Introduction to Population Genetics

Fernando Racimo

Adelaide, January 2018

Today

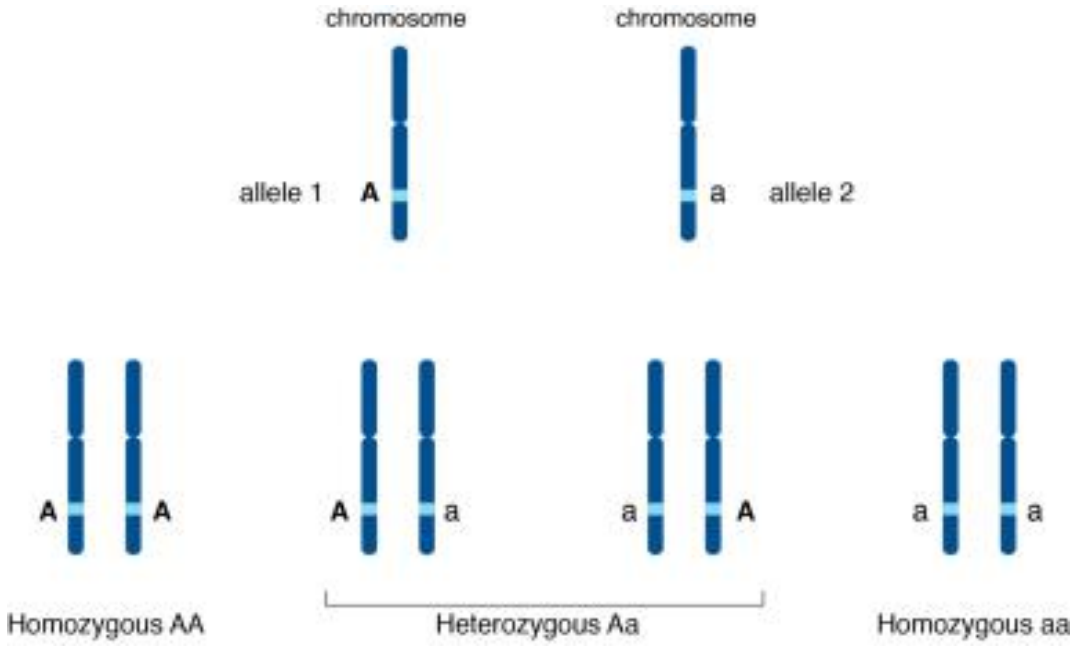
- Terminology
- Wright-Fisher model and genetic drift
- Kingman's coalescent as an approximation to the W-F model
- The infinite sites model
- Population size changes

Today

- **Terminology**
- Wright-Fisher model and genetic drift
- Kingman's coalescent as an approximation to the W-F model
- The infinite sites model
- Population size changes

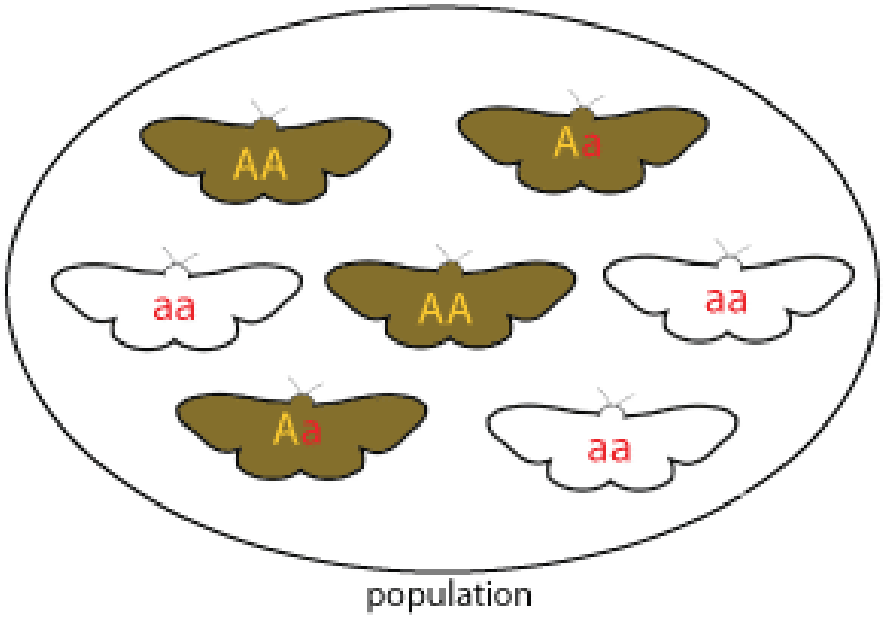
Some terminology

- **Allele**: one of two or more alternative forms of a genetic locus that reside at the same place on a chromosome.
- **Genotype**: the set of alleles present at a genetic locus in an organism (two alleles if the organism is diploid).



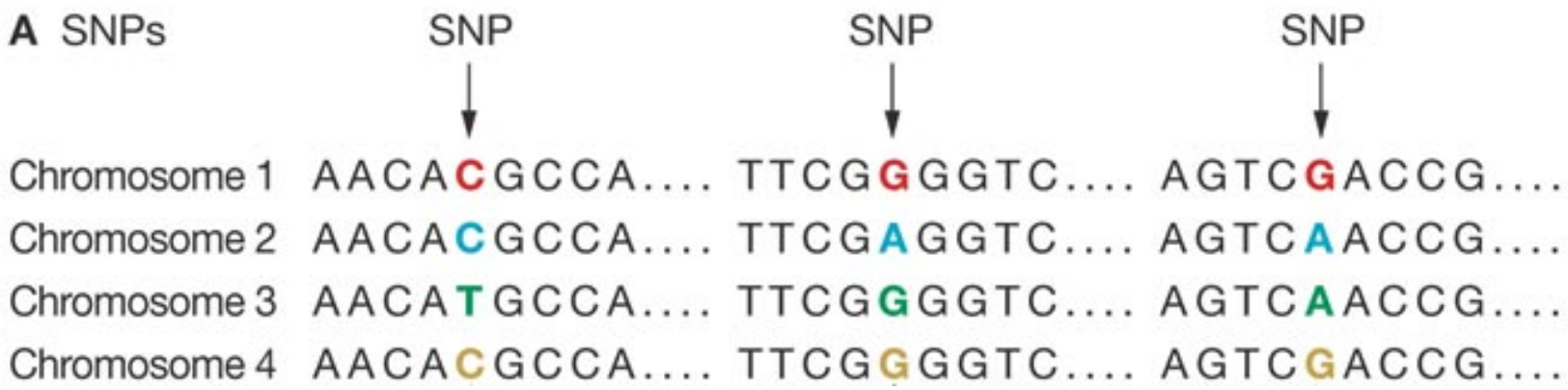
Some terminology

- **Allele frequency**: relative frequency of an allele in a population, expressed as the fraction of all chromosomes that carry that allele.
- **Genotype frequency**: relative frequency of a genotype in a population, expressed as the fraction of all individuals that carry that genotype.



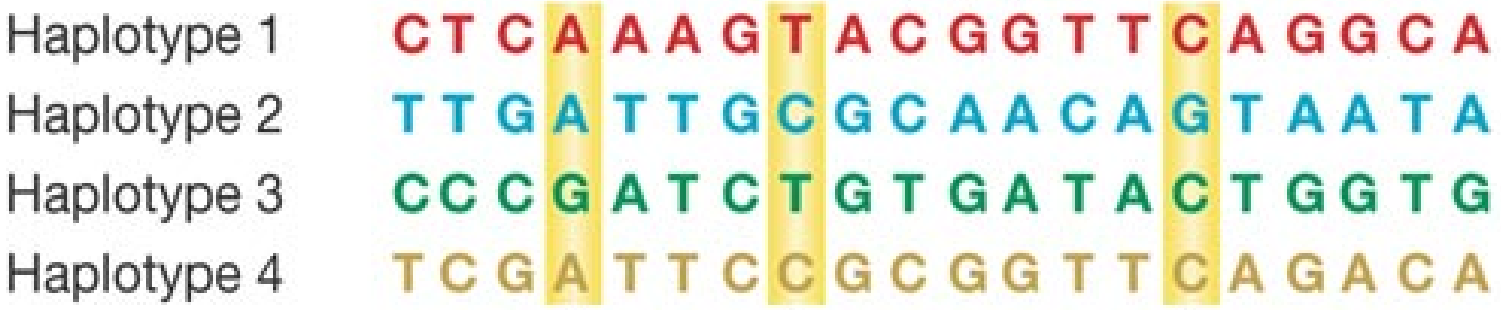
Some terminology

- **Polymorphism**: a site with two or more alleles segregating in a population
- **SNP**: single nucleotide polymorphism - a polymorphism in which a single nucleotide (A, C, T or G) differs among different members of the population
- **InDel**: an insertion or a deletion in the genome.
- **Polymorphisms** can be SNPs, InDel variants or larger structural variants (translocations, copy number variants, etc.)



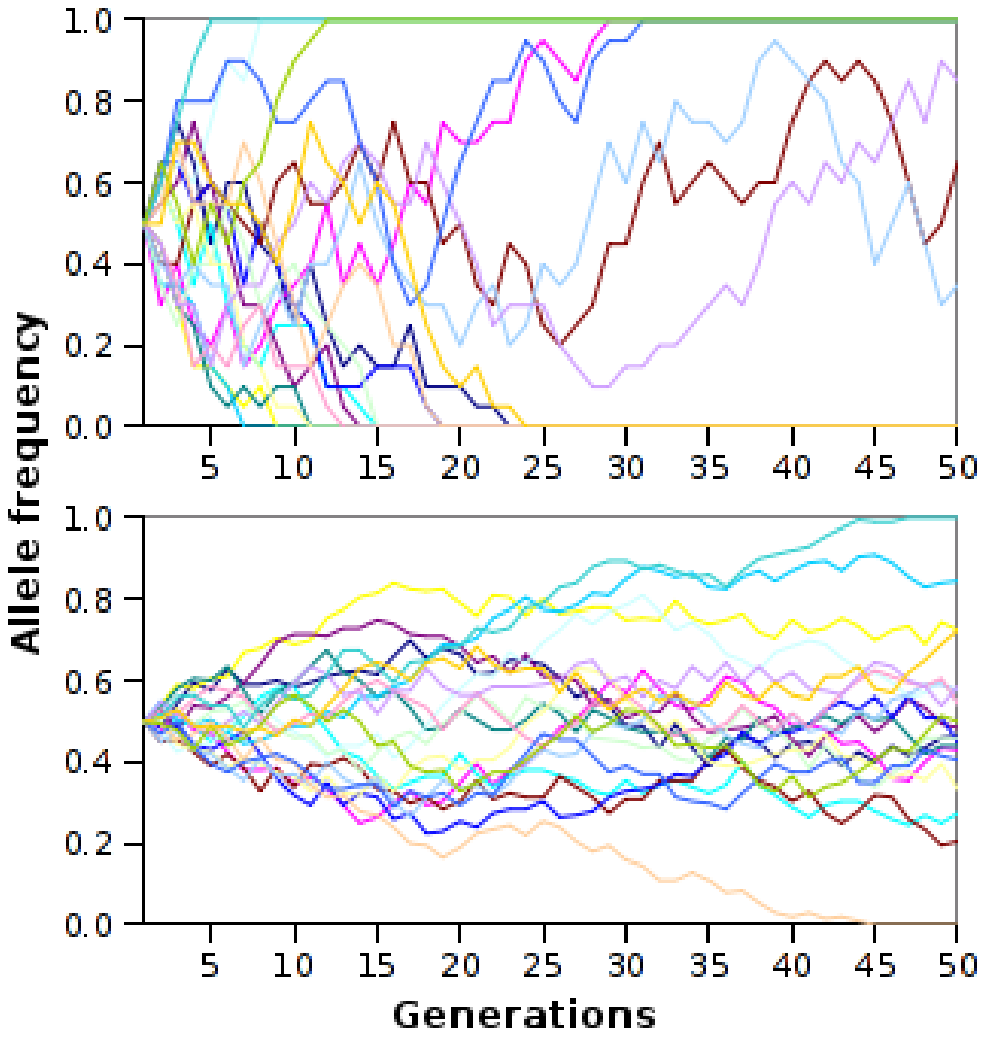
Some terminology

- **Haplotype**: set of alleles that tend to be inherited together, because their close physical proximity makes recombination among them unlikely.
- For now, we will treat alleles as independent. In other words, we'll assume that each site is far enough apart in the genome that it **segregates independently** of all other sites
- We will relax this assumption in today's afternoon lecture.

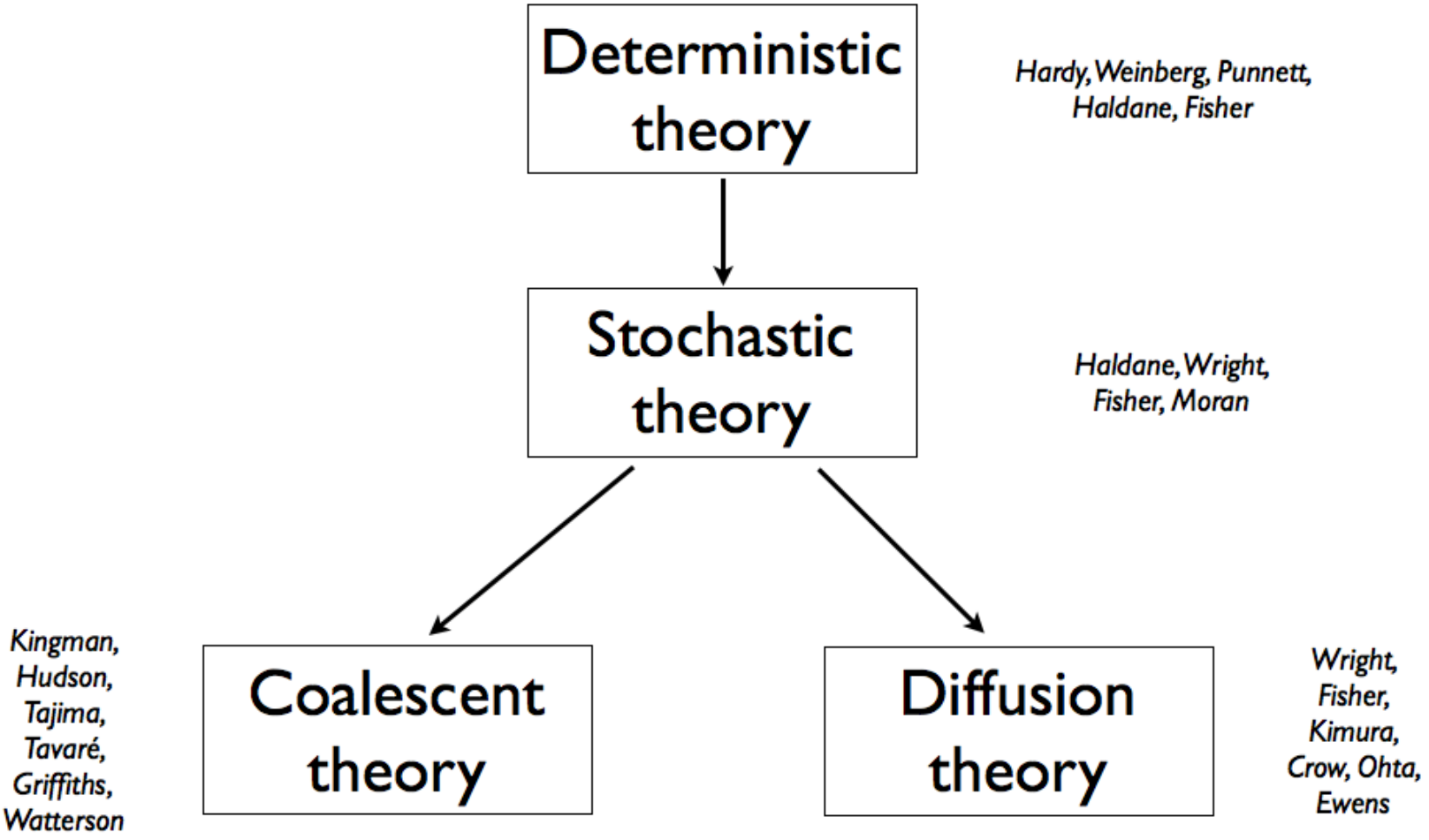


Some terminology

- **Population genetics**: the study of the distribution and evolution of allele frequencies in populations, over space and time.

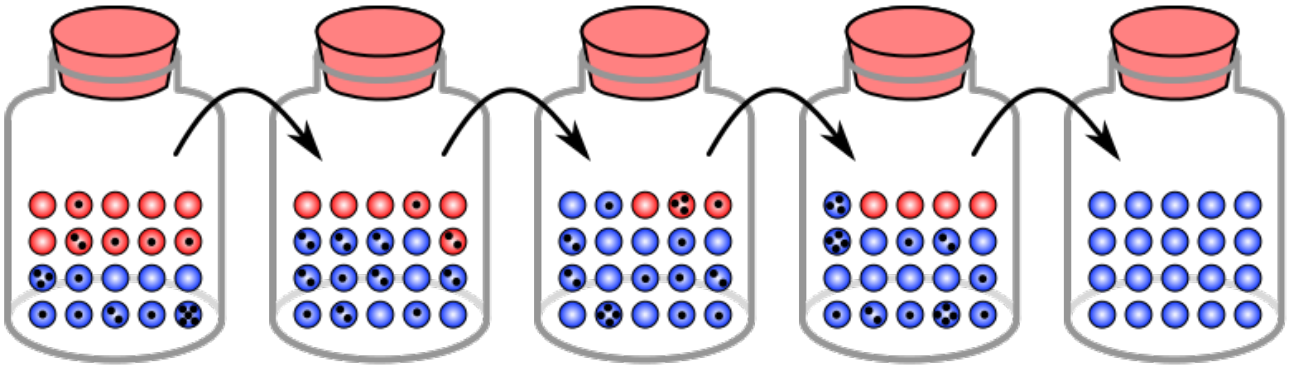


Evolution of population genetics



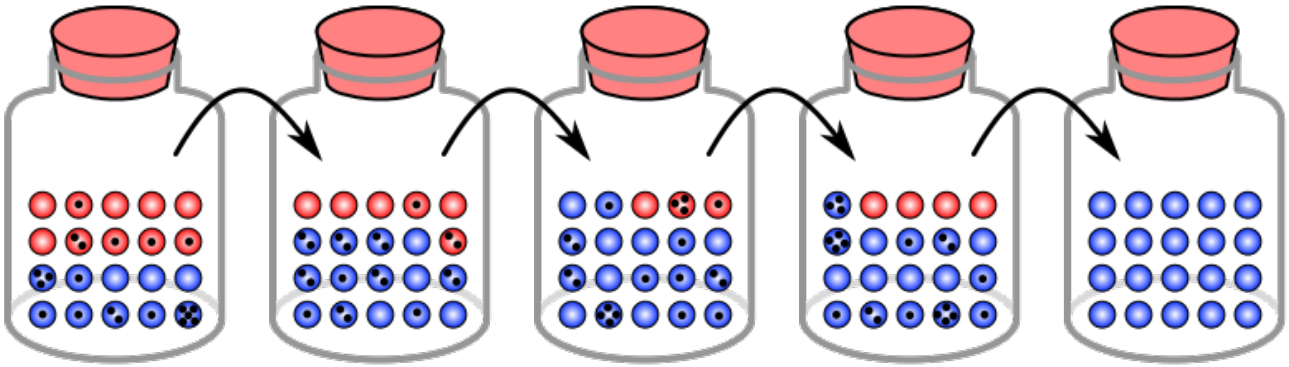
The discrete-time Wright-Fisher Model

- We start with $n = 20$ marbles, 10 of which are blue.



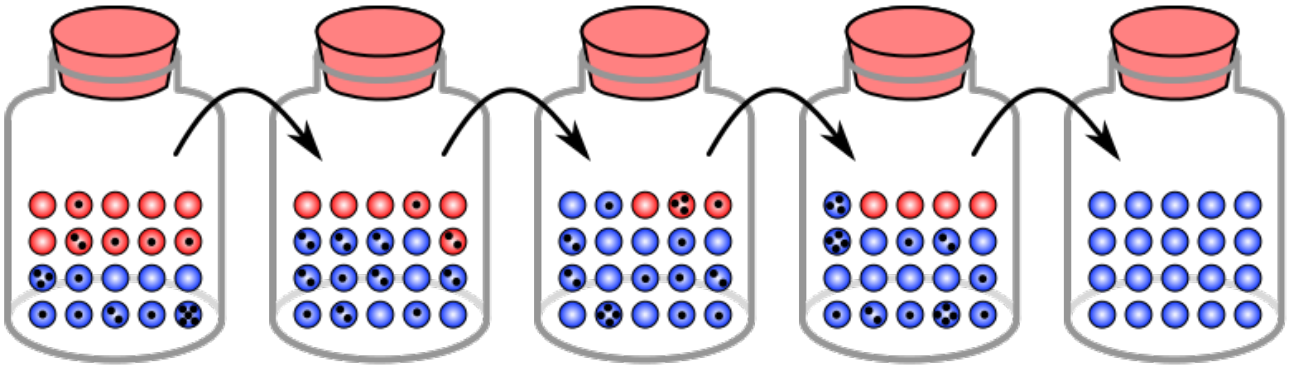
The discrete-time Wright-Fisher Model

- We start with $n = 20$ marbles, 10 of which are blue.
- We will sample with replacement to fill up the next jar.



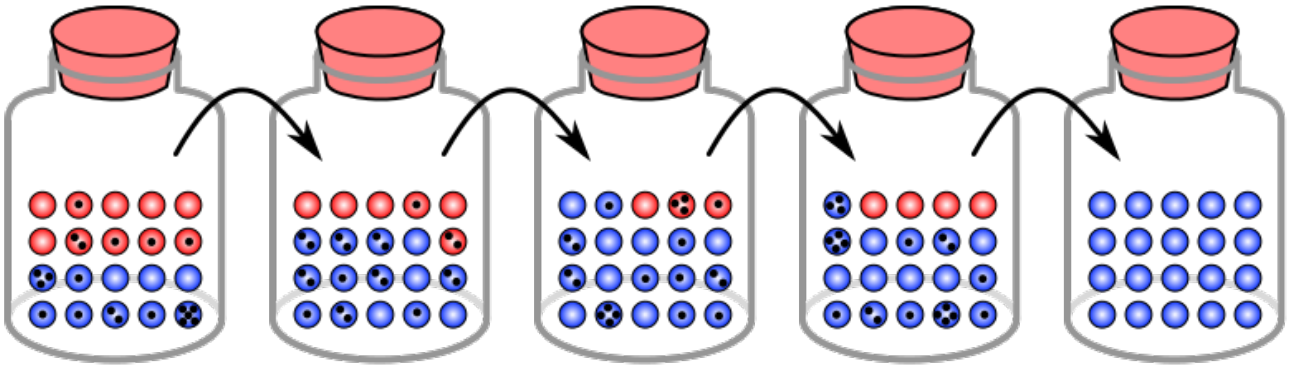
The discrete-time Wright-Fisher Model

- We start with $n = 20$ marbles, 10 of which are blue.
- We will sample with replacement to fill up the next jar.
- We assume the total number of marbles in the jar stays constant



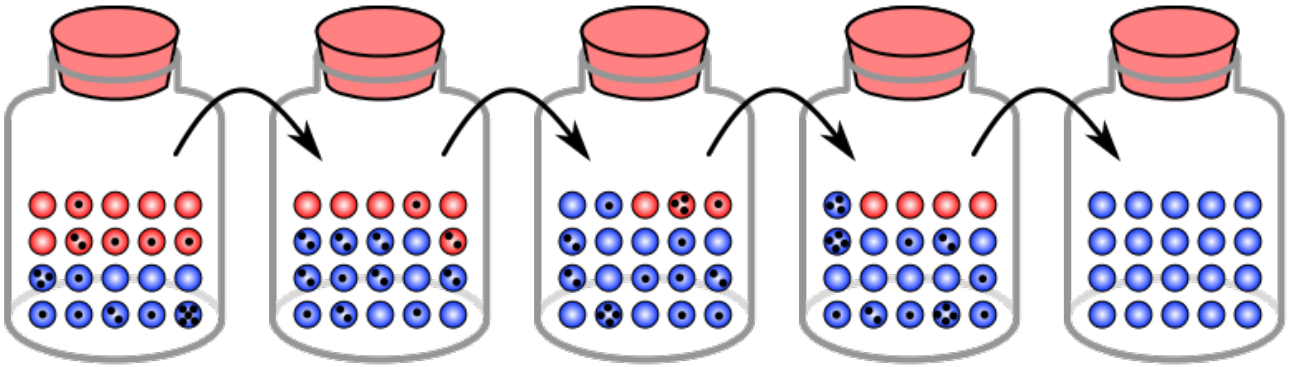
The discrete-time Wright-Fisher Model

- We start with $n = 20$ marbles, 10 of which are blue.
- We will sample with replacement to fill up the next jar.
- We assume the total number of marbles in the jar stays constant
- We are interested in the number of blue marbles at time t



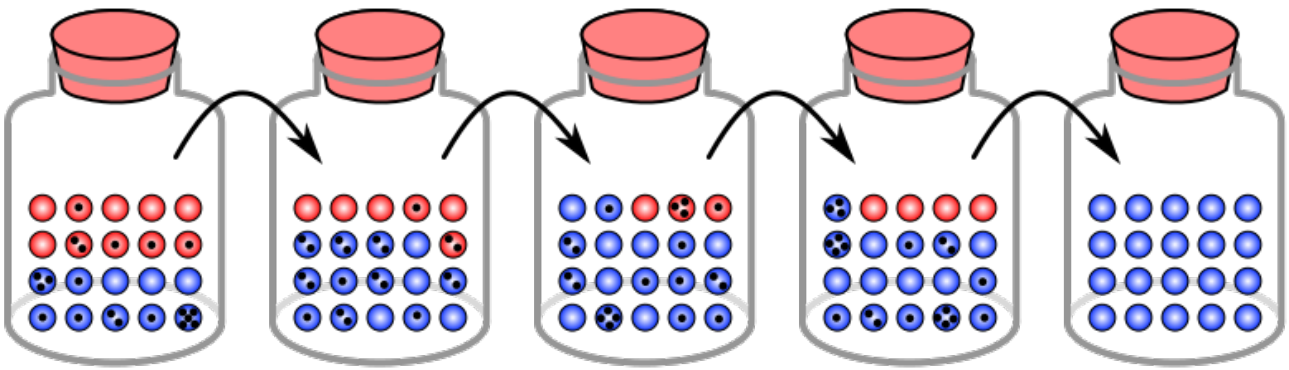
The discrete-time Wright-Fisher Model

- We start with $n = 20$ marbles, 10 of which are blue.
- We will sample with replacement to fill up the next jar.
- We assume the total number of marbles in the jar stays constant
- We are interested in the number of blue marbles at time t
- Let f(t) = frequency of blue marbles at time t



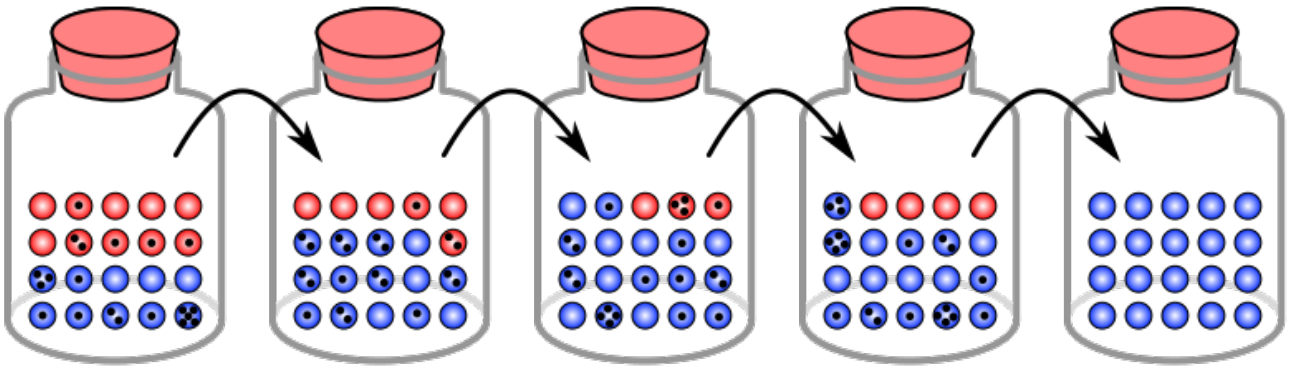
The discrete-time Wright-Fisher Model

- We start with $n = 20$ marbles, 10 of which are blue.
- We will sample with replacement to fill up the next jar.
- We assume the total number of marbles in the jar stays constant
- We are interested in the number of blue marbles at time t
- Let f(t) = frequency of blue marbles at time t
- P[no. blue marbles at $t_i = k \mid f(t_1), f(t_2), ..., f(t_{i-1})] =$
  P[no. blue marbles at $t_i = k \mid f(t_{i-1})]$

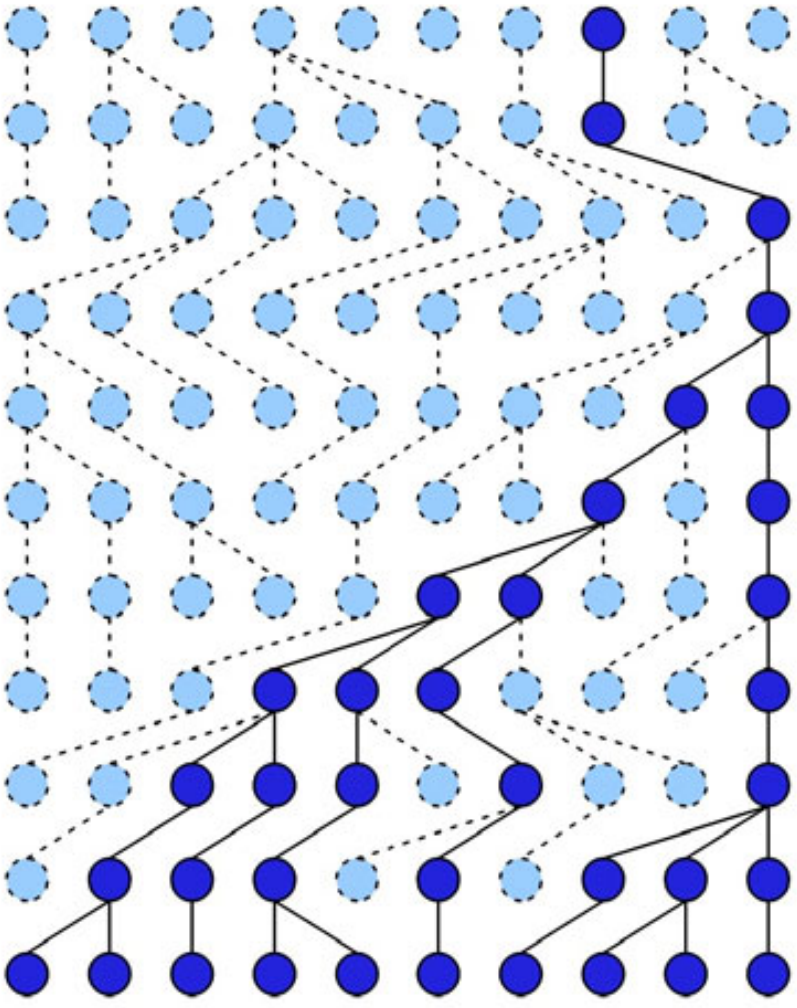


The discrete-time Wright-Fisher Model

- The number of blue marbles given that we know how many marbles we had in the blue jar follows a binomial distribution
- P[no. blue marbles = k at $t_i | f(t_{i-1}))] = \binom{n}{k} f(t_{i-1})^k (1 - f(t_{i-1}))^{n-k}$



The discrete-time Wright-Fisher Model

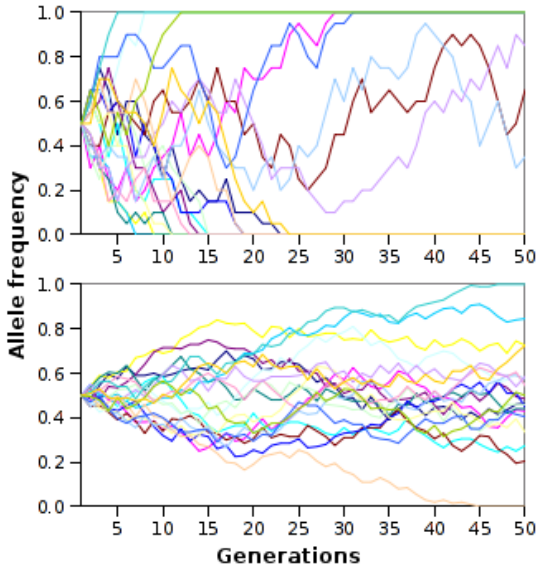


Assumptions of the W-F model

- Constant population size ($= 2N$)
- Individuals reproduce asexually and randomly (no population structure)
- No selection
- No migration
- Non-overlapping generations
- We'll be able to get rid of some of these assumptions later...

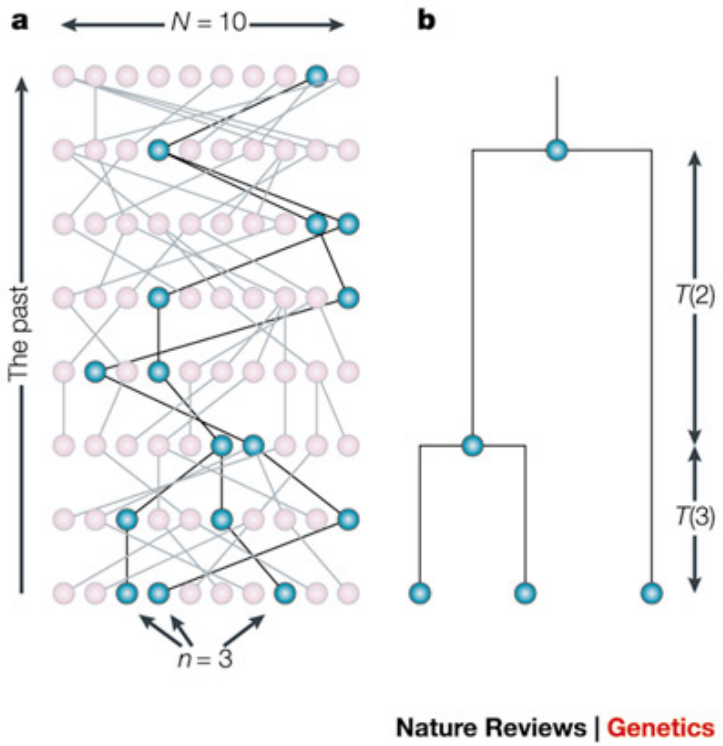
Genetic drift

- Under the pure Wright-Fisher model, allele frequencies evolve according to **genetic drit**.
- Genetic drift is the change in allele frequencies over time due to random sampling.
- Alleles survive, go extinct or get fixed purely by chance events.
- No allele has a special advantage over others at the same locus.
- The smaller the population, the stronger the drift.



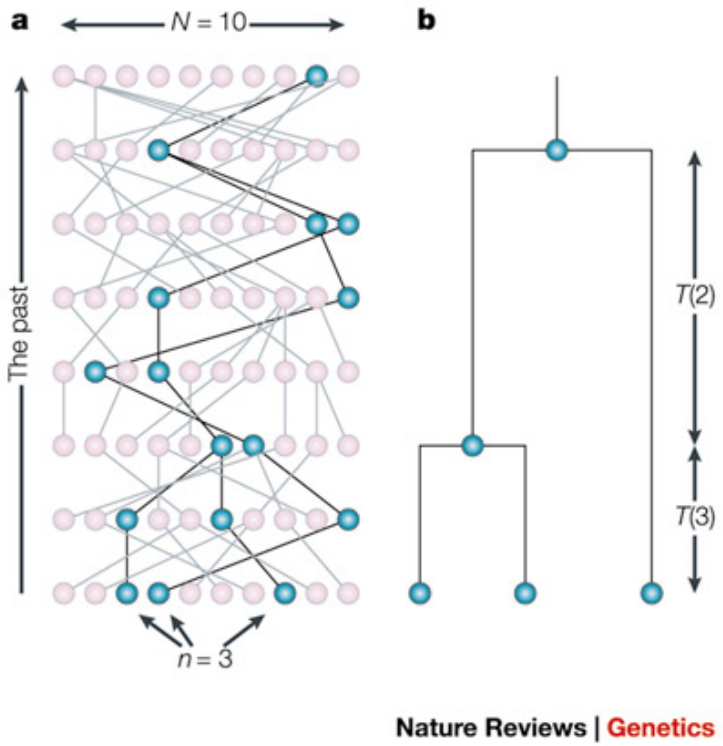
Coalescence terminology

- When two individual sampled gene copies have the same parent in a particular generation, we say that the **ancestral lineages** representing these two individuals have **coalesced**.



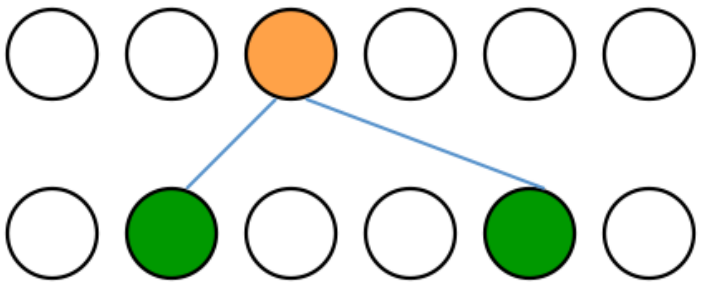
Coalescence terminology

- When two individual sampled gene copies have the same parent in a particular generation, we say that the **ancestral lineages** representing these two individuals have **coalesced**.
- That parent is the **most recent common ancestor (TMRCA)** of the two samples



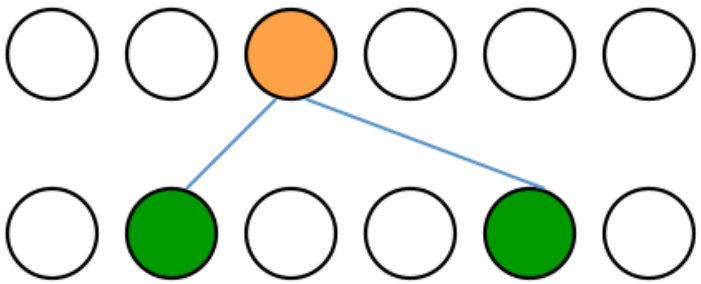
Coalescence in a sample of two sequences (n=2)

- P[2 gene copies have the same parent in the previous generation] =



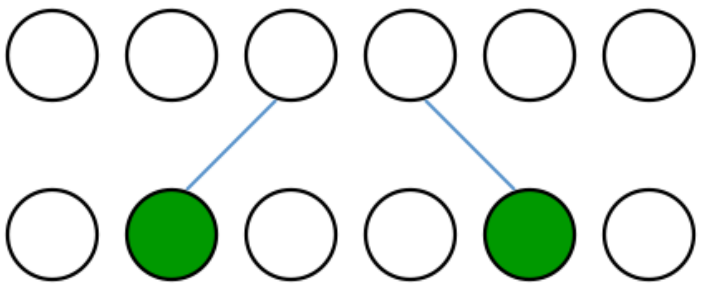
Coalescence in a sample of two sequences (n=2)

- P[2 gene copies have the same parent in the previous generation] =
- 2N * 1/(2N) * 1/(2N) = **1/(2N)**



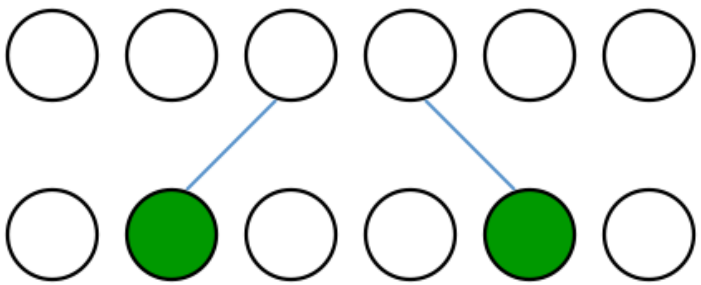
Coalescence in a sample of two sequences (n=2)

- P[2 gene copies do NOT have the same parent in the previous generation] =



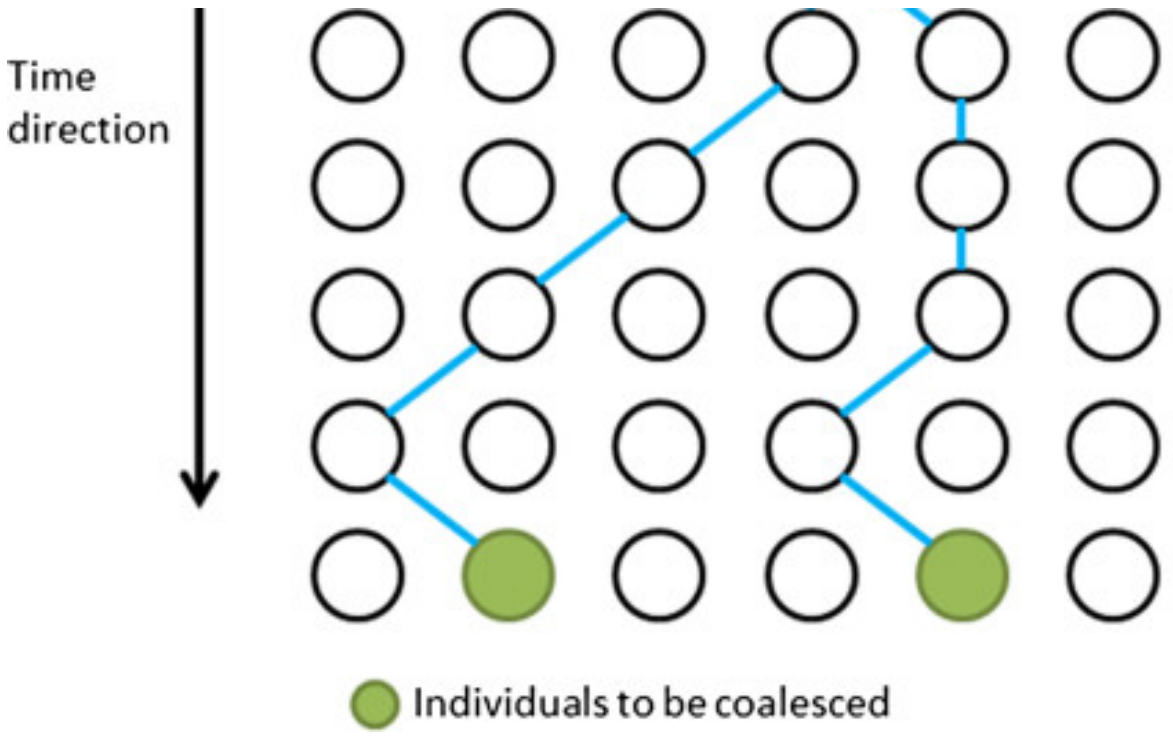
Coalescence in a sample of two sequences (n=2)

- P[2 gene copies do NOT have the same parent in the previous generation] =
- **1 - 1/(2N)**



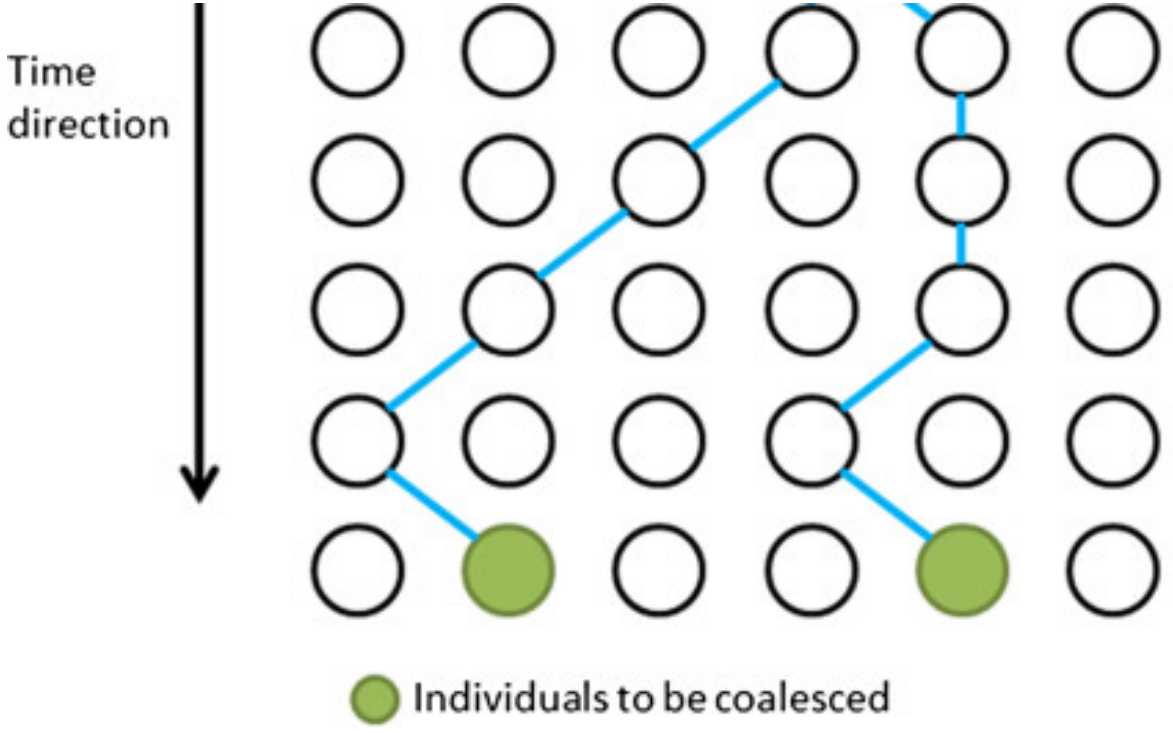
Coalescence in a sample of two sequences (n=2)

- P[2 gene copies **do not find** a common ancestor in **r generations**] =



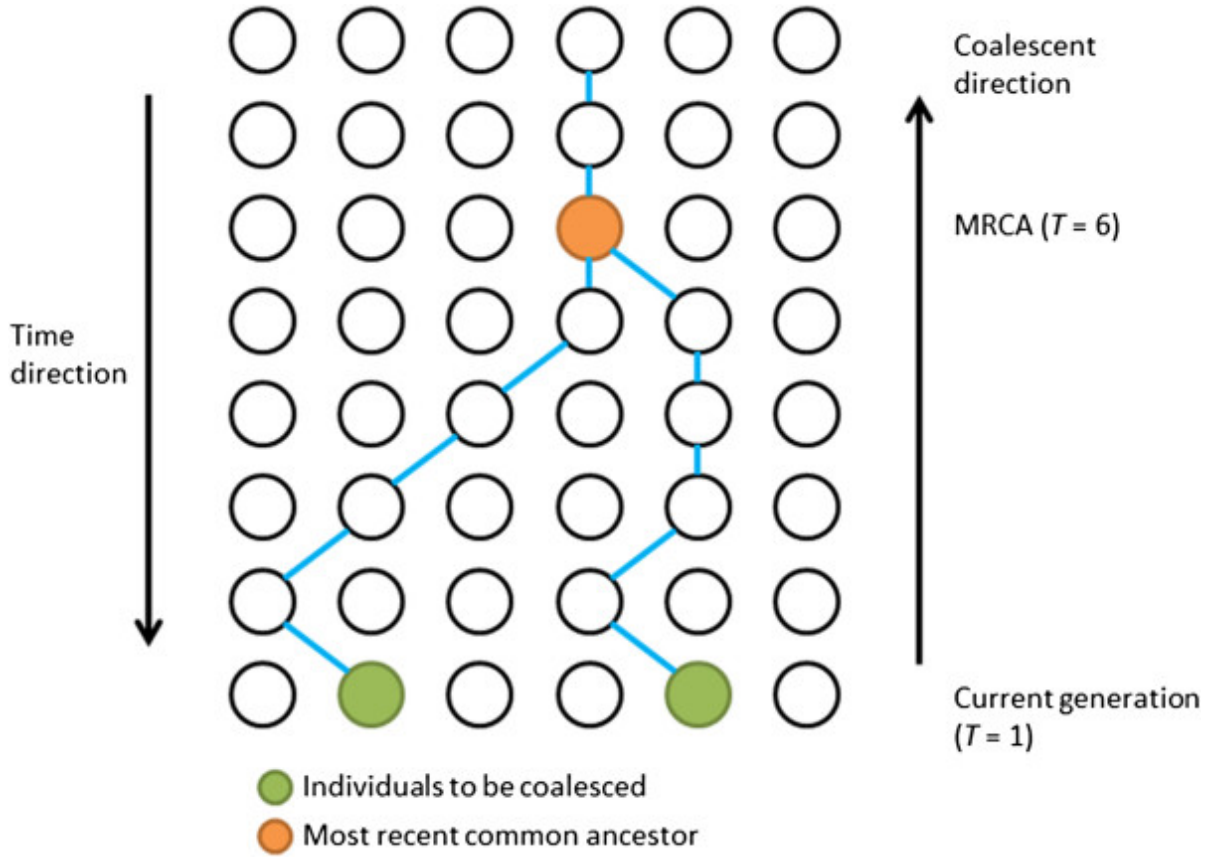
Coalescence in a sample of two sequences (n=2)

- P[2 gene copies **do not find** a common ancestor in **r generations**] =
- $(\mathbf{1} - \mathbf{1}/(\mathbf{2N}))^{\mathbf{r}}$



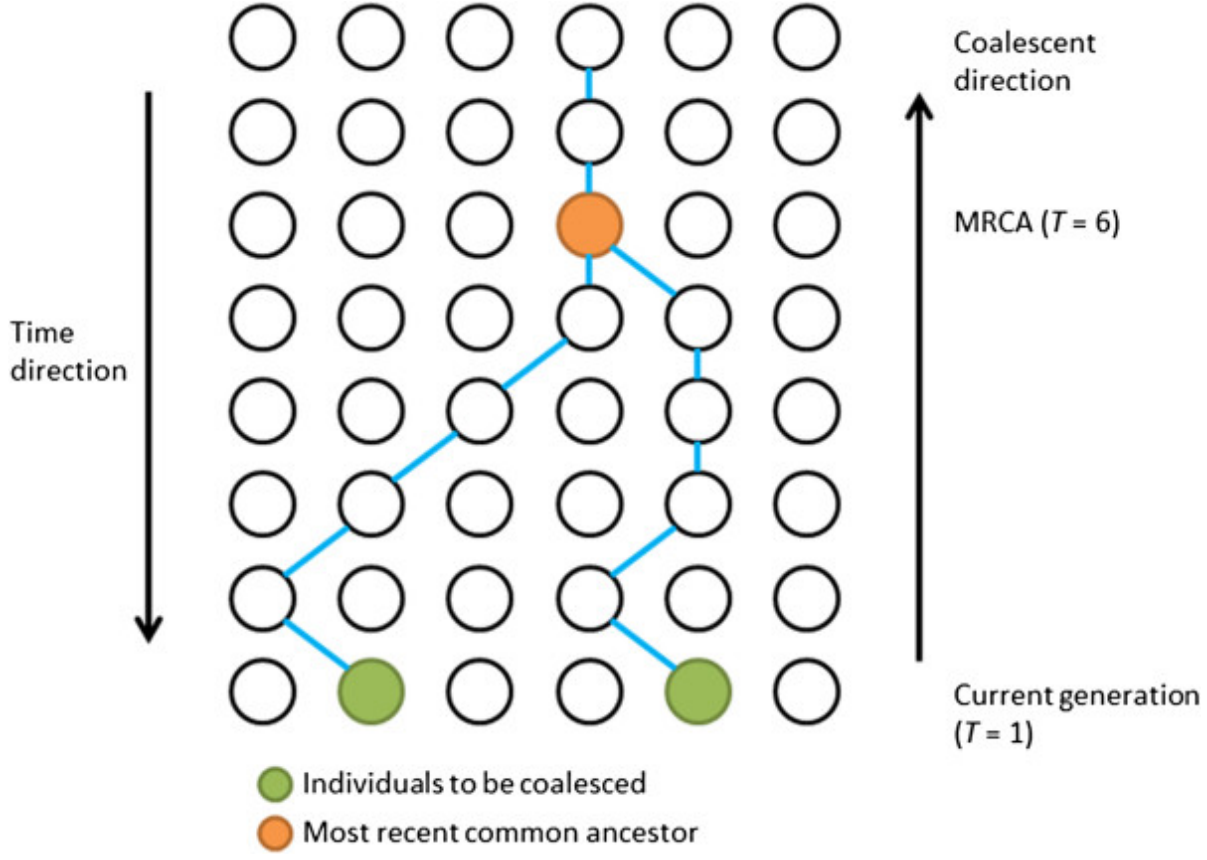
Coalescence in a sample of two sequences (n=2)

- P[2 gene copies find a common ancestor in generation r] =



Coalescence in a sample of two sequences (n=2)

- P[2 gene copies find a common ancestor in generation r] =
- $(\mathbf{1} - \mathbf{1}/(\mathbf{2N}))^{\mathbf{r}-\mathbf{1}} * (\mathbf{1}/(\mathbf{2N}))$

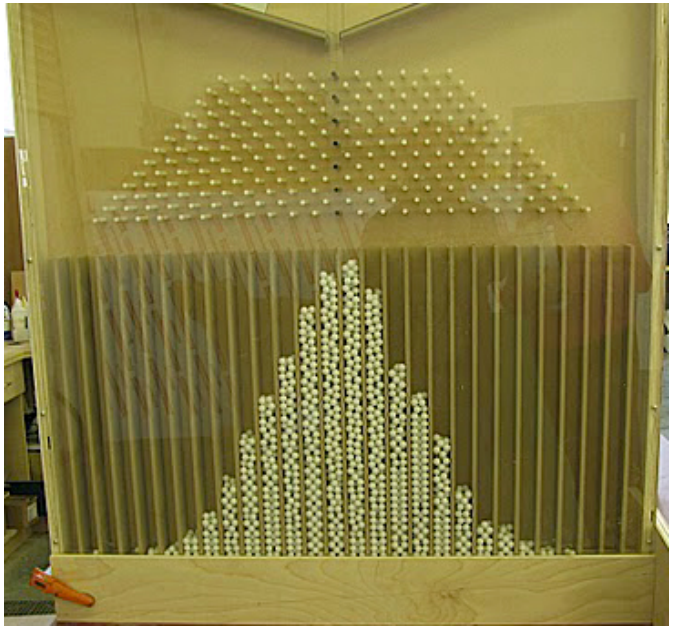


Wright-Fisher Exercises

- Follow the instructions in the Wright-Fisher exercise prompt:
- `https://github.com/FerRacimo/DemographicCourseAdelaide2018/blob/master/WrightFisherTutorial.md`

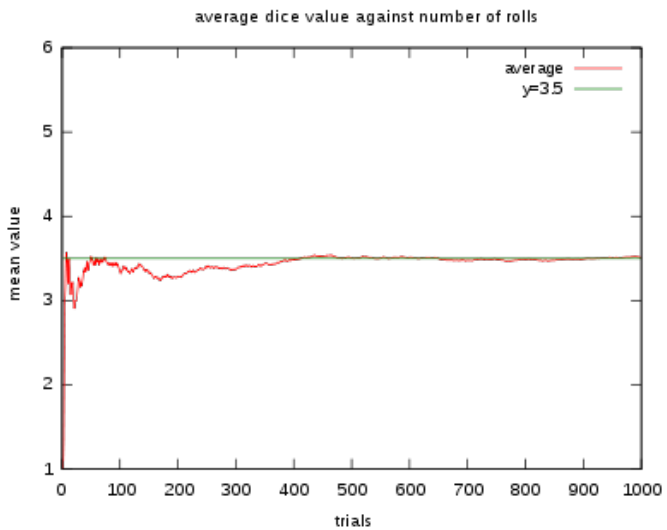
A digression: probability distributions

- In statistics (and population genetics), we often talk about **probability distributions**, which are descriptions of the way we think particular processes or random variables behave **over many repetitions**
- Examples: Normal distribution (continuous variable), binomial distribution (discrete variable)
- Two important properties of a probability distribution of a variable X are its **expected value** (E[X]) and **variance** (Var[X]).



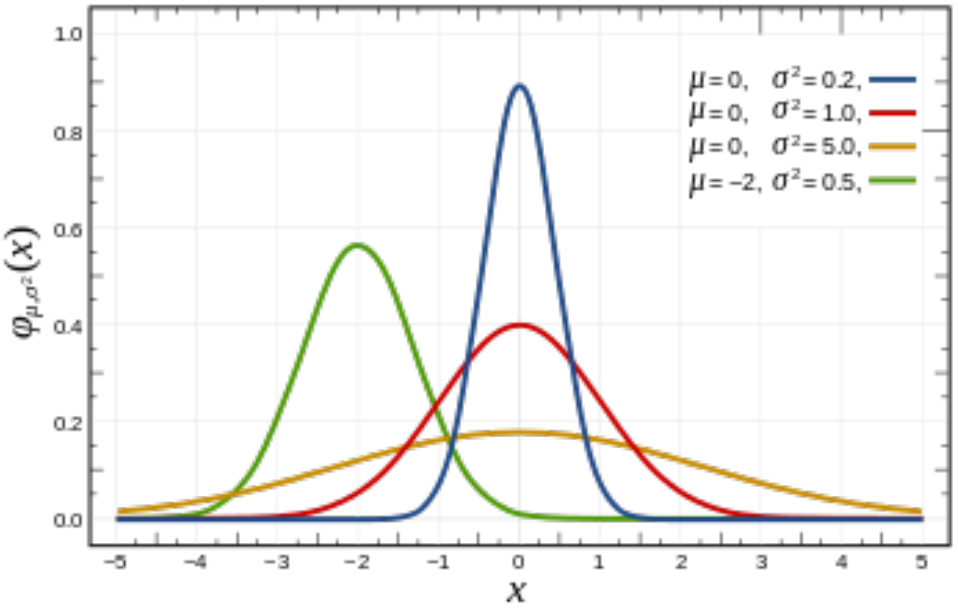
Expected value

- The expected value is the sum (or integral) over each possible value a variable can take, weighted by the probability that it will take that value
- It is usually represented with the greek letter μ
- The mean over many trials is an approximation of the expected value of a variable
- The mean over **infinite** trials is equal to the expected value of a variable
- Example: random variable = number obtained when rolling a dice



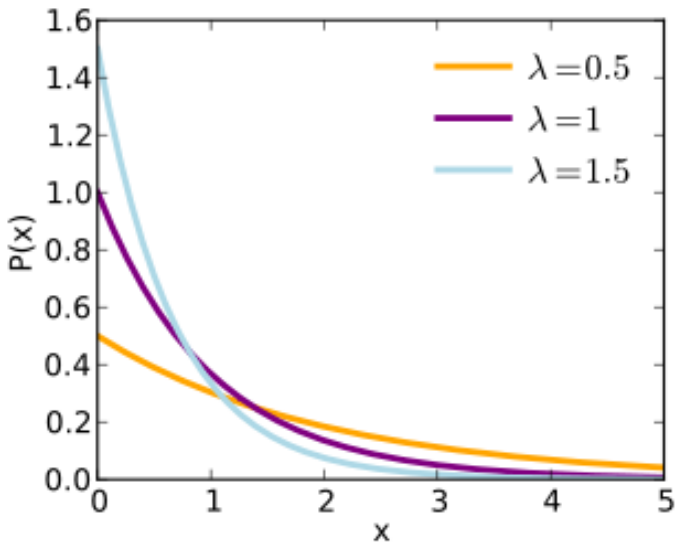
Variance

- The variance is the expected value of the square of the difference between a random varaible and its expected value:
  $Var[X] = E[(X - E[X])^2]$
- It is usually represented as σ^2
- The variance represents the **amount of variation** of the value of a random variable



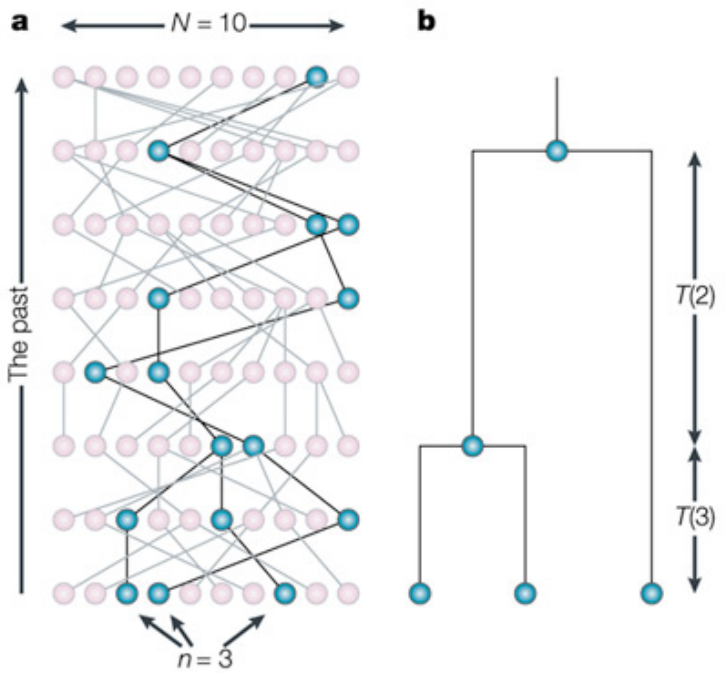
The exponential distribution

- The exponential distribution is a probability distribution used to model waiting times: $P[X > t] = e^{-\lambda t}$
- Examples: time till the next bus arrives; time till somebody calls me; time till my car breaks down
- One only needs a single parameter to describe an exponential distribution, the rate at which events occur: λ.
- The expected waiting time is the inverse of the rate (higher rate = smaller waiting time), so E[X] $= 1/\lambda$
- Var[X] $= 1/(\lambda^2)$



Wright-Fisher Model backwards in time

- Problem: many of the simplest questions under the W-F model become intractable for large populations over long time scales
- Cumbersome to keep track of all alleles at each time step
- For example: given that we sample 3 individuals in a population of size 10, what is the expected time till all 3 individuals find a common ancestor?



A solution: coalescent theory

- A way to infer how genetic genealogies behave over long time periods.
- "Bottom-up" approach
- The basis of many simulation tools in pop gen: ms, msms, ms prime, FastSimCoal, etc.
- The basis of many inference tools in pop gen: neutrality tests, PSMC, MSMC, Bayesian skyline, etc.
- **Key idea:** we only keep track of the genealogy of alleles that we have **sampled** in the present

The Kingman coalescent approximation

- P[2 gene copies **do not find** a common ancestor in **r generations**] =

The Kingman coalescent approximation

- P[2 gene copies **do not find** a common ancestor in **r generations**] =
- $(1 - 1/(2N))^r$

The Kingman coalescent approximation

- P[2 gene copies **do not find** a common ancestor in **r generations**] =
- $(1 - 1/(2N))^r$
- In units (t) of 2N generations, this is equivalent to $(1 - 1/(2N))^{2Nt}$

The Kingman coalescent approximation

- P[2 gene copies **do not find** a common ancestor in **r generations**] =
- $(1 - 1/(2N))^r$
- In units (t) of 2N generations, this is equivalent to $(1 - 1/(2N))^{2Nt}$
- For large population size N (i.e. in the limit as $N \to \infty$):

The Kingman coalescent approximation

- P[2 gene copies **do not find** a common ancestor in **r generations**] =
- $(1 - 1/(2N))^r$
- In units (t) of 2N generations, this is equivalent to $(1 - 1/(2N))^{2Nt}$
- For large population size N (i.e. in the limit as $N \to \infty$):
- $(1 - 1/(2N))^{2Nt} \to e^{-t}$

The Kingman coalescent approximation

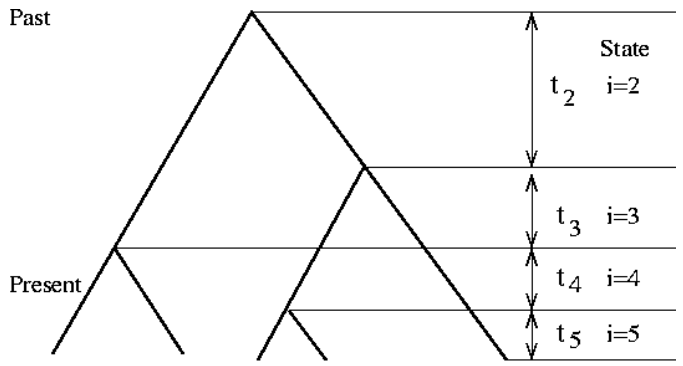
- P[2 gene copies **do not find** a common ancestor in **r generations**] =
- $(1 - 1/(2N))^r$
- In units (t) of 2N generations, this is equivalent to $(1 - 1/(2N))^{2Nt}$
- For large population size N (i.e. in the limit as $N \to \infty$):
- $(1 - 1/(2N))^{2Nt} \to e^{-t}$
- This means that, for large enough N, the time till coalescence (in units of 2N generations) follows an **exponential distribution with mean 1**.

The Kingman coalescent approximation

- P[2 gene copies **do not find** a common ancestor in **r generations**] =
- $(1 - 1/(2N))^r$
- In units (t) of 2N generations, this is equivalent to $(1 - 1/(2N))^{2Nt}$
- For large population size N (i.e. in the limit as $N \to \infty$):
- $(1 - 1/(2N))^{2Nt} \to e^{-t}$
- This means that, for large enough N, the time till coalescence (in units of 2N generations) follows an **exponential distribution with mean 1**.
- The mean coalescence time for two sequences is therefore 1 (in units of 2N generations) or 2N (in units of generations).

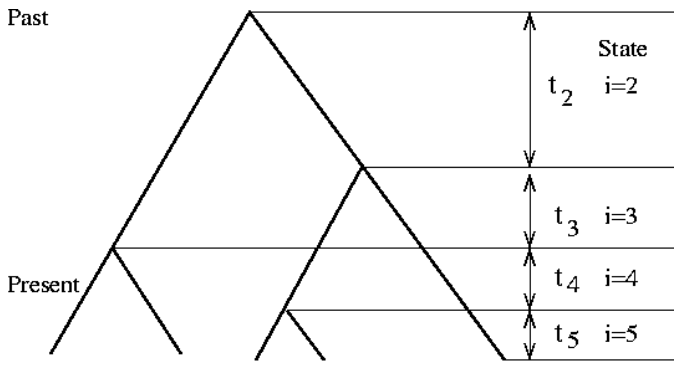
The Kingman coalescent approximation

- More generally, **the mean time till a coalescence event happens in a sample of n sequences** is $1/\binom{n}{2}$ (in units of 2N generations).



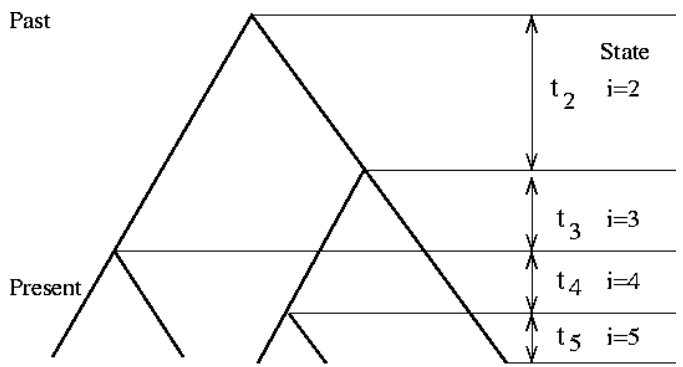
The Kingman coalescent approximation

- More generally, **the mean time till a coalescence event happens in a sample of n sequences** is $1/\binom{n}{2}$ (in units of 2N generations).
- As we go backwards in time, we will have less and less sequences (smaller n), so **the time till the next coalescent event will be larger**.



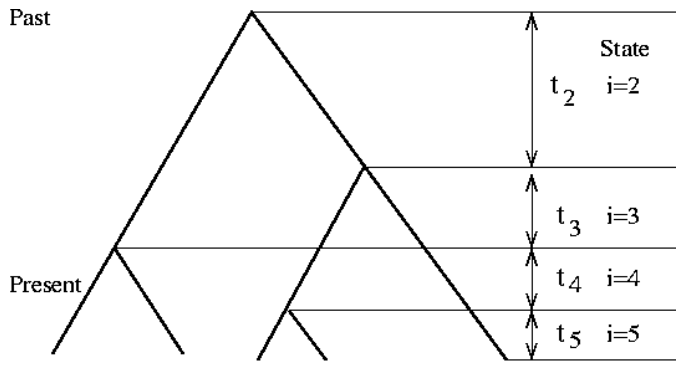
The Kingman coalescent approximation

- More generally, **the mean time till a coalescence event happens in a sample of n sequences** is $1/\binom{n}{2}$ (in units of 2N generations).
- As we go backwards in time, we will have less and less sequences (smaller n), so **the time till the next coalescent event will be larger**.
- For $n = 4$, E[time till 4 sequences become 3] $= 1/\binom{4}{2} = 1/6$



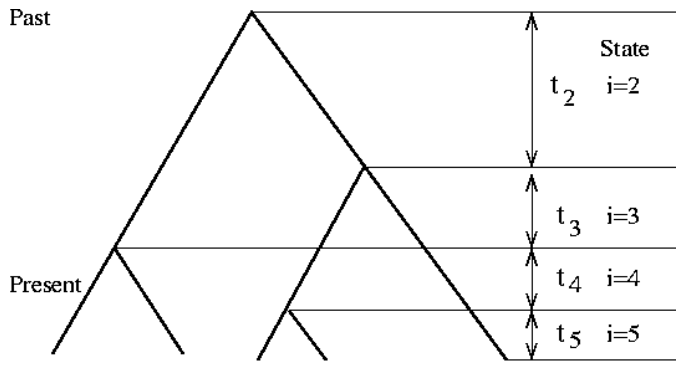
The Kingman coalescent approximation

- More generally, **the mean time till a coalescence event happens in a sample of n sequences** is $1/\binom{n}{2}$ (in units of 2N generations).
- As we go backwards in time, we will have less and less sequences (smaller n), so **the time till the next coalescent event will be larger**.
- For $n = 4$, E[time till 4 sequences become 3] $= 1/\binom{4}{2} = 1/6$
- For $n = 3$, E[time till 3 sequences become 2] $= 1/\binom{3}{2} = 1/3$



The Kingman coalescent approximation

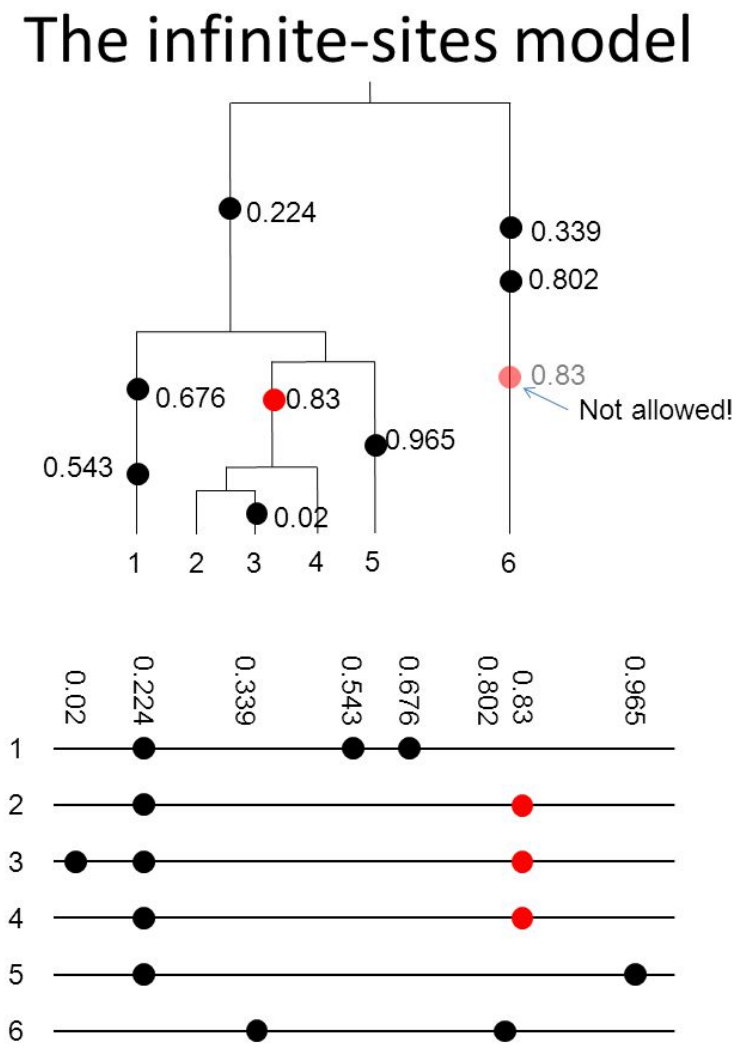
- More generally, **the mean time till a coalescence event happens in a sample of n sequences** is $1/\binom{n}{2}$ (in units of 2N generations).
- As we go backwards in time, we will have less and less sequences (smaller n), so **the time till the next coalescent event will be larger**.
- For $n = 4$, E[time till 4 sequences become 3] $= 1/\binom{4}{2} = 1/6$
- For $n = 3$, E[time till 3 sequences become 2] $= 1/\binom{3}{2} = 1/3$
- For $n = 2$, E[time till 2 sequences become 1] $= 1/\binom{2}{2} = 1$



The infinite sites model

- One way of introducing mutations in a genealogy is the **infinite sites model**.
- Assume we have a sequence with infinite number of sites (for example, a real line)
- This means no two mutations can occur in the same position of our sequence.
- In other words, each mutation creates a new segregating site.

The infinite sites model

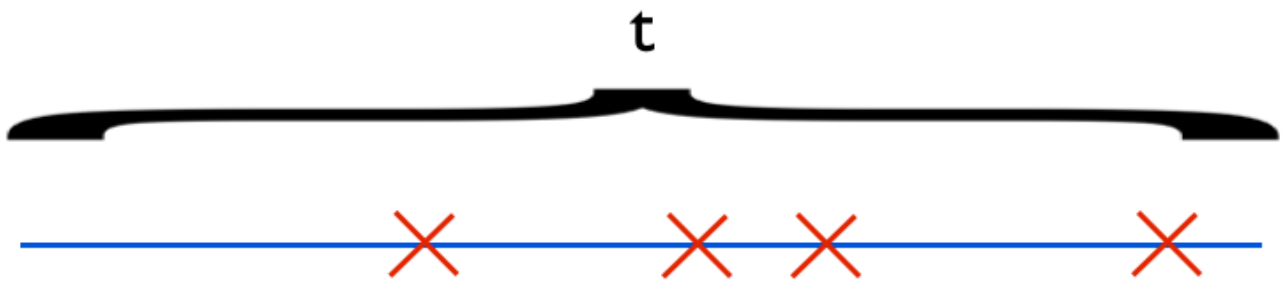


The infinite sites model: advantages and caveats

- Allows us to assume that each segregating site is due to a single mutation in the past
- Ignores double-substitutions and back-mutations
- Valid as long as the sequence we are studying is **long** and the **mutation rate is low**
- Good for short time-scales (population genetics) but not long time-scales (phylogenetics)

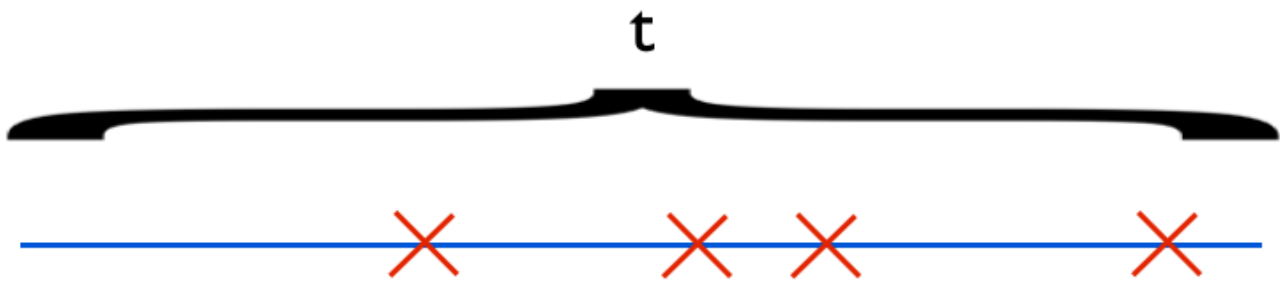
Introducing mutations to the coalescent

- Mutations occur at rate u per generation



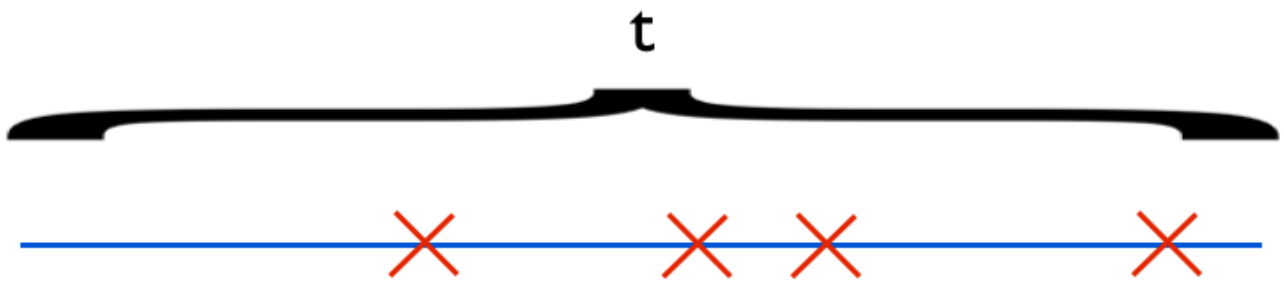
Introducing mutations to the coalescent

- Mutations occur at rate u per generation
- In other words, we expect u*r mutations in r generations



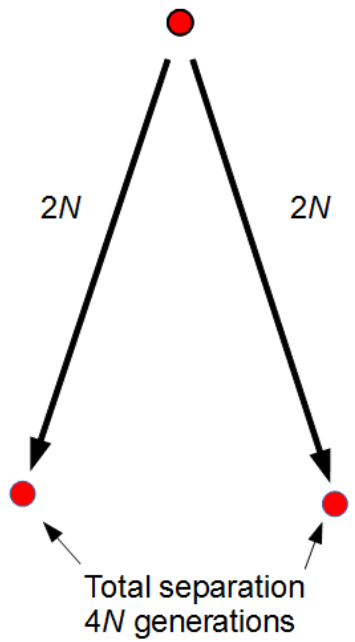
Introducing mutations to the coalescent

- Mutations occur at rate u per generation
- In other words, we expect u*r mutations in r generations
- If we measure time in units t of 2N generations, we expect 2N*t*u mutations in t units of time



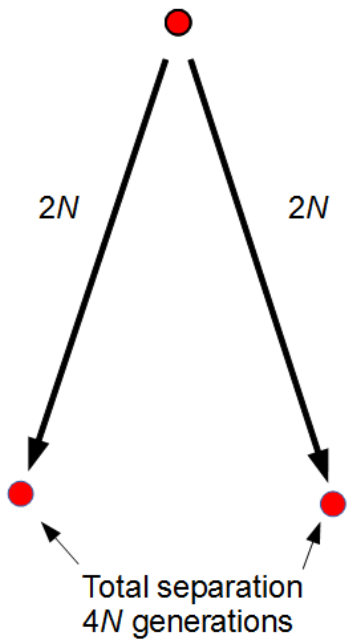
Introducing mutations to the coalescent

- In a sample of size two, the total expected length (t) is 4N generations, or 2 units of coalescent time



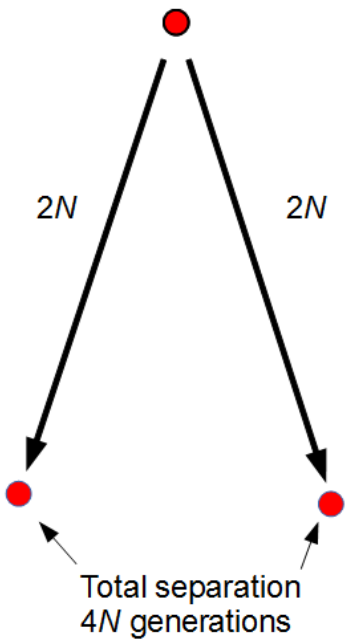
Introducing mutations to the coalescent

- In a sample of size two, the total expected length (t) is 4N generations, or 2 units of coalescent time
- Therefore, the total expected number of mutations is 2N*t*u = 4N*u



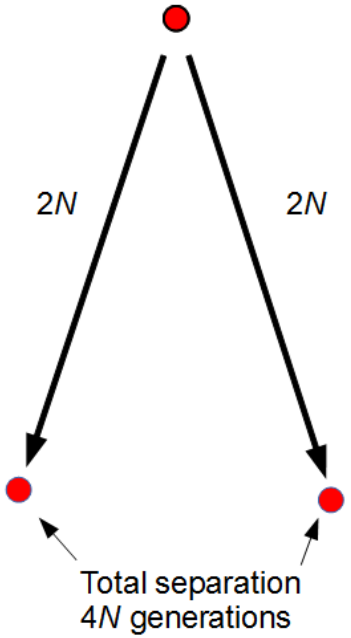
Introducing mutations to the coalescent

- In a sample of size two, the total expected length (t) is 4N generations, or 2 units of coalescent time
- Therefore, the total expected number of mutations is 2N*t*u = 4N*u
- 4N*u is also conveniently labeled as θ



Introducing mutations to the coalescent

- In a sample of size two, the total expected length (t) is 4N generations, or 2 units of coalescent time
- Therefore, the total expected number of mutations is 2N*t*u = 4N*u
- 4N*u is also conveniently labeled as θ
- In other words, mutations occur at rate $\theta/2$ per unit of coalescent time

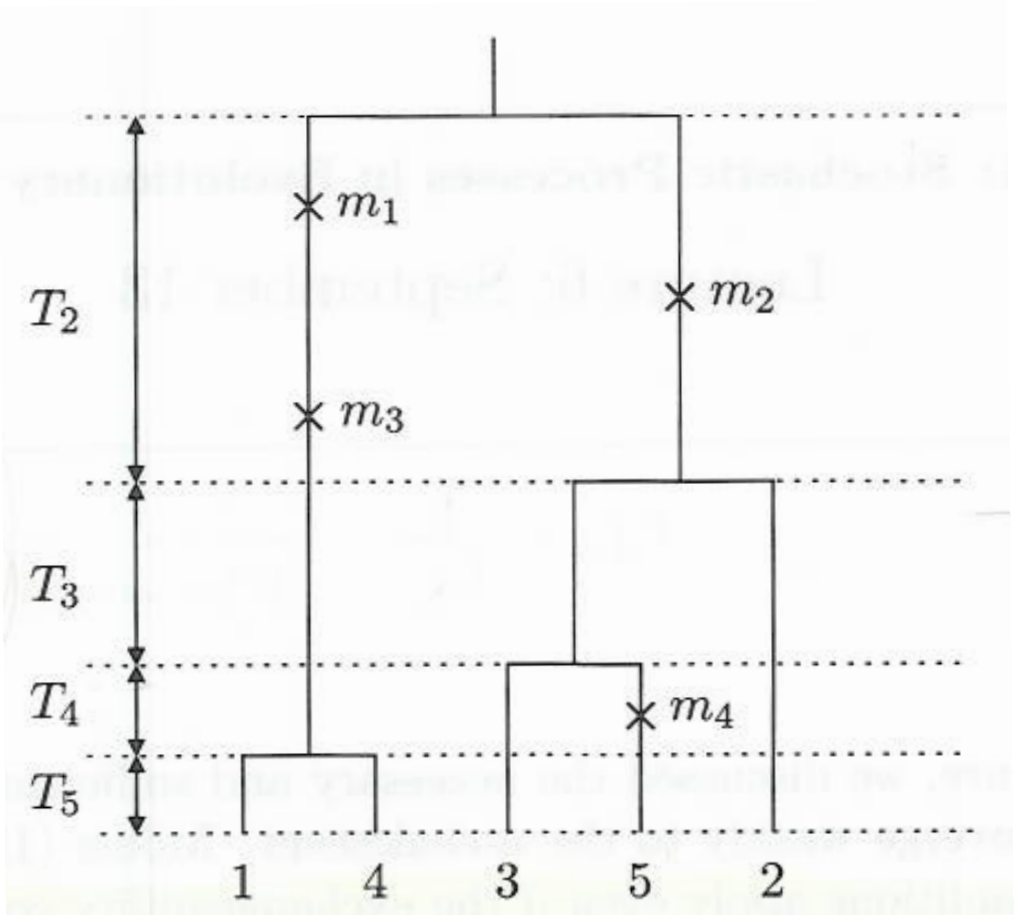


Estimators of theta

- Recall that $\theta = 4 * N * u$
- Tajima (1989) showed that we can use the expectation of particular statistics as a function of θ to test for deviations from the neutral coalescent model (constant demography + no selection).
- It is important to remember that deviations from the neutral coalescent model could have many causes.
- The tests we'll review are limited in their distinction of these causes

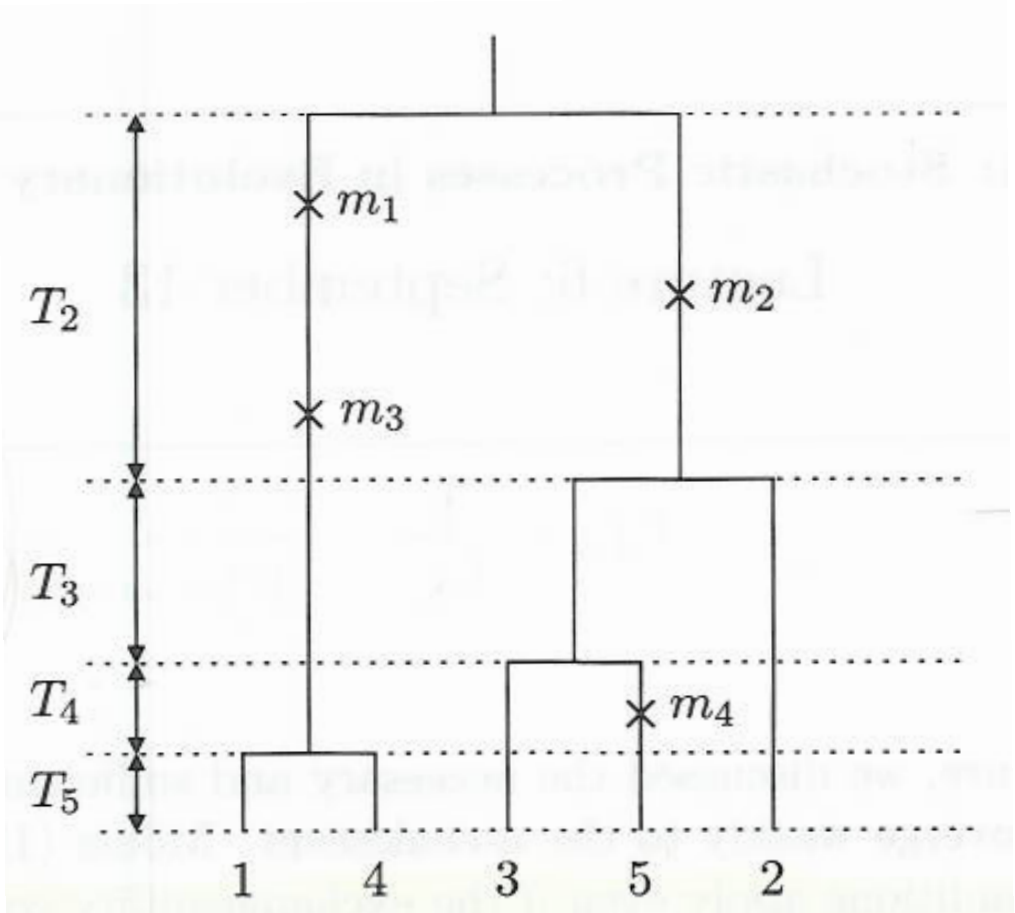
Watterson's Estimator

- One statistic comes from the number of segregating sites in the entire tree.



Watterson's Estimator

- One statistic comes from the number of segregating sites in the entire tree.
- Recall that under the infinite sites model: no. segregating sites (S) = no. mutations (M).



Watterson's Estimator

- $E[S] = E[M] = \frac{\theta}{2} * E[Length\ of\ \ tree]$

Watterson's Estimator

- $E[S] = E[M] = \frac{\theta}{2} * E[Length\ of\ \ tree]$
- $E[Length\ of\ \ tree] = \sum_{k=2}^{n} k * E[T_k]$

Watterson's Estimator

- $E[S] = E[M] = \frac{\theta}{2} * E[Length\ of\ \ tree]$
- $E[Length\ of\ \ tree] = \sum_{k=2}^{n} k * E[T_k]$
- If we assume a neutral coalescent model (as we've always done so far):

Watterson's Estimator

- $E[S] = E[M] = \frac{\theta}{2} * E[Length\ of\ \ tree]$
- $E[Length\ of\ \ tree] = \sum_{k=2}^{n} k * E[T_k]$
- If we assume a neutral coalescent model (as we've always done so far):
- $k * E[T_k] = \frac{k}{\binom{k}{2}} = \frac{2}{k-1}$

Watterson's Estimator

- $E[S] = E[M] = \frac{\theta}{2} * E[Length\ of\ \ tree]$
- $E[Length\ of\ \ tree] = \sum_{k=2}^{n} k * E[T_k]$
- If we assume a neutral coalescent model (as we've always done so far):
- $k * E[T_k] = \frac{k}{\binom{k}{2}} = \frac{2}{k-1}$
- Therefore, $E[S] = \frac{\theta}{2} * \sum_{k=2}^{n} \frac{2}{k-1} = \theta * \sum_{k=1}^{n-1} \frac{1}{k}$

Watterson's Estimator

- $E[S] = E[M] = \frac{\theta}{2} * E[Length\ of\ \ tree]$
- $E[Length\ of\ \ tree] = \sum_{k=2}^{n} k * E[T_k]$
- If we assume a neutral coalescent model (as we've always done so far):
- $k * E[T_k] = \frac{k}{\binom{k}{2}} = \frac{2}{k-1}$
- Therefore, $E[S] = \frac{\theta}{2} * \sum_{k=2}^{n} \frac{2}{k-1} = \theta * \sum_{k=1}^{n-1} \frac{1}{k}$
- This gives us an **estimator for** θ when we observe S segregating sites:

Watterson's Estimator

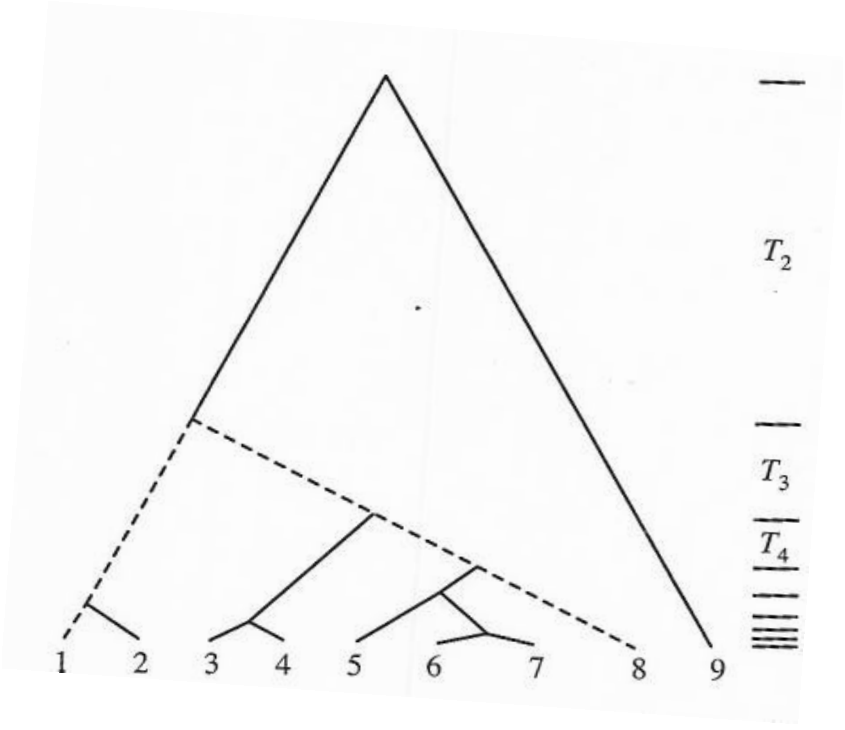
- $E[S] = E[M] = \frac{\theta}{2} * E[Length\ of\ \ tree]$
- $E[Length\ of\ \ tree] = \sum_{k=2}^{n} k * E[T_k]$
- If we assume a neutral coalescent model (as we've always done so far):
- $k * E[T_k] = \frac{k}{\binom{k}{2}} = \frac{2}{k-1}$
- Therefore, $E[S] = \frac{\theta}{2} * \sum_{k=2}^{n} \frac{2}{k-1} = \theta * \sum_{k=1}^{n-1} \frac{1}{k}$
- This gives us an **estimator for** θ when we observe S segregating sites:
- $\hat{\theta}_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$

Watterson's Estimator

- $E[S] = E[M] = \frac{\theta}{2} * E[Length\ of\ \ tree]$
- $E[Length\ of\ \ tree] = \sum_{k=2}^{n} k * E[T_k]$
- If we assume a neutral coalescent model (as we've always done so far):
- $k * E[T_k] = \frac{k}{\binom{k}{2}} = \frac{2}{k-1}$
- Therefore, $E[S] = \frac{\theta}{2} * \sum_{k=2}^{n} \frac{2}{k-1} = \theta * \sum_{k=1}^{n-1} \frac{1}{k}$
- This gives us an **estimator for** θ when we observe S segregating sites:
- $\hat{\theta}_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$
- This is called **Watterson's estimator**.

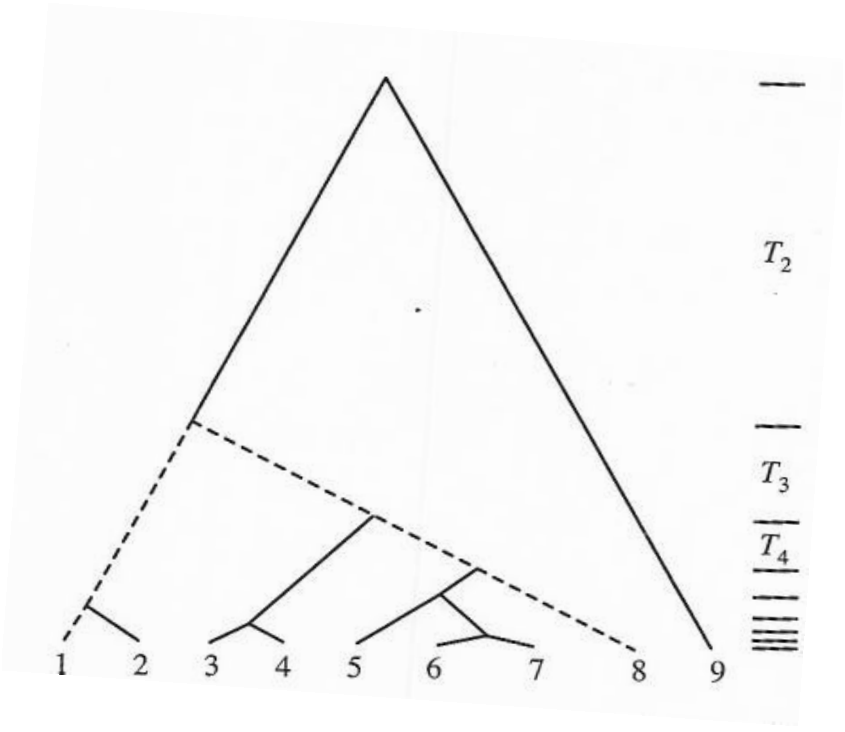
Tajima's Estimator

- Rather than looking at the number of segregating sites S in a sample of n sequences, we can look at the average number of pairwise differences between any two sequences from the sample.



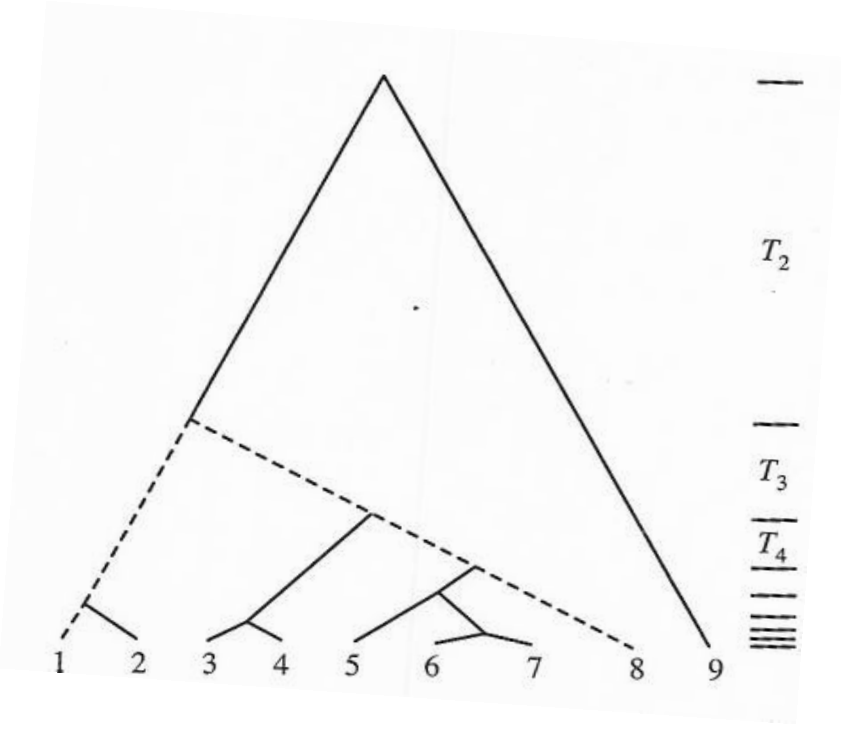
Tajima's Estimator

- Rather than looking at the number of segregating sites S in a sample of n sequences, we can look at the average number of pairwise differences between any two sequences from the sample.
- Let $\pi_{i,j}$ = number of differences between sequence i and sequence j



Tajima's Estimator

- Rather than looking at the number of segregating sites S in a sample of n sequences, we can look at the average number of pairwise differences between any two sequences from the sample.
- Let $\pi_{i,j}$ = number of differences between sequence i and sequence j
- Let π = average number of pairwise differences



Tajima's Estimator

- To obtain π, we need to average over every possible combination of different sequences.

Tajima's Estimator

- To obtain π, we need to average over every possible combination of different sequences.
- There are $\binom{n}{2}$ possible combinations, so...

Tajima's Estimator

- To obtain π, we need to average over every possible combination of different sequences.
- There are $\binom{n}{2}$ possible combinations, so...
- $\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \pi_{i,j}$

Tajima's Estimator

- To obtain π, we need to average over every possible combination of different sequences.
- There are $\binom{n}{2}$ possible combinations, so...
- $\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \pi_{i,j}$
- Therefore, $E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[\pi_{i,j}]$

Tajima's Estimator

- To obtain π, we need to average over every possible combination of different sequences.
- There are $\binom{n}{2}$ possible combinations, so...
- $\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \pi_{i,j}$
- Therefore, $E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[\pi_{i,j}]$
- We know that: $E[\pi_{i,j}] = \frac{\theta}{2} E[2 * T_{i,j}] = \frac{\theta}{2} E[2 * T_2]$ for all i and j

Tajima's Estimator

- To obtain π, we need to average over every possible combination of different sequences.
- There are $\binom{n}{2}$ possible combinations, so...
- $\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \pi_{i,j}$
- Therefore, $E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[\pi_{i,j}]$
- We know that: $E[\pi_{i,j}] = \frac{\theta}{2} E[2 * T_{i,j}] = \frac{\theta}{2} E[2 * T_2]$ for all i and j
- So we end up with: $E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\theta}{2} * 2 * E[T_2]$

Tajima's Estimator

- Under the neutral model, $E[T_2] = 1$ coalescent unit.

Tajima's Estimator

- Under the neutral model, $E[T_2] = 1$ coalescent unit.
- So our complicated expression gets reduced to:

Tajima's Estimator

- Under the neutral model, $E[T_2] = 1$ coalescent unit.
- So our complicated expression gets reduced to:
- $E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \theta * 1 = \frac{\binom{n}{2}}{\binom{n}{2}} \theta = \theta$

Tajima's Estimator

- Under the neutral model, $E[T_2] = 1$ coalescent unit.
- So our complicated expression gets reduced to:
- $E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \theta * 1 = \frac{\binom{n}{2}}{\binom{n}{2}} \theta = \theta$
- Thus, we have a **second estimator for** θ, if we know the average number of pairwise differences:

Tajima's Estimator

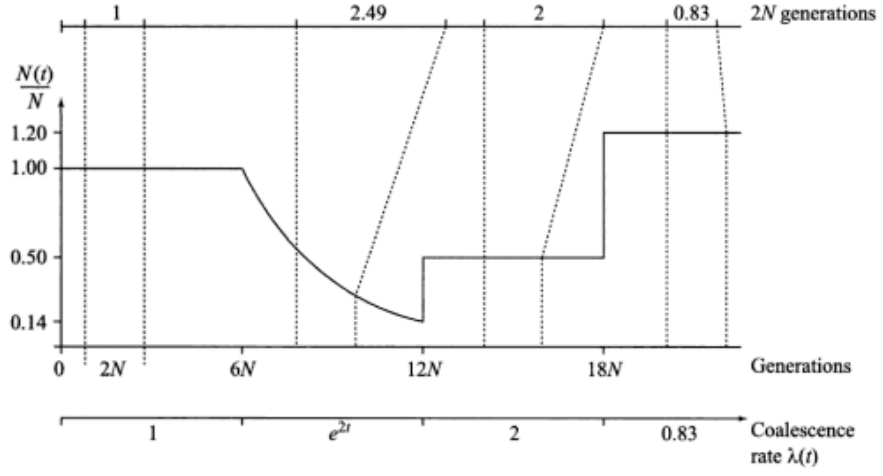
- Under the neutral model, $E[T_2] = 1$ coalescent unit.
- So our complicated expression gets reduced to:
- $E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \theta * 1 = \frac{\binom{n}{2}}{\binom{n}{2}} \theta = \theta$
- Thus, we have a **second estimator for** θ, if we know the average number of pairwise differences:
- $\hat{\theta}_T = \pi$

Tajima's Estimator

- Under the neutral model, $E[T_2] = 1$ coalescent unit.
- So our complicated expression gets reduced to:
- $E[\pi] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \theta * 1 = \frac{\binom{n}{2}}{\binom{n}{2}} \theta = \theta$
- Thus, we have a **second estimator for** θ, if we know the average number of pairwise differences:
- $\hat{\theta}_T = \pi$
- This is called **Tajima's estimator**.

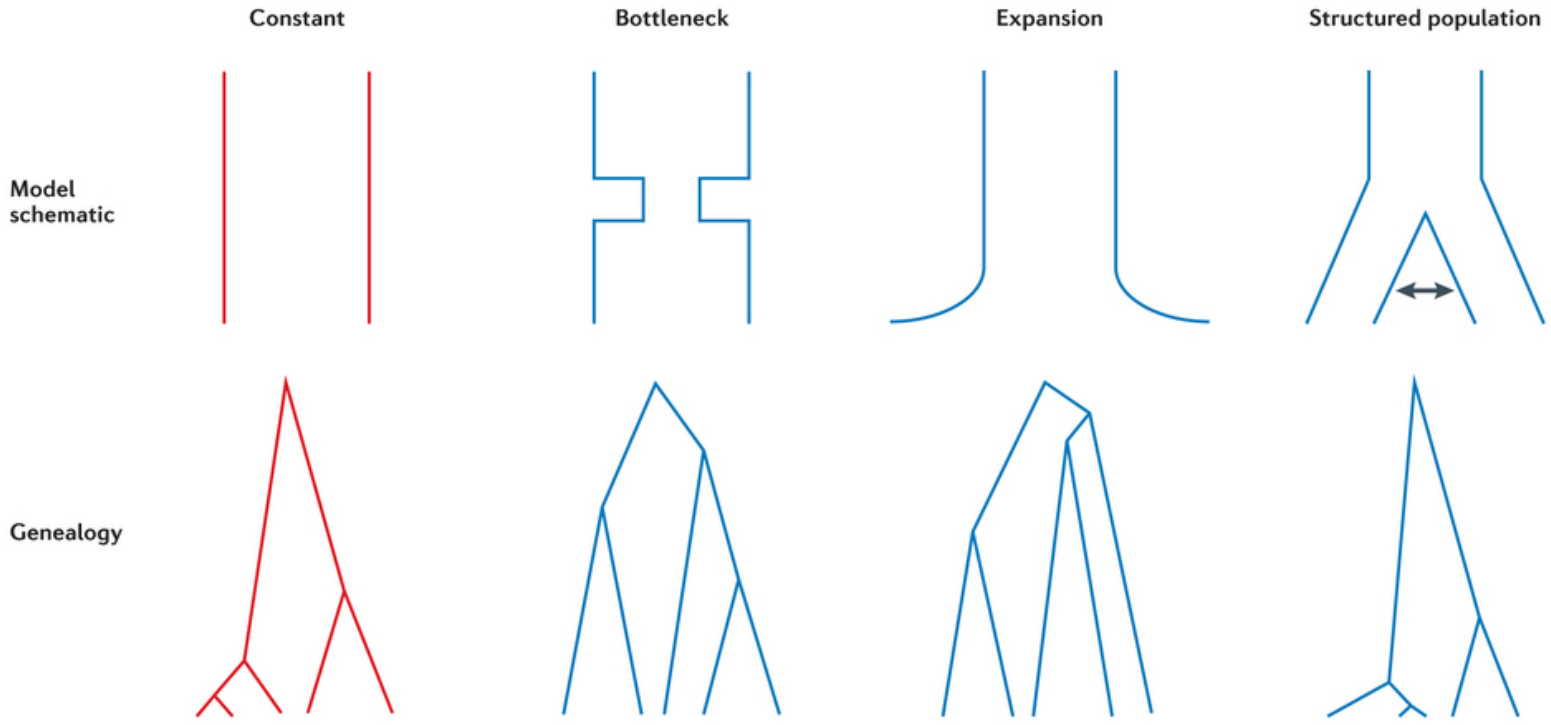
Demographic changes and the coalescent

- Recall that the rate of coalescence is in units 2N (the population size)
- Therefore, this rate depends on 2N
- When the population size is large, lineages are less likely to "find" each other, so the rate is low (less coalescences) and the expected time till coalescence is large.
- When the population size is small, lineages are more likely to "find" each other, so the rate is high (more coalescences) and the expected time till coalescence is small.



Demographic changes and the coalescent

- The coalescent tree contains information about the demographic history of our sample

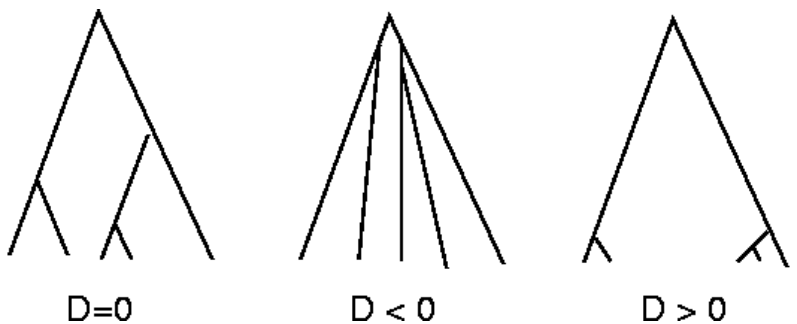


Tajima's D

- Under the neutral W-F model, Tajima's estimator for θ ($\hat{\theta}_T$ or π) should be equal to Waterson's estimator for θ ($\hat{\theta}_W$), because they are both estimating the same quantity (θ).
- However, if a coalescent tree is not evolving according to a neutral W-F model, these two estimators may be different.

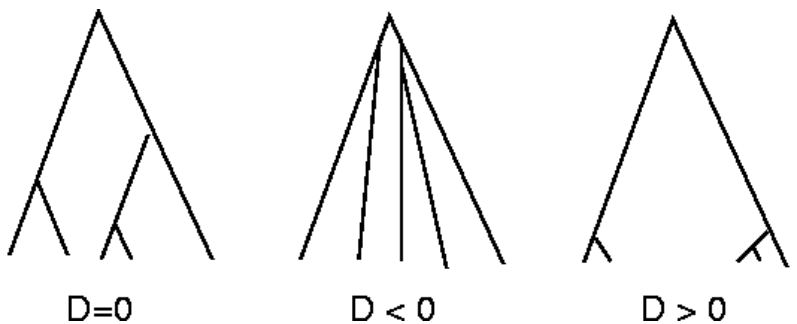
Tajima's D

- For example, if the terminal branches are too long, the number of singleton mutations will be large. Singletons contribute strongly to $\hat{\theta}_W$ but not so much to $\hat{\theta}_T$, so $\hat{\theta}_W > \hat{\theta}_T$.



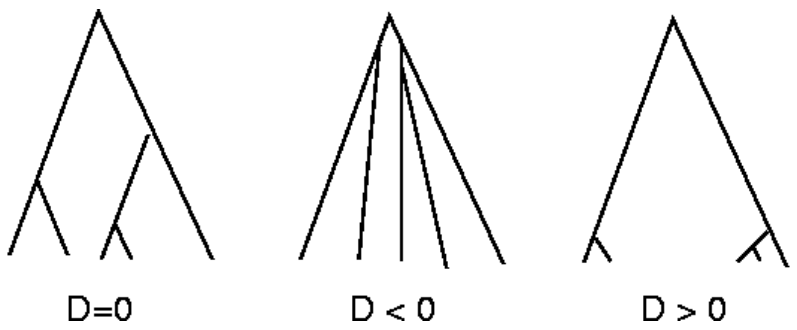
Tajima's D

- For example, if the terminal branches are too long, the number of singleton mutations will be large. Singletons contribute strongly to $\hat{\theta}_W$ but not so much to $\hat{\theta}_T$, so $\hat{\theta}_W > \hat{\theta}_T$.
- If a tree has very long internal branches, the reverse will happen and $\hat{\theta}_W < \hat{\theta}_T$.



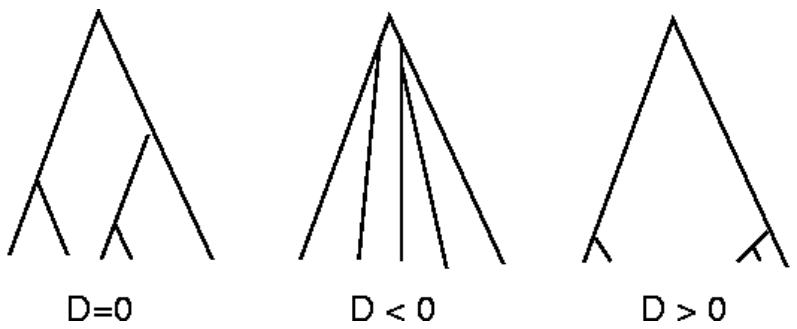
Tajima's D

- For example, if the terminal branches are too long, the number of singleton mutations will be large. Singletons contribute strongly to $\hat{\theta}_W$ but not so much to $\hat{\theta}_T$, so $\hat{\theta}_W > \hat{\theta}_T$.
- If a tree has very long internal branches, the reverse will happen and $\hat{\theta}_W < \hat{\theta}_T$.
- Tajima's D measures the difference between these estimators, scaled by a variance:



Tajima's D

- For example, if the terminal branches are too long, the number of singleton mutations will be large. Singletons contribute strongly to $\hat{\theta}_W$ but not so much to $\hat{\theta}_T$, so $\hat{\theta}_W > \hat{\theta}_T$.
- If a tree has very long internal branches, the reverse will happen and $\hat{\theta}_W < \hat{\theta}_T$.
- Tajima's D measures the difference between these estimators, scaled by a variance:
- $D = (\hat{\theta}_T - \hat{\theta}_W)/\sqrt{V(\hat{\theta}_T - \hat{\theta}_W)}$



Tajima's D

- There may be multiple reasons why the value of D may not be 0.

	Whole genome effect	Local effect
Long external branches (Tajima's $D < 0$)	Population growth Very severe bottleneck	Directional selection
Long internal branches (Tajima's $D > 0$)	Population subdivision Less severe bottleneck	Balancing selection Recent population mixing

Coalescent Exercises

- Follow the instructions in this prompt:
- `https://github.com/FerRacimo/DemographicCourseAdelaide2018/blob/master/CoalTutorial.md`