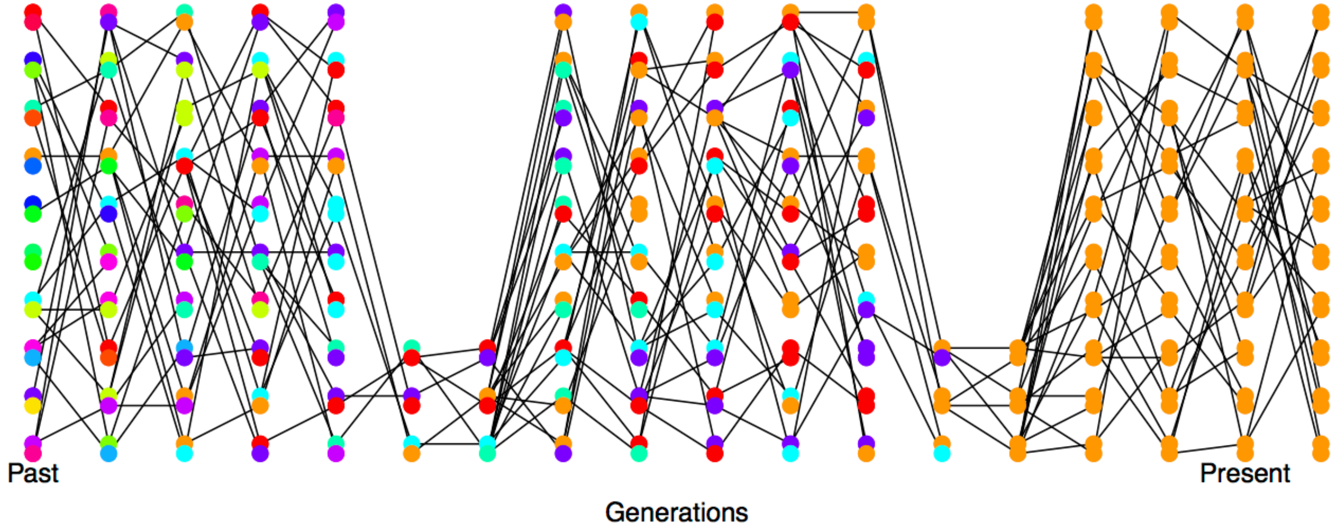
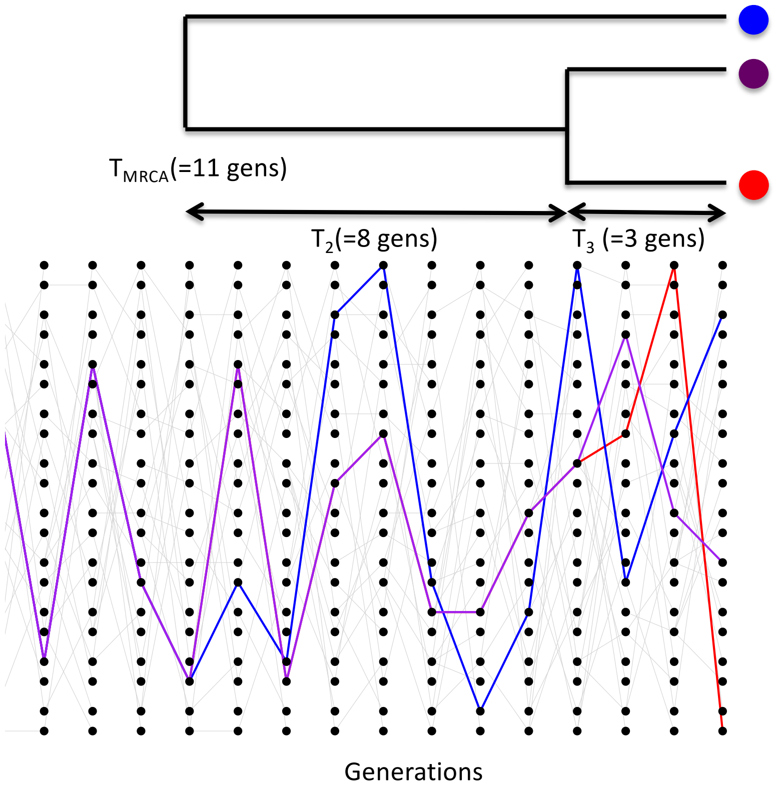
Coop, Chapter 4: Intro.-4.1

Genetic Drift and Neutral Diversity

Introduction and Loss of heterozygosity due to drift



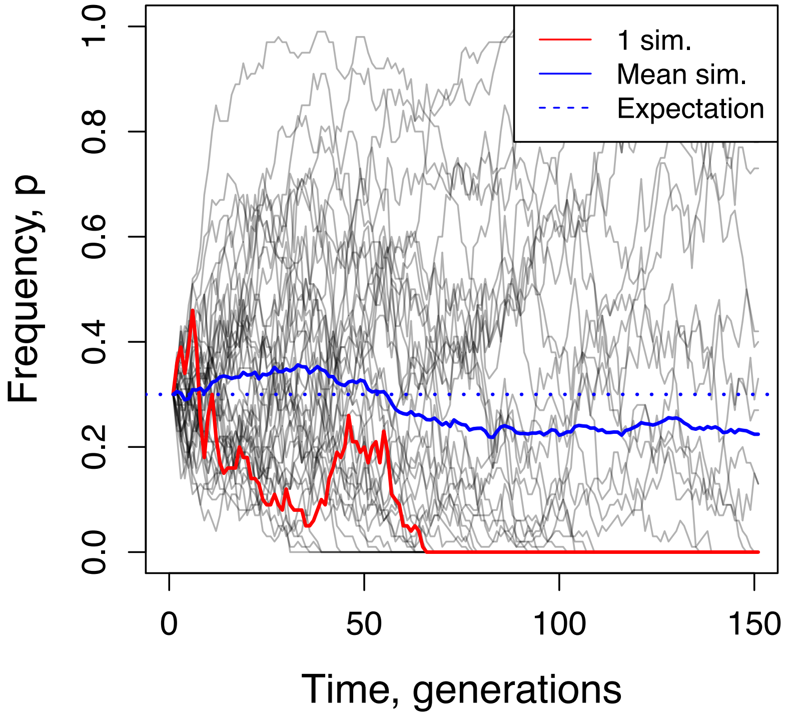


Introduction

- While evolutionary processes such as natural selection, mutation, and gene flow may seem more exciting or intuitively important, genetic drift alone can explain a lot of the variation we see across populations

- Genetic drift occurs because more or less copies of an allele can be transmitted across generations just due to chance

- While genetic drift can affect allele frequencies across the genome, it is particularly influential at neutral loci that do not discernably affect fitness



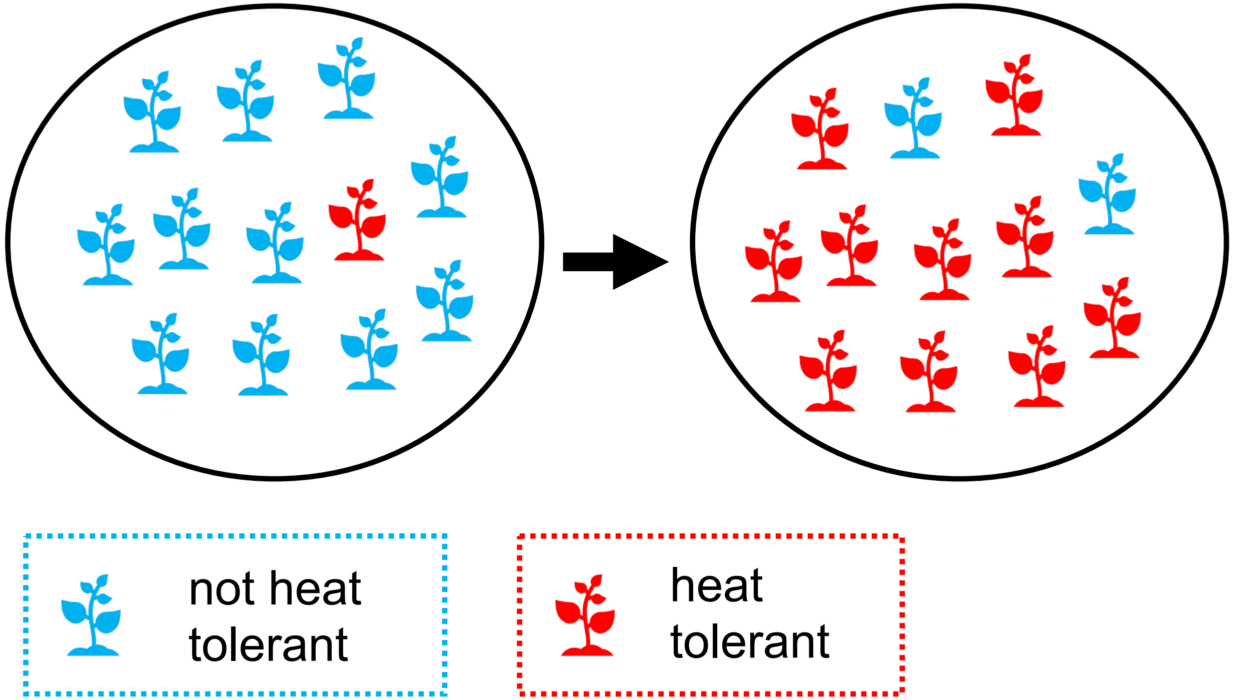
Introduction

The Neutral Theory of Molecular Evolution was proposed by Motoo Kimura in the 1960's:

- Patterns of polymorphism within species and substitution across species can be largely explained by neutral alleles subject to drift
- The vast majority of new mutations are neutral or highly deleterious (disrupt protein function)
- Deleterious alleles are removed by selection too quickly to meaningfully contribute to variation

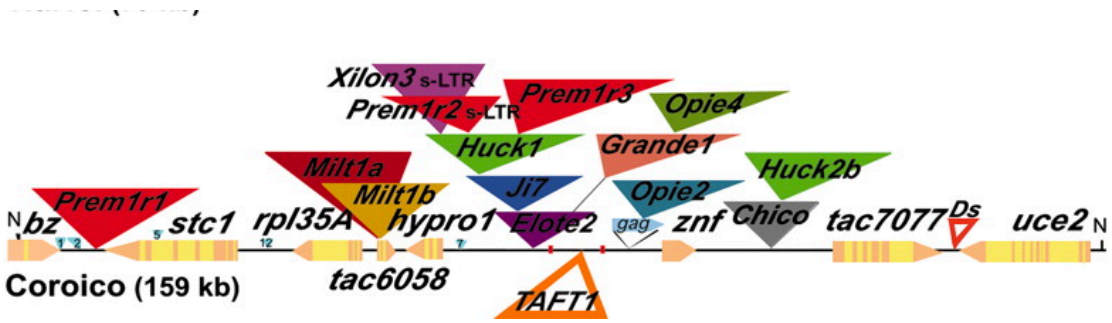
Motoo Kimura

Introduction



- But what about adaptation?
- Proponents of the Neutral Theory did not deny adaptation, but thought that beneficial alleles were rare and did not explain the bulk of variation in genomes
- Several clear examples of neutral variation can be found within genomes

Introduction



The Bronze Locus in maize with numerous transposable element insertions

Wang and Dooner 2006

```
>ATGGAGAACGATGAACTCAGCCCAGAAGCCAGCTAA
>ATGGAGAATGATGAACTCAGCCCAGAAGCCAGCTAA
>ATGGAAAATGATGAACTCAGCCCAGAAGCCAGCTAA
>ATGGAAAACGATGAACTCAGCACAGAAGCCAGCTAA
>ATGGAGAACGATGAACTCAGCACAGAAGCTAGCTAA
```

Synonymous: An allele that encodes the same amino acid in a protein

- Many genomes are primarily comprised of non-coding DNA (transposable elements, tandem repeats, old viruses, pseudogenes, etc…)
- Synonymous changes that don’t affect amino acids
- Nonsynonymous changes that don’t dramatically alter protein properties
- Nonsynonymous changes that do alter the phenotype, but the phenotype does not affect fitness

Introduction

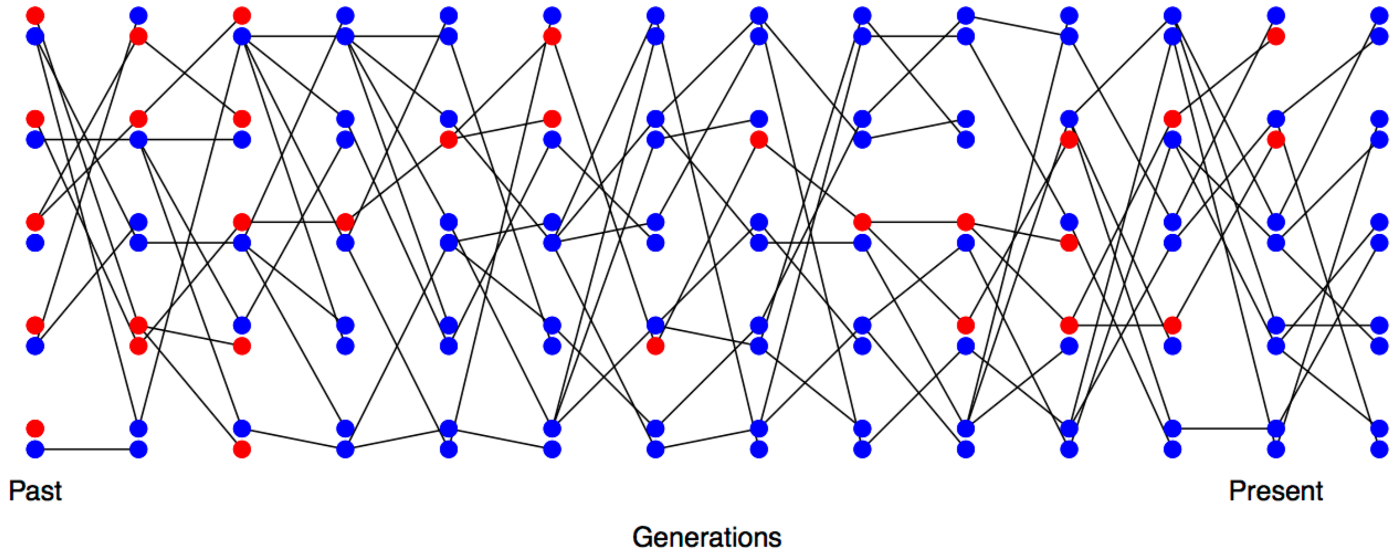
- The Neutral Theory has been supported by the high amount of polymorphism seen within and across species and the molecular clock which we’ll explore later

- However, for explaining other aspects of variation across populations, the Neutral Theory is clearly wrong

- The Neutral Theory can also serve as a useful null model which can be rejected when evidence for, for example, natural selection is overwhelming

4.1 Loss of heterozygosity due to drift

- Over time and without mutations, genetic drift will slowly remove variation from populations, with alleles moving to high or low frequency and being fixed or lost
- We can track, for example, the fate of the red and blues alleles across generations in this figure
- While in the first generation these five diploid individuals are all heterozygous, after 14 generations, the population is homozygous blue



4.1 Loss of heterozygosity due to drift

- Let’s consider the heterozygosity in a population at time t (H_t) and how this changes in the subsequent generation (H_{t+1})
- We have a diploid population with N individuals or $2N$ alleles
- The probability that our two alleles in generation t + 1 have the same parental allele is thus $1/(2N)$
- The probability that they have different parental alleles is 1- $1/(2N)$
- From equation 4.1, we can see that there is a slight loss in heterozygosity across generations:

$$H_{t+1} = \frac{1}{2N} \times 0 + \left(1 - \frac{1}{2N}\right) H_t \qquad (4.1)$$

4.1 Loss of heterozygosity due to drift

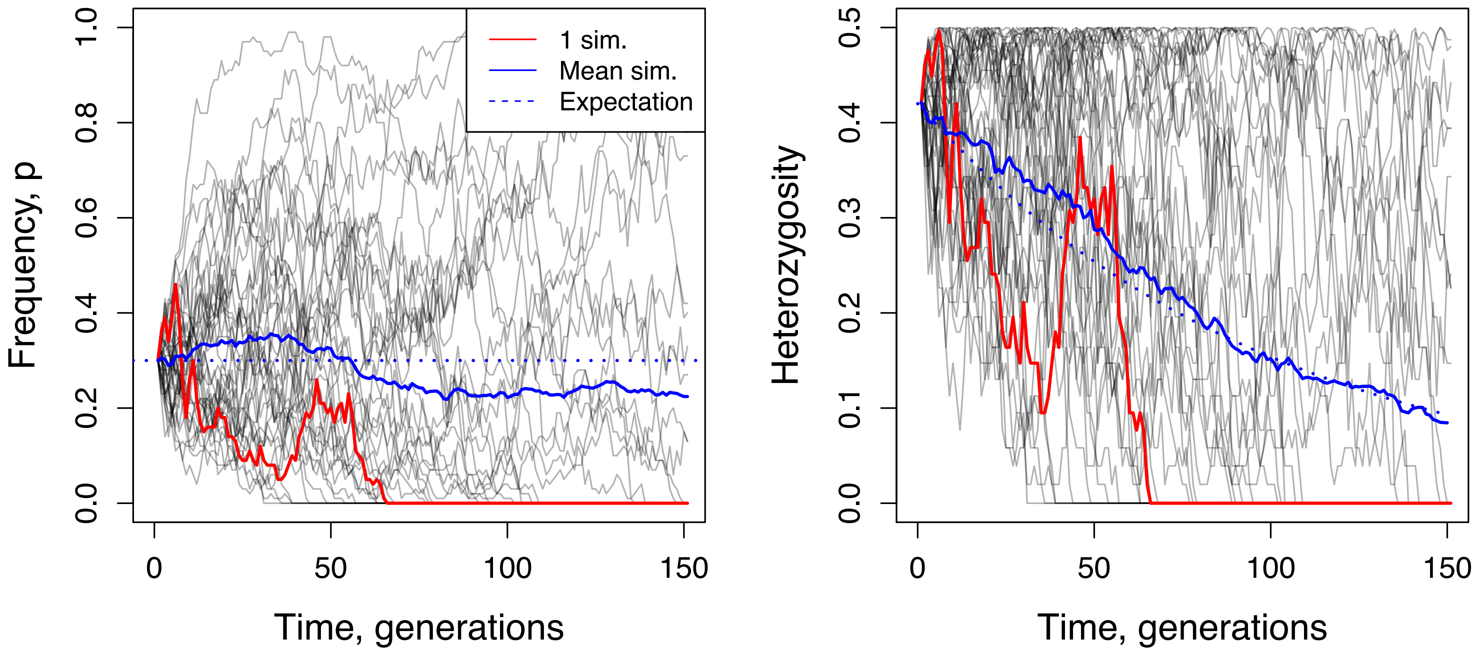
- Equation 4.1 can be simplified and generalized across any number of generations as:

$$H_t = \left(1 - \frac{1}{2N}\right)^t H_0 \qquad (4.2)$$

- If we assume that 1/(2*N*) is very small we can, as we did with LD decay in Chapter 3, approximate the geometric decay with an exponential:

$$H_t = H_0 e^{-t/(2N)} \qquad (4.3)$$

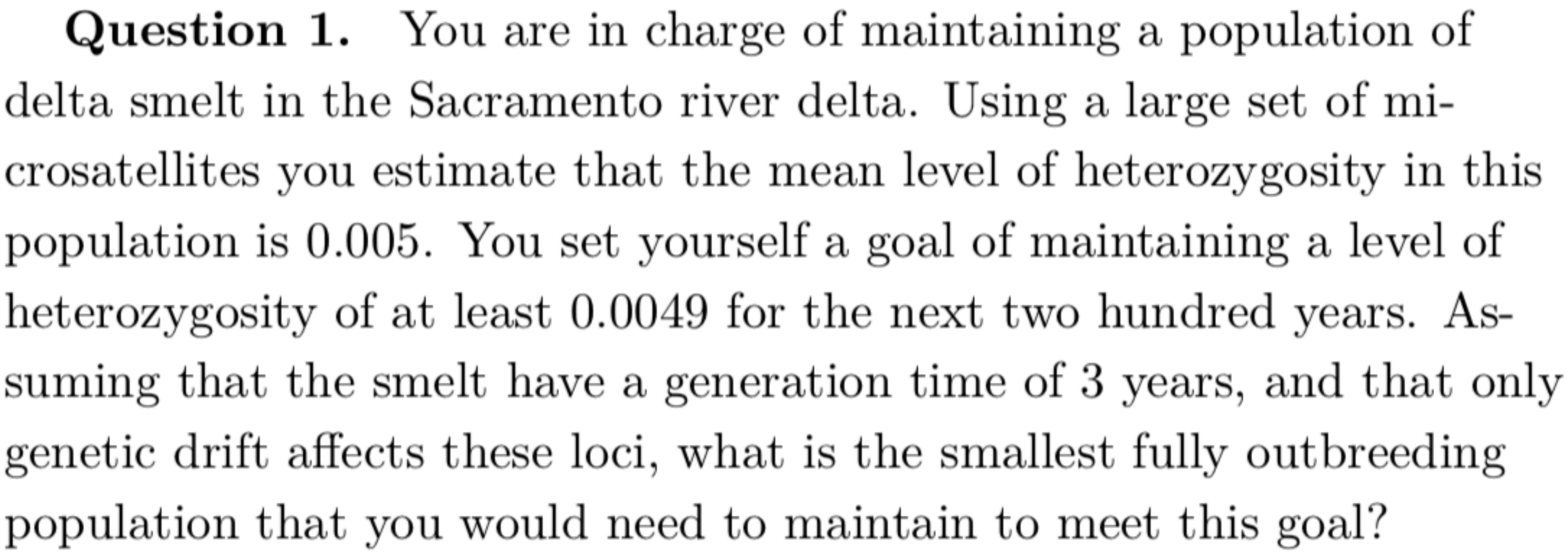
4.1 Loss of heterozygosity due to drift



- 40 independent alleles drifting in populations of 50 individuals with starting frequency of 0.3
- Some drift up, some drift down, but overall the frequency is ~0.3
- Heterozygosity, however, is slowly lost at a rate close to equations 4.2/4.3

4.1 Loss of heterozygosity due to drift

Let's try our hand at a problem:



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?

$$H_t = H_0 e^{-t/(2N)} \tag{4.3}$$

4.1 Loss of heterozygosity due to drift

Let’s try our hand at a problem:

H_t = 0.0049 H_0 = 0.005 t = generations = 200/3 = 66.67

$$H_t = H_0 e^{-t/(2N)} \tag{4.3}$$

$$0.0049 = 0.005(e^{-66.67/(2N)})$$

$$0.0049/0.005 = e^{-66.67/(2N)}$$

$$0.98 = e^{-66.67/(2N)}$$

$$\ln(0.98) = \ln(e^{-66.67/(2N)})$$

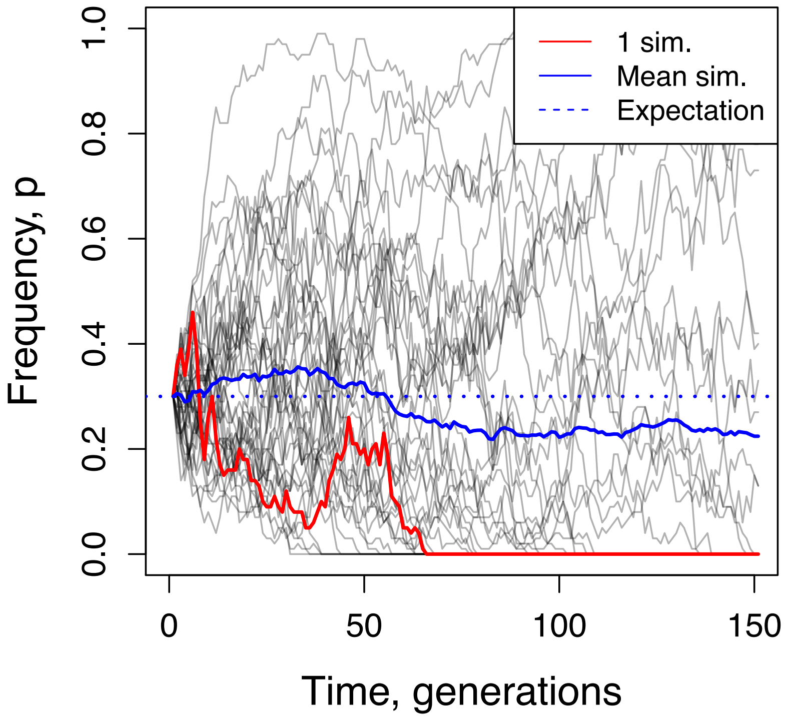
$$-0.02 = -66.67/(2N)$$

$$2N = -66.67/-0.02$$

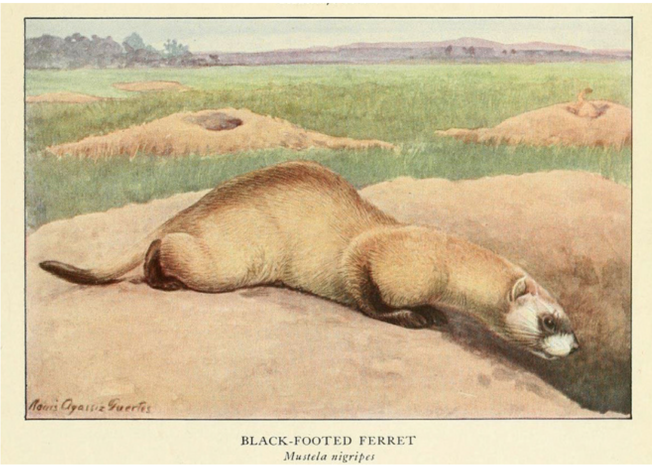
$$N = 1667$$

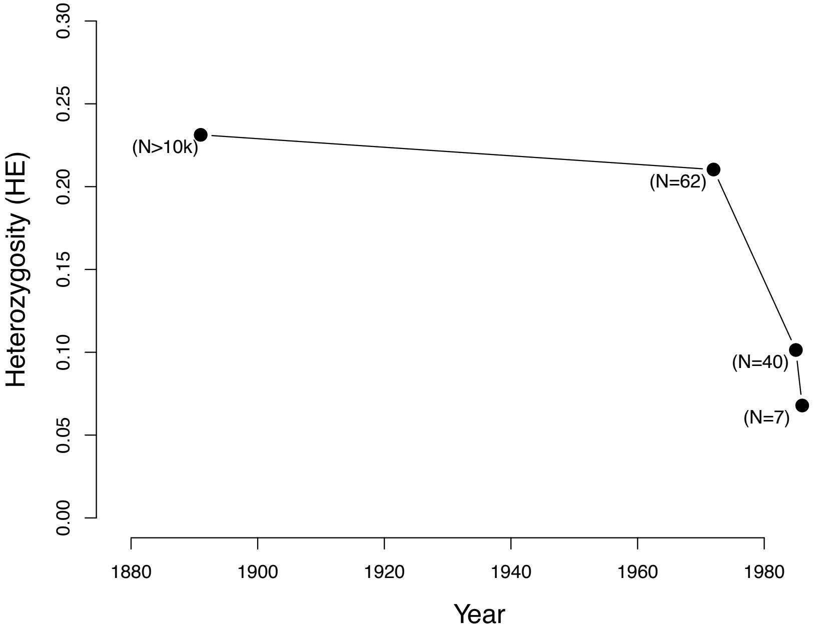
4.1 Loss of heterozygosity due to drift

- While we are clearly seeing a reduction in heterozygosity over time due to drift, our Hardy-Weinberg proportions do hold from one generation to the next
- Random samples from a finite population size explain the gradual loss of heterozygosity and change in allele frequency over time



4.1 Loss of heterozygosity due to drift





- Black-footed ferret is a good example of decline in heterozygosity due to small sample size

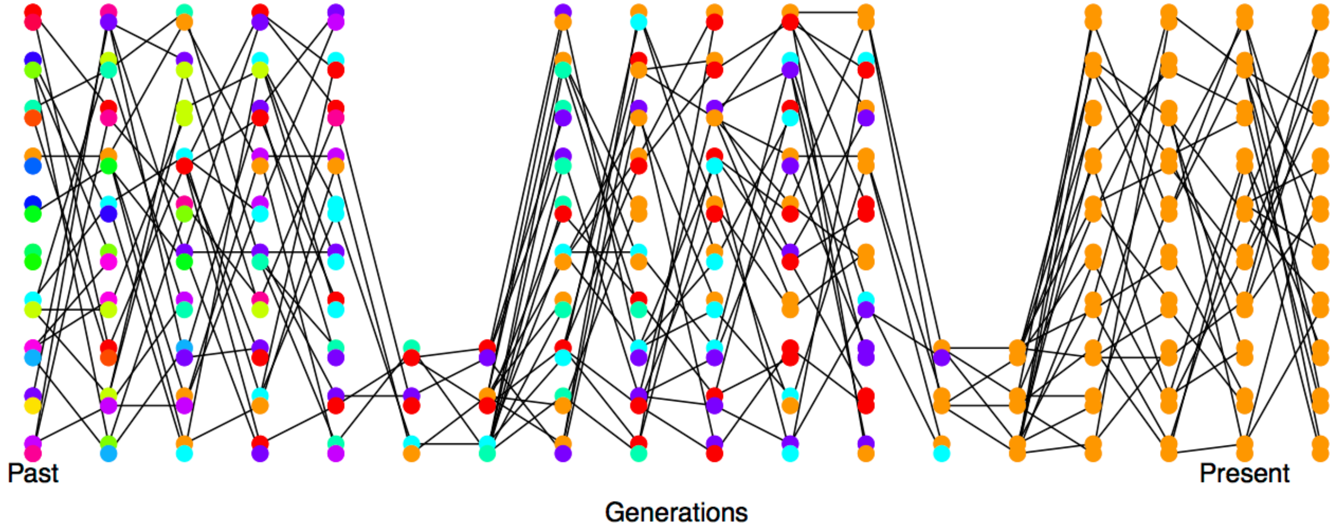
- Dramatic population decline to 7 individuals during the 20th century due to habitat destruction and disease

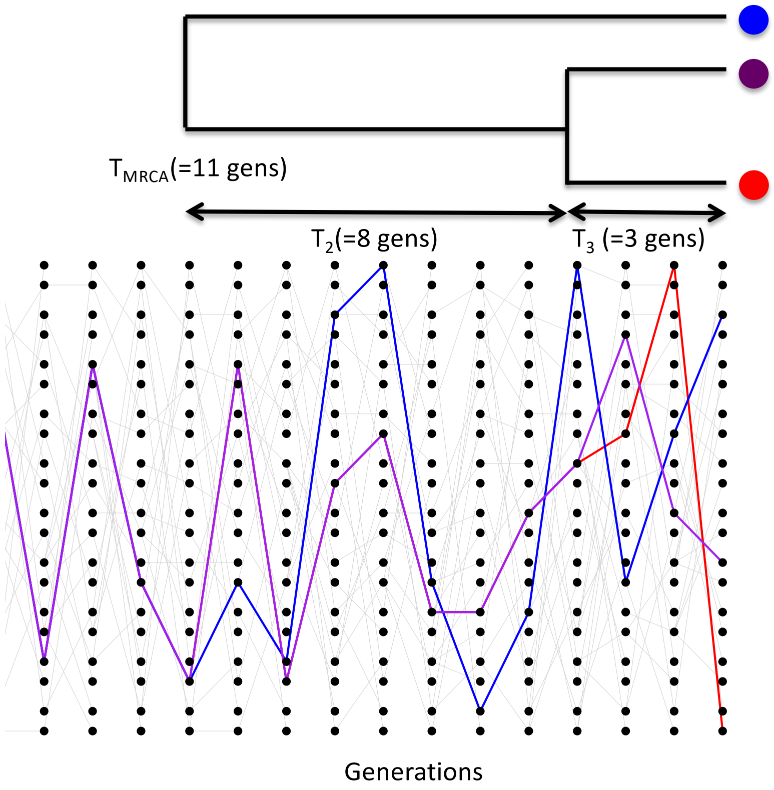
- Population has recovered, but heterozygosity remains low due to bottleneck to *N* = 7

Coop, Chapter 4: 4.1.1-4.1.2

Genetic Drift and Neutral Diversity

Levels of genetic diversity maintained by a balance between mutation and drift



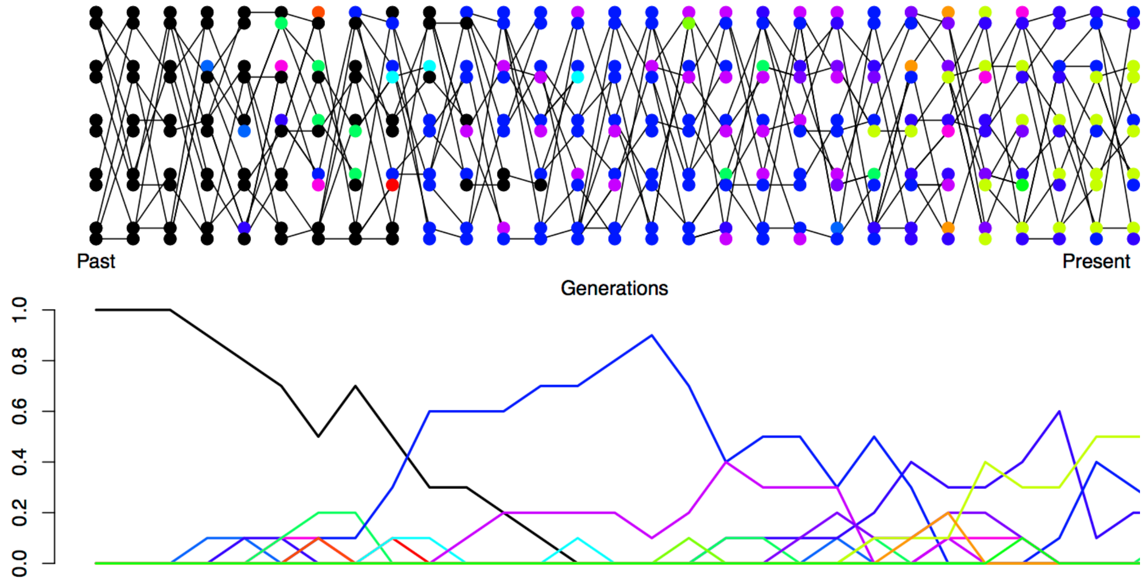


4.1.1 Mutation and Drift Balance

- While previously we've assumed drift has been the only evolutionary force affecting variation, let's now consider the balance between drift (removing variation) and mutation (adding variation)

- In the figure we have five diploid individuals and allow mutations (switch to different color dot) to occur between generations

- This is a high mutation rate sufficient to retain variation in such a small population



4.1.1 Mutation and Drift Balance

- To consider how mutation can balance genetic drift, we'll need to know the rate at which it introduces novel variation into a population

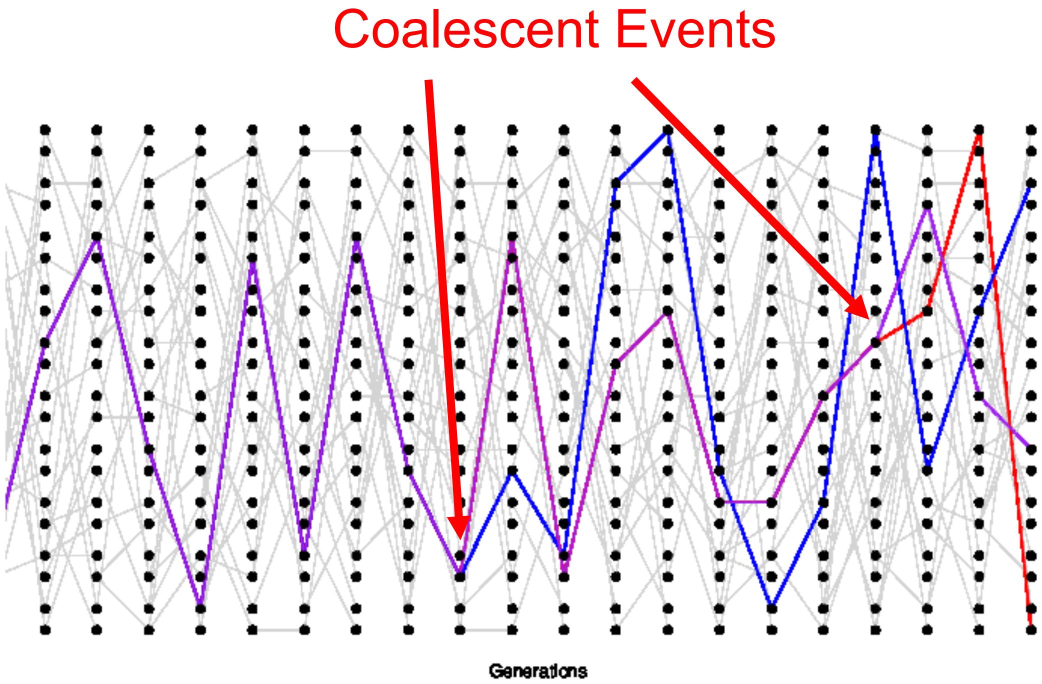
- The overall mutation rate per generation is referred to as μ, and we can divide this into the fraction of deleterious (C) mutations that are quickly removed by selection and neutral mutations ($1-C$)

- The neutral mutation rate is thus $(1-C)\mu$

Cross 2017, *Science* magazine News

4.1.1 Mutation and Drift Balance

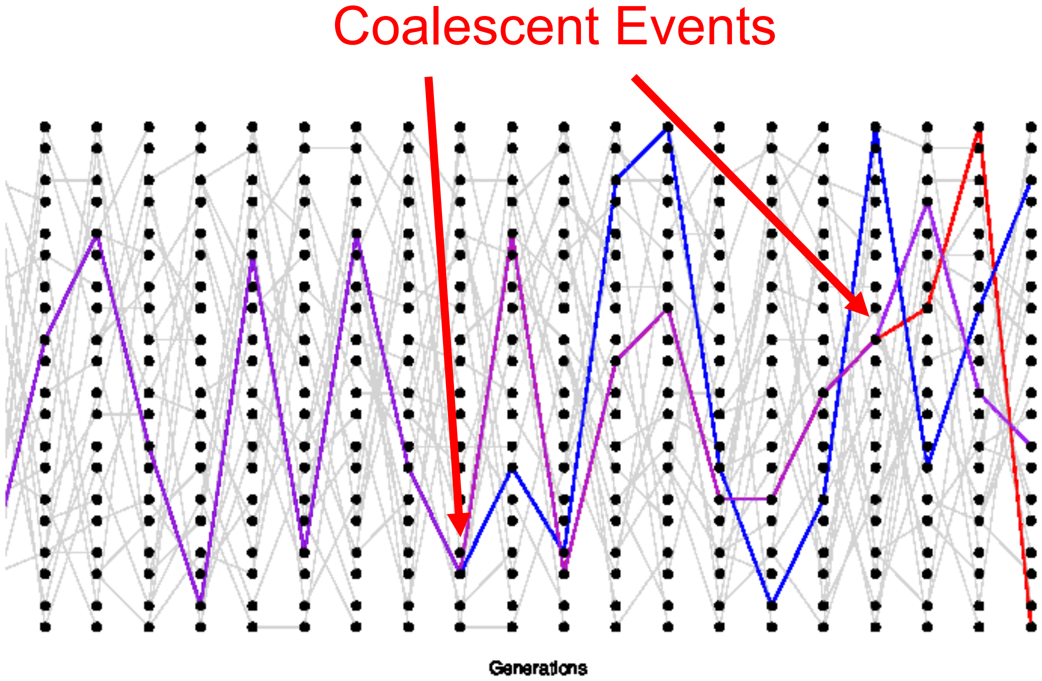
- To think about mutation-drift balance, let’s use a “backward-in-time” approach
- We can say that two alleles that have the same parental allele in a previous generation have “coalesced”
- The probability that alleles coalesce in a previous generation is 1/(2N) and the probability that they do not coalesce is 1-1/(2N)



4.1.1 Mutation and Drift Balance

- We'll also need to consider the probability that a mutation changes the state of the transmitted allele (μ) and the probability that no mutation occurs $(1 - \mu)$

- We'll assume that when a new mutation occurs, it creates a new allelic type that is not already present within the population (the infinitely-many-alleles model)



Coalescent Events

4.1.1 Mutation and Drift Balance

- We can now develop a model in which we determine both 1) when our two alleles last shared a common ancestor; and 2) whether the alleles are identical due to a lack of mutation

- For example, we can determine the probability that two randomly sampled alleles coalesced two generations ago and are identical:

4 meioses

did not coalesce

$$\left(1-\frac{1}{2N}\right)\frac{1}{2N}(1-\mu)^4 \qquad (4.4)$$

coalesced

did not mutate

4.1.1 Mutation and Drift Balance

- We can more generally summarize the probability that our alleles coalesced at generation $t + 1$ (thinking back in time) with no mutation as:

$$P(\text{coal. in t+1 \& no mutations}) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^{t}(1-\mu)^{2(t+1)} \quad (4.5)$$

and assuming that $t + 1 \approx t$

$$P(\text{coal. in t+1 \& no mutations}) \approx \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^{t}(1-\mu)^{2t} \quad (4.6)$$

4.1.1 Mutation and Drift Balance

- In practice, we will not know if our alleles coalesce in generation 2 or generation 20 or generation 20,000,000, so we can calculate the probability that they coalesce in any generation and have no mutations as:

$$\begin{aligned} P(\text{coal. in any generation \& no mutations}) \approx & P(\text{coal. in } t=1 \text{ \& no mutations}) + \\ & P(\text{coal. in } t=2 \text{ \& no mutations}) + \ldots \\ = & \sum_{t=1}^{\infty} P(\text{coal. in } t \text{ generations \& no mutation}) \end{aligned} \tag{4.7}$$

4.1.1 Mutation and Drift Balance

- By making some assumptions, that $\frac{1}{2N} \ll 1$ and $\mu \ll 1$, and by once again approximating geometric decay as exponential decay (see Coop textbook for details), and then approximating the summation with an integral, we end up with:

$$\frac{1}{2N}\int_0^{\infty} e^{-t(2\mu+1/(2N))}dt = \frac{1/(2N)}{1/(2N)+2\mu} = \frac{1}{1+4N\mu} \qquad (4.11)$$

- This general equation give us the probability that our two alleles coalesce before mutating, in other words, that they are homozygous

4.1.1 Mutation and Drift Balance

- The complementary probability, that our alleles are non-identical (heterozygous) is just 1 – our probability of being homozygous:

$$H = \frac{4N\mu}{1+4N\mu} \tag{4.12}$$

- The parameter $4N\mu$ is known as the population-scaled mutation rate and will come up several times in this book, so we will give it its own special name:

$$\theta = 4N\mu \tag{4.13}$$

4.1.1 Mutation and Drift Balance

- A take-home message from this equation:

$$H = \frac{4N\mu}{1 + 4N\mu} \tag{4.12}$$

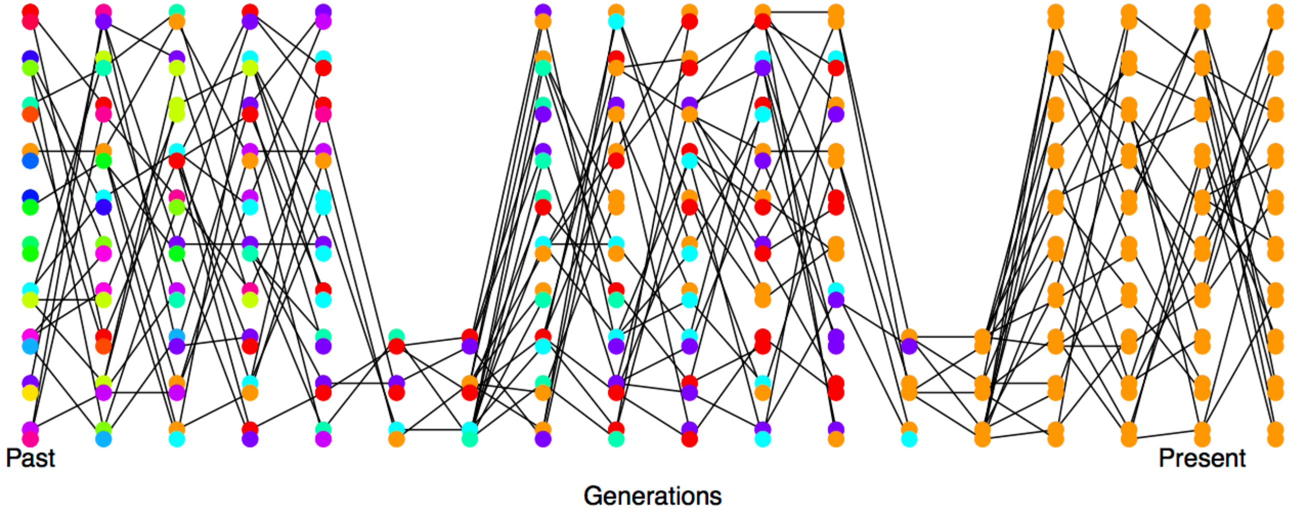
Generally, the larger the population size (N), the greater the extent of neutral polymorphism

4.1.2 The Effective Population Size

- Populations are rarely consistent in size over time and rarely have equal contributions to reproduction

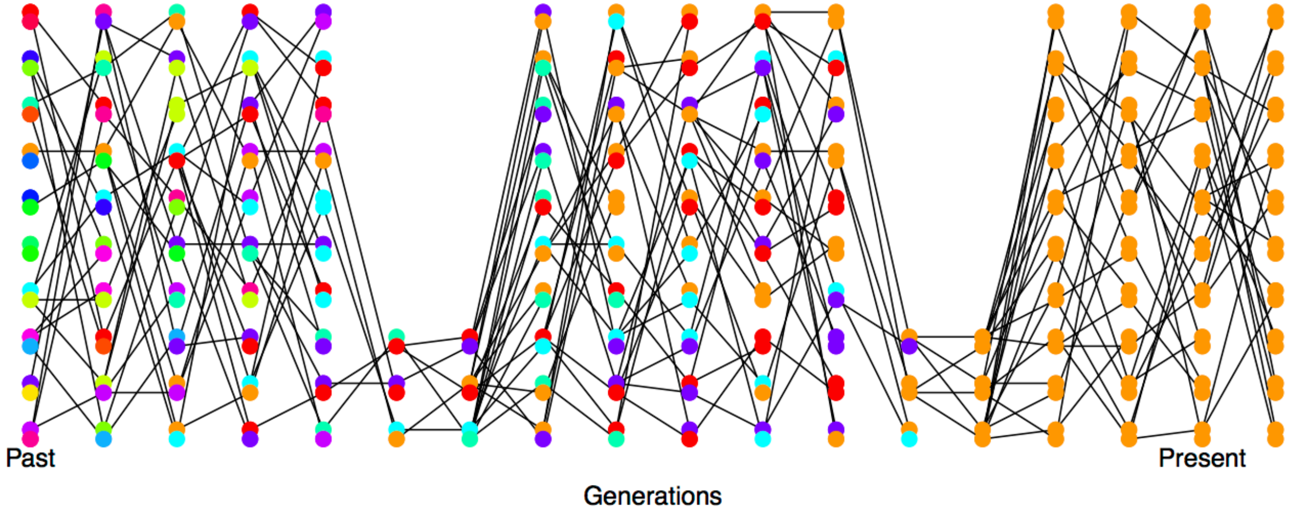
- This means that the effects of genetic drift may be more profound than would be clearly evident based on the current population size

- Consider this figure with two dramatic population bottlenecks as an example: the current population census size is high, but diversity is quite low



4.1.2 The Effective Population Size

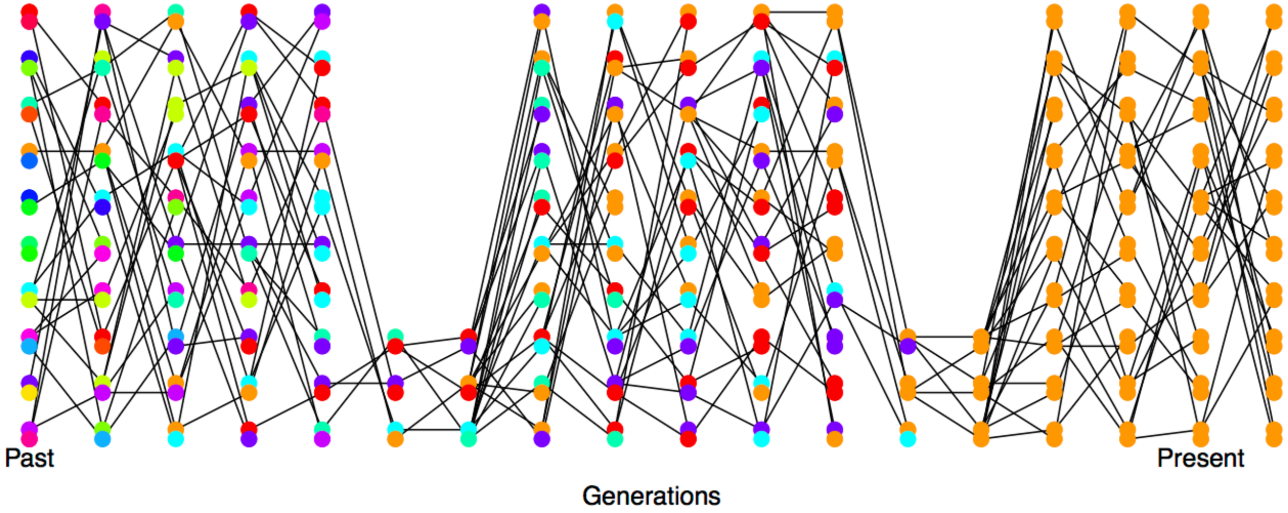
- To deal with this discrepancy, population geneticists often invoke the concept of "effective population size" or N_e
- N_e is the idealized constant population size that matches the extent of drift in the population
- When population sizes vary rapidly, the harmonic mean of population size over time may be a better approximation than census size



4.1.2 The Effective Population Size

- The harmonic mean is very affected by small values

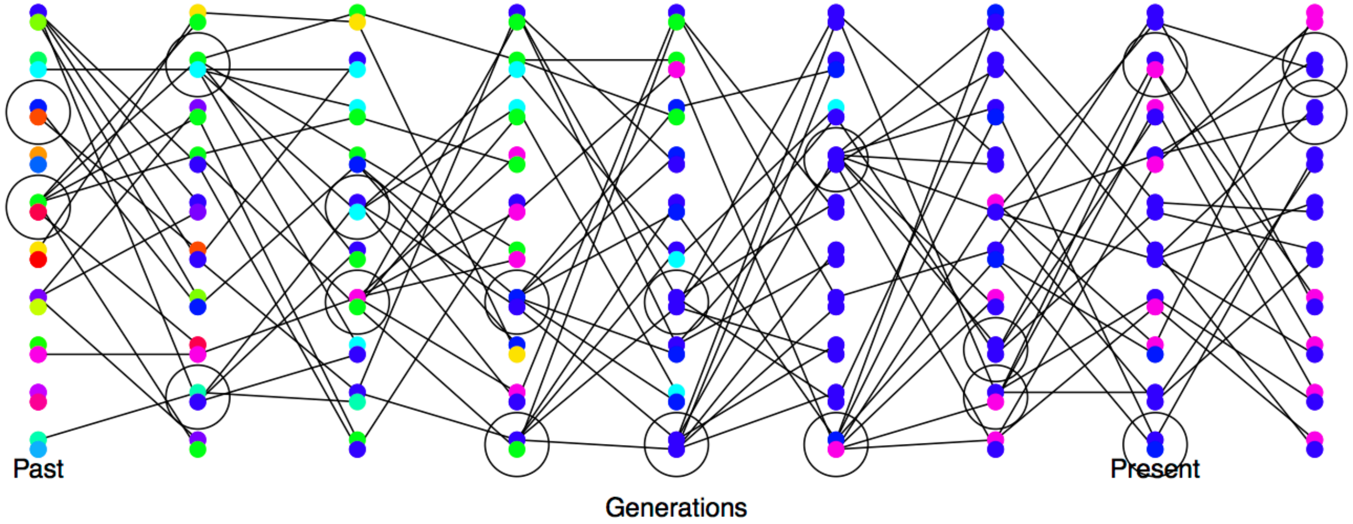
- If the census size of a population was 1,000,000 for 99% of its history, but shrank to 1,000 for 1% of its history, N_e would be much closer to 1,000 than 1,000,000



4.1.2 The Effective Population Size

- Even if the population size does not vary substantially over time, variation in reproductive success can cause discrepancies between the census size and N_e

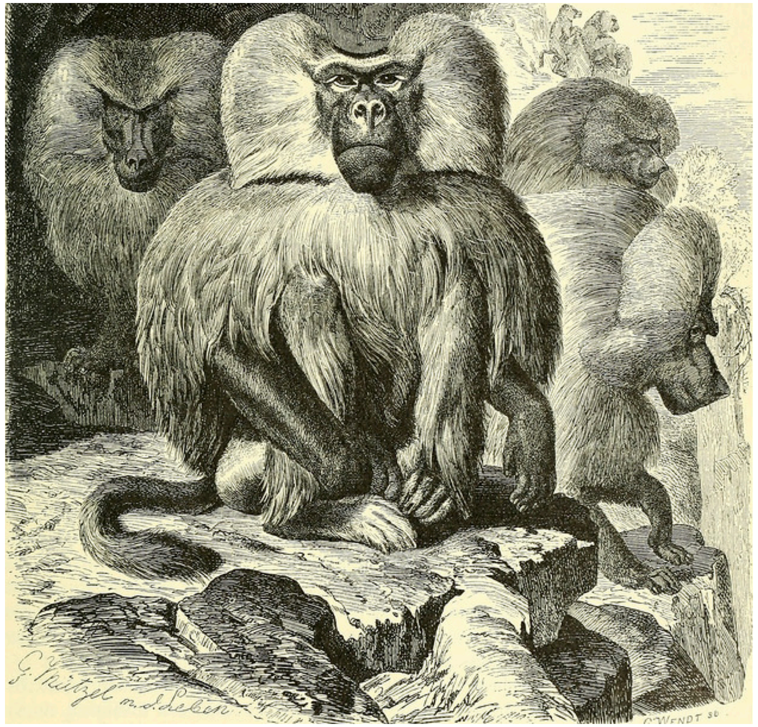
- The rate of drift will reflect the small number of individuals that are able to reproduce



4.1.2 The Effective Population Size

- For example, in many species, like the Hamadryas baboon, $N_{\mathrm{M}} < NF$, and few males have the opportunity to mate

- When reproductive success is very skewed in one sex, the effective population size is much less than the census size

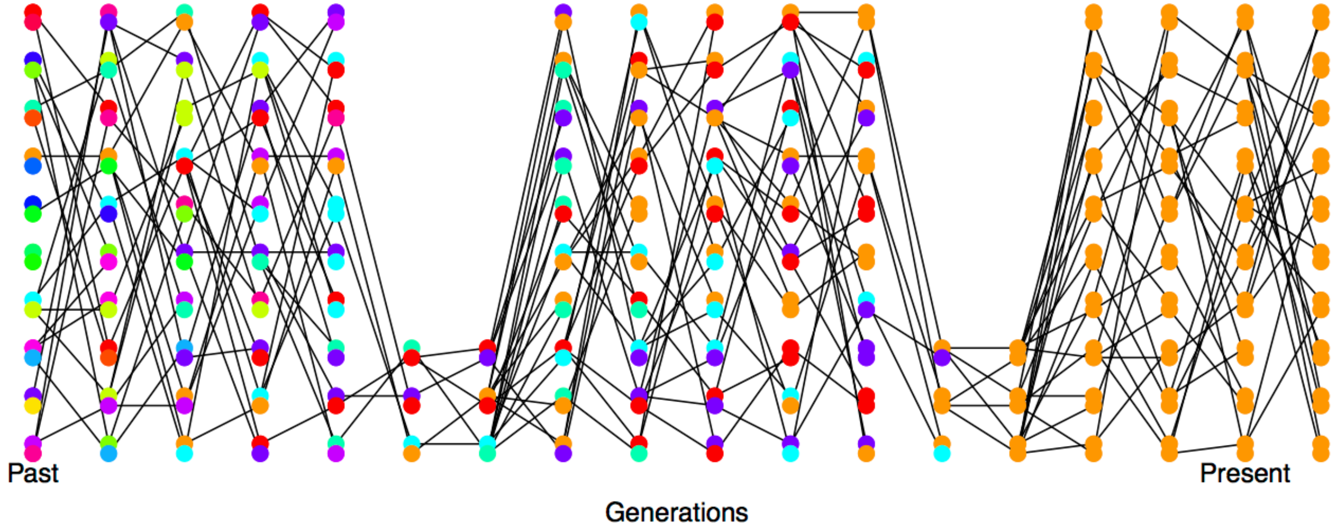


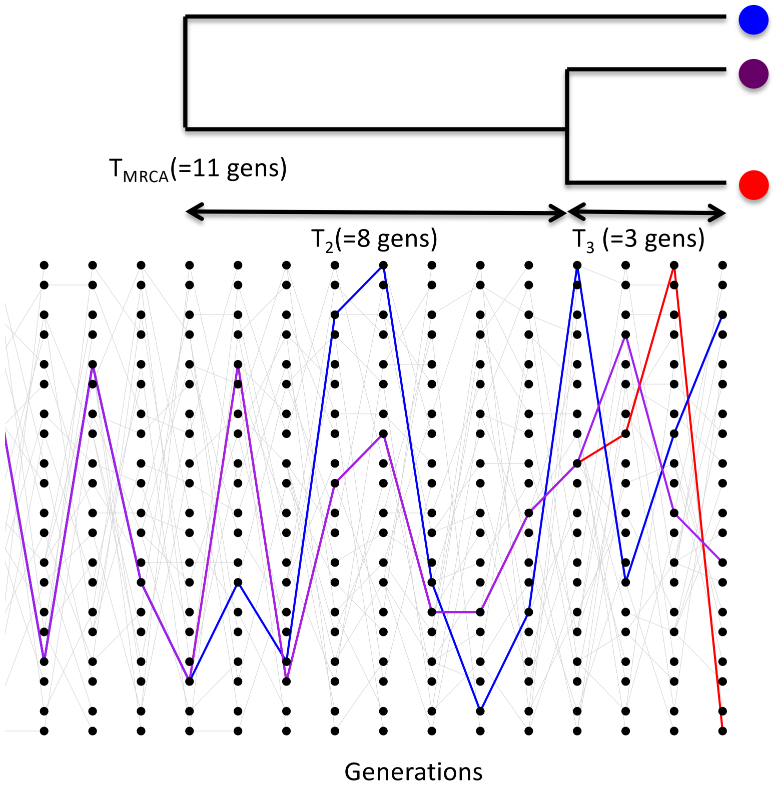
Male Hamadryas Baboon

Coop, Chapter 4: 4.2-4.3

Genetic Drift and Neutral Diversity

The Coalescent and patterns of neutral diversity &
The Coalescent process of a sample of alleles



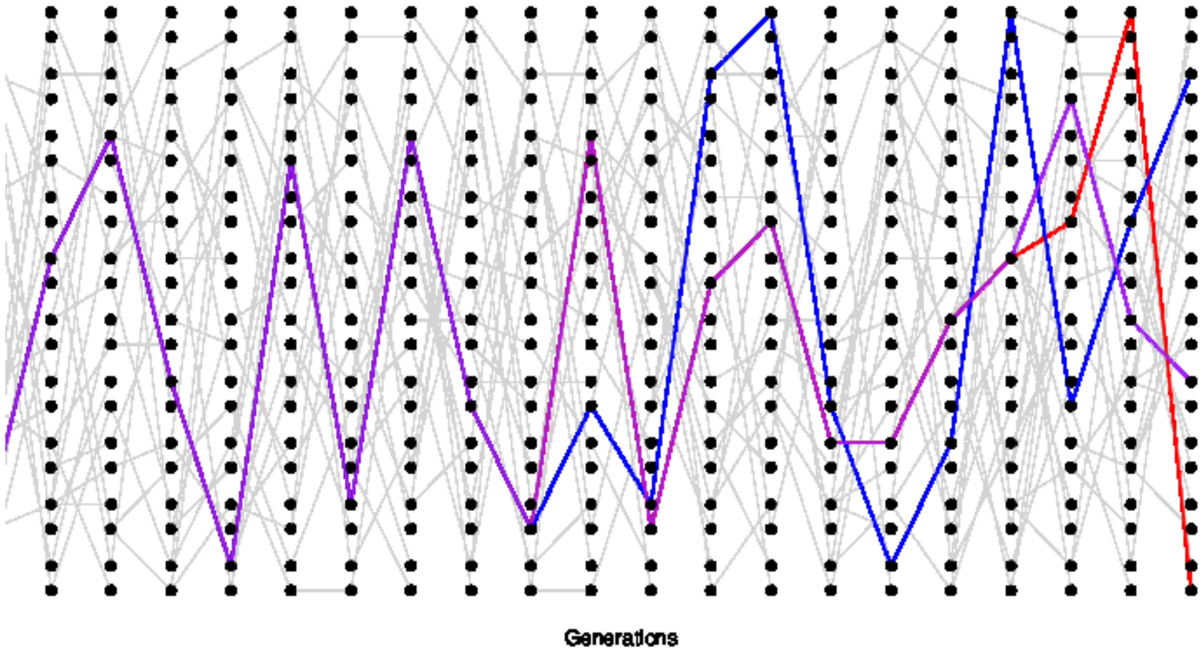


4.2 The Coalescent and patterns of neutral diversity

- As discussed in previous sections, it's helpful to first think about the time to the most recent common ancestor (coalescence) and then think about the impact of that time on diversity

- We can summarize the coalescence process as the probability that a pair of alleles has failed to coalesce in *t* generations and then coalesce in *t* + 1 generations:

$$P(T_2 = t + 1) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^t \qquad (4.20)$$



4.2 The Coalescent and patterns of neutral diversity

$$P(T_2 = t+1) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^t \qquad (4.20)$$

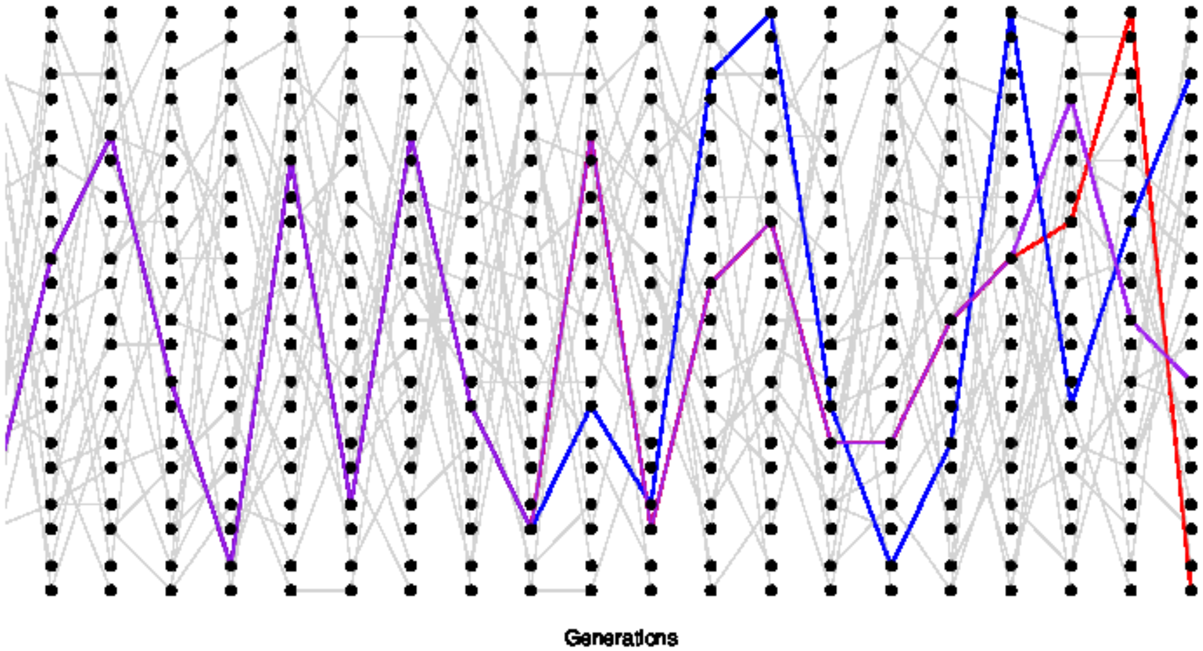
- For example, the probability that alleles coalesce 3 generations back is the probability that they fail to coalesce in the last two generations but then do in the third generation back:

$$\left(1 - \frac{1}{2N}\right) \times \left(1 - \frac{1}{2N}\right) \times \left(\frac{1}{2N}\right)$$

1st generation back

2nd generation back

3rd generation back

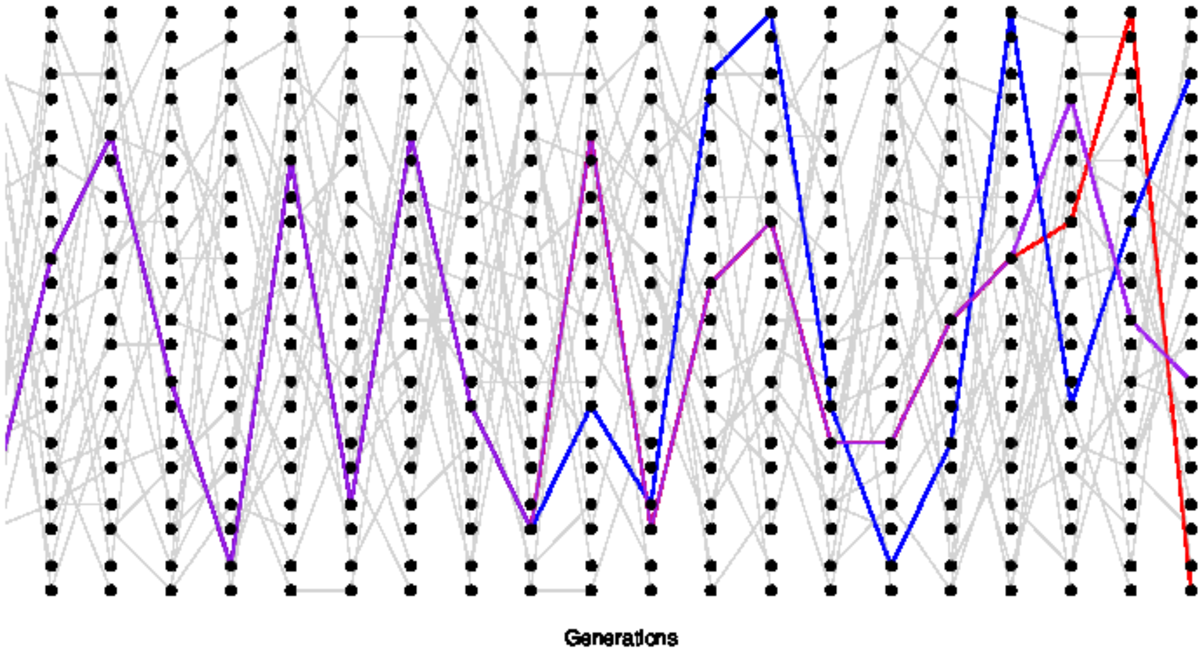


4.2 The Coalescent and patterns of neutral diversity

$$P(T_2 = t + 1) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^t \qquad (4.20)$$

- The form of equation 4.20 tells us that the coalescent time of our sequences is a geometrically distributed random variable with a probability of success of:
$$p = \frac{1}{2N}$$

- We can think of the waiting time for two alleles to coalesce to be similar to waiting for a heads to come up in a toin coss, but the probability is $p = \frac{1}{2N}$ rather than 0.5

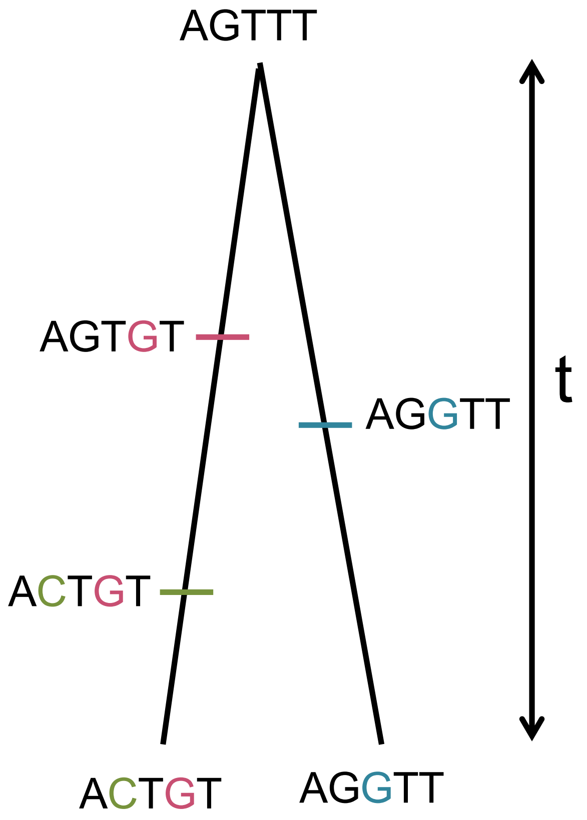


4.2 The Coalescent and patterns of neutral diversity

- The expected coalescent time can then be calculated as the mean of a geometric random variable which is $\frac{1}{p}$:

$$\mathbb{E}(T_2) = 2N \qquad (4.21)$$

- And once we know coalescent time, we can consider mutations in this context
- If alleles coalesce *t* generations in the past, there are *2t* generations in which a mutation could occur
- And if mutation rate is μ, then the number of expected mutations is $2t\mu$
- Putting this together with our expected coalescent time, we can expect $4N\mu$ mutations to occur (with assumption of infinitely many alleles/sites)



4.2 The Coalescent and patterns of neutral diversity

- Thinking back to our summaries of nucleotide diversity in Chapter 2, remember that we calculated π as the average pairwise differences between sequences
- Given our expectation for mutations prior to coalescence, we can say:

$$\mathbb{E}(\pi) = 4N\mu = \theta \tag{4.23}$$

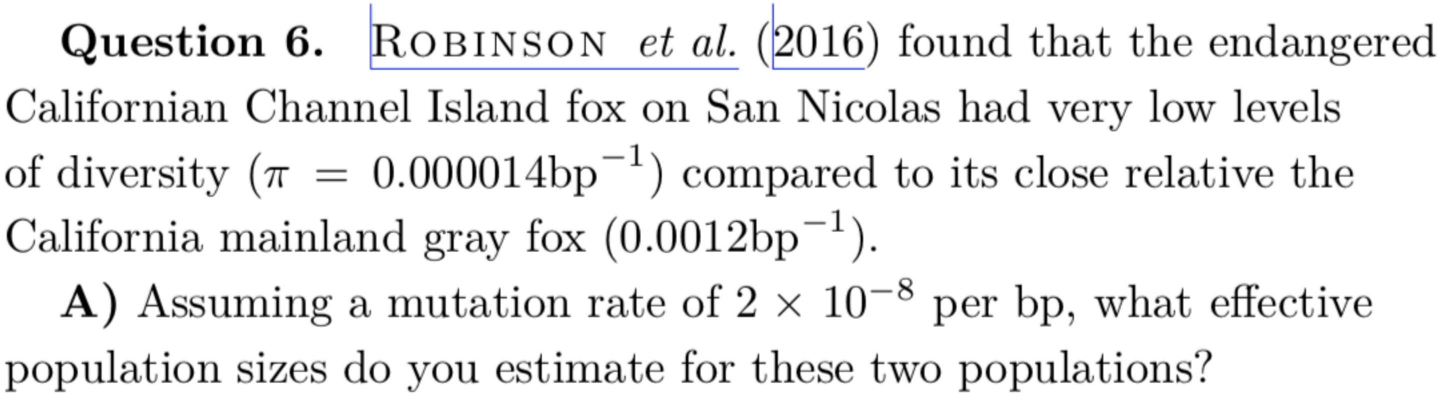
- This means that we can get an empirical estimate (based on sequence data we collect from some species) of θ from π which we will call $\hat{\theta}_\pi$
- Therefore, if we have an independent estimate of the mutation rate, μ, then we can use $\hat{\theta}_\pi = 4N\mu$ to get an estimate of the population size (N) which is the effective coalescent population size (N_e)
- Since this value averages over demographic history (bottlenecks, expansions) it may not be an accurate representation of population size at any given time

4.2 The Coalescent and patterns of neutral diversity

- Looking back, let's distinguish our coalescent expected heterozygosity, $H = \frac{4N\mu}{1+4N\mu}$, from our coalescent-based estimate of pairwise nucleotide diversity, $\hat{\theta}_\pi = 4N\mu$
- Our heterozygosity is the probability that two alleles drawn at random are different from each other, but our nucleotide diversity is the average number of differences between sequences
- Nucleotide diversity is therefore more useful because it is a measure of the number of differences in a sequence, not just whether differences exist
- When $\hat{\theta}_\pi$ is small (a short sequence or low diversity), it is similar to our coalescent expected heterozygosity

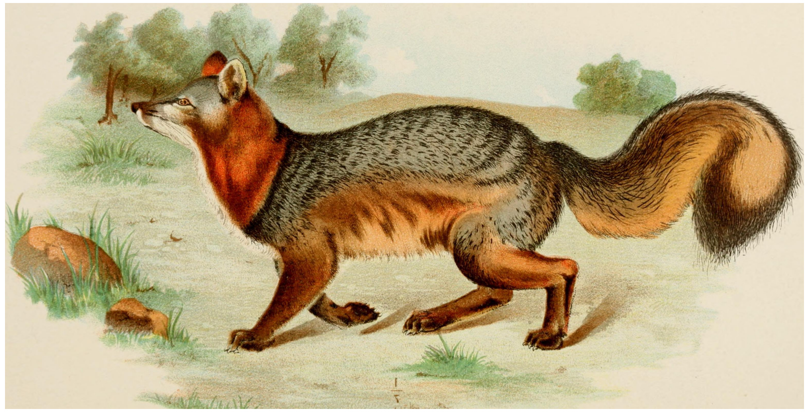
4.2 The Coalescent and patterns of neutral diversity

- Let’s try our hand at a problem:



Question 6. 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?



Channel Island Fox

$$\hat{\theta}_{\pi} = 4N\mu$$

$$0.000014 = 4N(2\times10^{-8})$$

$$700 = 4N$$

$$175 = N$$

Mainland Gray Fox

$$\hat{\theta}_{\pi} = 4N\mu$$

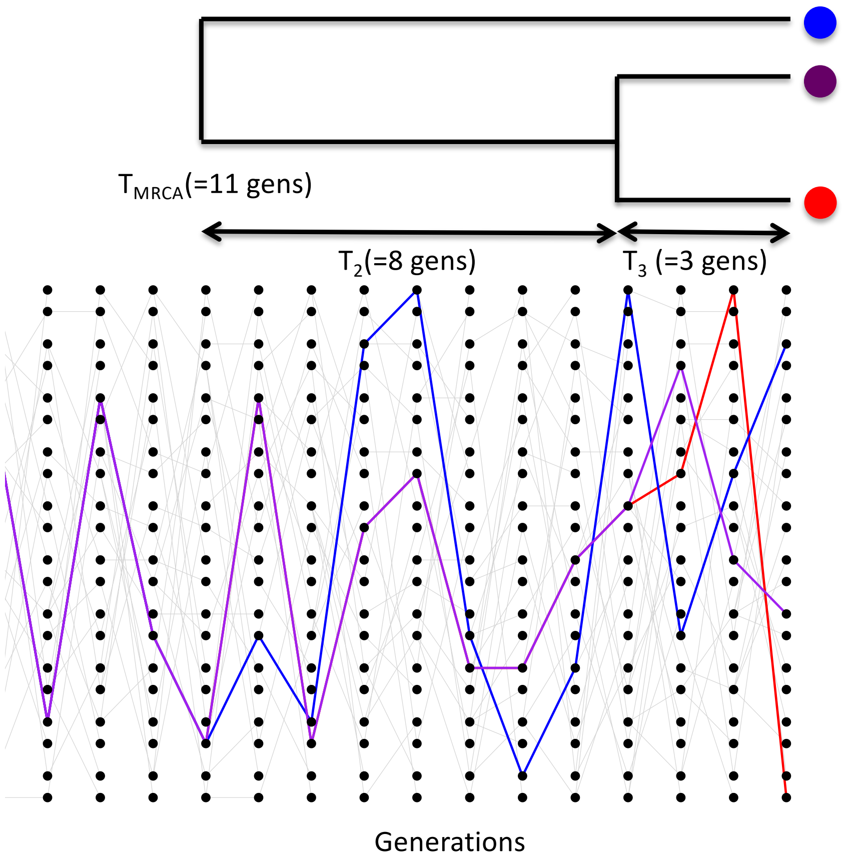
$$0.0012 = 4N(2\times10^{-8})$$

$$60{,}000 = 4N$$

$$15{,}000 = N$$

4.3 The coalescent process of a sample of alleles

- Up until now we've been discussing the simplified cases of pairs of alleles and average pairwise diversity, but we're often interested in diversity properties of many alleles drawn from a population
- This means we'll need to track the coalescence of many alleles back in time
- For example, in the figure, we're tracking coalescence of 3 alleles, with the first coalescence occurring 3 generations in the past and the second 11 generations in the past
- Time to the most recent common ancestor (T_{MRCA}) is $T_3 + T_2 = 11$ generations and the total time in the tree is 25 generations ($T_{tot} = 3T_3 + 2T_2$)



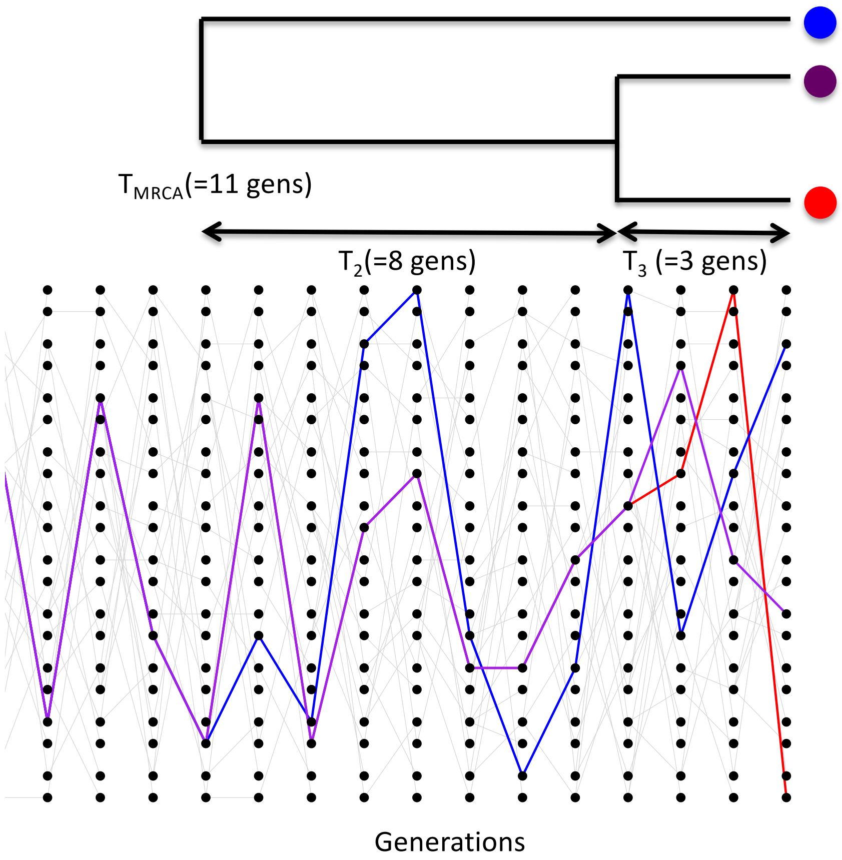
4.3 The coalescent process of a sample of alleles

- When we are considering the coalescence of more than 2 alleles, we'll track the history coalescence by coalescence

- With 3 alleles, we can modify our previous expectation of no coalescence in the previous generation to be:

$$\left(1-\frac{1}{2N}\right)^3 \approx \left(1-\frac{3}{2N}\right) \qquad (4.27)$$

Using what's known as a Taylor approximation when multiplying this out, ignoring values of $1/N^2$ that are very small

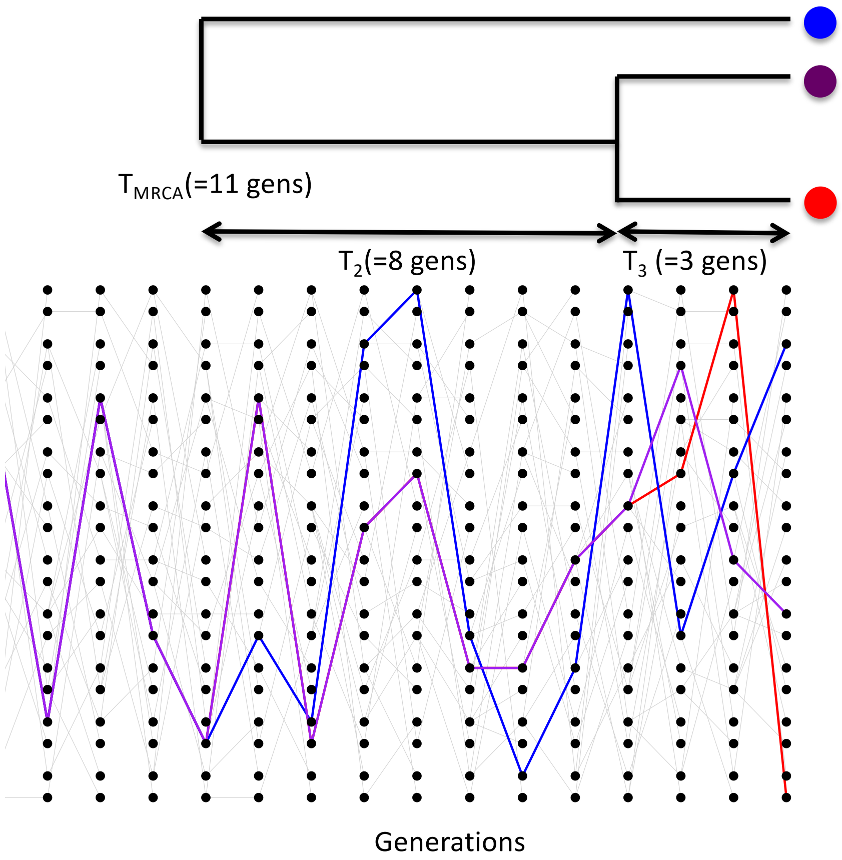


4.3 The coalescent process of a sample of alleles

- We can generalize this to any number of alleles by saying we sample i alleles in "i choose 2" or $\binom{i}{2}$ pairs
- Therefore the probability that no alleles coalesce in a sample of i alleles in the preceding generation is:

$$\left(1-\frac{1}{(2N)}\right)^{\binom{i}{2}} \approx \left(1-\frac{\binom{i}{2}}{2N}\right) \qquad (4.28)$$

- And the probability that they do coalesce is $\binom{i}{2}/2\mathrm{N}$



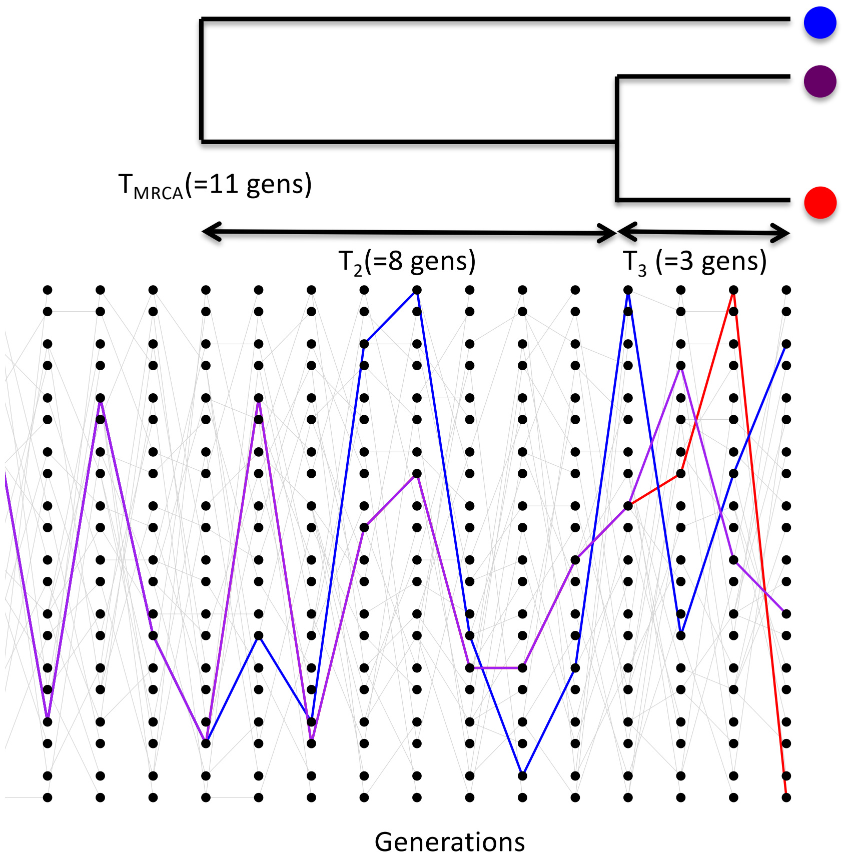
4.3 The coalescent process of a sample of alleles

- Using this notation, the time to the first coalescence in a sample of alleles is:

$$P(T_i = t+1) = \frac{\binom{i}{2}}{2N}\left(1 - \frac{\binom{i}{2}}{2N}\right)^t \qquad (4.29)$$

- After a pair of alleles coalesces or finds a common ancestor we merge these into this single ancestral allele and only consider it moving backwards, so our number of alleles becomes $i - 1$

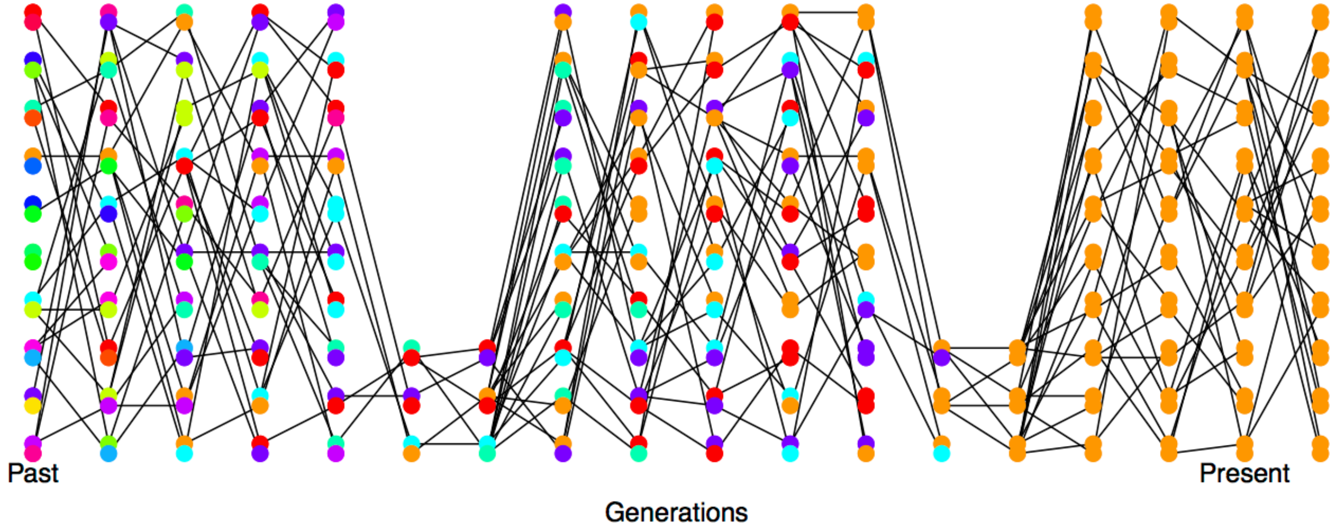
- This process continues until we coalesce back to a sample of 2 and then finally to the Most Recent Common Ancestor (MRCA)

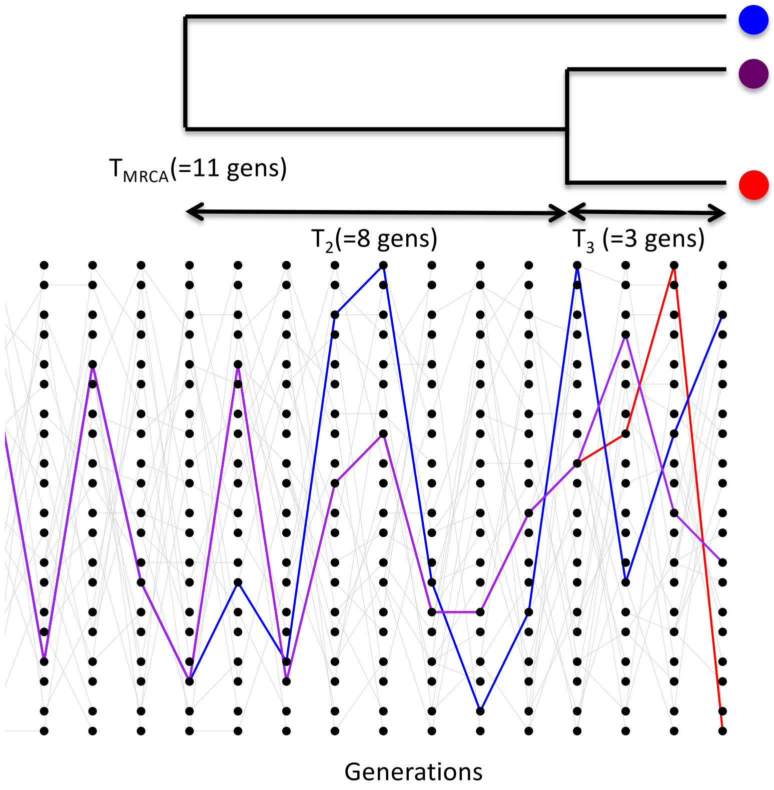


Coop, Chapter 4: 4.3.1

Genetic Drift and Neutral Diversity

Expected properties of coalescent genealogies and mutations





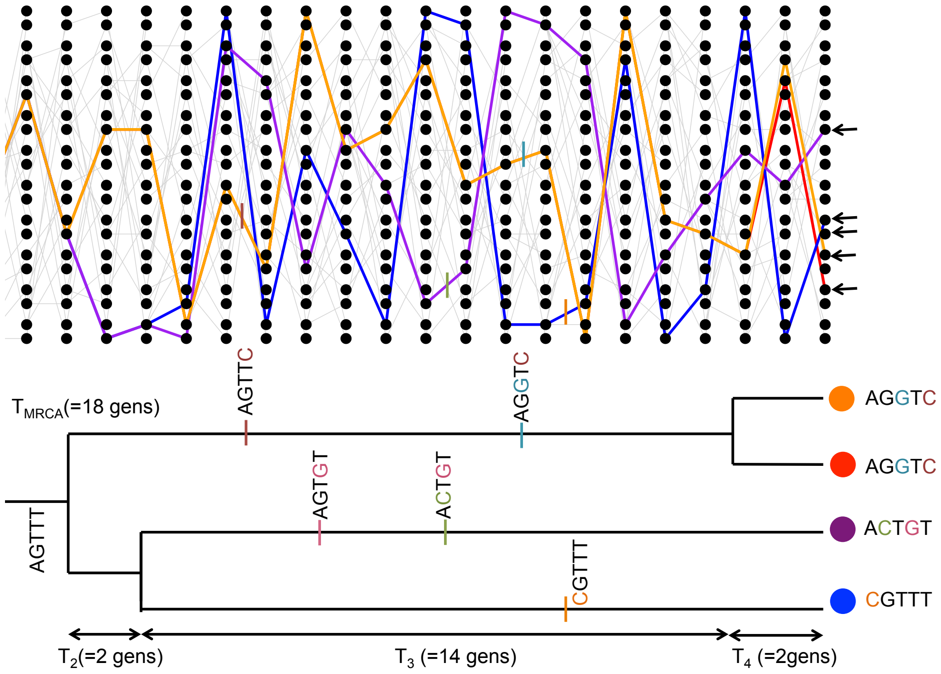
4.3.1 Coalescent genealogies and mutations

- A bit of math can help provide a simple expectation for the T_{MRCA} ; Let's work through this…

- First, let's consider the T_{MRCA} to be:

$$T_{MRCA} = \sum_{i=n}^{2} T_i \qquad (4.33)$$

where we are summing the time in generations from our full sample of alleles ($i = n$) to 2 remaining alleles after all other alleles coalesce



4.3.1 Coalescent genealogies and mutations

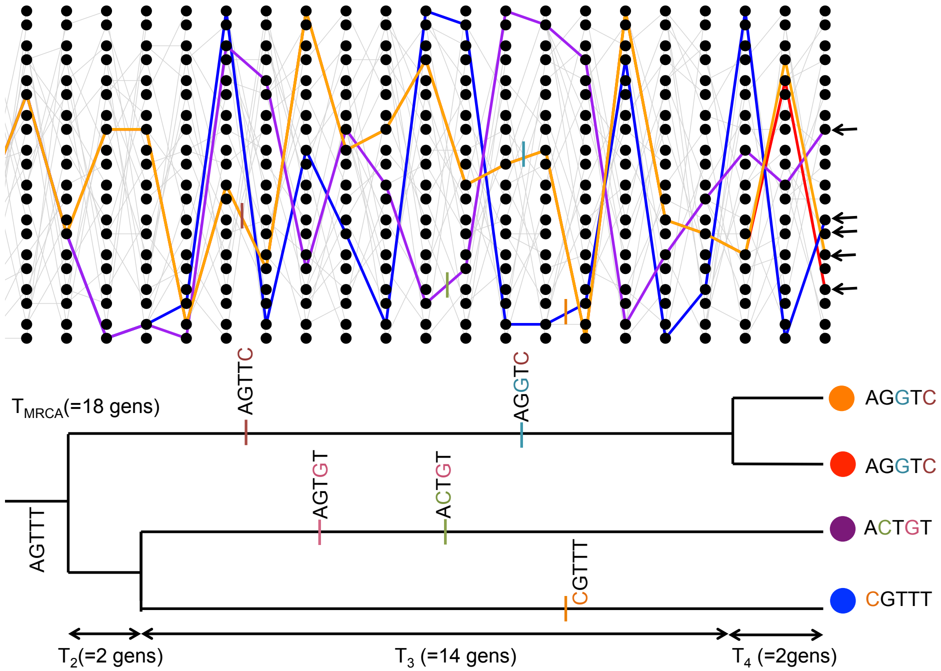
- Our coalescence time between each pair of alleles is independent, so our expected T_{MRCA} becomes:

$$\mathbb{E}(T_{MRCA}) = \sum_{i=n}^{2} \mathbb{E}(T_i) = \sum_{i=n}^{2} 2N/\binom{i}{2} \qquad (4.34)$$

- Some rearrangement of this equation yields the form:

$$\mathbb{E}(T_{MRCA}) = 4N\left(1 - \frac{1}{n}\right) \qquad (4.35)$$

- And this reveals that as our sample size (n) gets large, our T_{MRCA} is $\approx 4N$

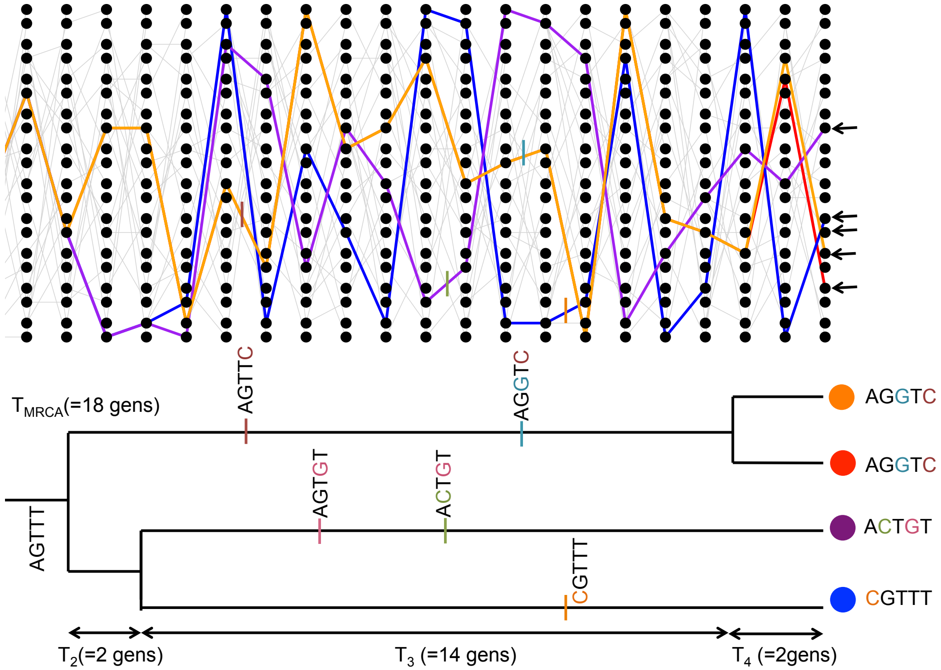


4.3.1 Coalescent genealogies and mutations

- While $4N$ is the approximate number of generations until the T_{MRCA}, there are many more generations cumulatively in the genealogy
- Mutations will occur on all lineages within the genealogy, so it is important to be able to derive an expectation for the total time (T_{tot})

$$T_{tot} = \sum_{i=n}^{2} iT_i \qquad (4.36)$$

- This means that each lineage (i) contributes T_i time in generations to the total time

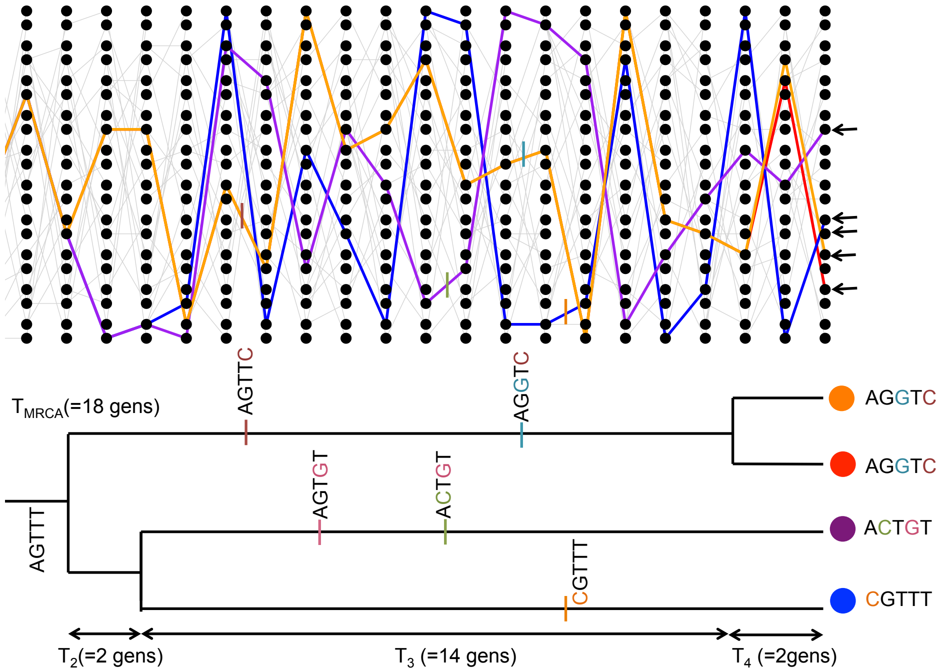


4.3.1 Coalescent genealogies and mutations

- The expectation for total time can then be found as:

$$\mathbb{E}(T_{tot}) = \sum_{i=n}^{2} i \frac{2N}{\binom{i}{2}} = \sum_{i=n}^{2} \frac{4N}{i-1} = \sum_{i=n-1}^{1} \frac{4N}{i} \qquad (4.37)$$

- From this expectation of T_{tot} we can learn that:
  - the total time scales linearly with the population size (N)
  - total time increases with sample size (n), but very slowly
  - with large samples, initial coalescence happens rapidly and addition of more individuals does little to add to total time in the tree



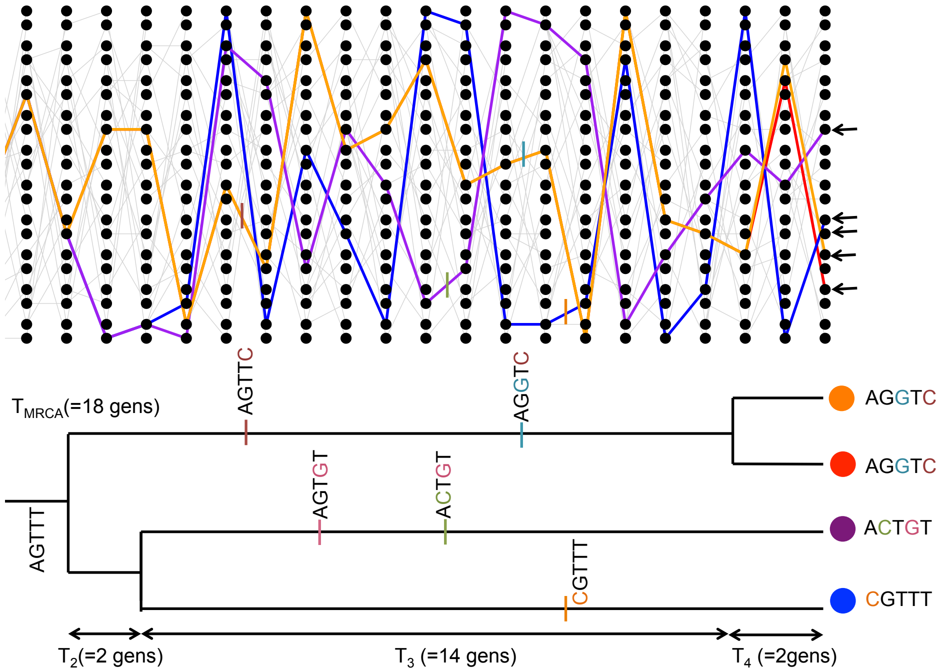
4.3.1 Coalescent genealogies and mutations

- Now that we have an expectation for T_{tot}, we can determine the number of mutations, or segregating sites (S) that are found within our samples.

- The expected number of segregating sites in a sample of size n is:

$$\mathbb{E}(S) = \mu \mathbb{E}(T_{tot}) = \sum_{i=n-1}^{1} \frac{4N\mu}{i} = \theta \sum_{i=n-1}^{1} \frac{1}{i} \qquad (4.38)$$

- Again, this value is growing very slowly with increasing sample size



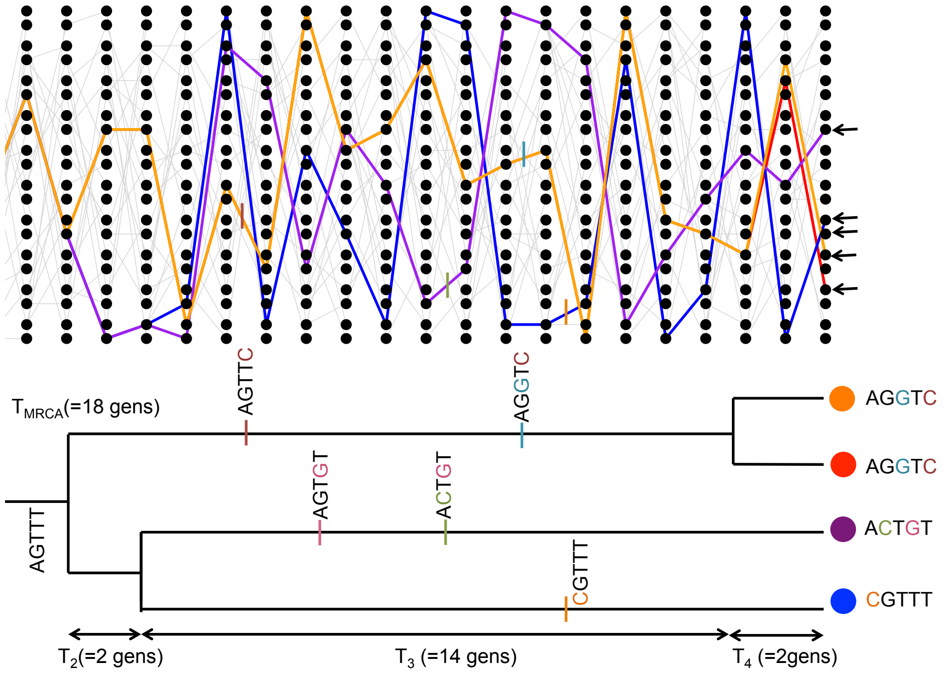
4.3.1 Coalescent genealogies and mutations

$$\mathbb{E}(S) = \mu \mathbb{E}(T_{tot}) = \sum_{i=n-1}^{1} \frac{4N\mu}{i} = \theta \sum_{i=n-1}^{1} \frac{1}{i} \qquad (4.38)$$

- This expectation of the number of segregating sites was used by Watterson to create another empirical estimate of θ

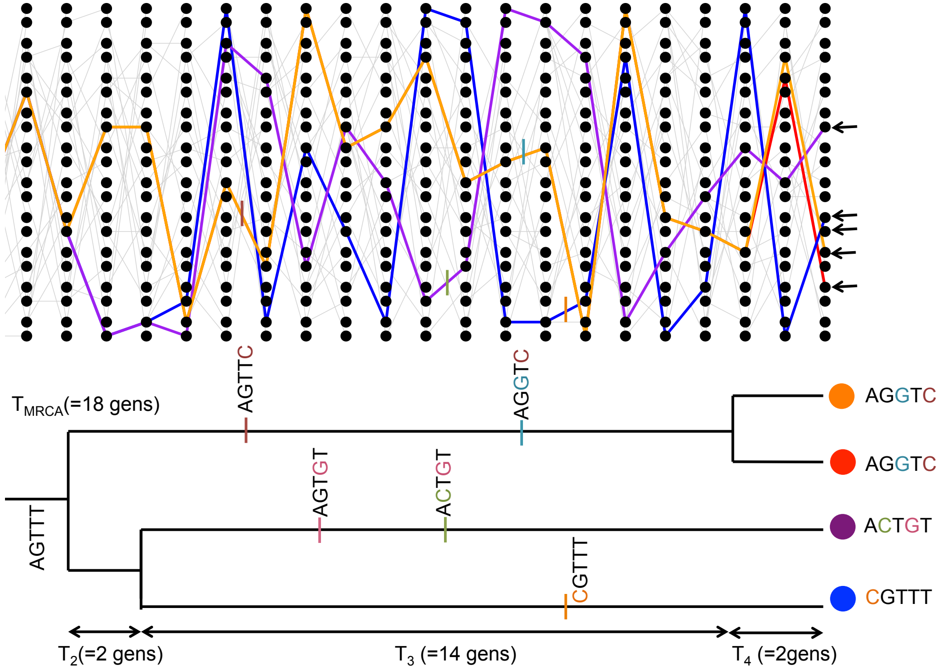
- If we substitute our empirical count of segregating sites in a sample, then:

$$\widehat{\theta}_W = \frac{S}{\sum_{i=n-1}^{1} 1/i} \qquad (4.39)$$



4.3.1 Coalescent genealogies and mutations

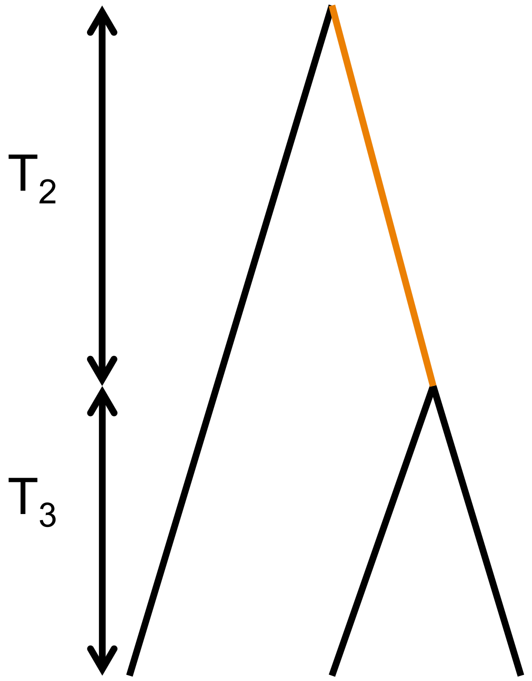
- Once we have our coalescent genealogy and have place our mutations, we can determine the number of times each mutant (derived allele) will occur within a sample
- A mutation that falls on a branch with i descendants will have a frequency of i
- For example, in the genealogy to the right, mutation AGTTC will have a frequency of 0.5 in this sample; it is a doubleton, occurring in 2 of the 4 individuals



4.3.1 Coalescent genealogies and mutations

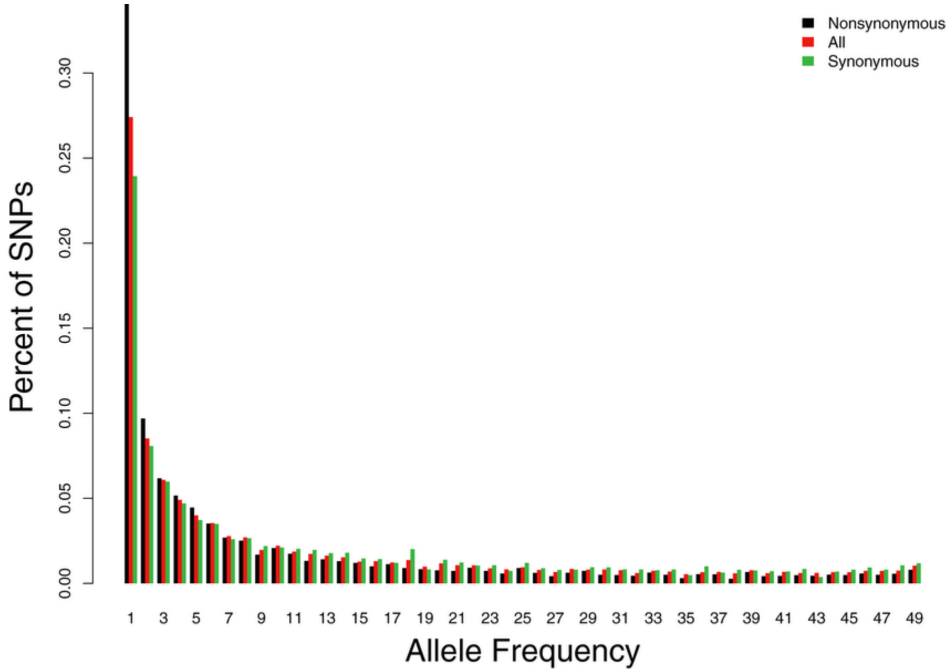
- To clarify, in the simple coalescent tree here with three samples, mutations that fall on the black branches will be singletons, but mutations that fall on the orange branch will be doubletons
- The total time in which a mutation creates a singleton will be $3T_3 + T_2$ and the total time in which a mutation creates a doubleton will be T_2
- Hudson (2015) wrote a simple proof to show that the relative frequency of singletons, doubletons, tripletons, etc… would be:

$$\mathbb{E}(S_i) = \frac{\theta}{i} \tag{4.41}$$



4.3.1 Coalescent genealogies and mutations

- This means there are twice as many singletons as doubletons, three times as many singletons as tripletons, etc…
- Empirical data back up this expectation that singletons are much more common than mutations/derived alleles at higher frequency
- Another important thing to know about singletons is that they are younger than mutations/derived alleles at higher frequency
- Based on our neutral expectation of the frequency of singletons, doubletons, tripletons, etc… we can construct a neutral site frequency spectrum (SFS)



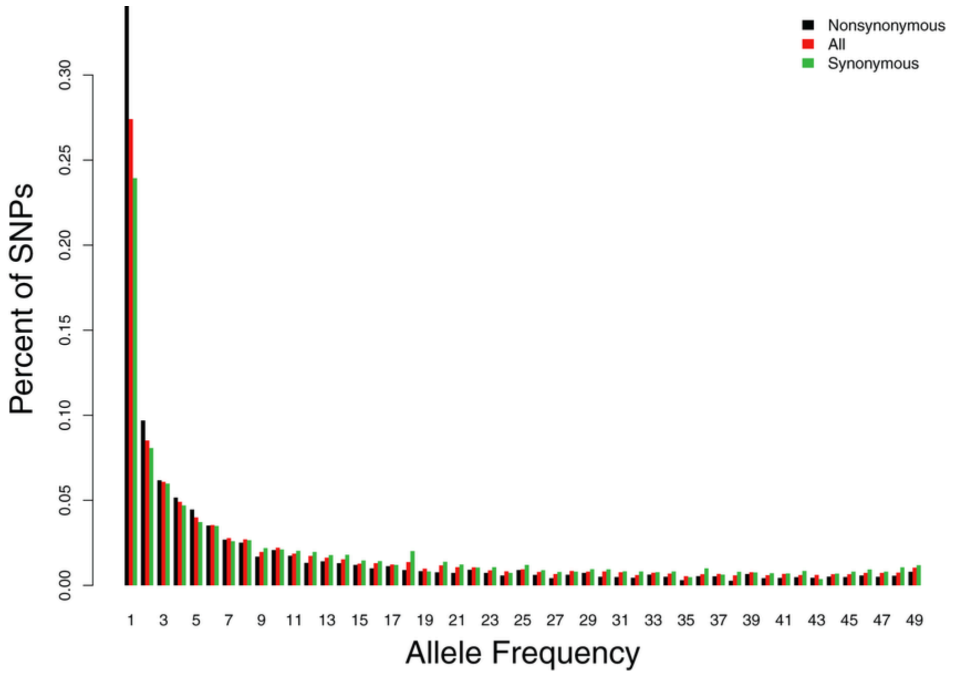
4.3.1 Coalescent genealogies and mutations

- Population geneticists often compare an empirical site frequency spectrum (one they generate with experimental data) to the neutral expectation to see if they are significantly different and if a neutral null model can be rejected

- These tests can detect sudden population size changes or natural selection

- One of earliest tests that summarized deviation from neutrality in the SFS was Tajima's D:

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{C} \qquad (4.43)$$

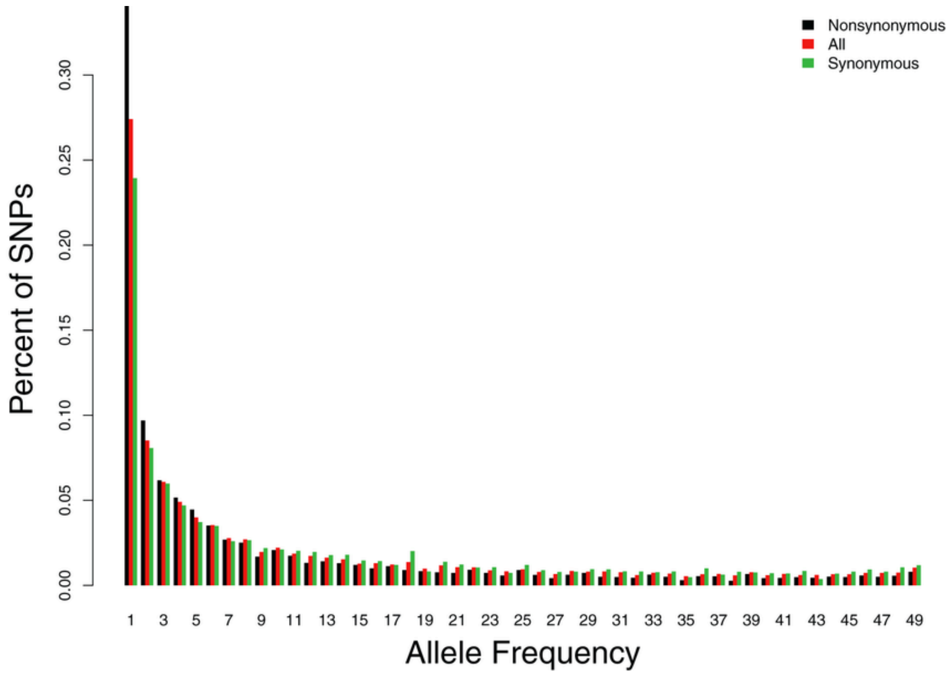


4.3.1 Coalescent genealogies and mutations

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{C} \qquad (4.43)$$

- An excess of rare variants (e.g., singletons) in the empirical data relative to the neutral expectation will result in a negative value for Tajima’s D

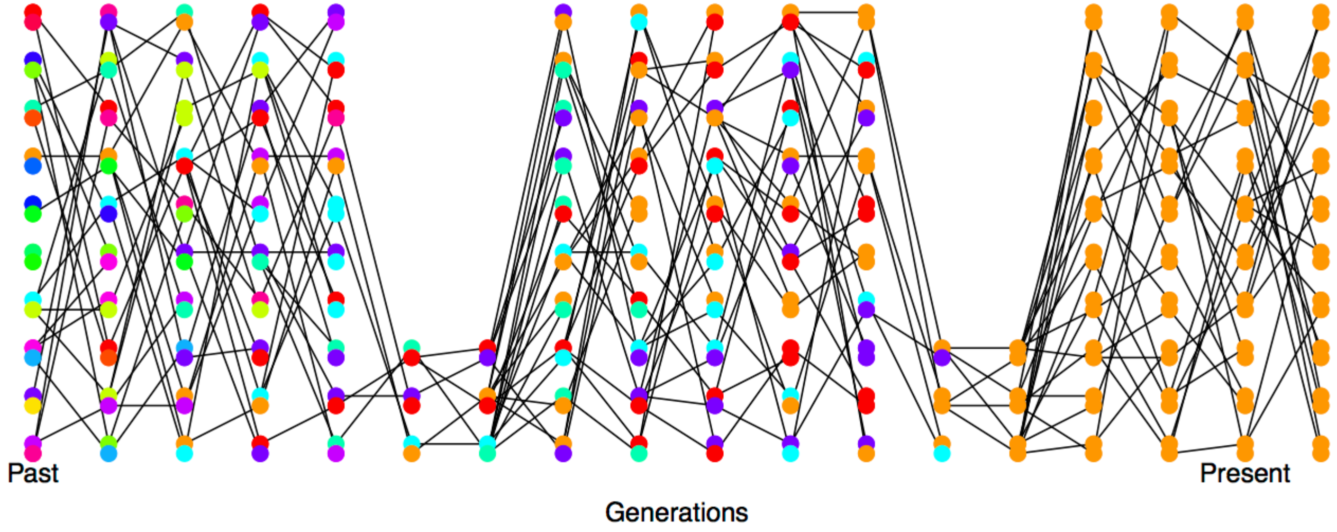
- An excess of intermediate-frequency variants in the empirical data relative to the neutral expectation will result in a positive value for Tajima’s D

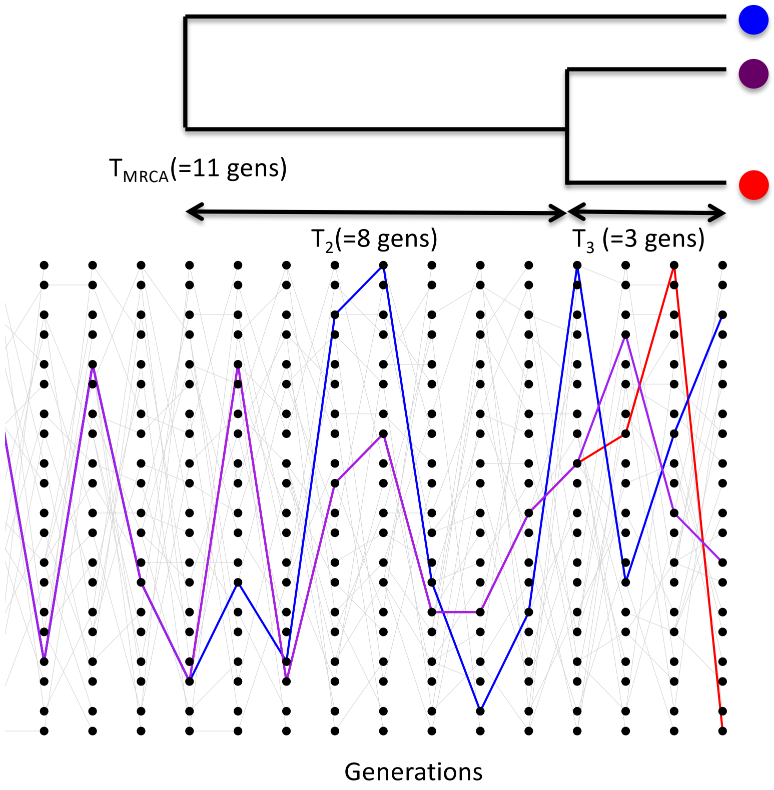


Coop, Chapter 4: 4.3.2

Genetic Drift and Neutral Diversity

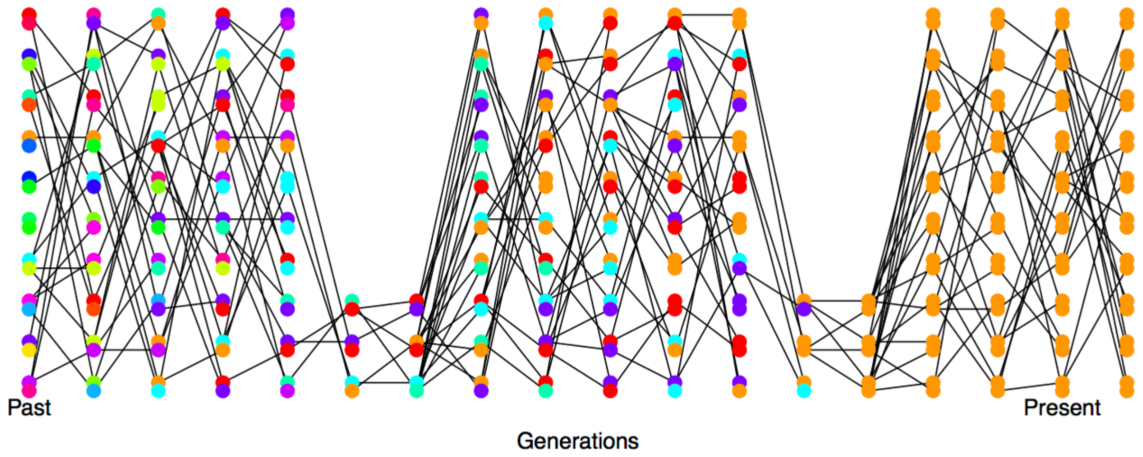
Demography and the coalescent





4.3.2 Demography and the Coalescent

- We've seen in previous sections that the rate of loss of heterozygosity due to drift depends on the population size
- With the coalescent, we also know that if the population size in generation i is N_i, then the probability that a pair of lineages coalesces is $1/2N_i$; if the population is small, then lineages will coalesce more quickly
- We can average over fluctuations in population size by using N_e rather than N, but longer-term, systematic changes will cause deviations away from expectations based on the neutral coalescent

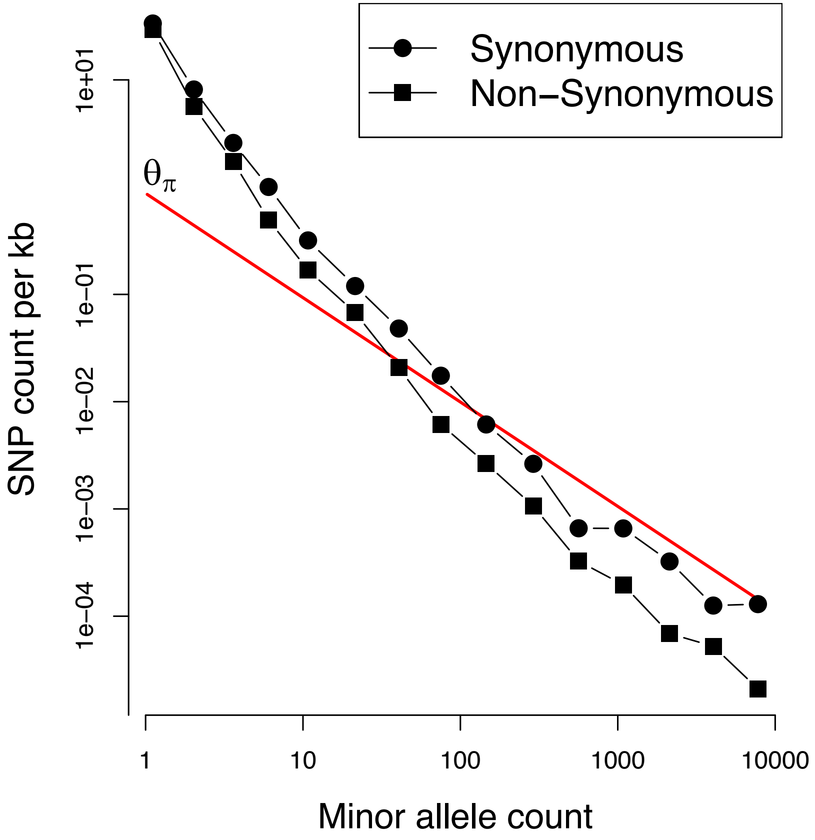


4.3.2 Demography and the Coalescent

- Take, for example, data from 202 genes in a large sample of humans ($n = 14{,}002$)

- The expectation for allele frequencies under the neutral coalescent is shown with the red line and the empirical data are in black for both synonymous and non-synonymous sites

- There are many more rare alleles in the empirical human data than we would expect, but common alleles roughly match the neutral expectation

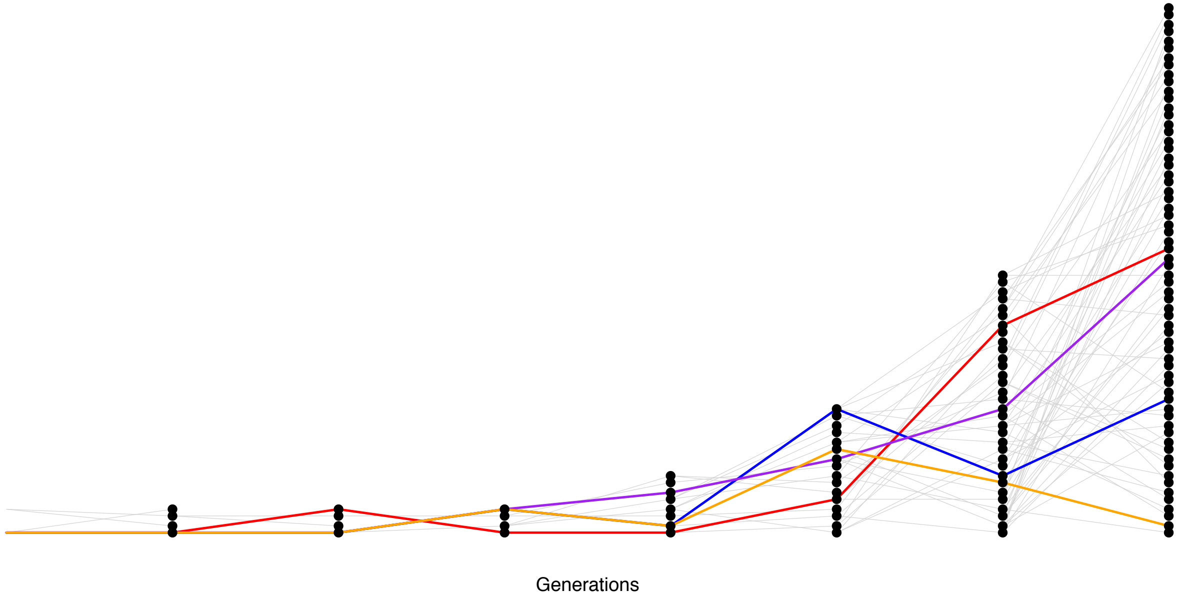


4.3.2 Demography and the Coalescent

- These patterns likely reflect the recent explosive population growth in humans over the last 1,000-10,000 years to a global population of > 7 billion

- The genetic diversity in humans is much smaller than would be expected based on this large census size due to our smaller ancestral population

- In an expanding population, most of the coalescence events happen further back in time in the tree

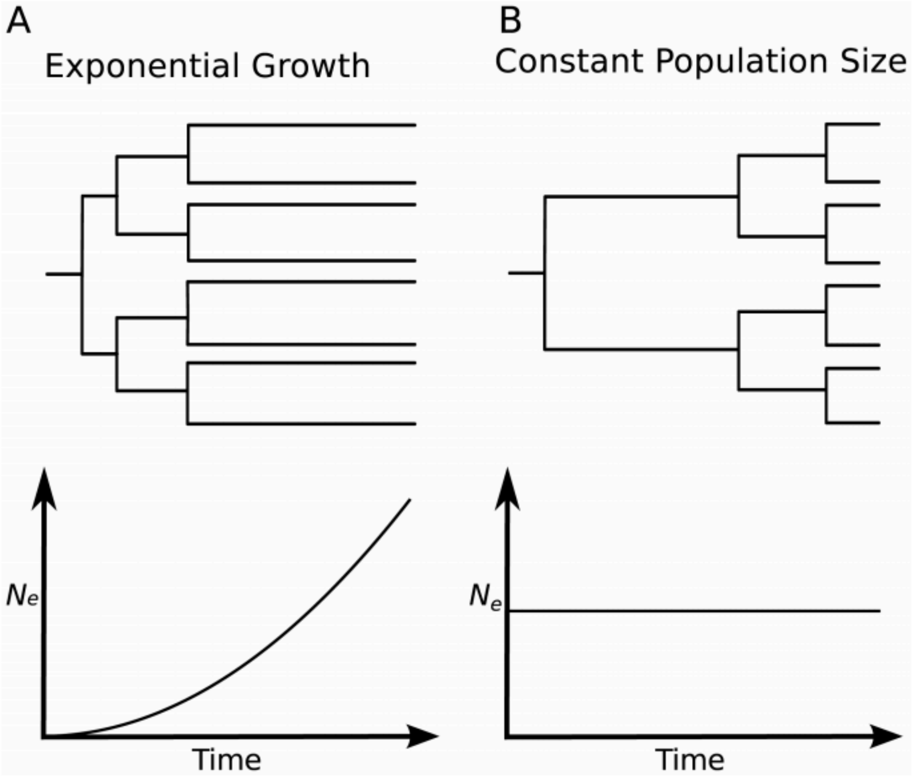


4.3.2 Demography and the Coalescent

- Relative to the neutral coalescent, with expanding populations, lineage time is compressed further back in the tree where older, common mutations arise

- Branches toward the present where rare mutations arise are longer than constant-sized populations

- This explains why we see an excess of singletons in human populations



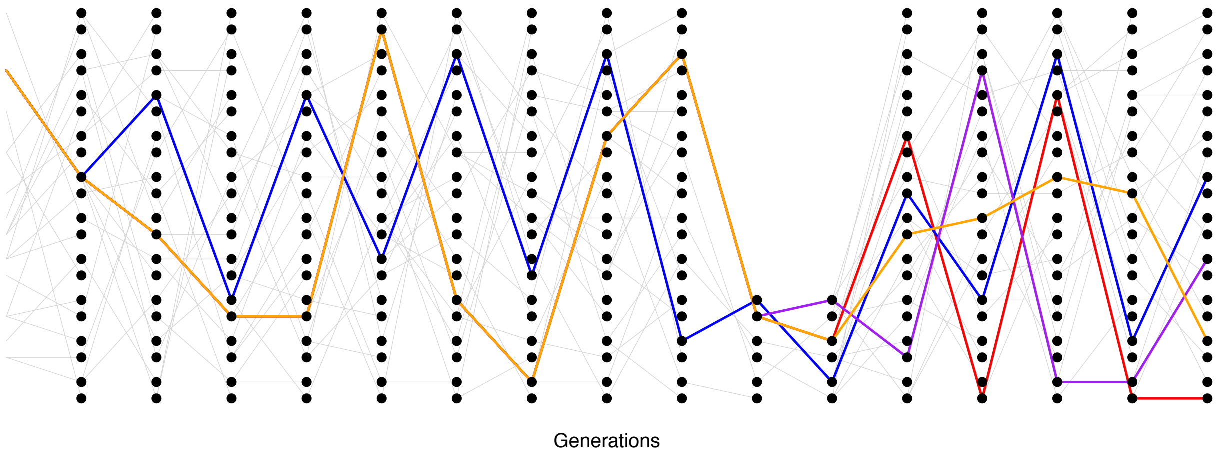
http://evol.bio.lmu.de/_teaching/evogen/EvolGenet_L5_Coalescent.pdf

4.3.2 Demography and the Coalescent

- Population bottlenecks are another demographic deviation from expectations under the neutral coalescent

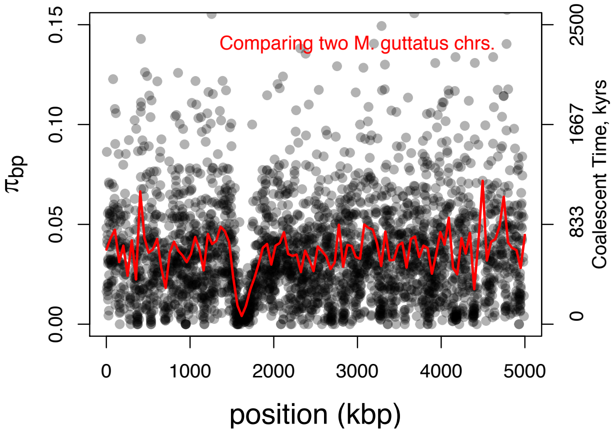
- When looking back in time at patterns, very rapid coalescence occurs during the bottleneck

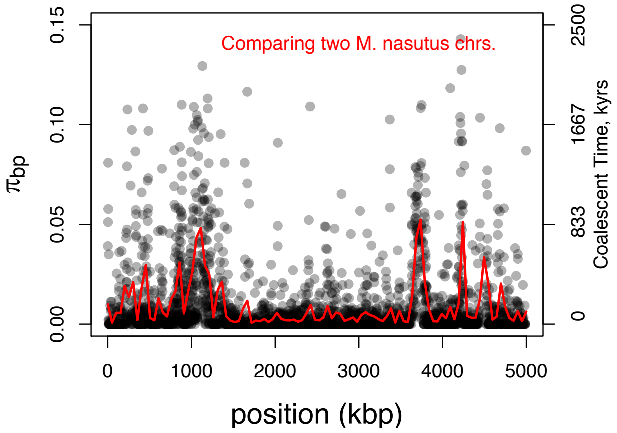
- If the bottleneck is strong enough, all lineages coalesce and the SFS a few generations later looks a lot like population expansion (many rare alleles)



4.3.2 Demography and the Coalescent

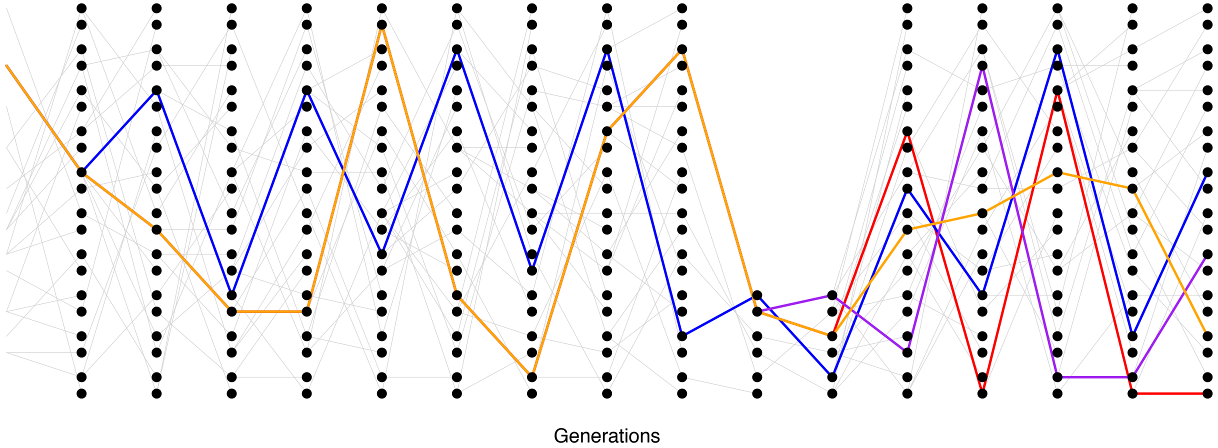
- If multiple lineages survive the bottleneck, then, within the population, there will be a subset of lineages with very deep coalescent time
- For example *Mimulus nasutus* is a selfing species recently derived from *M. guttatus; M. nasutus* has recently gone through a bottleneck
- While low nucleotide diversity is observed across the majority of *M. nasutus* chromosomes, high diversity regions can be found where multiple lineages made it through the bottleneck

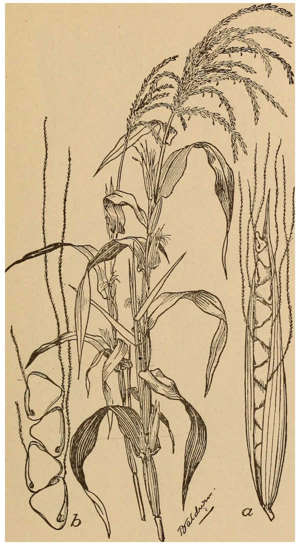




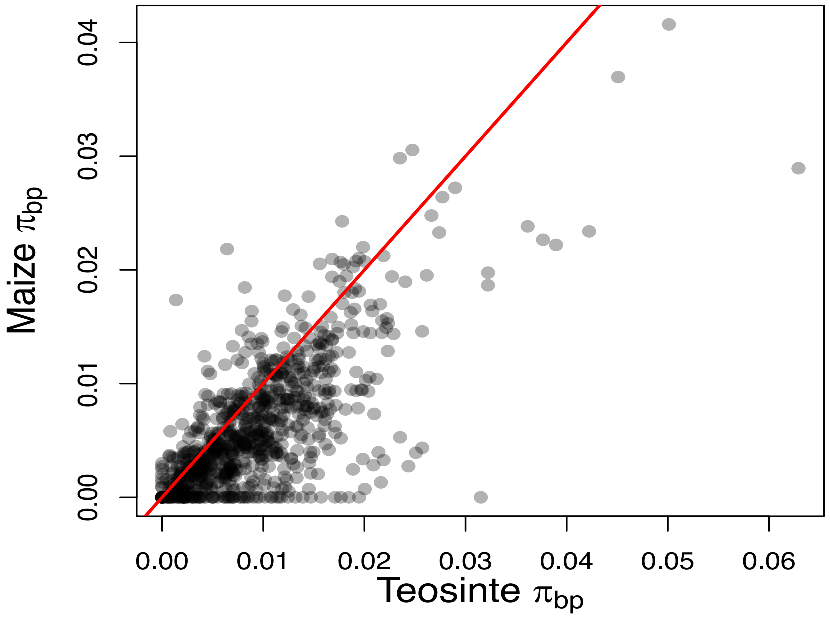
4.3.2 Demography and the Coalescent

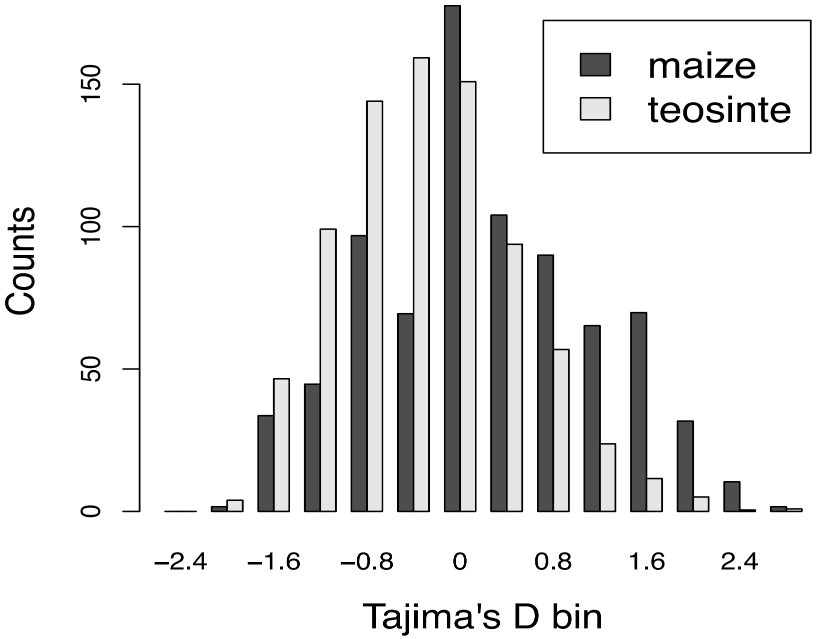
- Maize is a good example of a species that went through a recent, mild bottleneck (caused by domestication from the wild plant teosinte)
- Multiple lineages survived the bottleneck and these have deep coalescence times like the orange and blue lineages in the figure to the right
- This causes an excess of older, more common alleles relative to the neutral expectation and therefore shifts Tajima's D to positive values





4.3.2 Demography and the Coalescent





- Nucleotide diversity measured by θ_π is lower in maize than teosinte due to the genetic bottleneck

- Tajima's *D* values are shifted toward more positive values in maize relative to teosinte because this was a more mild bottleneck and multiple, old lineages survived