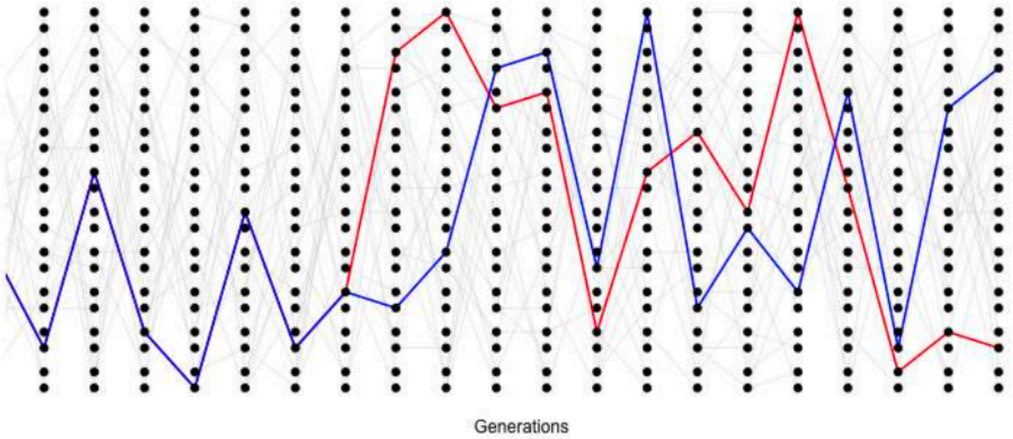
Lecture 6: The coalescent process for a sample of alleles

Population genetic PCB4553/6685

The coalescent

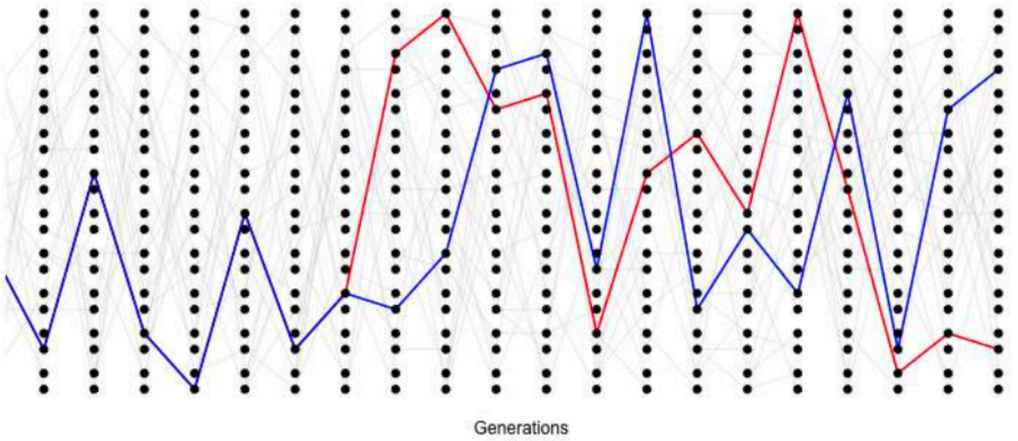
- The coalescent process describes how alleles in a population may have been traced back to the ancestral allele
- Looking back in time

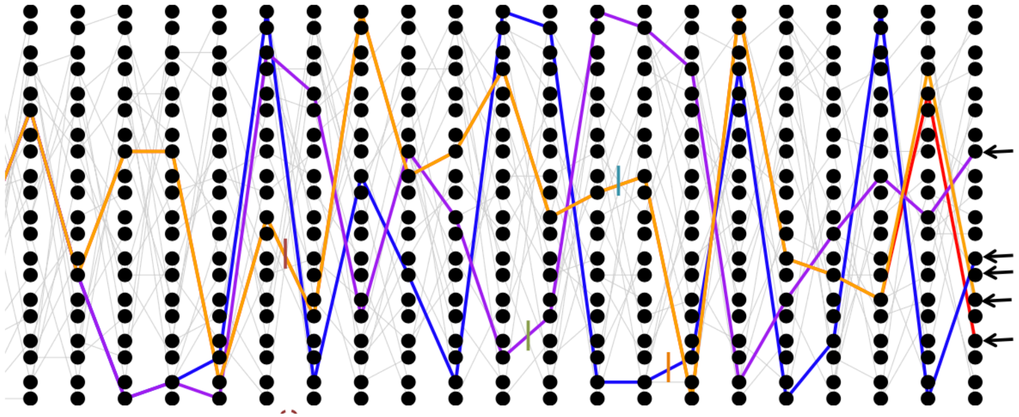


- Usually we are not just interested in pairs of alleles, or the average pairwise diversity.

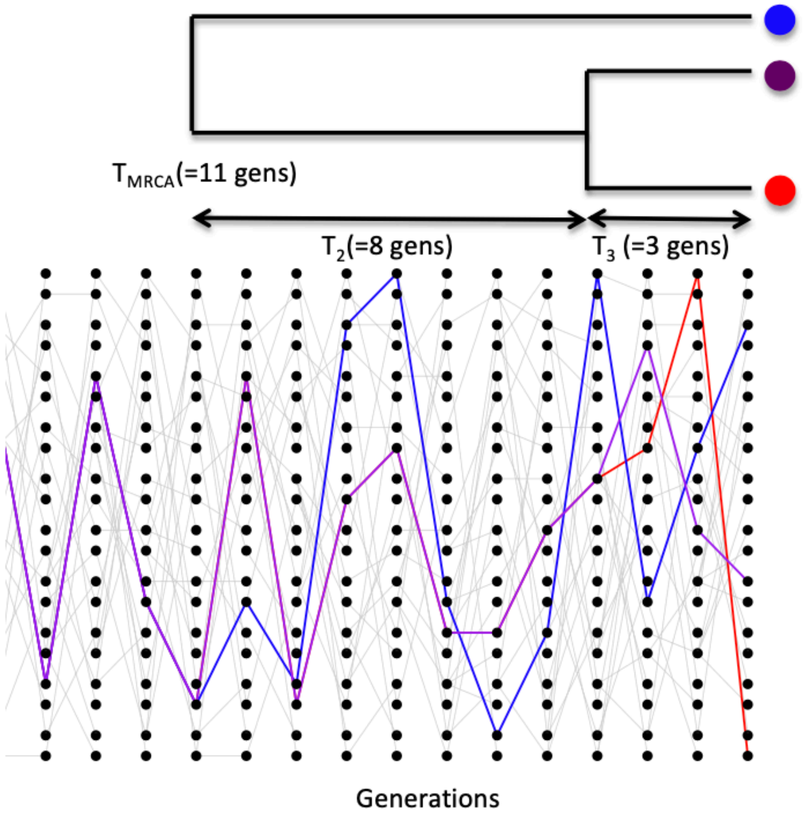
- Generally we are interested in the properties of diversity in samples of a number of alleles drawn from the population.

- Instead of just following a pair of lineages back until they coalesce, we can follow the history of **a sample** of alleles back through the population.

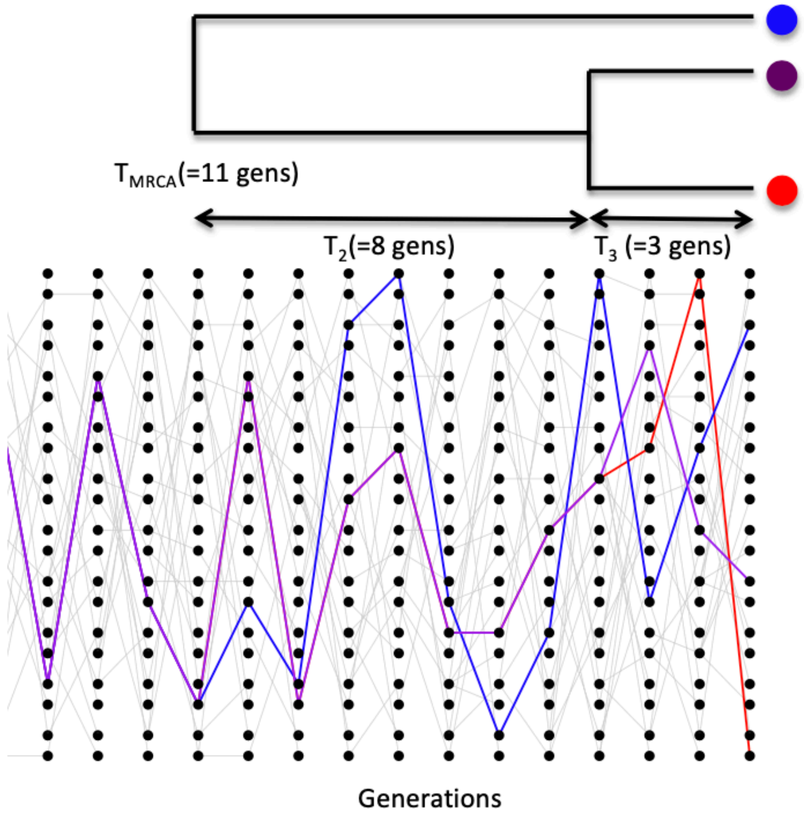




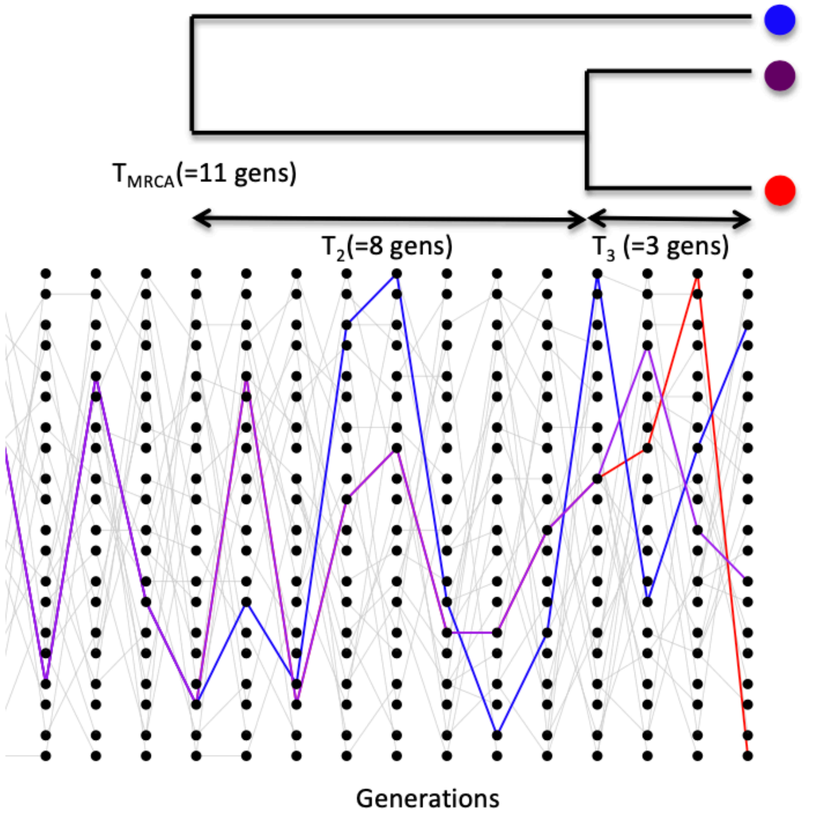
- P of Coalescent for a pair of alleles = $1/2N$
- P of simultaneous coalescent of more than two alleles
  $= \mathcal{O}(\frac{1}{(2N)^2})$
  - Negligible for large N
- Safe to consider pairwise coalescent if sample size $i \ll N$



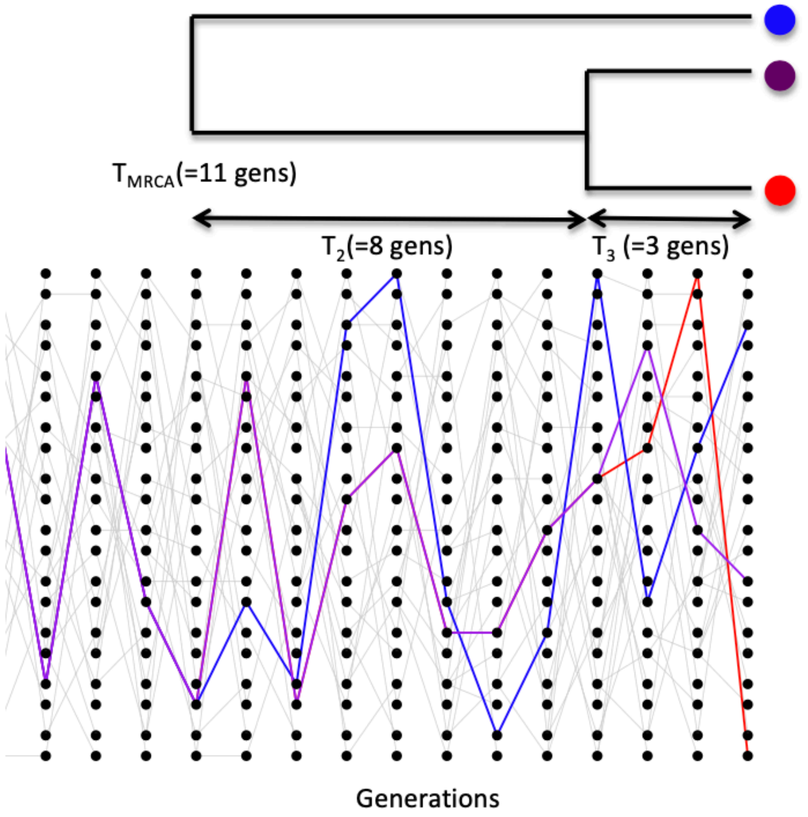
- Consider a sample of three alleles
- There are $\binom{3}{2} = 3$ possible pairs of alleles
- probability that no pair finds a common ancestor in the preceding generation is
- $\left(1 - \frac{1}{2N}\right)^3$



- More generally, for a sample of i alleles
- Probability that no pair of alleles in a sample of size i coalesces in the preceding generation is
- $\left(1-\frac{1}{2N}\right)^{\binom{i}{2}}$
- $\left(1-\frac{1}{2N}\right)^{\binom{i}{2}} \approx 1-\frac{\binom{i}{2}}{2N}$



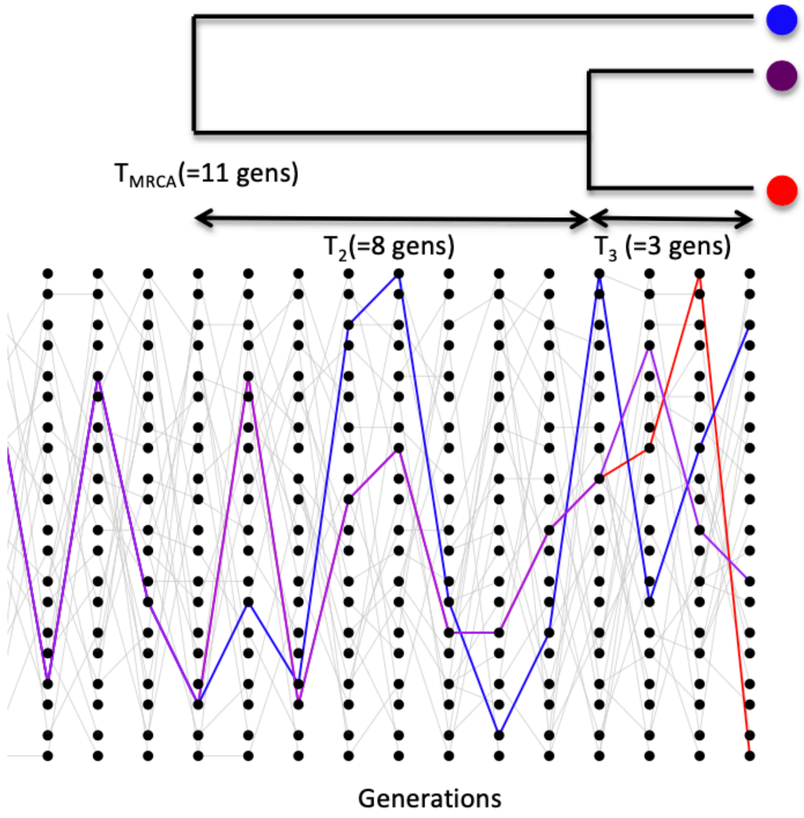
- Another way to derive this is
- The probability that no coalescent occurs is
- $(1-\frac{1}{2N})(1-\frac{2}{2N})(1-\frac{3}{2N})\cdots(1-\frac{i-1}{2N})$
- $\approx e^{-\frac{1}{2N}\times\sum_{j=1}^{i-1} j}$
- $= e^{-\frac{1}{2N}\binom{i}{2}}$
- $\approx (1-\frac{1}{2N})^{\binom{i}{2}}$



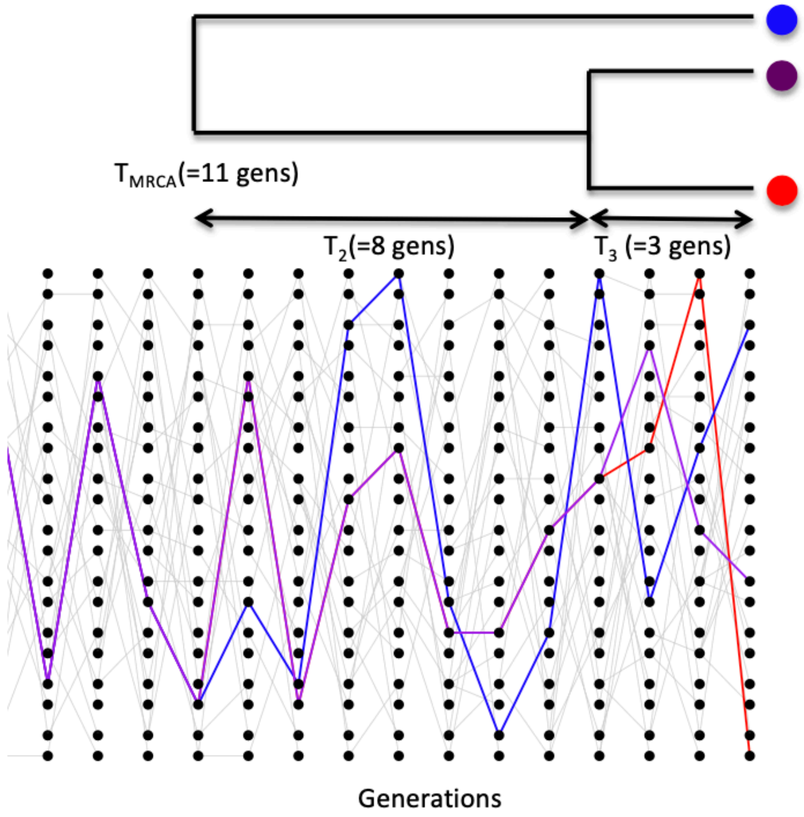
- We have assumed that only coalescent between pairs of alleles are possible
- When there are i alleles, the probability that we wait until the $t+1$ generation before any pair of alleles coalesce is

$$\mathbb{P}(T_i = t+1) = \frac{\binom{i}{2}}{2N}\left(1 - \frac{\binom{i}{2}}{2N}\right)^t$$

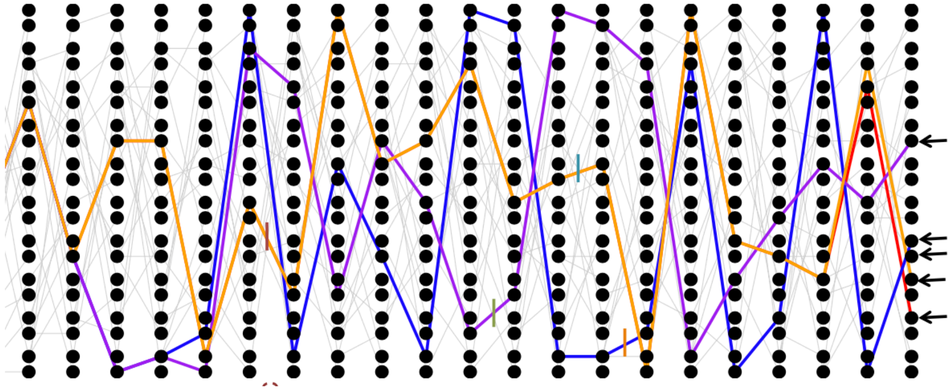
- Waiting time to the first coalescent event while there are i lineages is a geometrically distributed with $p = \binom{i}{2}/2N$
- $T_i \sim \text{Geo}(\binom{i}{2}/2N)$



- Waiting time to the first coalescent event while there are i lineages is a geometrically distributed with $p = \binom{i}{2}/2N$
  - $T_i \sim \text{Geo}(\binom{i}{2}/2N)$
- The mean waiting time till any of pair within our sample coalesce
  - $\mathbb{E}(T_i) = 1/p = 2N/\binom{i}{2}$
- Using continuous time approximation
  - $T_i \sim \text{Exp}(\binom{i}{2}/2N)$
  - $\mathbb{E}(T_i) = 1/p = 2N/\binom{i}{2}$

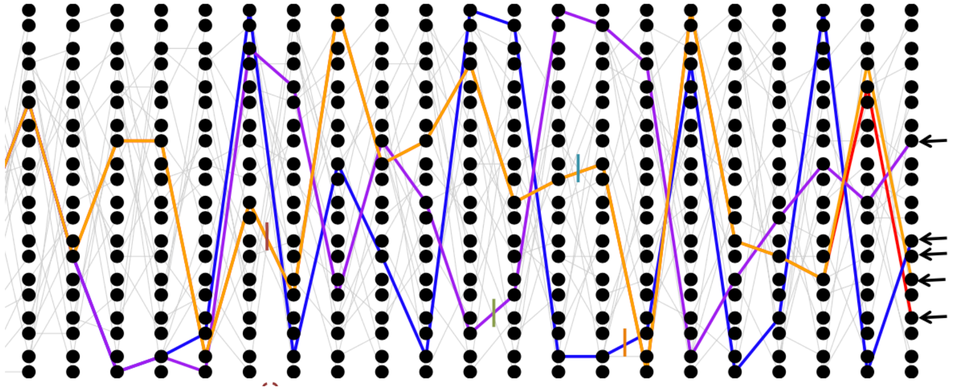


- After a pair of alleles in our sample of i alleles coalesces
- we then switch to having to follow i – 1 alleles back in time.
- Then when a pair of these i – 1 alleles coalesce, we then only have to follow i – 2 alleles back.
- This process continues until we coalesce back to a sample of two, and from there to a single most recent common ancestor (MRCA).



Algorithm to simulate a coalescent genealogy

1. Set $i = n$.
2. Simulate a random variable to be the time T_i to the next coalescent event from $T_i \sim \text{Exp}\left(\binom{i}{2}/2N\right)$
3. Choose a pair of alleles to coalesce at random from all possible pairs.
4. Set $i = i - 1$
5. Continue looping steps 2-4 until $i = 1$, i.e. the most recent common ancestor of the sample is found.

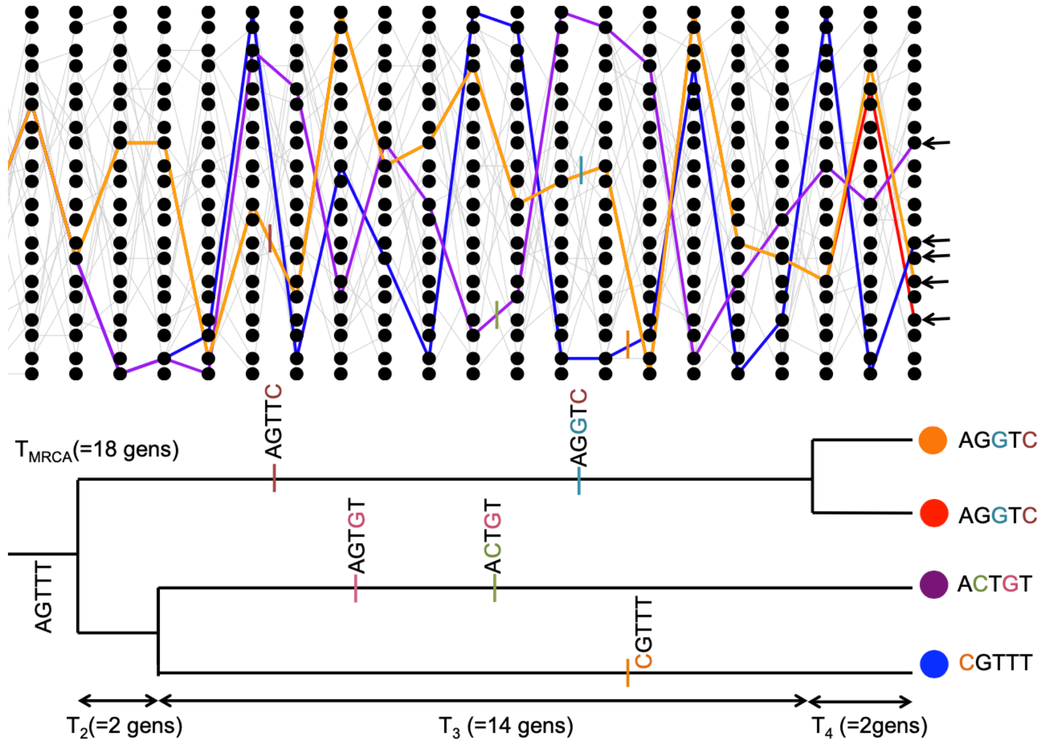


Time to MRCA

- Time to most recent common ancestor (MRCA):
- $T_{\text{MRCA}} = \sum_{i=2}^{n} T_i$
- As our coalescent times for different i are independent, the expected time to the most recent common ancestor is

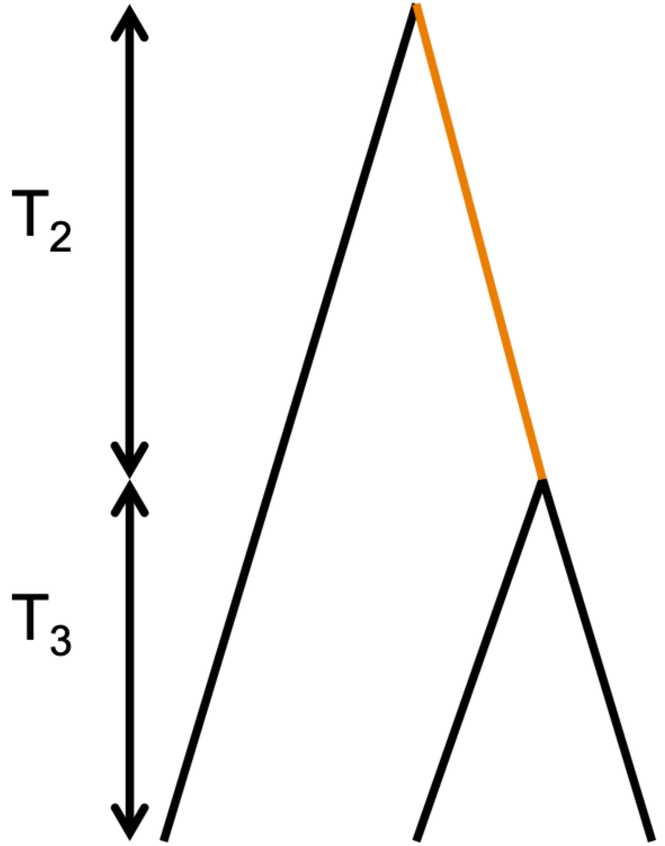
$$\mathbb{E}(T_{MRCA}) = \sum_{i=n}^{2} \mathbb{E}(T_i) = \sum_{i=n}^{2} 2N / \binom{i}{2}$$

$$\mathbb{E}(T_{MRCA}) = 4N\left(1 - \frac{1}{n}\right)$$



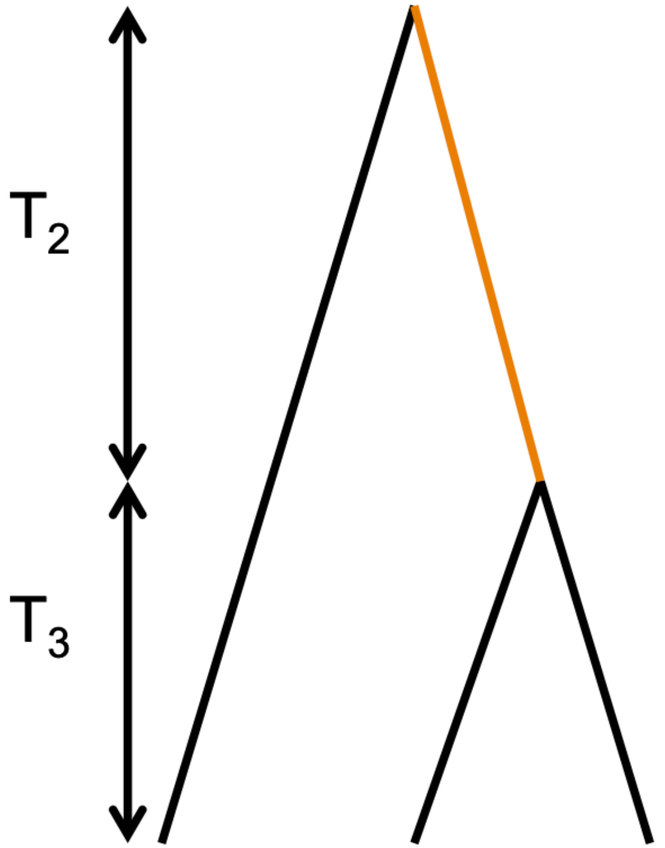
The expected total time in a genealogy and the number of segregating sites

- Total amount of time in the genealogy of the sample, or the sum of all the *branch lengths* on the genealogical tree
- $T_{\text{tot}} = \sum_{i=n}^{2} iT_i$
- Expected total time in a genealogy:
- $$\mathbb{E}(T_{tot}) = \sum_{i=n}^{2} i\frac{2N}{\binom{i}{2}} = \sum_{i=n}^{2} \frac{4N}{i-1} = \sum_{i=n-1}^{1} \frac{4N}{i}$$



Adding mutations

- Mutations follow a Poisson (point) process
- Use the infinite site assumption
  - Number of mutations = number of segregating sites
- So the total number of segregating sites in our sample (S) is Poisson with mean μT_{Tot}
- Thus the expected number of segregating sites in a sample of size n is
- $$\mathbb{E}(S) = \mu \mathbb{E}(T_{tot}) = \sum_{i=n-1}^{1} \frac{4N\mu}{i} = \theta \sum_{i=n-1}^{1} \frac{1}{i}$$



- Expected number of segregating sites in a sample of size n is

$$\mathbb{E}(S) = \mu \mathbb{E}(T_{tot}) = \sum_{i=n-1}^{1} \frac{4N\mu}{i} = \theta \sum_{i=n-1}^{1} \frac{1}{i}$$

- We can use this formula to derive another estimate of the population scaled mutation rate θ, by setting our observed number of segregating sites in a sample (S) equal to this expectation

$$\widehat{\theta}_W = \frac{S}{\sum_{i=n-1}^{1} 1/i}$$

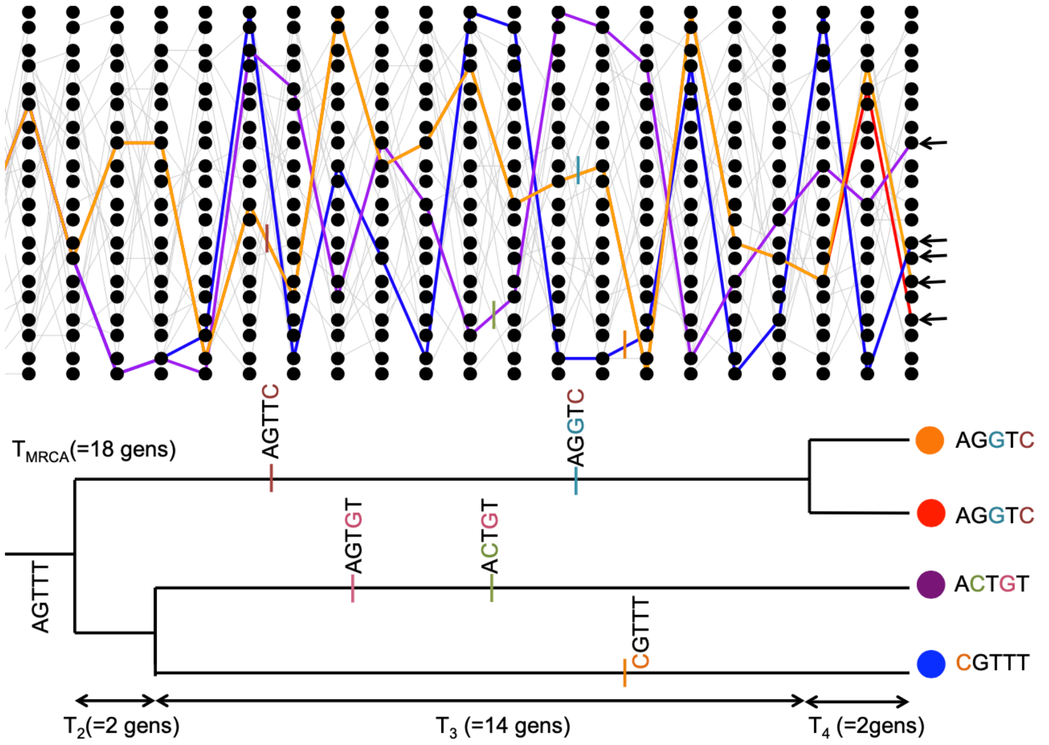
Site frequency spectrum

- What is the site frequency spectrum?
  - **Allele frequency spectrum**: distribution of the allele frequencies of a given set of loci (often SNPs) in a population or sample
  - For every locus, count number of minor allele
  - Make histogram with bin_width =1
  - Example: allele frequency spectrum is (4, 2, 1, 0, 1)

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8
Sample 1	0	1	0	0	0	0	1	0
Sample 2	1	0	1	0	0	0	1	0
Sample 3	0	1	1	0	0	1	0	0
Sample 4	0	0	0	0	1	0	1	1
Sample 5	0	0	1	0	0	0	1	0
Sample 6	0	0	0	1	0	1	1	0
Total	1	2	3	1	1	2	5	1

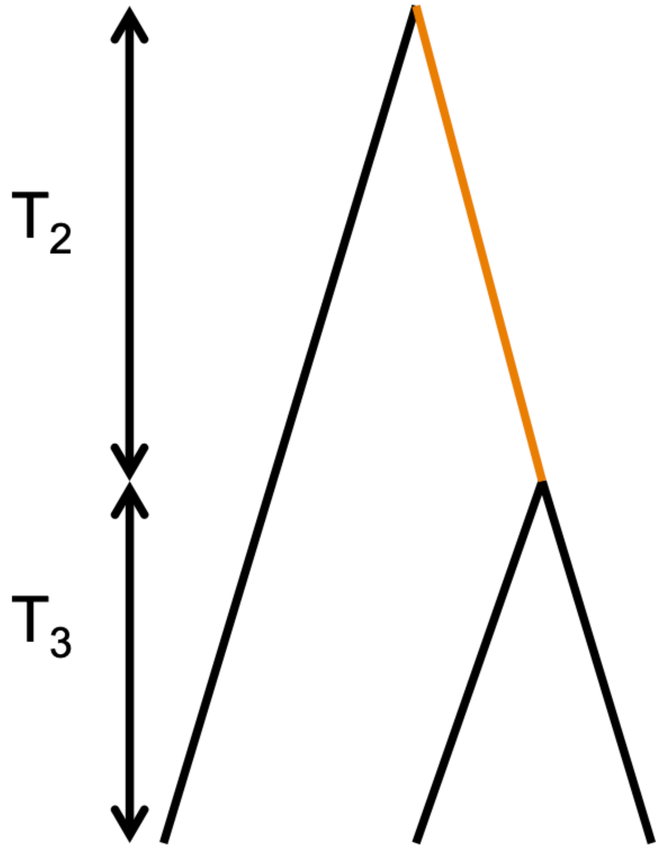
Site frequency spectrum

- We can use our coalescent process to find the expected number of derived alleles present i times out of a sample size n
- e.g. how many singletons (i = 1) do we expect to find in our sample?



Site frequency spectrum

- $n = 3$
- The total number of generations where a singleton mutation could arise is $3T_3 + T_2$
- The total number of generations where a doubleton mutation could arise is T_2
- So our expected number of singletons
- $$\mathbb{E}(S_i) = \mu\left(3\mathbb{E}(T_3) + \mathbb{E}(T_2)\right) = \mu\left(3\frac{2N}{3} + 2N\right) = \theta$$
- Expected number of doubletons:
- $$\mathbb{E}(S_i) = \theta/2$$



Site frequency spectrum

- Neutral site frequency spectrum in the general case
  $\mathbb{E}(S_i) = \frac{\theta}{i}$

- The probability of observing a derived allele segregating at frequency i/n

$$\mathbb{P}(i|0 < i < n) = \frac{\mathbb{E}(S_i)}{\sum_{j=1}^{n-1} \mathbb{E}(S_j)} = \frac{1/i}{\sum_{j=1}^{n-1} 1/j}$$

Tajima's D

- Test whether an observed site frequency spectrum conforms to its neutral, constant-size expectations.
- Tajima's D:
- $$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{C}$$
- C is the square-root of an estimator of the variance of this difference under the constant population size, neutral model.
- D have mean zero and variance 1 under the null mode

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8
Sample 1	0	1	0	0	0	0	1	0
Sample 2	1	0	1	0	0	0	1	0
Sample 3	0	1	1	0	0	1	0	0
Sample 4	0	0	0	0	1	0	1	1
Sample 5	0	0	1	0	0	0	1	0
Sample 6	0	0	0	1	0	1	1	0
Total	1	2	3	1	1	2	5	1

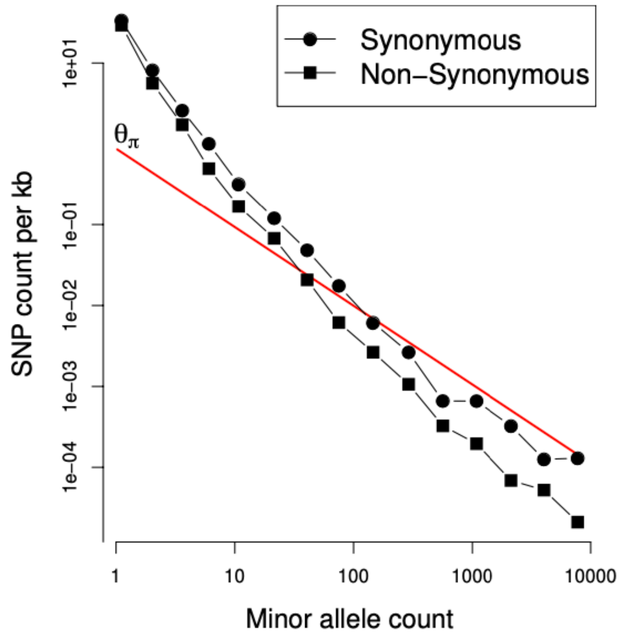
• • •

Tajima’s D

- Tajima’s D:
- $$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{C}$$
- An excess of rare alleles results in a negative Tajima’s D
  - each additional rare allele increases the number of segregating sites by 1, but only has a small effect on the number of pairwise differences between samples.
- Positive Tajima’s D reflects an excess of intermediate frequency alleles relative to the constant-size, neutral expectation.
  - Alleles at intermediate-frequency increase pairwise diversity more per segregating site than typical, thus increasing θ more than θ .

Demography and the coalescent

- Background:
  - Real populations fluctuate in size
  - We can potentially accommodate rapid random fluctuations in population size by simply using the effective population size N_e in place of N.
  - However, longer-term, more systematic changes in population size will distort the coalescent genealogies, and hence patterns of diversity, in more systematic ways.

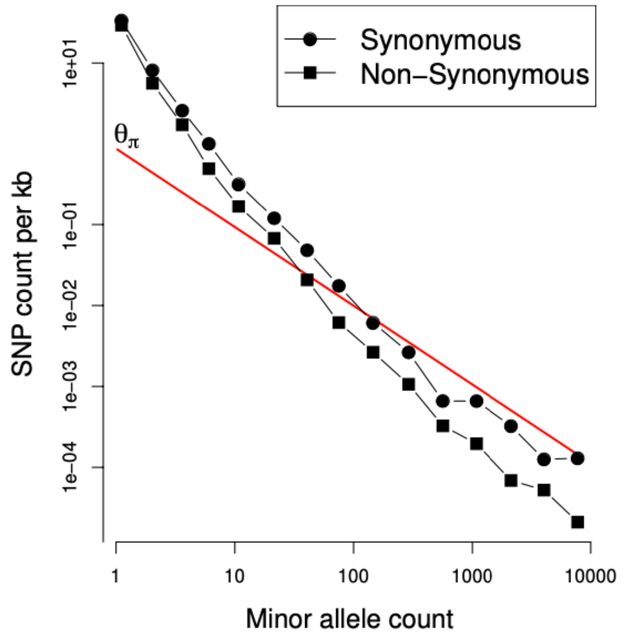


Expected and observed site frequency spectrum for 202 genes from *14,002* people of *European ancestry* (28,004 alleles).

The red line gives the neutral, constant population size estimate of the site frequency spectrum, our equation (4.42), using a θ estimated from π.

Recent population expansion

- Neutral frequency spectrum, $\mathbb{E}(S_i) = \theta/i$, is shown as a red line.
- The plot is on log-log scale
- There are vastly more rare alleles than expected under our neutral, constant size model, but the neutral prediction and reality agree somewhat more for alleles that are more common.
- Why is this?

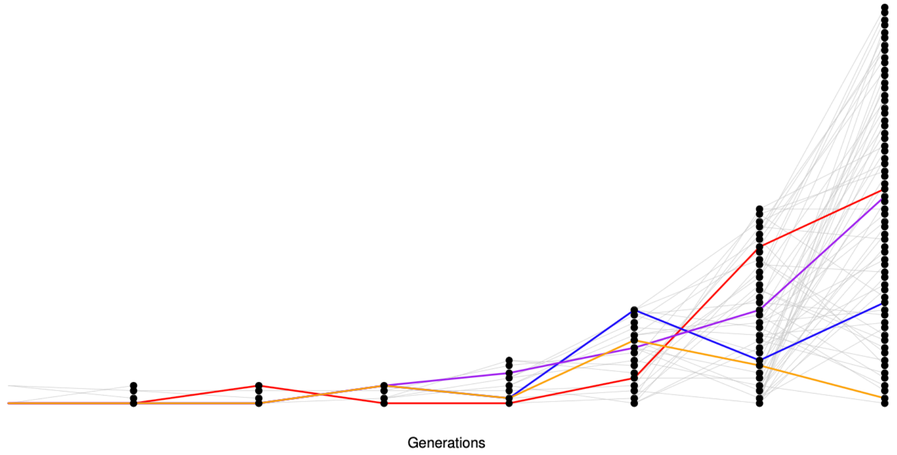


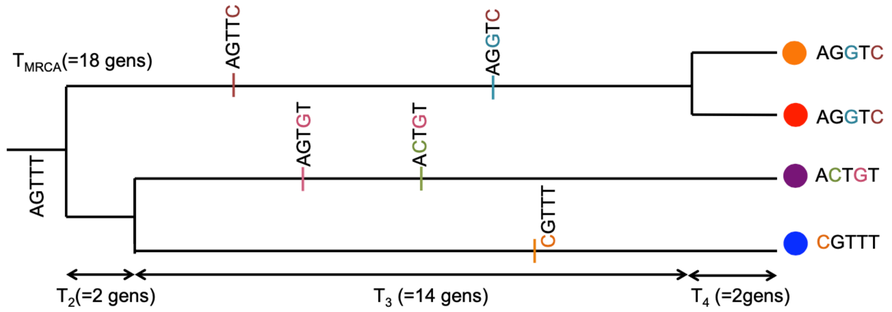
Expected and observed site frequency spectrum for 202 genes from *14,002* people of *European ancestry* (28,004 alleles).

The red line gives the neutral, constant population size estimate of the site frequency spectrum, our equation (4.42), using a θ estimated from π.

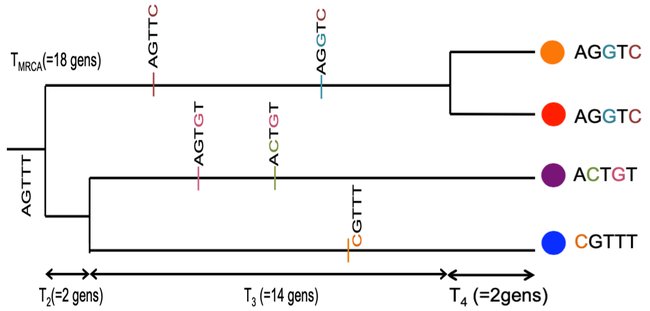
Recent population expansion

- Likely the result of the very recent explosive growth in human populations.
- If the population has grown rapidly, then the pairwise-coalescent rate in the past may be much higher than the coalescent rate closer to the present.
- **Deep branches in the genealogy are shorter than recent beaches, leading to enrichment of rare alleles**
- Recent population expansion leads to much lower genetic diversity in the population than expected under the census population size.
- Humans: N = 7 billion today, due to very rapid population growth over the past thousand to tens of thousands of years.
- Our level of genetic diversity is very much lower than you'd predict given our census size, reflecting our much smaller ancestral population.



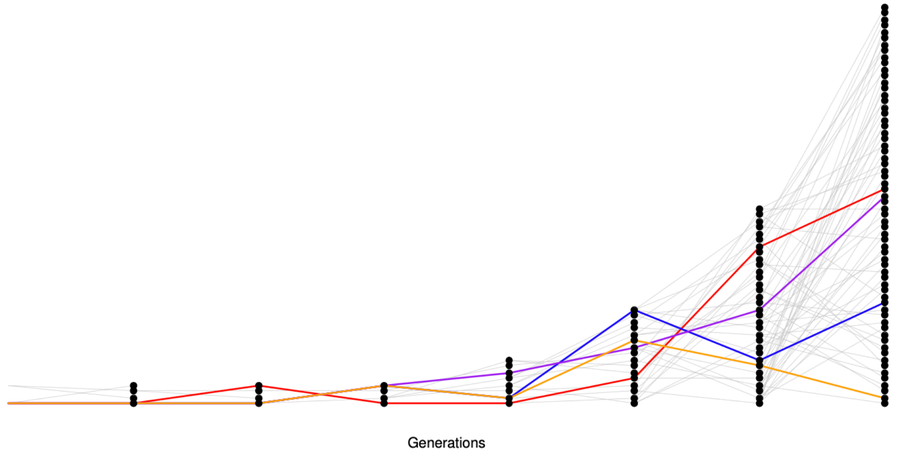


Constant pop size



Expanding pop size

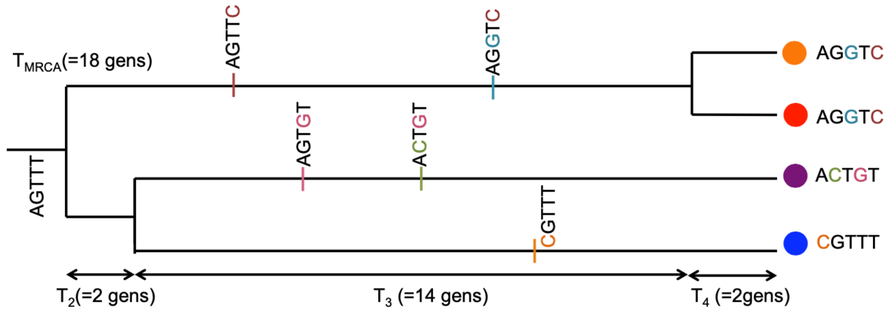
Recent population expansion

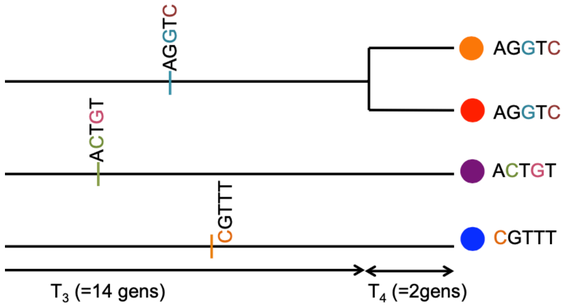


- Deeper coalescent branches are much more squished together in time compared to those in a constant-sized population.

- Mutations on deeper branches are the source of alleles at more intermediate frequencies, and so there are even fewer intermediate-frequency alleles in growing populations.

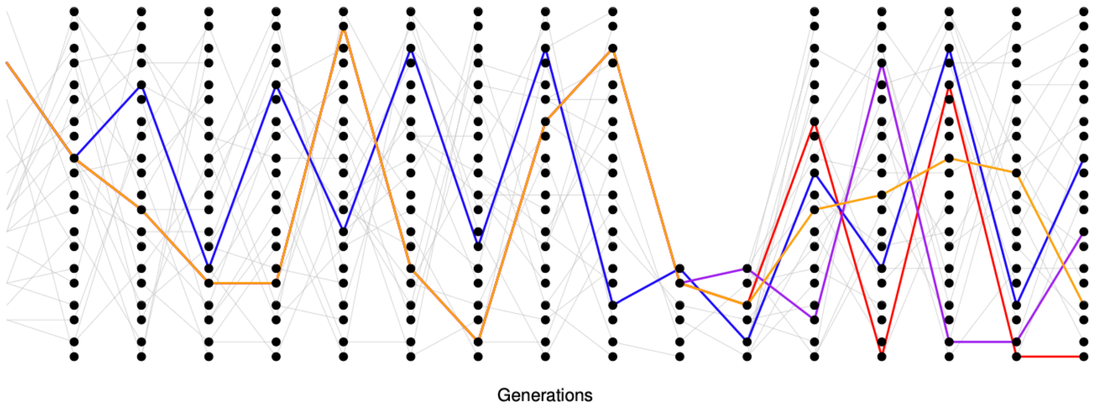
- This why there are so many rare alleles, especially singletons, in this large sample of Europeans.





Population bottleneck

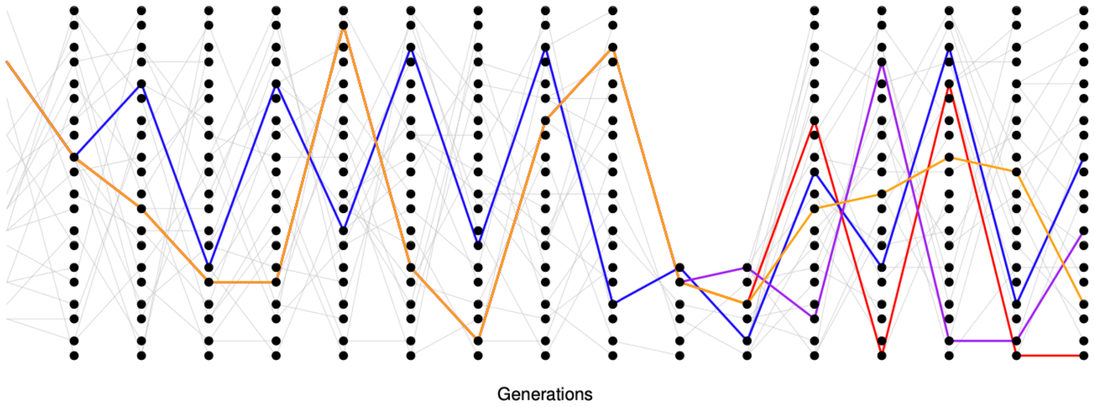
- In a bottleneck, the population size crashes dramatically, and subsequently recovers.
- For example, our population may have had size N_{Big} and crashed down to N_{Small}.
- For recent bottlenecks ($\ll N_{\text{Big}}$ generations in the past), many lineages will not have coalesced before reaching the bottleneck, moving backward in time.
- During the bottleneck lineages coalesce at a much higher rate, such that many of our lineages will coalesce if the bottleneck lasts long enough ($\sim N_{\text{Small}}$ generations).



Population bottleneck

- If the bottleneck is very strong, then all lineages will coalesce during the bottleneck, and the resulting in an excess of rare alleles.

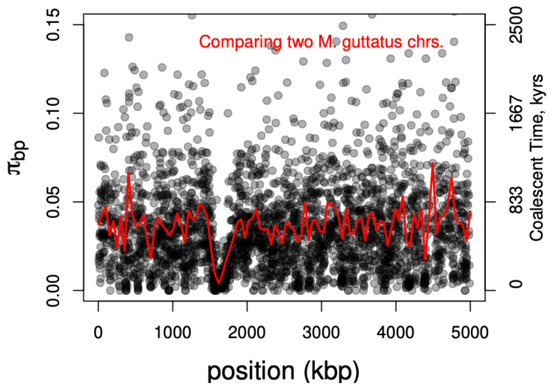
- But if some pairs of lineages escape coalescing during the bottleneck, they will coalesce much more deeply in time

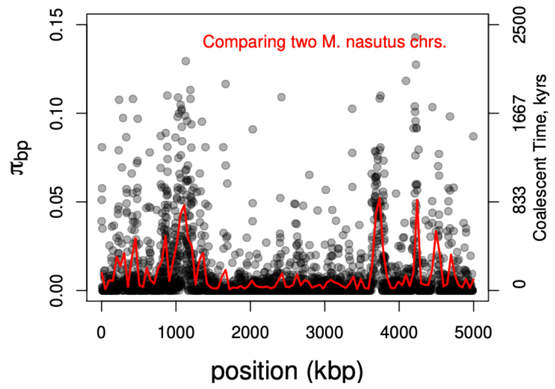


Example: *Mimulus nasutus*

Figure 4.25: Yellow Monkeyflower *M. guttatus*.
Choix des plus belles fleurs et des plus beaux fruits. Pierre-Joseph Redouté. (1833). Contributed to Flickr by Swallowtail Garden Seeds. Public Domain.

- *Mimulus nasutus* is a selfing species that arose recently from an out-crossing progenitor *M. guttatus,* and experienced a strong bottleneck.
- *M. guttatus* π = 4% at synonymous sites
- *M. nasutus* π = 1%.
- *M. guttatus* diversity are fairly uniformly high.
- *M. nasutus* chromosomes, diversity is mostly low because the pair of lineages generally coalesce recently.
- Yet in a few places we see levels of diversity comparable to *M. guttatus,* due to deep coalescent

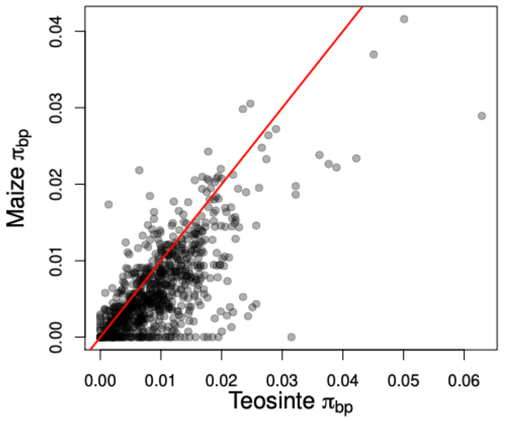


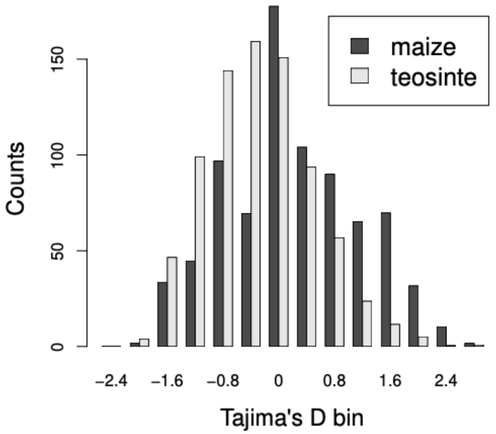


Example: Maize

Figure 4.27: Teosinite (*Zea mays* ssp. *mexicana*)
American grasses (1897). Scribner, FL Image from the Biodiversity Heritage Library. Contributed by Smithsonian Libraries. Not in copyright.

- Mutations that arise on deeper lineages will be at intermediate frequency in our sample, and so mild bottlenecks can lead to an excess of intermediate frequency alleles compared to the standard constant-size model.
- This can skew Tajima's D towards positive values and away from its expectation of zero.
- Maize (Zea mays subsp. mays) was domesticated from its wild progenitor teosinte (Zea mays subsp. parviglumis) roughly **ten thousand** years ago.
- Maize $N_e \approx 10^4$
- Teosinte $N_e \approx 10^5$
- Bottleneck associated with domestication has resulted in a loss of genetic diversity in maize compared to teosinte, and the polymorphism that remains is somewhat skewed towards intermediate frequencies resulting in more positive values of Tajima's D.





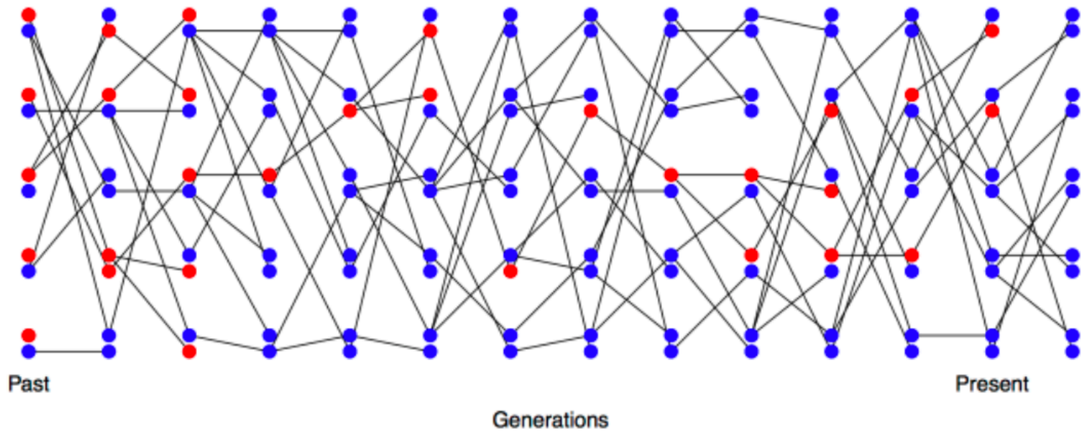
Chapter Review

Question 1.

You are in charge of maintaining a population of delta smelt in the Sacramento River delta. Using a large set of microsatellites you estimate that the mean level of heterozygosity in this population is 0.005. You set yourself a goal of maintaining a level of heterozygosity of at least 0.0049 for the next two hundred years. Assuming that the smelt have a generation time of 3 years, and that only genetic drift affects these loci, what is the smallest fully outbreeding population that you would need to maintain to meet this goal?

Loss of heterozygosity due to drift

- The probability that our two alleles have the same parental allele in the proceeding generation is $\frac{1}{2N}$
- The probability that they have different parental alleles is $1 - \frac{1}{2N}$
- Apply the law of total probability, the *expected* heterozygosity in generation $t + 1$ is:
  - $H_{t+1} = \frac{1}{2N} \times 0 + (1 - \frac{1}{2N}) \times H_t$
- We can write down a recursive formula
  - $H_t = (1 - \frac{1}{2N})^t H_0 \approx e^{-\frac{t}{2N}} H_0$



Question 4.

You are studying a population of 500 male and 500 female Hamadryas baboons. Assume that all of the females but only 1/10 of the males get to mate. What is the effective population size for the autosome?

Variance in reproductive success

- Variance in reproductive success will also affect our effective population size. Even if our population has a large constant size N individuals, if only small proportion of them get to reproduce, then the rate of drift will reflect this much smaller number of reproducing individuals.

- NF of females get to reproduce and NM males get reproduce.

- our probability of coalescence in the preceding generation is

- $$\frac{1}{4}\left(\frac{1}{2N_M}\right)+\frac{1}{4}\left(\frac{1}{2N_F}\right)$$

- $$N_e = \frac{4N_F N_M}{N_F + N_M}$$

Question 6.

One of the highest levels of genetic diversity is seen in the diploid split-gill fungus, *Schizophyllum commune*. Populations in the USA have a sequence-level heterozygosity of 0.13 per synonymous base (Baranova *et al.*, 2015). Baranova *et al.* sequenced parents and multiple offspring to estimate that $\mu = 2 \times 10^{-8} bp^{-1}$ per generation. What is your estimate of the effective population size of *S. commune*?

- Hint:

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

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

Question 7.

Robinson *et al.* (2016) found that the endangered Californian Channel Island fox on San Nicolas had very low levels of diversity ($\pi = 0.000014\text{bp}^{-1}$) compared to its close relative the California mainland gray fox (0.0012bp^{-1}).

A) Assuming a mutation rate of 2×10^{-8} per bp, what effective population sizes do you estimate for these two populations?

B) Why is the effective population size of the Channel Island fox so low? [Hint: quickly google Channel island foxes to read up on their history, also to see how ridiculously cute they are.]

Question 9.

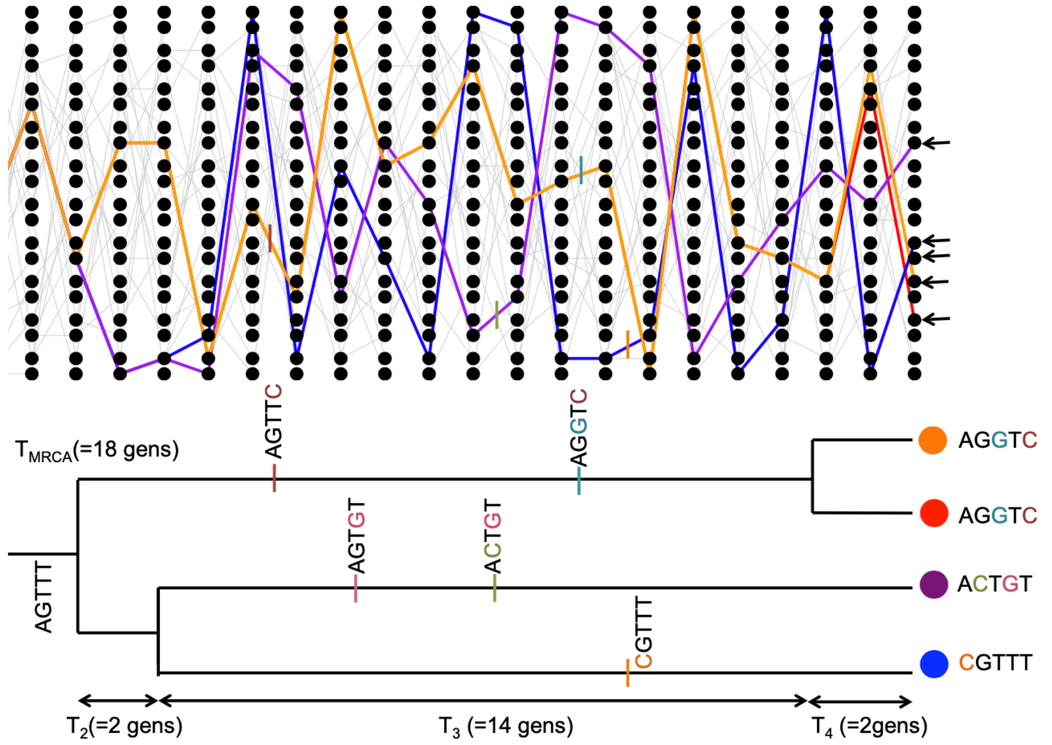
Assume an autosomal effective population of 10,000 individuals (roughly the long-term human estimate) and a generation time of 30 years. What is the expected time to the most recent common ancestor of a sample of 20 people? What is this time for a sample of 500 people?

Time to MRCA

- Time to most recent common ancestor (MRCA):
- $T_{\text{MRCA}} = \sum_{i=2}^{n} T_i$
- As our coalescent times for different i are independent, the expected time to the most recent common ancestor is

$$\mathbb{E}(T_{MRCA}) = \sum_{i=n}^{2} \mathbb{E}(T_i) = \sum_{i=n}^{2} 2N / \binom{i}{2}$$

$$\mathbb{E}(T_{MRCA}) = 4N\left(1 - \frac{1}{n}\right)$$



Question 10.

There are two possible tree shapes that could relate four samples. Draw both of them and separately colour (or otherwise mark) the branches by where singletons, doubletons, and tripleton derived alleles could arise.

Question 11.

Voight *et al.* (2005) sequenced 40 autosomal regions from 15 diploid samples of Hausa people from Yaounde, Cameroon. The average length of locus they sequenced for each region was 2365bp. They found that the average number of segregating sites per locus was $S = 11.1$ and the average $\pi = 0.0011$ per base over the loci. Is Tajima's D positive or negative? Is a demographic model with a bottleneck or growth more consistent with this result?

- $$\widehat{\theta}_W = \frac{S}{\sum_{i=n-1}^{1} {}^{1}/_{i}}$$
- $\widehat{\theta_\pi} = \pi$

Question 12.

Based on museum samples from $\sim$ 1800, you estimate that the average heterozygosity in Northern Elephant Seals was 0.0304 across many loci. Based on further samples, you estimate that in 1960 this had dropped to 0.011. Elephant Seals have a generation time of 8 years.

What effective population size do you estimate is consistent with this drop?

Question 13.

A) Why are large populations expected to harbor more neutral variation?

B) What is the effective population size? Is it usually higher or lower than the census population size?

C) Why does the effective population size differ across the autosomes, Y chromosome, and mtDNA?

Is it possible to have a N_e larger than the census population size?

Question 14.

You sequence a genomic region of a species of Baboon. Out of 100 thousand basepairs, on average, 200 differ between each pair of sequences. Assume a per base mutation rate of 1×10^{-8} and a generation time of ten years.

A) What is the effective population size of these Baboons?

B) What is the average coalescent time (in years) of a pair of sequences in this species?