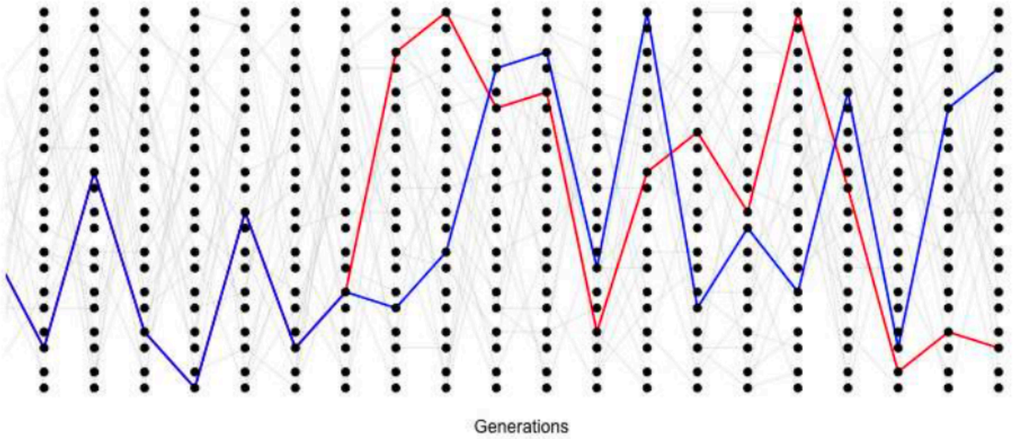
Lecture 5: The Coalescent and patterns of neutral diversity

Population genetic PCB4553/6685

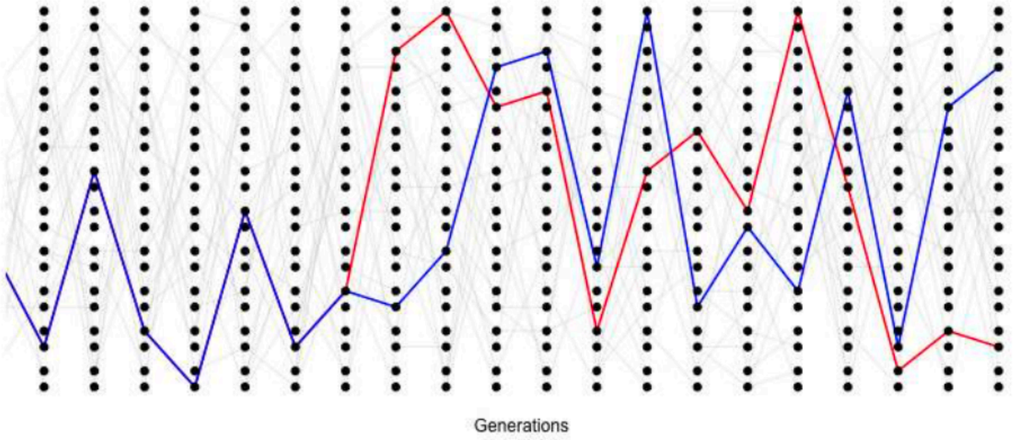
The coalescent

- The coalescent process describes how alleles in a population may have been traced back to the ancestral allele
- Looking back in time



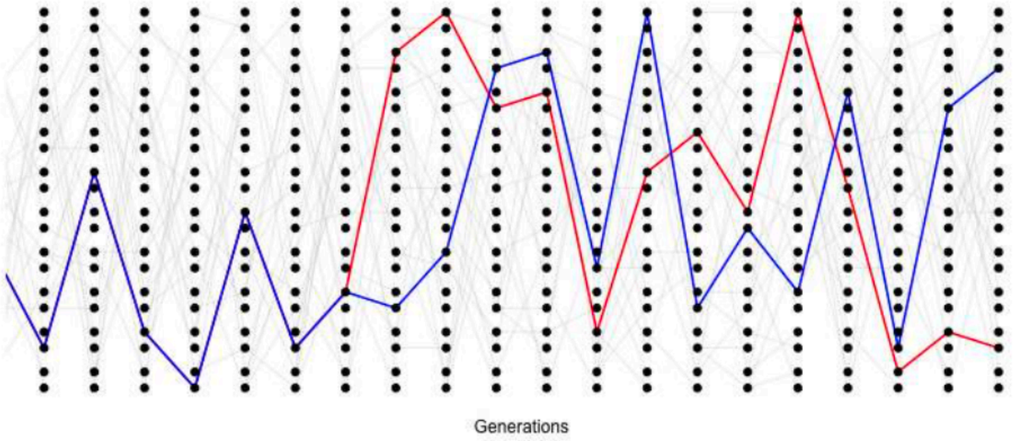
The coalescent

- Today (1/23): coalescent for a pair of alleles
- Thursday (1/25): coalescent for a sample of alleles



Time to common ancestor for a pair of alleles

- Assume a Wright-Fisher population with size N
- Probability that a pair of alleles were derived from different ancestral alleles in the previous generation
- i.e. they failed to *coalesce* in the previous generation is
  - $1 - \frac{1}{2N}$
- What is the distribution of the number of generations to coalescent?



Time to common ancestor for a pair of alleles

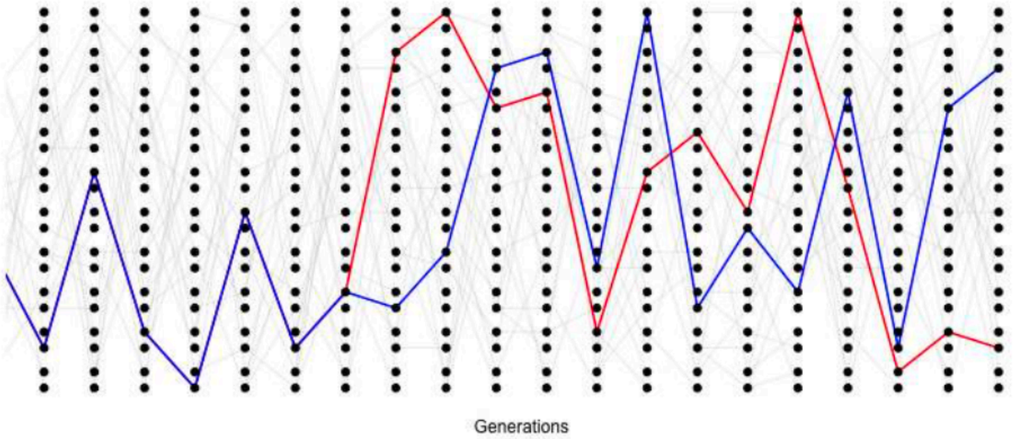
- Probability that a pair of alleles have failed to coalesce in t generations and then coalesce in the $t+1$ generation back is

- $P(T_2 = t+1) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^t$

- T_2 follows geometric distribution with $p = 1/2N$

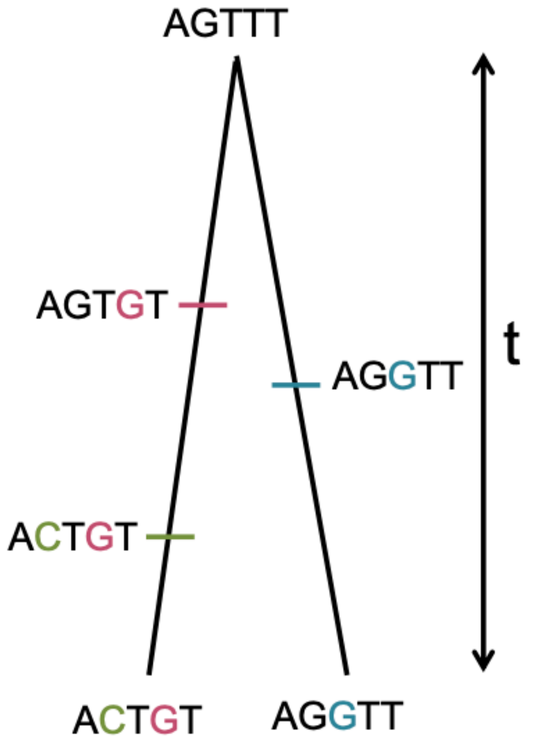
- $\mathbb{E}(T_2) = 2N$

- $\text{VAR}(T_2) = \frac{1-p}{p^2} \approx 4N^2$



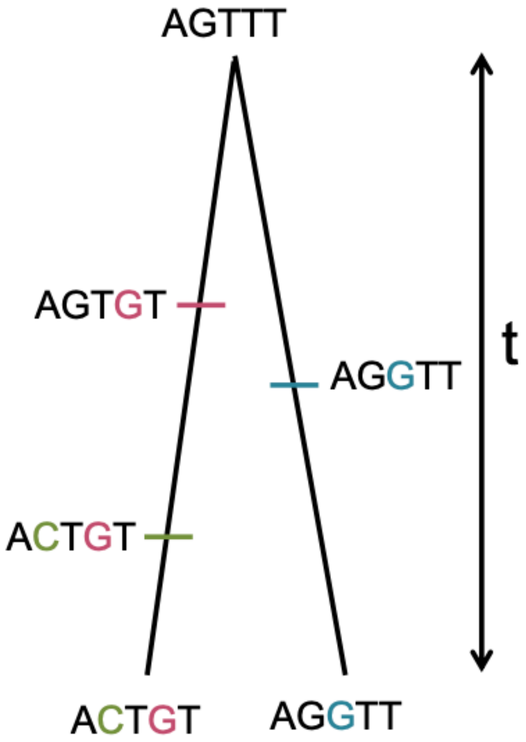
Adding mutations to the coalescent

- Conditional on a pair of alleles coalescing t generations ago, there are $2t$ generations in which a mutation could occur
- The number of mutations s_2 since they last shared a common ancestor is binomially distributed
- $P(S_2 \mid T_2 = t) = \binom{2t}{j} \mu^j (1-\mu)^{2t-j}$
- $\mathbb{E}(S_2 \mid T_2 = t) = 2t \times \mu$



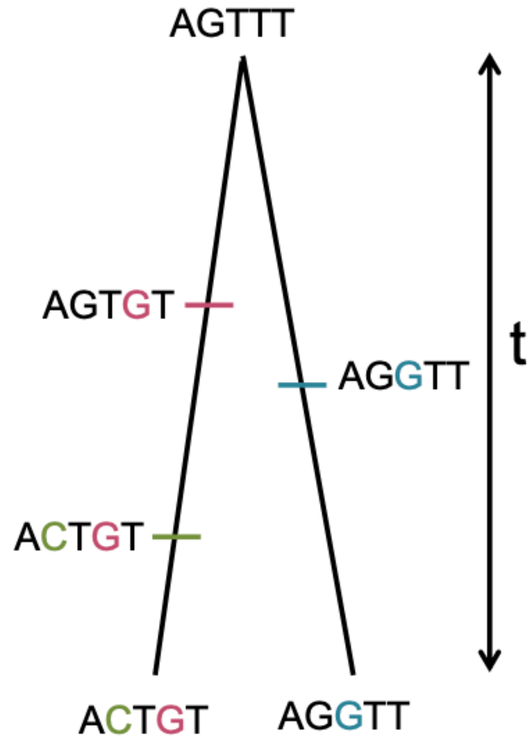
Levels of heterozygosity maintained by mutation-drift balance

- We'd like to derive the expected heterozygosity rate under random mating, and mutation-drift balance
- Expected heterozygosity (H) = 1 - P of two random alleles ar identical
- Only way for two alleles to be identical is that they share a common ancestor and have not since accumulated any mutations



Levels of heterozygosity at the mutation-drift balance

- P of a random pair of alleles coalesce in the preceding generation
  $$= \frac{1}{2N}$$
- P of them fail to coalesce $= 1 - \frac{1}{2N}$
- P a mutation occur in a generation $= \mu$
- P no mutation occurring $= 1 - \mu$
- Putting these together:
  - P of two randomly sampled alleles coalesce 2 generations the past and are identical
    $$= \left(1 - \frac{1}{2N}\right) \frac{1}{2N} (1 - \mu)^4$$



Levels of heterozygosity at the mutation-drift balance

- More generally, the probability that our alleles coalesce in generation t + 1 and are identical due to no mutation

$$\mathbb{P}(\text{coal. in t+1 \& no mutations}) = \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{t}(1-\mu)^{2(t+1)}$$

$$\mathbb{P}(\text{coal. in t+1 \& no mutations}) \approx \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{t}(1-\mu)^{2t}$$

Levels of heterozygosity at the mutation-drift balance

- Two alleles could coalesce in generation $t = 1,2,3\cdots$

- Apply law of total probability

$$\begin{aligned}\mathbb{P}(\text{coal. in any generation \& no mutations}) \approx & \mathbb{P}(\text{coal. in } t=1 \text{ \& no mutations}) + \\ & \mathbb{P}(\text{coal. in } t=2 \text{ \& no mutations}) + \ldots \\ = & \sum_{t=1}^{\infty} \mathbb{P}(\text{coal. in } \ t \text{ generations \& no mutation})\end{aligned}$$

- Assuming that 1/(2N) ≪ 1 and μ ≪ 1. This allows us to approximate the geometric decay as an exponential decay

$$\begin{aligned}\mathbb{P}(\text{coal. in t+1 \& no mutations}) &\approx \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{t}(1-\mu)^{2t} \\ &\approx \frac{1}{2N}e^{-t/(2N)}e^{-2\mu t} \\ &= \frac{1}{2N}e^{-t(2\mu+1/(2N))}\end{aligned}$$

- Approximate sum with integral:

$$\frac{1}{2N}\int_0^{\infty} e^{-t(2\mu+1/(2N))}dt = \frac{1/(2N)}{1/(2N)+2\mu}$$

- equilibrium heterozygosity in a population at equilibrium between mutation and drift

$$H = \frac{2\mu}{1/(2N)+2\mu} = \frac{4N\mu}{1+4N\mu}$$

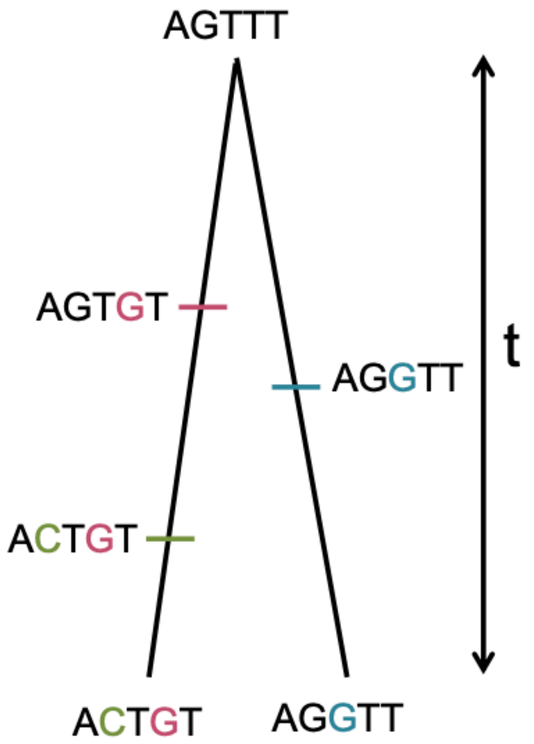
$$\theta = 4N\mu$$

Expected number of segregating sites

- Conditional on a pair of alleles coalescing t generations ago, there are $2t$ generations in which a mutation could occur
- Expected number of mutations separating two alleles drawn at random from the population

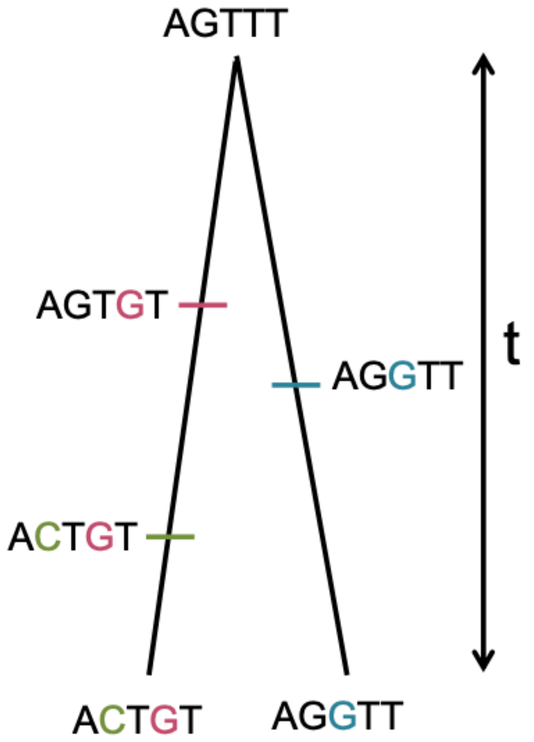
$$\begin{aligned}\mathbb{E}(S_2) &= \sum_{t=0}^{\infty} \mathbb{E}(S_2|T_2 = t)P(T_2 = t) \\ &= \sum_{t=0}^{\infty} 2\mu t P(T_2 = t) \\ &= 2\mu\mathbb{E}(T_2) \\ &= 4\mu N\end{aligned}$$

- Law of total expectation



Expected number of segregating sites

- Infinite many site assumption
  - Mutation is rare enough that it never happens at the same basepair twice, i.e. no multiple hits, such that we get to see all of the mutation events that separate our pair of sequences.
- number of mutations between a pair of sites
  =
  observed number of differences between a pair of sequences π
- $\mathbb{E}(\pi) = 4N\mu = \theta$
- Empirical estimate of θ
  - $\pi = \widehat{\theta_\pi} = 4N_e\mu$
- $2N_e$ = effective coalescent population size



Continuous-time approximation

- We can approximate the geometric distribution with exponential distribution:
  - $P(T_2 = t + 1) = \frac{1}{2N}(1 - \frac{1}{2N})^t$
  - $P(T_2 = t + 1) \to \frac{1}{2N} e^{-\frac{1}{2N}t}$
  - Roughly this can be seen as the first order Taylor expansion of the exponential function

Limiting distribution of coalescent time

- Set $\lambda = \frac{1}{2N}$
- Dividing the interval $[0,t]$ in to $n \times t$ small intervals, each with width $1/n$
- Assume the coalescent now happens in a small interval at rate $= \lambda \times 1/n$
- Probability that the outcome doesn't happen between 0 and t
  $= (1 - \frac{\lambda}{n})^{n \times t}$
- Now let a generation be divided into increasingly smaller pieces, i.e. $n \to \infty$
- Use the fact that $(1 + \frac{x}{n})^n$ approximates e^x as x goes to zero
- P(no coalescent until time t) $\to e^{-\lambda t}$
- $P(T < t) = 1 - e^{-\lambda t}$, **Cumulative distribution function (CDF)**
- $f(t) = \lambda e^{-\lambda t}$, **Density function (PDF)**

Limiting distribution of the number of mutations

- What happens to the distribution of the number of mutations in the continuous time limit?
- Recall the binomial distribution of mutations
- $P(S_2 \,|\, T_2 = t) = \binom{2t}{j} \mu^j (1-\mu)^{2t-j}$
- Dividing the interval $[0,t]$ in to $n \times t$ small intervals, each with width $\delta = 1/n$
- Let the probability of a mutation occurring in the interval $[t, t+\delta]$ be μ/n
- $P(S_2 \,|\, T_2 = t) = \binom{2t \times n}{j} \left(\frac{\mu}{n}\right)^j \left(1 - \frac{\mu}{n}\right)^{2t \times n - j}$
- $P(S_2 \,|\, T_2 = t) \to \frac{(2\mu t)^j}{j!} e^{-2\mu t}$