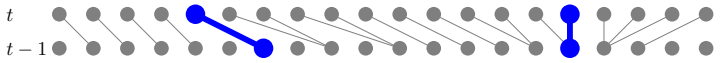
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in last generation is

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



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

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



Population model

Wright

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

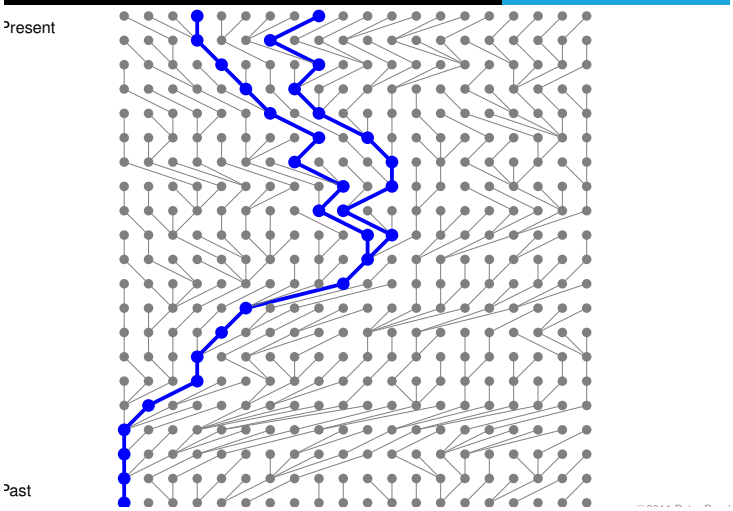
$$\mathrm{P}(\tau|N) = \left(1 - \frac{1}{2N}\right)^{\tau} \left(\frac{1}{2N}\right)$$

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:

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

Population model

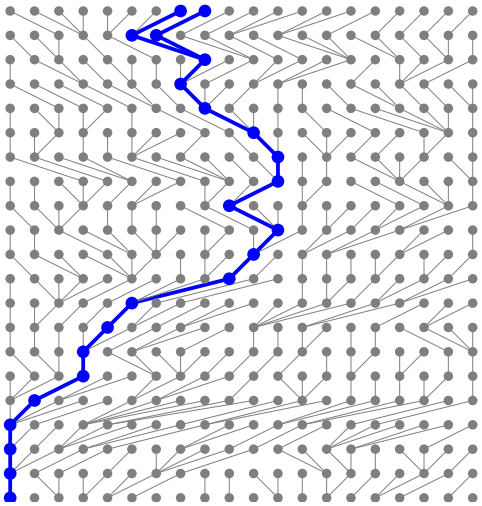
Wright-Fisher



Population model

Wright-Fisher

Present

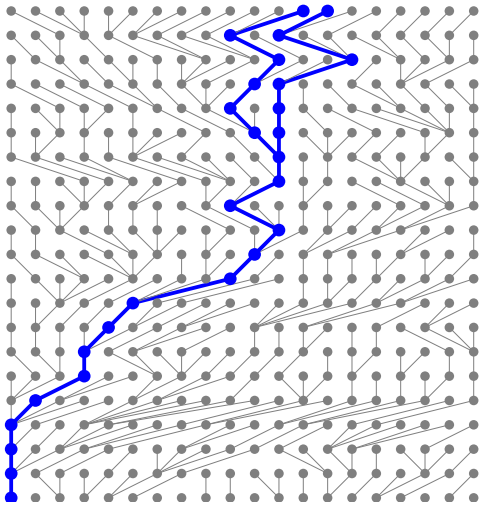


Past

Population model

Wright-Fisher

Present

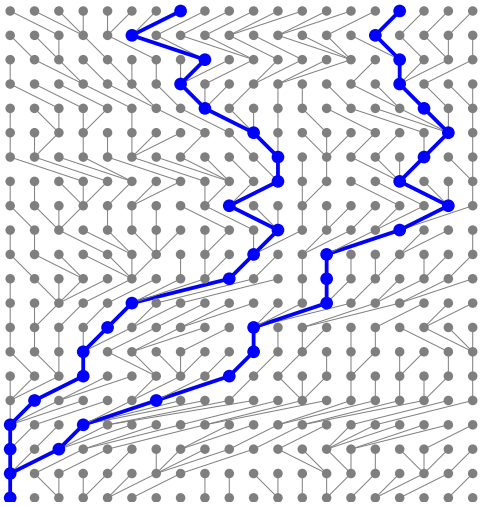


Past

Population model

Wright-Fisher

Present

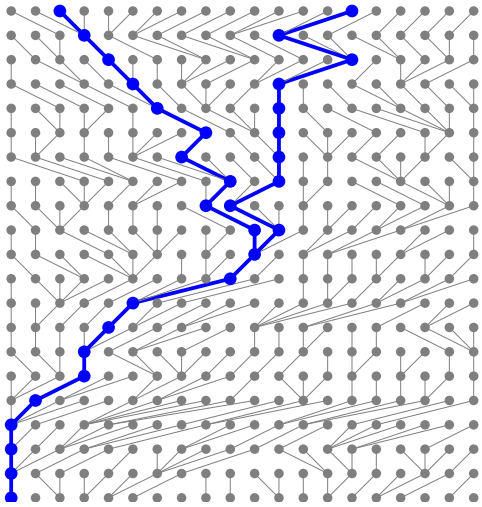


Past

Population model

Wright-Fisher

Present

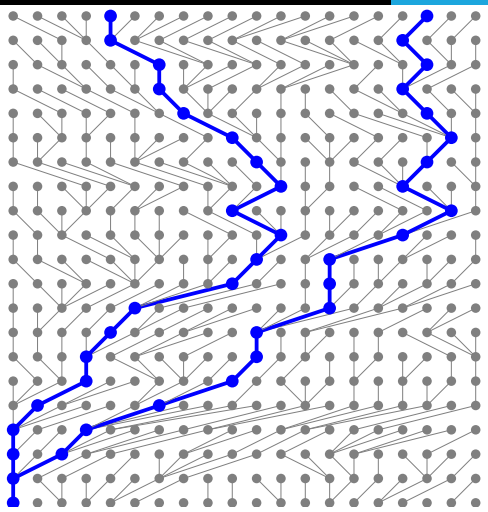


Past

Population model

Wright-Fisher

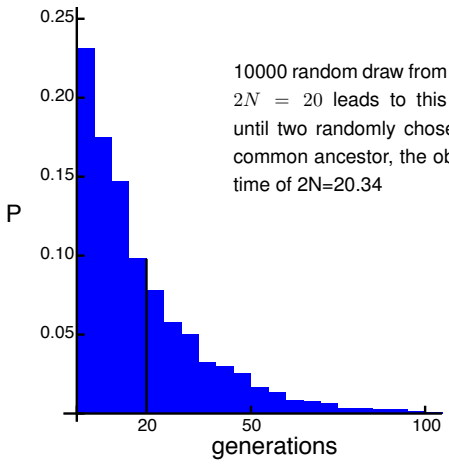
Present



Past

Probability Distribution

2N=20



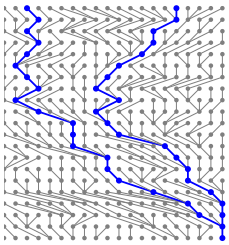
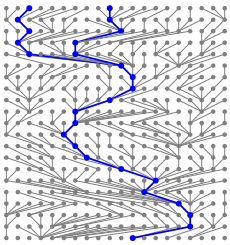
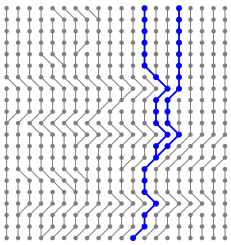
10000 random draw from a population with size $2N = 20$ leads to this distribution of times until two randomly chosen individuals have a common ancestor, the observed mean waiting time of 2N=20.34

Observations

Coalescence of two

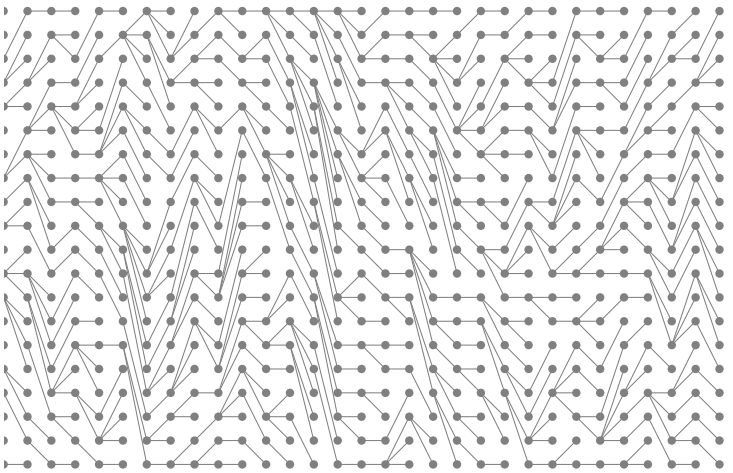
- For the time of coalescence in a sample of two we wait on average $2N$ generations assuming it is a Wright-Fisher population
- The geometric distribution used assumes discrete non-overlapping generations
- Real populations do not necessarily behave like a Wright-Fisher (the 'ideal' population)
- We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

Other population models

Wright-Fisher	Canning	Moran
		
$\sigma^2_{\text{offspring}} \simeq 1$	$\sigma^2_{\text{offspring}} = x$	$\sigma^2_{\text{offspring}} = \frac{2}{2N}$
$\mathbb{E}(\tau) = 2N$	$\mathbb{E}(\tau) = 2N/x$	$\mathbb{E}(\tau) = \frac{1}{2}(2N)^2$
generation time $g = 1$	$g = 1$	$g = 2N$
$\tau_{\text{relative}} = 1$	$\tau_{\text{relative}} = 1/x$	$\tau_{\text{relative}} = \frac{1}{2}(2N)$

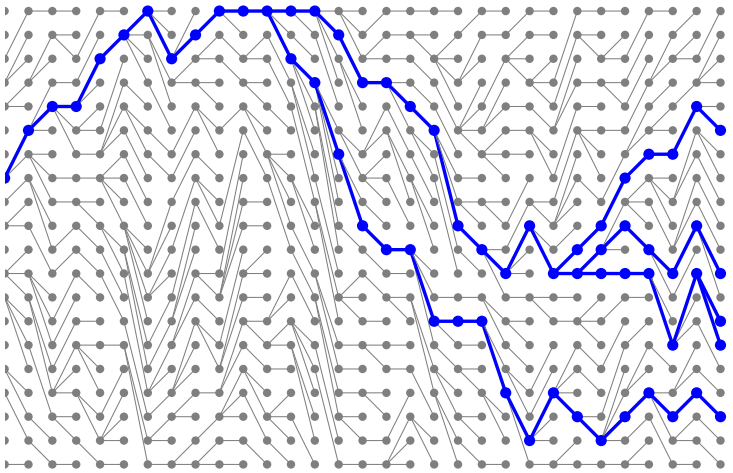
Sample larger than TWO

Wright-Fisher



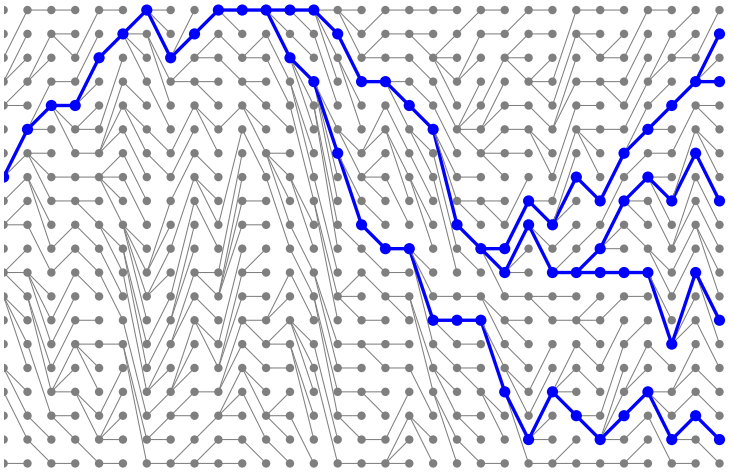
Sample larger than TWO

Wright-Fisher



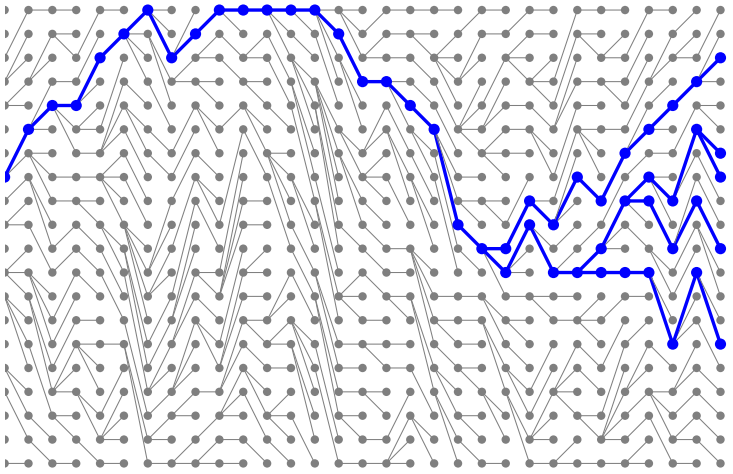
Sample larger than TWO

Wright-Fisher



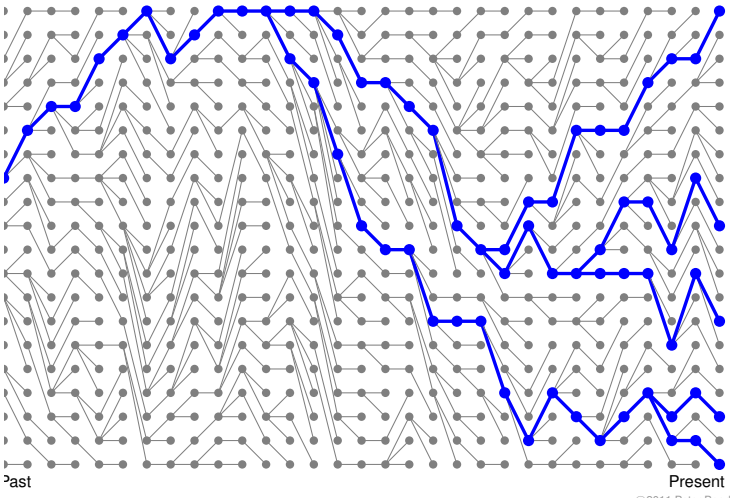
Sample larger than TWO

Wright-Fisher



Sample larger than TWO

Wright-Fisher



Samples larger than two

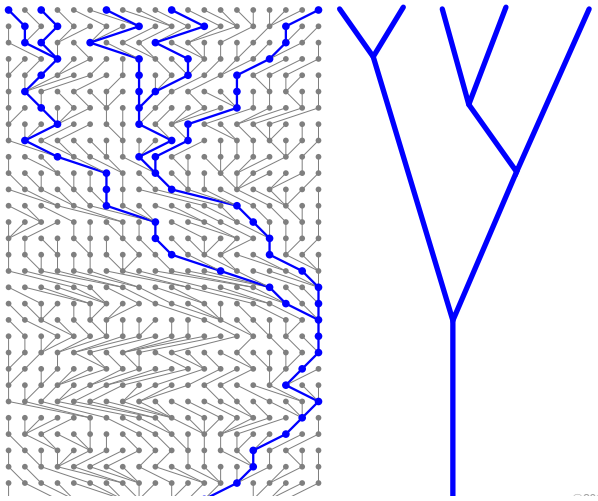
Sir J. F. C. Kingman described in 1982 the n-coalecent. He shows the behavior of a sample of size n, and its probability structure.

General findings:

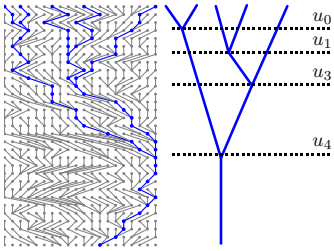
$$\text{coalescence rate} = \binom{n}{2} = \frac{n(n-1)}{2}$$

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

Samples larger than two

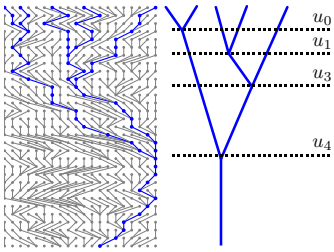


Samples larger than two



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.

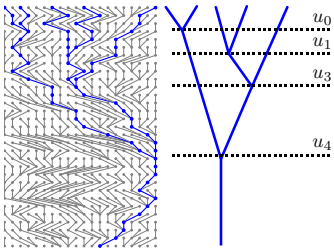
Samples larger than two



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.

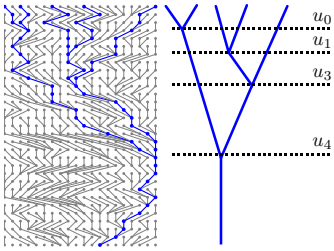
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:

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

with the scaled coalescence rate

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

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.

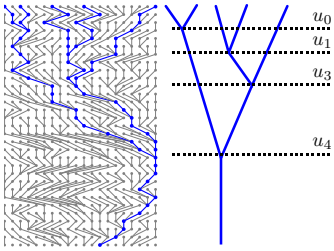
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:

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

with the scaled coalescence rate

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

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.

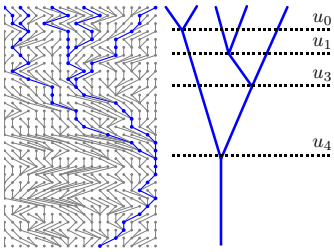
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:

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

with the scaled coalescence rate

$$\lambda = \binom{k}{2}\frac{1}{2N} + O(\frac{1}{N^2})$$

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:

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

with the scaled coalescence rate

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

Chance of coalescence in a particular generation

The chance that no lineages coalesce

$$1 - \left[1 \times (1-\frac{1}{2N}) \times (1-\frac{2}{2N}) \times ... \times (1-\frac{k-1}{2N})\right]$$

After some reshuffling

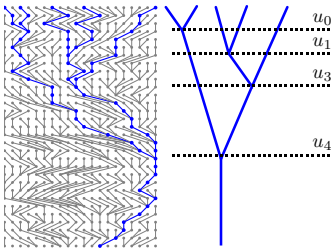
$$1 - \left[1 - \frac{k(k-1)}{2(2N)} + O(\frac{1}{N^2})\right] \simeq \frac{k(k-1)}{4N}$$

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

N	0	1	>1
100	0.79560747	0.18744678	0.01694575
1000	0.97771632	0.02209806	0.00018562
10000	0.99775217	0.00224595	0.00000187

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

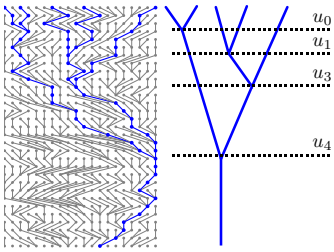


If we know the relationships among all individuals we can calculate the probability for each of the particular coalescence event.

With probability $\mathrm{P}(u_j|N)$ a coalescent event happens, but we still do not know which pair of individuals is involved, we pick a random pair with probability

$$\frac{1}{\binom{k}{2}},$$

Samples larger than two



If we know the relationships among all individuals we can calculate the probability for each of the particular coalescence event.

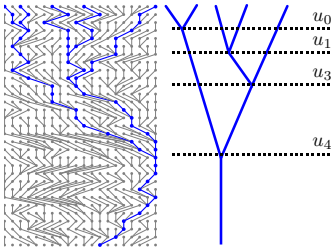
With probability $\mathrm{P}(u_j|N)$ a coalescent event happens, but we still do not know which pair of individuals is involved, we pick a random pair with probability

$$\frac{1}{\binom{k}{2}},$$

therefore

$$\mathrm{P}(u_j|N, i_1, i_2) = \mathrm{P}(u_j|N)\frac{1}{\binom{k}{2}}$$

Samples larger than two



If we know the relationships among all individuals we can calculate the probability for each of the particular coalescence event.

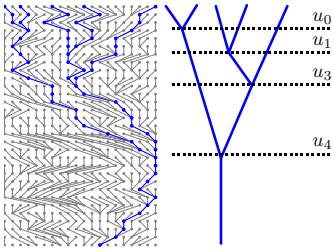
With probability $\mathrm{P}(u_j|N)$ a coalescent event happens, but we still do not know which pair of individuals is involved, we pick a random pair with probability

$$\frac{1}{\binom{k}{2}},$$

therefore

$$\mathrm{P}(u_j|N,i_1,i_2) = \left[e^{-u_j\frac{k(k-1)}{4N}}\frac{k(k-1)}{4N}\right]\frac{2}{k(k-1)}$$

Samples larger than two



If we know the relationships among all individuals we can calculate the probability for each of the particular coalescence event.

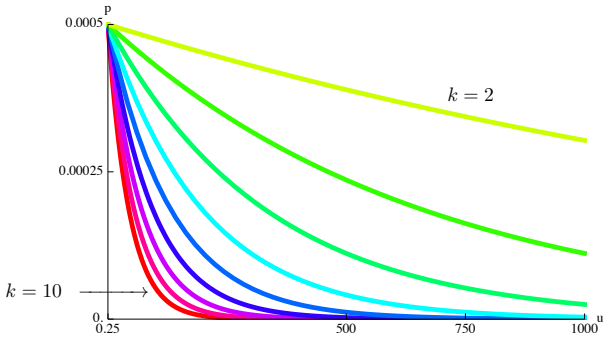
With probability $\mathrm{P}(u_j|N)$ a coalescent event happens, but we still do not know which pair of individuals is involved, we pick a random pair with probability

$$\frac{1}{\binom{k}{2}},$$

therefore

$$\mathrm{P}(u_j|N,i_1,i_2) = e^{-u_j\frac{k(k-1)}{4N}}\frac{2}{4N}$$

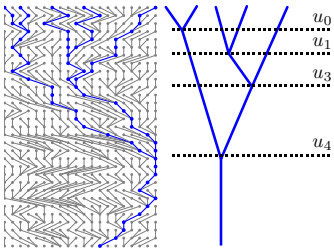
Samples larger than two



$$\mathrm{P}(u_j|N,i_1,i_2) = e^{-u_j\frac{k(k-1)}{4N}}\frac{2}{4N}$$

Samples larger than two

the coalescent

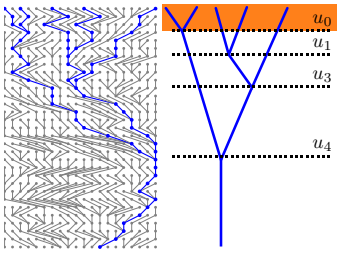


We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

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

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:

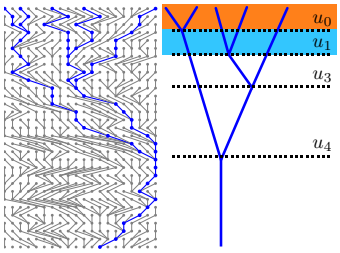
$$\mathrm{P}(G|N) = \mathrm{P}(u_0|N,i_1,i_2)$$

$$\times$$

)

Samples larger than two

the coalescent



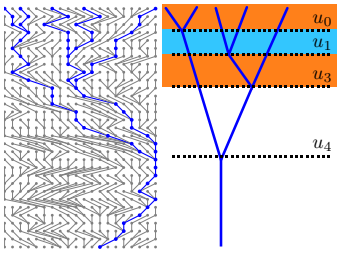
We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

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

$$\times \mathrm{P}(u_1|N,i_3,i_4)$$

Samples larger than two

the coalescent

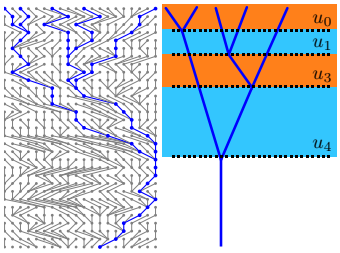


We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

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

Samples larger than two

the coalescent

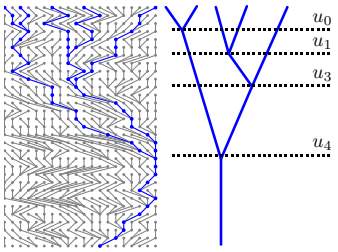


We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

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

Samples larger than two

the coalescent



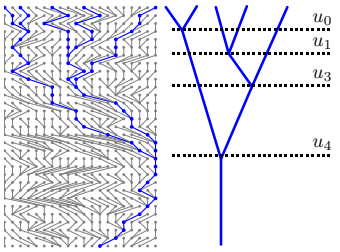
We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

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

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

Samples larger than two

the coalescent



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

The expectations of the probability is the sum of the expectations for each interval. Each interval has expectation

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

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

$$\mathbb{E}(\tau_{\mathrm{MRCA}}) = \sum_{j=0}^{J} \frac{4N}{k_j(k_j-1)}$$

where J is the number of time intervals u_j. In the limit this is

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

What is it good for?

Coalescence

If we know the genealogy G with certainty then we can 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?

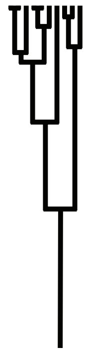
Using an oracle

If we know the genealogy G with certainty then we can 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?

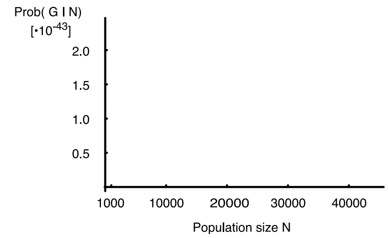
Using an oracle

If we know the genealogy G with certainty then we can 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.



Population size estimation

using an oracle

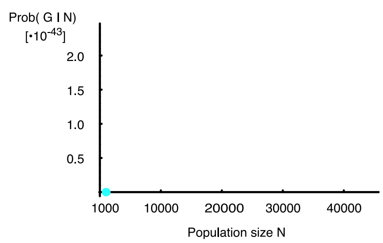


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle

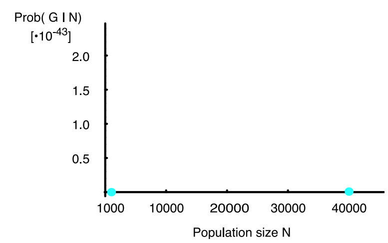


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle

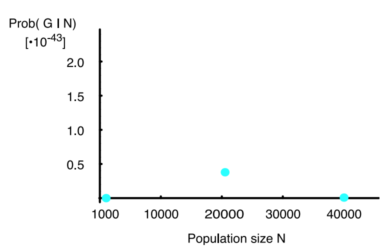


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle

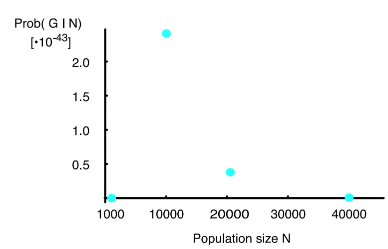


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle

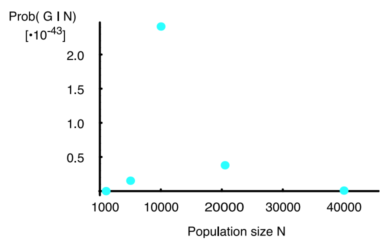


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle

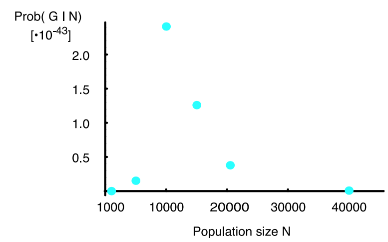


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle

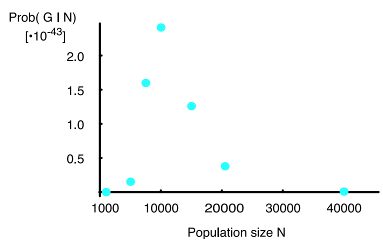


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle

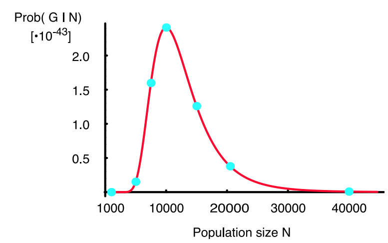


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle

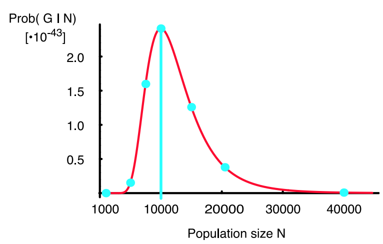


If an oracle gives us the true relationship tree G then we can calculate the population size N.

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

Population size estimation

using an oracle



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

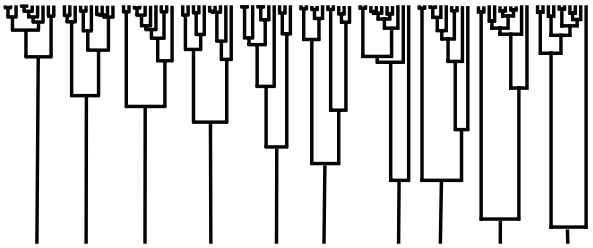
$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right)\frac{2}{4N}$$

Population size estimation

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

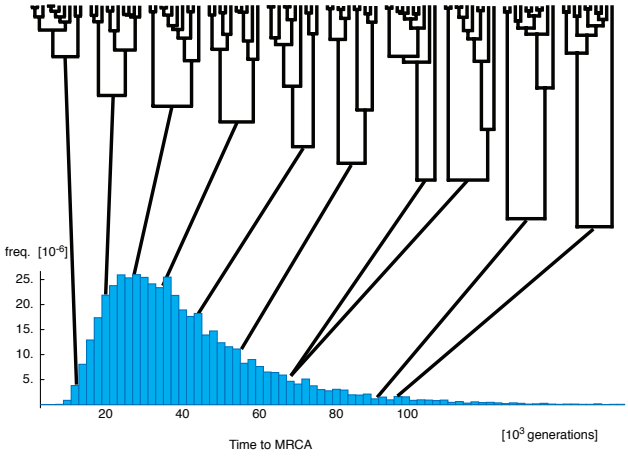
- There is no oracle to gives us clear information!
- We do not record genealogies, our data is genetic!
- What about the variability of the coalescence process?

Variability of the coalescent process



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

Variability of the coalescent process



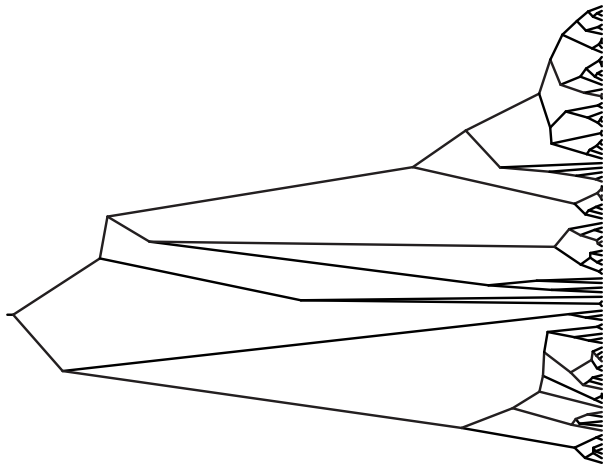
Kingman's n-coalescent is an approximation

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

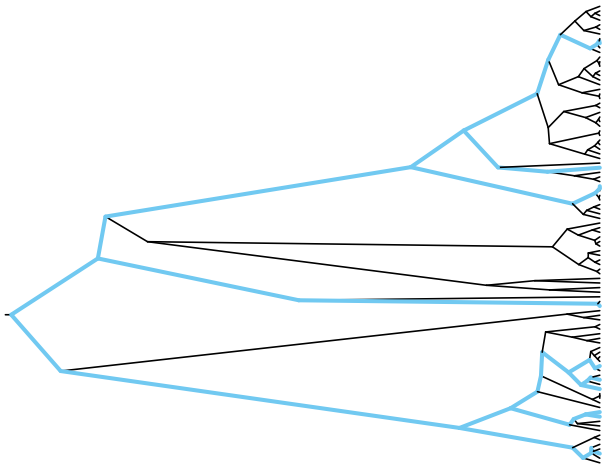
$$n << N$$

- 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.

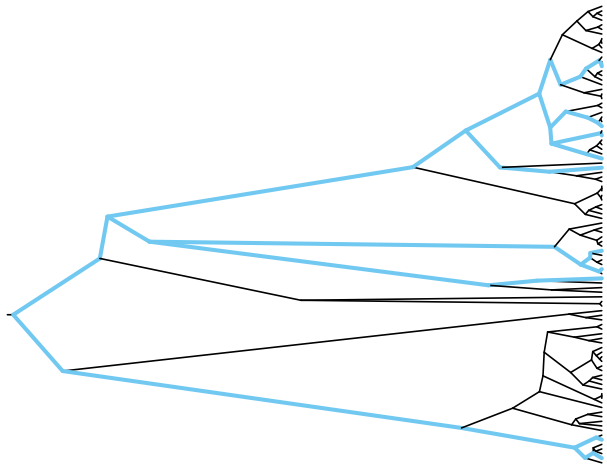
Kingman's n-coalescent is an approximation



Kingman's n-coalescent is an approximation

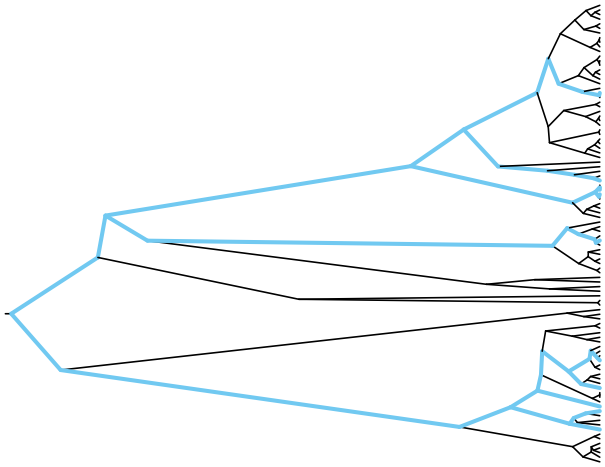


Kingman's n-coalescent is an approximation



Kingman's n-coalescent is an approximation

Sample size



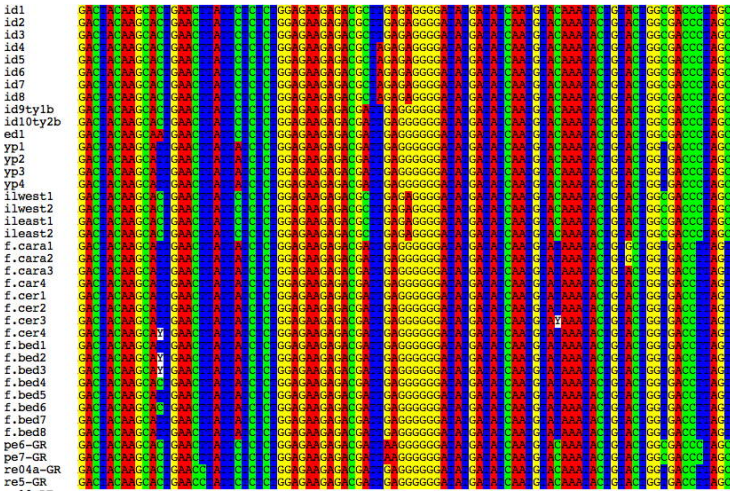
Observations

Coalescence

- 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:



Genealogy and data

our data looks like this:

Genetic data and the coalescent

- Finite populations loose 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 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 and the following settings 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 our N should be marked and we call it the *effective* population size N_e, and consider Θ the mutation-scaled effective population size.

Historical humpback whale population size

Historical humpback whale population size

using the data 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
$N_{♀} = \frac{\Theta}{2\mu}$	31,854	with $\mu = 2.0\times10^{-8}\mathrm{bp}^{-1}\mathrm{year}^{-1}$ and a generation time of 12 years
$N_e = N_{♀} + N_{♂}$	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)

Genetic data and the coalescent

Watterson's θ

Using the infinite sites model we use the number of variable sites S 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

Construction of a versatile estimator

Modern inference

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

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

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

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

where $\mathrm{p}(\mathrm{D}|\Theta)$ is the likelihood of the parameters.

(almost) Felsenstein equation

aka Likelihood calculation

$$p(D|\boldsymbol{\Theta}, G) = \mathrm{p}(\mathrm{G}|\boldsymbol{\Theta})\mathrm{p}(\mathrm{D}|\mathrm{G})$$

$p(G|\boldsymbol{\Theta})$

The probability of a genealogy given parameters.

$p(\mathrm{D}|\mathrm{G})$

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

Felsenstein equation

aka Likelihood calculation

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

$p(G|\Theta)$ 

The probability of a genealogy given parameters.

$\mathrm{p(D|G)}$ 

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

Felsenstein equation

aka Likelihood calculation

$$p(D|\Theta) = \sum_G \mathrm{p(G|\Theta)p(D|G)dG}$$

$p(G|\Theta)$ 

The probability of a genealogy given parameters.

$\mathrm{p(D|G)}$ 

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

Problem with integration formula

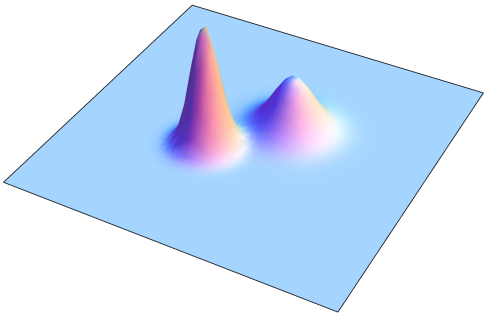
Tips	Labeled histories
3	3
4	18
5	180
6	2700
7	56700
8	1587600
9	57153600
10	2571912000
15	6958057668962400000
20	564480989588730591336960000000
30	436846661310306951246468019862076389144064000000000000
40	3027333829948007356546303364551457200042939432053862501707 8...
50	3.28632×10^{112}
100	1.37416×10^{284}

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

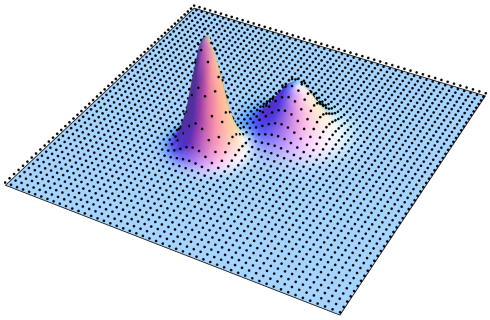
The number of possible genealogies is very large and for realistic data sets, programs need to use Markov chain Monte Carlo methods.

For reference: Florida Lotto 6 out of 53: 22957480

Naive integration approach



Naive integration approach



Markov chain Monte Carlo

MCMC

Metropolis recipe

0. first state

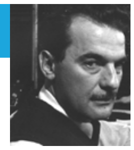
1. perturb old state and calculate probability of new state

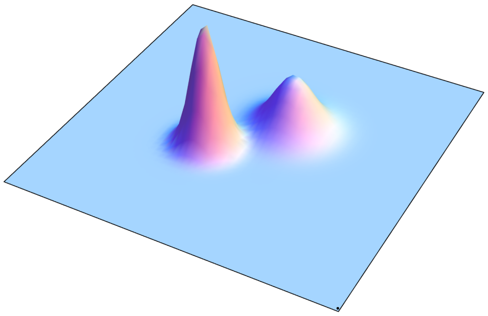
2. test if new state is better than old state: accept if ratio of new and old is larger than a random number between 0 and 1.

3. move to new state if accepted otherwise stay at old state

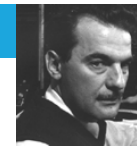
4. go to 1

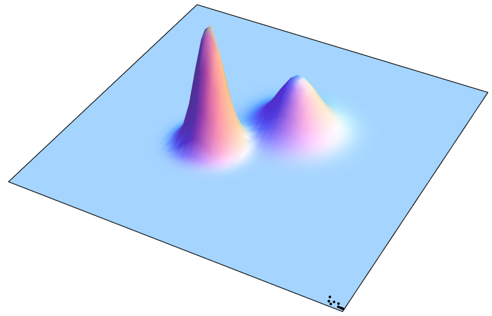
Metropolis-Hastings algorithm



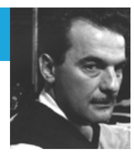


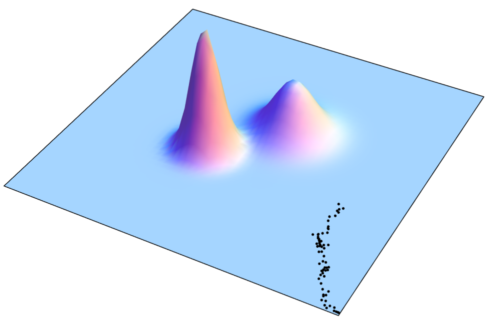
Metropolis-Hastings algorithm



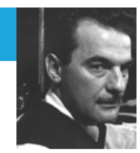


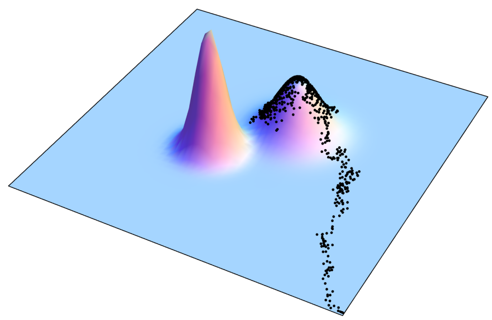
Metropolis-Hastings algorithm



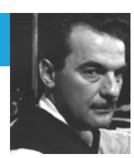


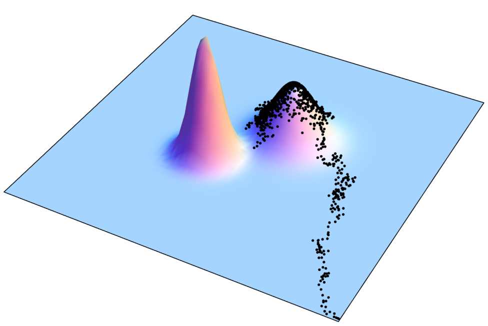
Metropolis-Hastings algorithm



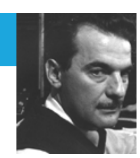


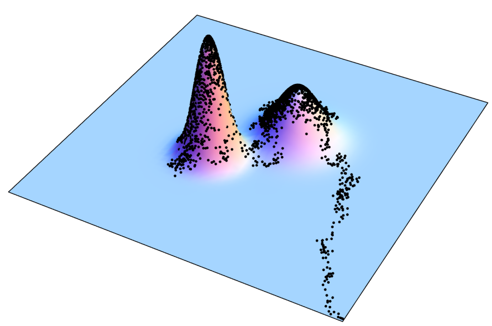
Metropolis-Hastings algorithm



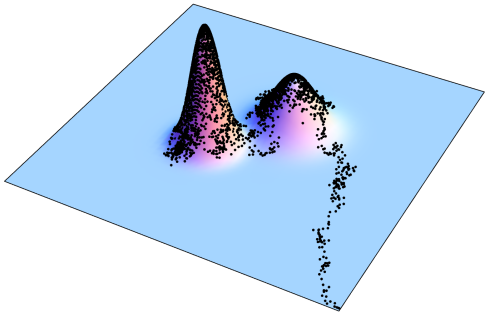


Metropolis-Hastings algorithm

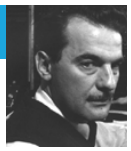


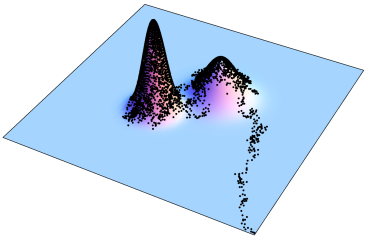


Metropolis-Hastings algorithm



Metropolis-Hastings algorithm





- *Irreducibility*: the Markov chain must be able to reach all interesting parts of the distribution.
- *Recurrence*: all interesting parts must be reached (in principle) infinitely often if the chain is run infinitely long.
- *Convergence*: the sample mean must converge to the expectation.

A closer look at the MCMC approximation

ML

We want to approximate the likelihood

$$L(\Theta) = p(D|\mathbf{\Theta}) = \int_G \mathrm{p(G|\mathbf{\Theta})p(D|G)dG}$$

by

$$L'(\Theta) \simeq \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{p(G_i|\Theta)}{p(G_i|\Theta_0)} \qquad G_i \sim p(G|\Theta_0)p(D|G)$$

where ℓ is the length of the run and

$$L'(\Theta) = \frac{L(\Theta)}{L(\Theta_0)}$$

but $L(\Theta_0)$ is not known $\rightarrow$ we calculate a scaled likelihood. With a Bayesian framework or a maximum likelihood framework this still leads to the same parameter estimates.

$$\underset{\Theta}{\operatorname{argmax}}\, L(\Theta) = \underset{\Theta}{\operatorname{argmax}}\, \frac{L(\Theta)}{L(\Theta_0)}$$

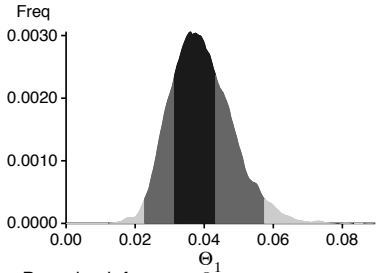
Because L' is scaled, do not be alarmed to see positive $\ln L'$.

Inference of population size

Nuu-Chah-Nulth

Inference of population size

Nuu-Chah-Nulth



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 84112; 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)]

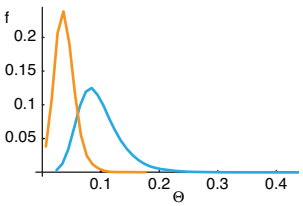
Bayesian inference: $\hat{\Theta} = 0.036$

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 $N = 8328$.

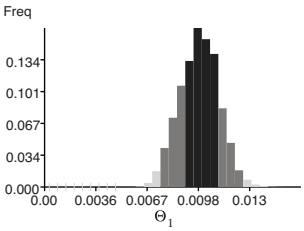
Multilocus locus inference

Fairy wren



Multilocus locus inference

Fairy wren



Extensions of the basic coalescence

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

Extensions of the basic coalescent

Growth

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

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

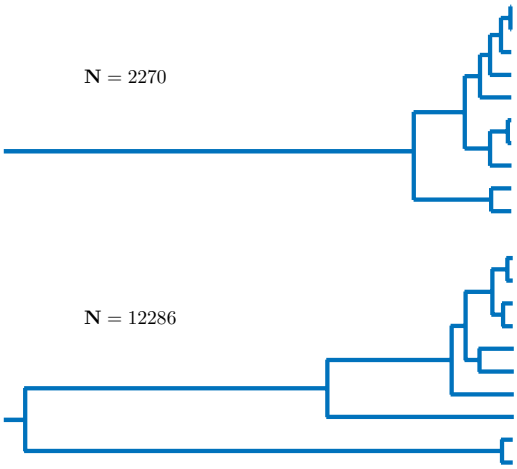
$$N_t = N_0 e^{-rt}$$

- 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

Growth



Extensions of the basic coalescent

Growth

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

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

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

- 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

Growth

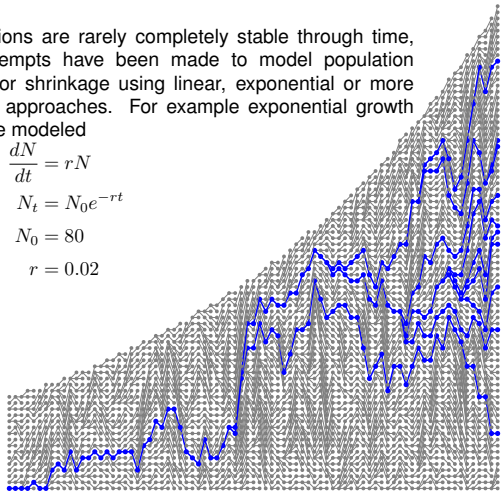
Populations are rarely completely stable through time, and attempts have been made to model population growth or shrinkage using linear, exponential or more general approaches. For example exponential growth could be modeled

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

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

$$N_0 = 80$$

$$r = 0.02$$



Extensions of the basic coalescent

Growth

For constant population size we found

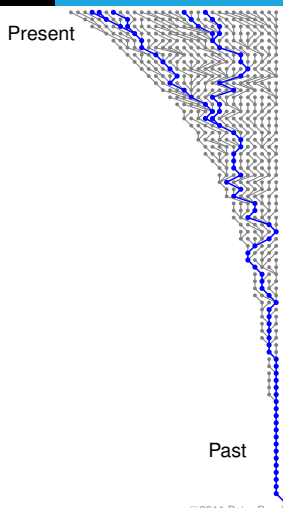
$$p(G|\Theta) = \prod_j e^{-u_j \frac{k(k-1)}{\Theta}} \frac{2}{\Theta}$$

Using the exponential growth formula we get

$$N_t = N_0 e^{-rt} \xrightarrow{\Theta = 4N\mu} \Theta_t = \Theta_0 e^{-(r/\mu)t}$$

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

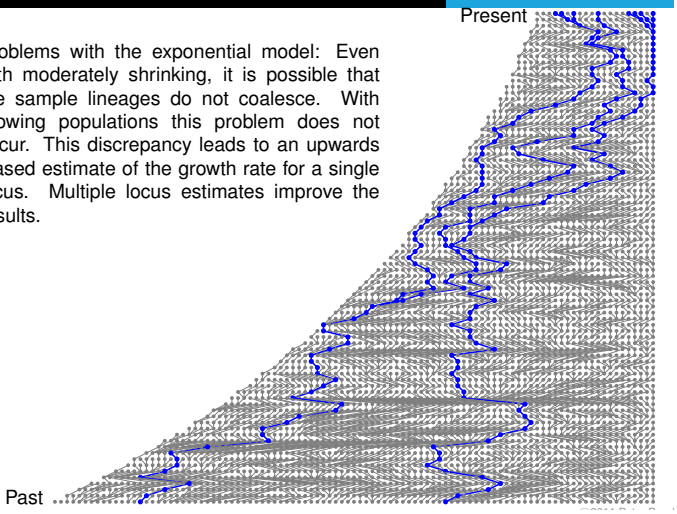
$$p(G|\Theta_0, g) = \prod_j e^{-(t_j - t_{j-1})\frac{k(k-1)}{\Theta_0 e^{-gt}}} \frac{2}{\Theta_0 e^{-gt}}$$



Extensions of the basic coalescent

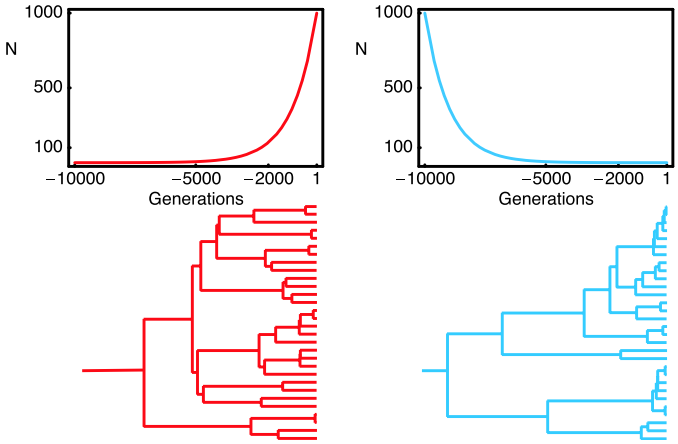
Growth

Problems with the exponential model: Even with moderately shrinking, 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.



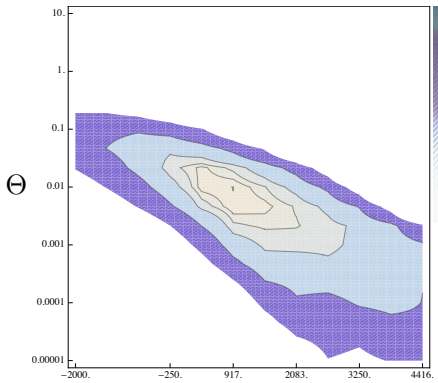
Extensions of the basic coalescent

Growth



Grow-A-Frog

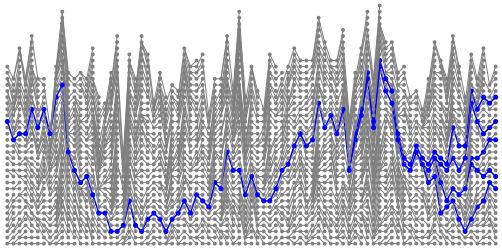
Expansion of *Pelophylax lessonae* in Europe



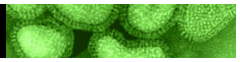
Extensions of the basic coalescent

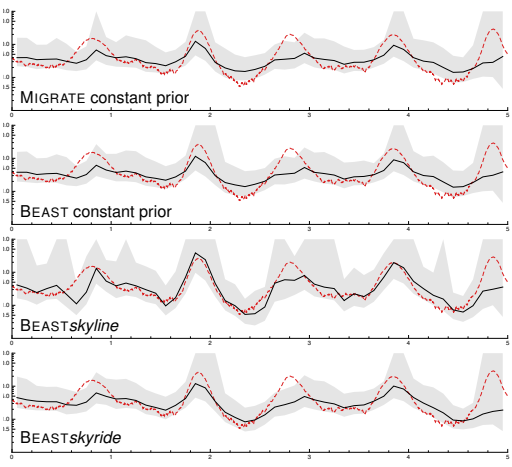
Fluctuations

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



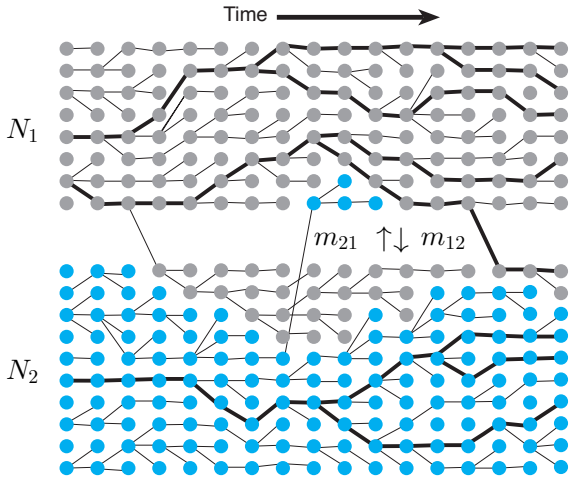
Extensions of the basic coalescent



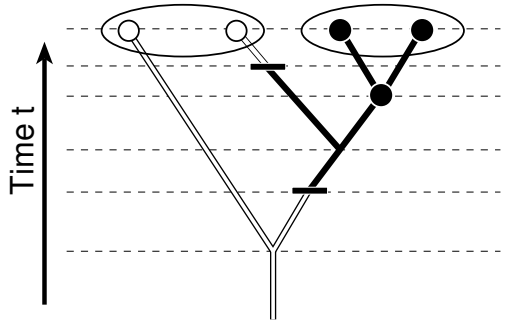


Comparison of the skyline plots of simulated influenza dynamics analyzed by MIGRATE and BEAST. The x-axis is the time in years and the y-axis is effective population size. The data are sequences from 250 individuals sampled at regular intervals over 5 years. The dashed curve is the actual population size deduced from the true genealogy; black lines are the mean results of MIGRATE or BEAST; gray area is the 95% credibility interval. BEAST *skyline* matches the actual population size better than all other methods. Simulation and graphs courtesy of Trevor Bedford, University of Michigan (unpublished).

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

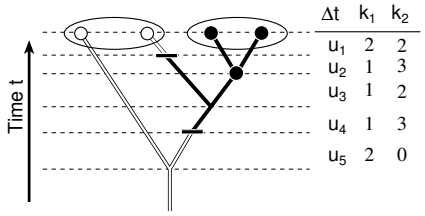
Extensions of the basic coalescent

Point probability

$$p(G|\underline{\mathbf{N}},\underline{\mathbf{m}}) = \prod_{t=1}^{T} \exp\left(-u_t\left[\sum_i^d \left(\frac{k_{ti}[k_{ti}-1]}{4N_i} + k_{ti}\sum_{j\neq i}^d m_{ji}\right)\right]\right) \delta_t$$

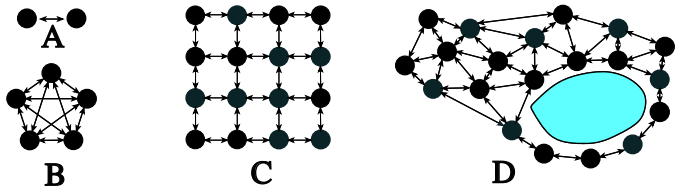
Waiting time

$$\delta_t = \begin{cases} \frac{1}{2N_i} & \text{if event is a coalescence in population } i \\ m_{ij} & \text{if event is a migration from } j \text{ to } i \end{cases}$$

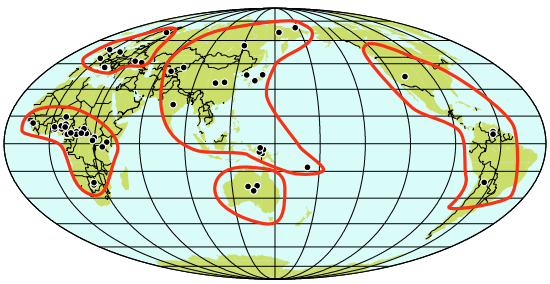


Δt	k_1	k_2
u_1	2	2
u_2	1	3
u_3	1	2
u_4	1	3
u_5	2	0

Structured populations

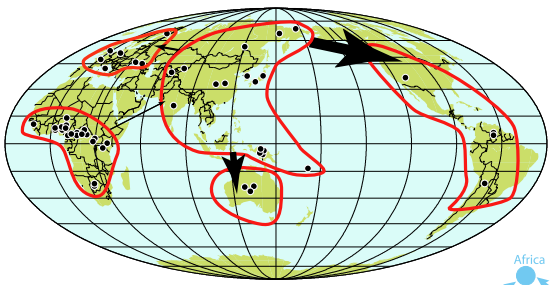


Structured populations

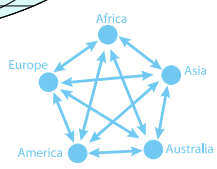


A total of 53 complete mtDNA sequences (∼ 16 kb):
Africa: 22, Asia: 17, Australia: 3, America: 4, Europe: 7.
Assumed mutation model: F84+Γ

Human migration

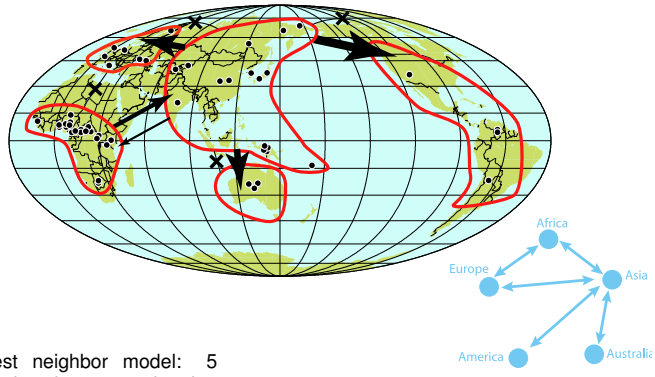


Full model: 5 population sizes + 20 migration rates



Structured populations

Migration

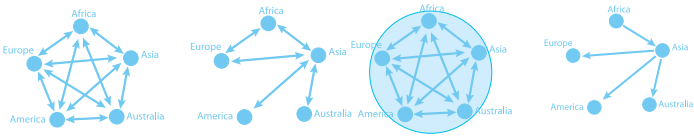


Nearest neighbor model: 5 population sizes + 10 migration rates

Structured populations

Migration

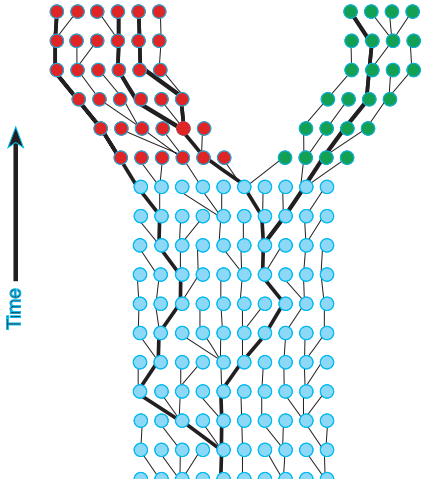
Model order and probability using Bayes factors



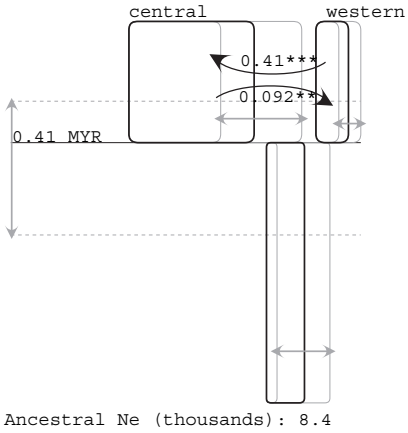
10^{-29} 10^{-29} 10^{-15} 1.0

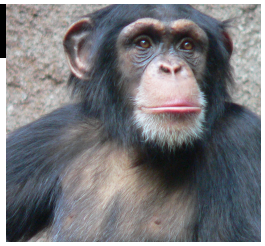
Extensions of the basic coalescent

Population splitting



Population splitting

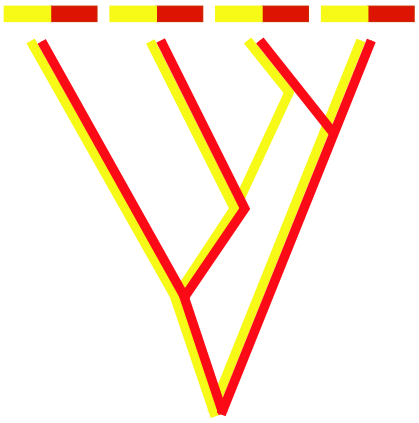




IM: isolation with migration; co-estimation of divergence parameters, population sizes and migration rates. Not all datasets can separate migration from divergence, and multiple loci are helpful.

Extensions of the basic coalescent

Recombination



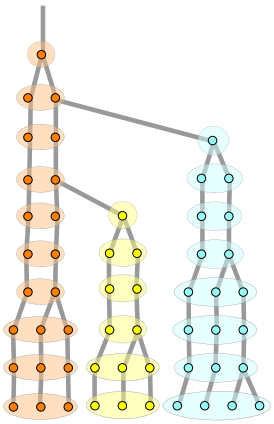
Extensions of the basic coalescent

Recombination

Example using LAMARC

Age of mutations

Genetree



Age of mutations

Genetree

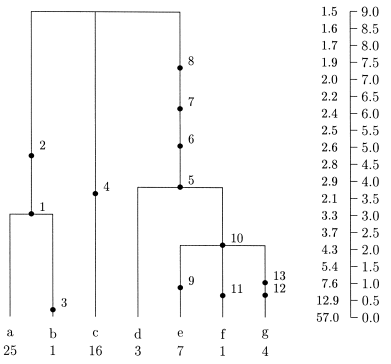
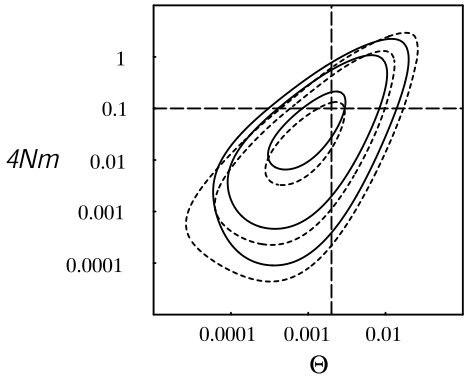


FIG. 3. *Melanesian β-globin tree. Time in units of* 100,000 *years.*

MIGRATE versus GENETREE

Comparison



Robustness of the coalescence

Population model

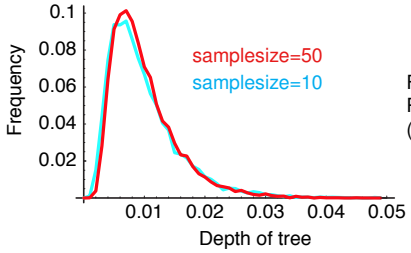
Violating assumptions

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

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

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.

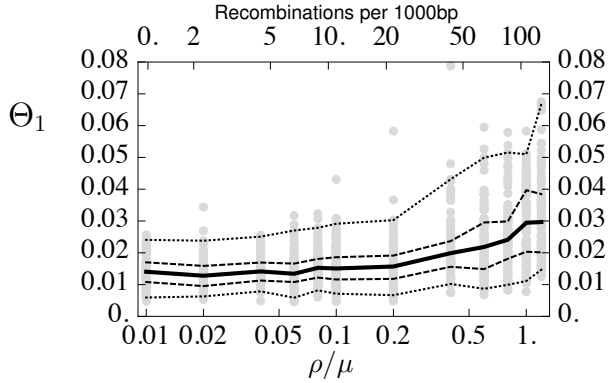


Felsenstein (2005)
Pluzhnikov and Donnelly (1996)

- Currently little is known whether this is true for inference migration rates, although we know that even with highly variable markers, such as microsatellites, we will not need huge numbers of individuals.

Ignored recombination

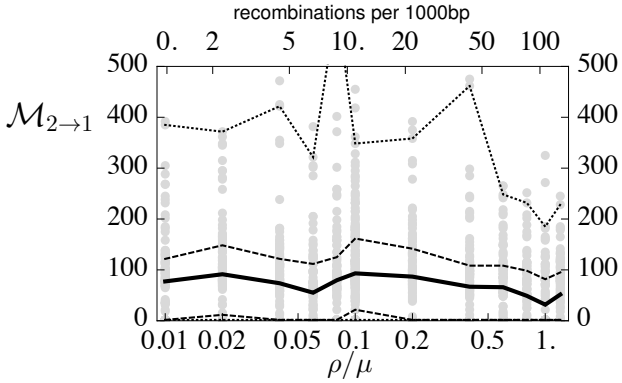
Violating assumptions



Effect of unrecognized recombination on estimates of mutation-scaled effective population size Θ and mutation-scaled immigration rates $\mathcal{M}$. Data for two populations was simulated using equal populations sizes of $\Theta = 0.01$ and symmetric $\mathcal{M} = 100$ using the software Netrecodon (Arenas and Posada 2010). A range of recombination rates were used to simulate 1000 sites for a total of 40 individuals so that the ratio of recombination rate to mutation rates, R, covers the range of 0.01 to 1.0, these settings resulted in sequences that had between 0 to more than 200 countable recombination events; the average number of recombination events is *#rec*/1000bp. For each R 100 datasets were simulated and analyzed with MIGRATE 3.2.13. The thick black lines are averages of the posterior modes (gray dots), the dark dashed line is the 50% support interval around the average and the fine dashed line is the 95% support interval.

Ignored recombination

Violating assumptions



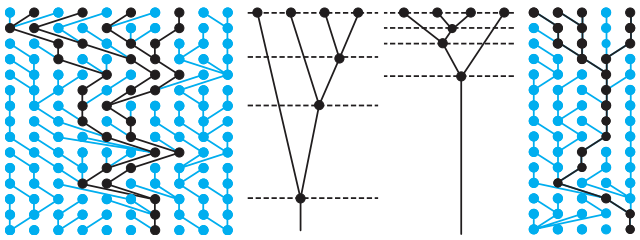
Effect of unrecognized recombination on estimates of mutation-scaled effective population size Θ and mutation-scaled immigration rates $\mathcal{M}$. Data for two populations was simulated using equal populations sizes of $\Theta = 0.01$ and symmetric $\mathcal{M} = 100$ using the software Netrecodon (Arenas and Posada 2010). A range of recombination rates were used to simulate 1000 sites for a total of 40 individuals so that the ratio of recombination rate to mutation rates, R, covers the range of 0.01 to 1.0, these settings resulted in sequences that had between 0 to more than 200 countable recombination events; the average number of recombination events is *#rec*/1000bp. For each R 100 datasets were simulated and analyzed with MIGRATE 3.2.13. The thick black lines are averages of the posterior modes (gray dots), the dark dashed line is the 50% support interval around the average and the fine dashed line is the 95% support interval.

Average of parameters over long time

Coalescent-based methods

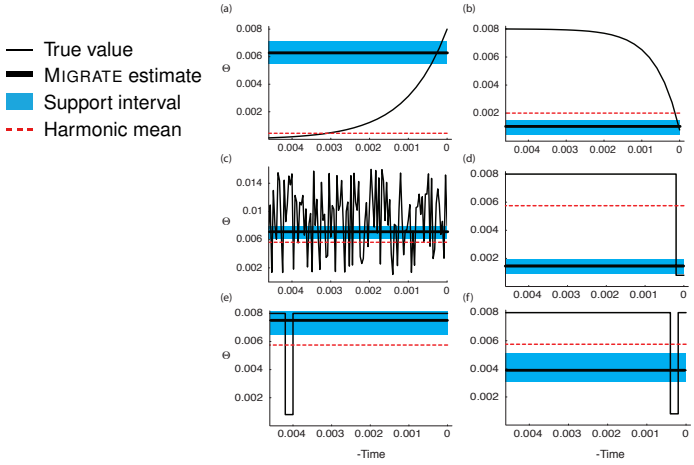
Researchers from the frequency-based camp claim that the coalescence-based methods are working on an evolutionary time-scale and therefore are not really usable in a conservation genetics or management context.

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

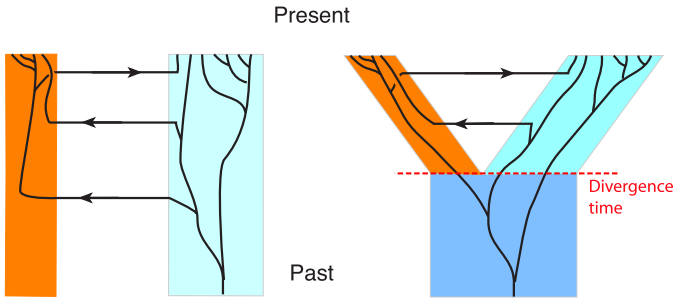


Average of parameters over long time

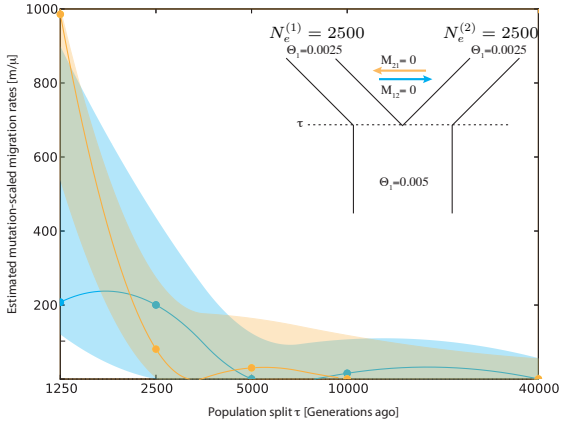
Coalescent-based methods



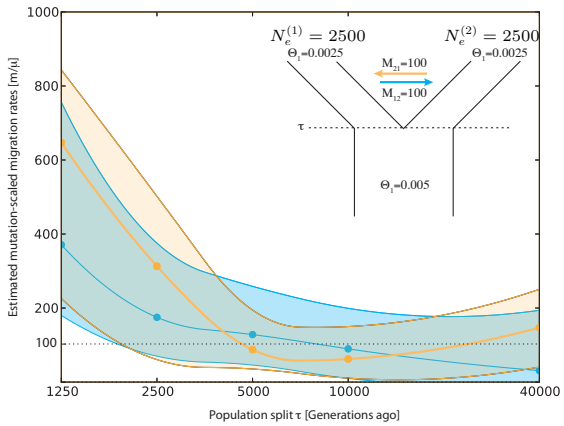
Ignored divergence



Ignored divergence



Ignored divergence



Ignored selection

The standard coalescent assumes neutral mutations and also exchangeable number of offspring, loci under selection will violate both tenets.

- A new mutation that has a positive effect will replace some of the variability present in the population. All linked sites will suffer a drop in effective population size.
- A new mutation that has a negative effect and will be most likely removed , also resulting in a reduction of variability (and population size)

This is used in genome-wide selection scans, but influence of population growth, population structure on such estimates are not studied.

Software

Program	Maximal # populations	Population sizes	Change through time	Migration rates	Divergence	Recombination rate	Serial Sampling
MIGRATE	>20	●	○	●	-	-	●
LAMARC	>20	●	●	●	-	●	-
IM	>10	●	○	●	●	-	-
BEAST	1	●	●	-	○	-	●
GENETREE	>10	●	●	●	-	?	-

Outlook

- Monday: MIGRATE; use to compare different migration hypotheses using Bayes factors. Date set size and speed.
- Tuesday: LAMARC; general use and recombination estimation

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; Sigurğardóttir 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.
Inference: