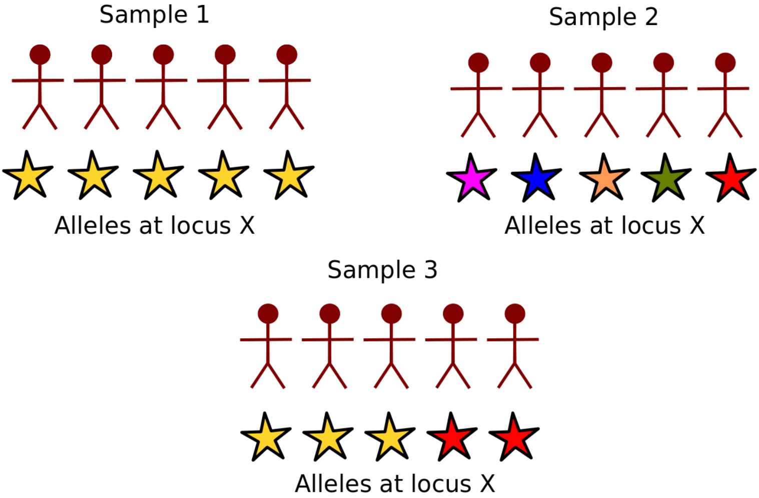
Introduction to the Coalescent

Mark Reppell

July 21, 2015

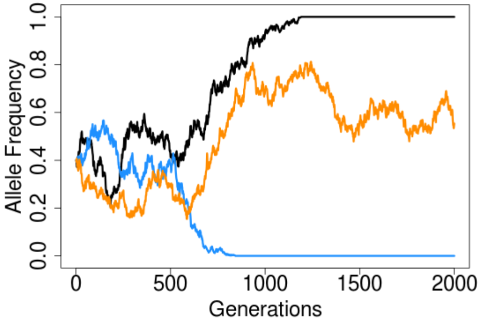
Why do we need models for how genes behave in populations?



Outline

- Genetic drift and the Wright-Fisher model
- What is the coalescent?
- Math behind the model
- Features of a genealogy
- Mutations and the infinite sites model

Genetic drift: the luck of the draw



Over time alleles may:

become fixed

remain segregating

become lost

Genetic drift refers to changes in allele frequencies over time as alleles **by chance** produce more/less offspring each generation

This contrasts with *natural selection* where alleles **systematically** produce more/less offspring each generation

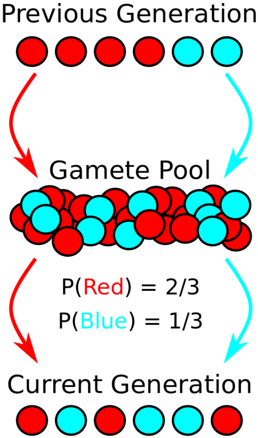
The Wright-Fisher model of genetic drift

In a population with size $2N$, two alleles are segregating at a genetic locus

Each new generation, every chromosome inherits one of the allele types, with probability p equal to $\frac{x}{2N}$ where x is the number of alleles of that type in the proceeding generation

At the population level the number of alleles of type i in a generation follows a binomial distribution, with p the allele frequency in the proceeding generation:

$$Prob(X_i = x) = \binom{2N}{x} p^x (1-p)^{2N-x}$$

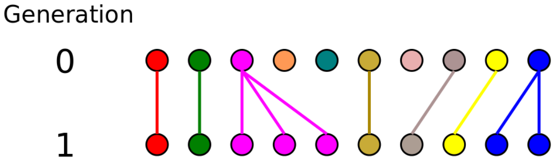


Assumptions of the Wright-Fisher model

- Discrete and non-overlapping generations
- Haploid chromosomes
- Randomly mating population with constant size
- No recombination or selection

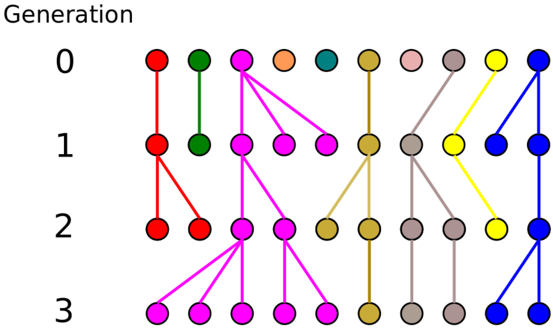
A sample evolving under Wright-Fisher

Every generation each sample has an ancestor in the preceding generation



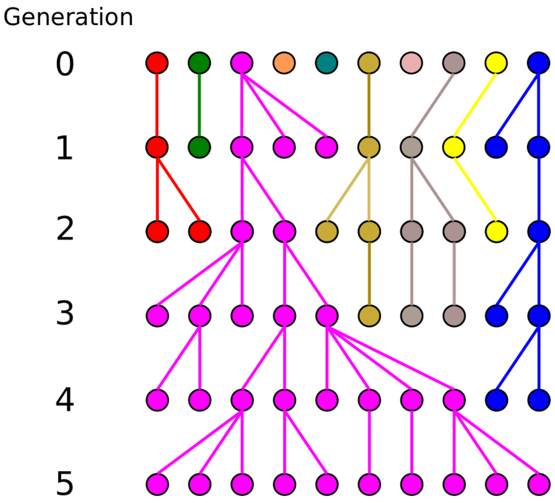
A sample evolving under Wright-Fisher

After several generations, all samples are descendants of a limited number of original individuals

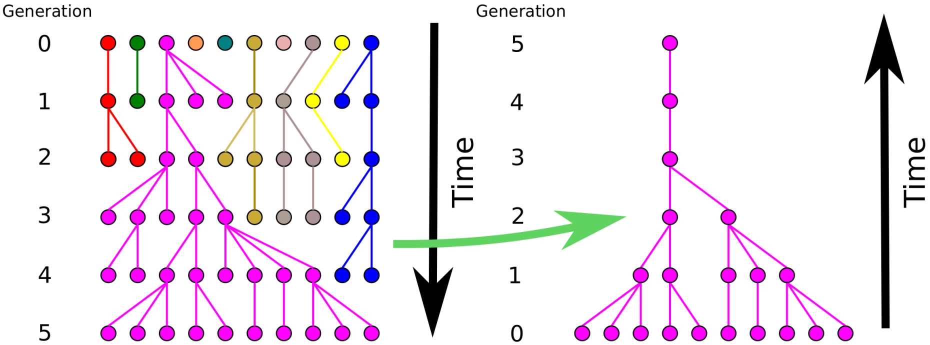


A sample evolving under Wright-Fisher

Eventually, all lineages will share a single common ancestor

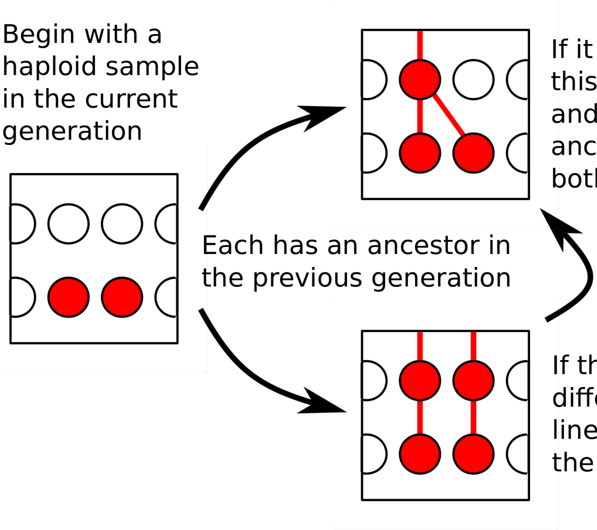


Reversing our perspective



In the present, we cannot observe lineages that have been lost from the population

The discrete time Kingman coalescent



Begin with a haploid sample in the current generation

Each has an ancestor in the previous generation

If it is the same ancestor, this is a **coalescence**, and all preceding ancestors are common to both samples

If the ancestors are different, separate lineages continue into the past

How many generations does it take to find a common ancestor?

In a population with size $2N$

$n = 2$

P(coalesce in next gen) = $\frac{1}{2N}$

P(don't coalesce in 1 gen) = $(1 - \frac{1}{2N})$

P(don't coalesce in j gen) = $(1 - \frac{1}{2N})^j$

P(j gens until coalescence) =

$$(1 - \frac{1}{2N})^{j-1}\frac{1}{2N}$$

$n \geq k > 2$

P(don't coalesce in 1 gen) =

$$\frac{(2N-1)}{2N}\frac{(2N-2)}{2N}\cdots\frac{(2N-k+1)}{2N}$$

$$= \prod_{i=1}^{k-1}(1-\frac{i}{2N}) = 1-\sum_{i=1}^{k-1}\frac{i}{2N}+O(\frac{1}{N^2})$$

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

Time until the first coalescence

For $n \geq k > 2$, from last slide:

P(don't coalesce in 1 gen) $\approx 1 - \binom{k}{2}\frac{1}{2N}$

so, with assumption of at most 1 coalescence in a generation:

P(coalesce in 1 gen) = $\binom{k}{2}\frac{1}{2N}$

and, for T_k^* time until first coalescent event:

$$\text{P}(T_k^* = j \text{ generations}) = \left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{j-1}\binom{k}{2}\frac{1}{2N}$$

The continuous time coalescent

While the Wright-Fisher works with discrete generations, it is computationally beneficial to work with continuous time

$$P(T_k^* > j \text{ gens}) = \left(1 - \frac{\binom{k}{2}}{2N}\right)^j$$

Scale by the average time for two lineages to find a common ancestor, $t = \frac{j}{2N}$:

$$P(T_k > t) = \left(1 - \frac{\binom{k}{2}}{2N}\right)^{2Nt} \rightarrow \lim_{2N\rightarrow\infty} \left(1 - \frac{\binom{k}{2}}{2N}\right)^{2Nt} = e^{-\binom{k}{2}t}$$

T_k is an exponential random variable, with rate $\binom{k}{2}$

The basic coalescent algorithm

1) Start with $k = n$

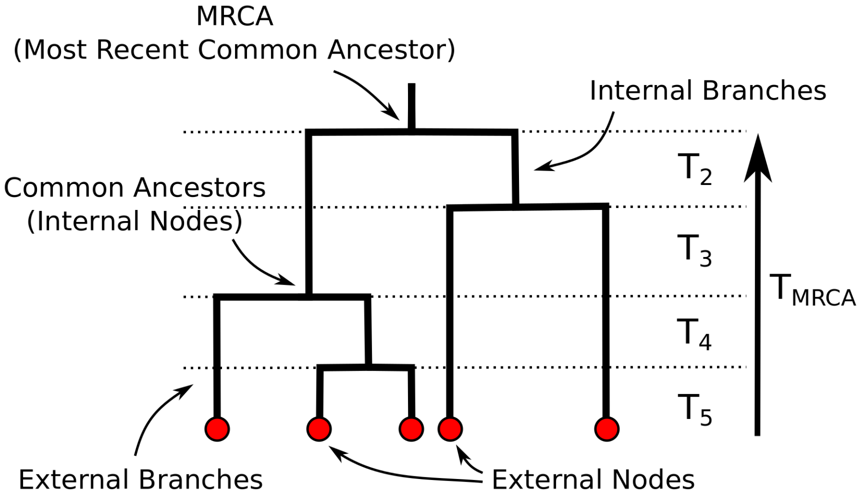
2) Simulate waiting time T_k to next event, $T_k \sim \mathrm{Exp}\left(\binom{k}{2}\right)$

3) Choose pair of lineages (i, j) uniformly among $\binom{k}{2}$ possible pairs

4) Merge i and j into single lineage, and decrease sample size by one, $k \to k - 1$

5) If $k \geq 2$ go to **2)**, otherwise stop

Anatomy of a coalescent genealogy



Coalescent times

Each coalescent time is independent of all other times

Majority of $E(T_{MRCA})$ is $E(T_2)$

Variance in times is small for big k and big for small k

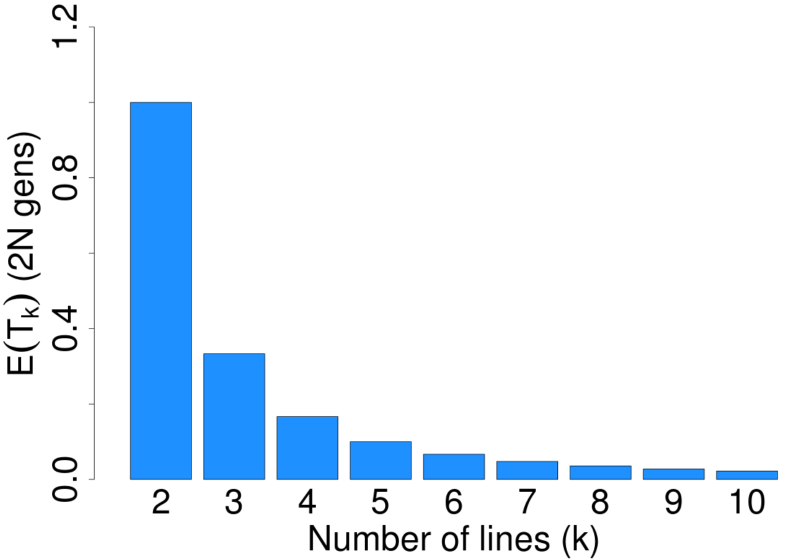
$$E(T_k) = \binom{k}{2} = \frac{2}{k(k-1)}$$

$$Var(T_k) = \left(\frac{2}{k(k-1)}\right)^2$$

$$E(T_2) = 1 = 2N \text{ gens}$$

$$Var(T_2) = 1$$

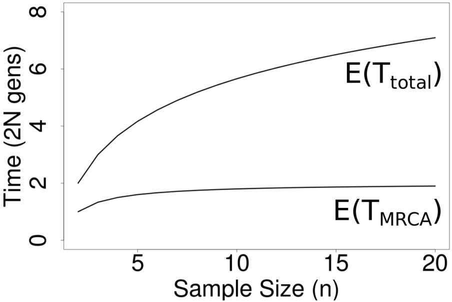
Expected coalescent times



Total tree length is the summed length of all the branches on the tree

$$T_{total} = \sum_{k=2}^{n} kT_k$$

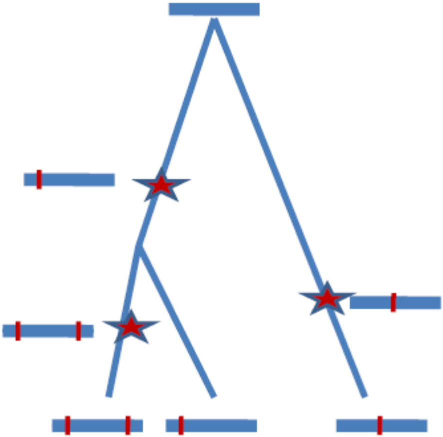
$$E(T_{total}) = \sum_{k=2}^{n} kE(T_k) = \sum_{k=2}^{n} k\frac{2}{k(k-1)} = \sum_{k=1}^{n-1} \frac{2}{k} \approx 2ln(n-1)$$



The infinite sites mutation model

Infinite sites is a mutation model where every polymorphism is the result of a single mutation event. Each mutation creates a new polymorphic site

Infinite sites is a reasonable assumption for long genetic sequences with low mutation rates



Theta

The parameter θ is the **scaled mutation rate** (also called the population mutation rate)

Usually defined as $\theta = 4N\mu b$, where $2N$ is the population size, μ is the per base per gen mutation rate, and b is the length in bases of the locus

θ can be interpreted as the expected number of mutations separating a sample of 2 sequences

(2 branches $\times 2N$ gen $\times b$ bases $\times \mu$ mutations/base*gen)

Mutations are modeled as a Poisson process

Along a coalescent genealogy, mutations are modeled as Poisson distributed with rate $\frac{\theta}{2}$ per **coalescent time unit** ($2N$ generations)

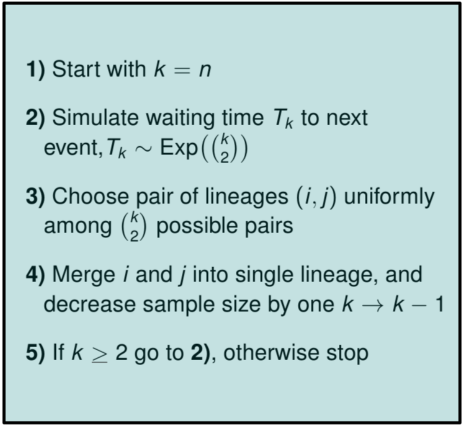
The number of mutations, x during time t:

$$P(X = x|t) = \frac{\left(\frac{\theta t}{2}\right)^x}{x!} e^{-\frac{\theta t}{2}} \quad E(X|t) = Var(X|t) = \frac{\theta t}{2}$$

With this definition the time between mutation events is exponentially distributed with rate $\frac{\theta}{2}$

Adding mutations to the coalescent algorithm

Previous algorithm



1) Start with $k = n$

2) Simulate waiting time T_k to next event, $T_k \sim \text{Exp}(\binom{k}{2})$

3) Choose pair of lineages (i, j) uniformly among $\binom{k}{2}$ possible pairs

4) Merge i and j into single lineage, and decrease sample size by one $k \to k - 1$

5) If $k \geq 2$ go to **2)**, otherwise stop

6) For each branch along genealogy, with length ℓ

> **a)** draw x mutations, $x \sim Pois(\frac{\theta\ell}{2})$
>
> **b)** select location of each mutation along sequence uniformly

Generally, sequence is treated as having length 1, so mutation locations follow Uniform(0,1)

The first lab

- Review working with the Linux command line
- Install the program *ms*
- Run a simple coalescent simulation and explore the output