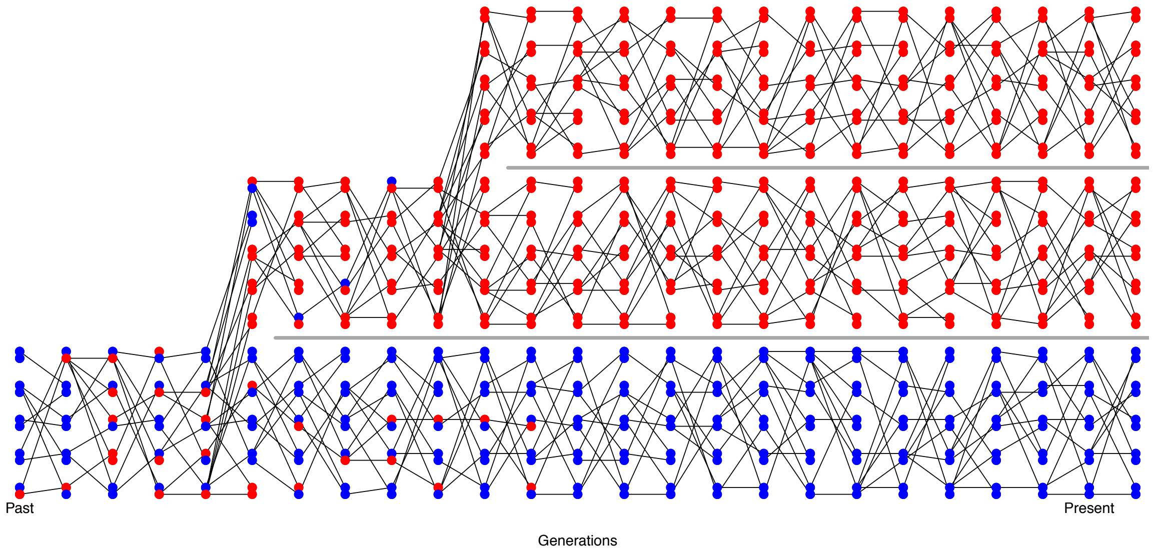
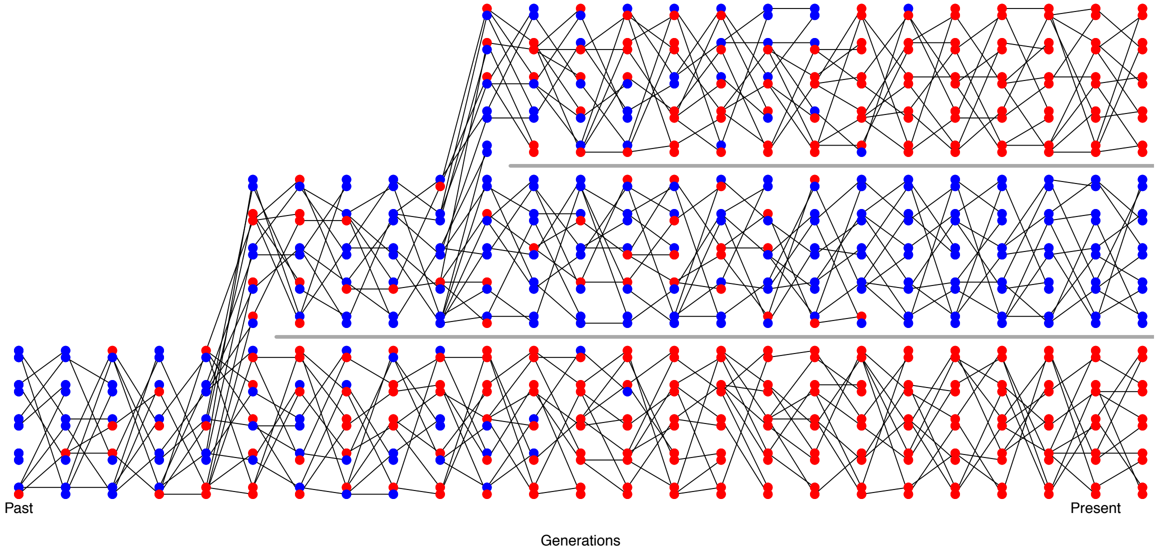
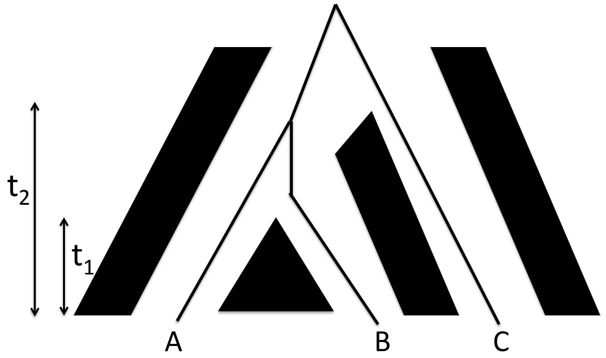
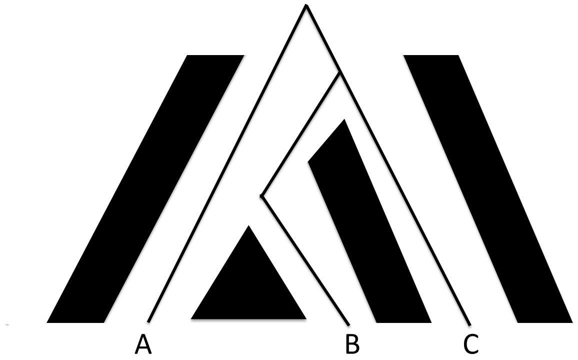
Coop, Chapter 6

Neutral Diversity and Population Structure







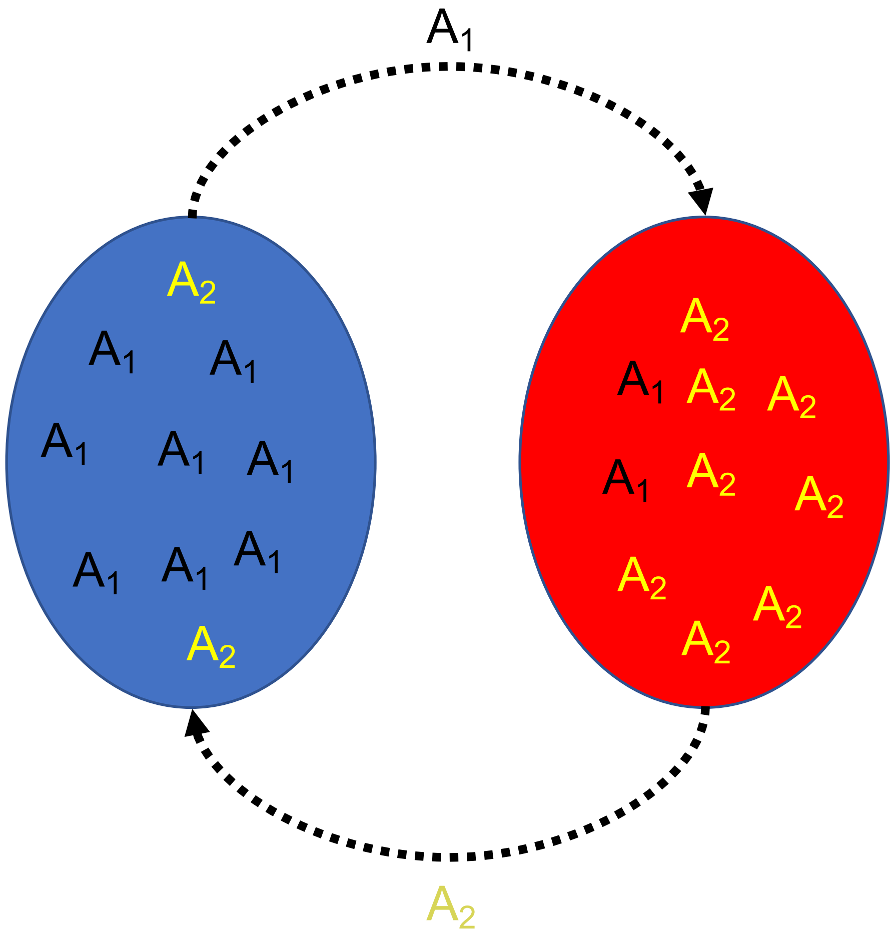


Introduction

- When genetic differentiation occurs between subpopulations, gene flow or migration can reduce differentiation

- So far we've been assuming that any pair of alleles are equally likely to coalesce

- When we have differentiated populations connected by migration, the assumption of equal probability of coalescence is violated



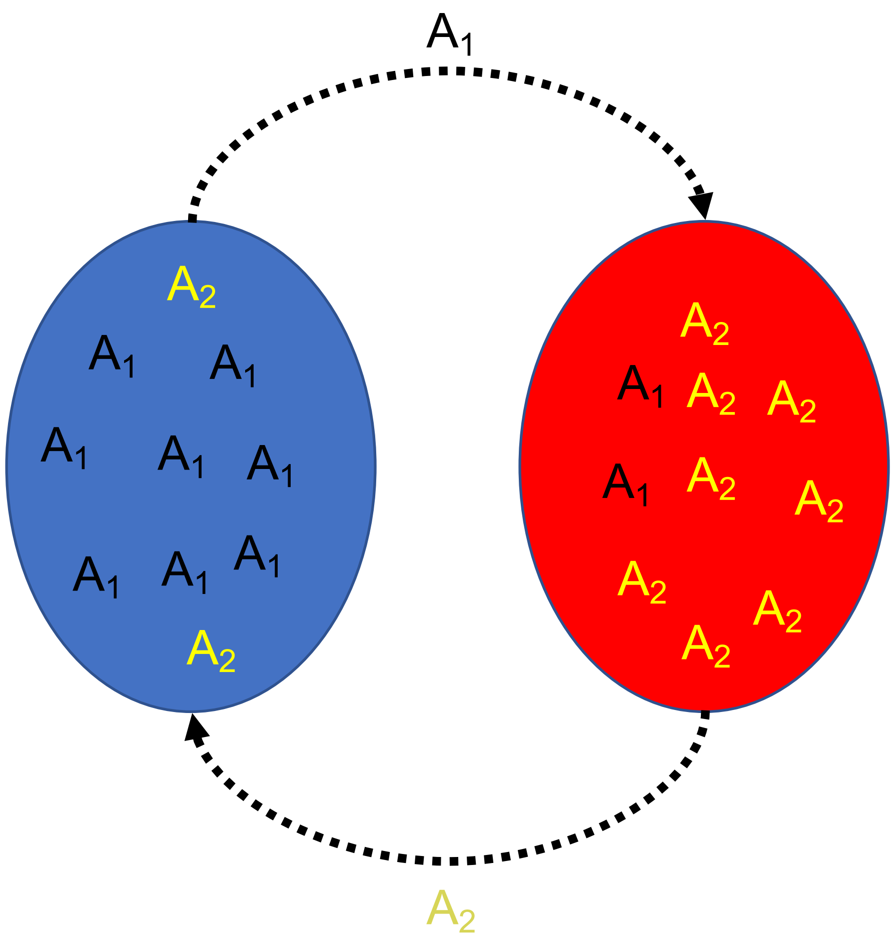
Introduction

- We will further develop models of population structure, incorporating migration, by using the F_{ST} statistic we first considered in Chapter 3:

$$F_{\mathrm{ST}} = \frac{H_T - H_S}{H_T} \qquad (6.1)$$

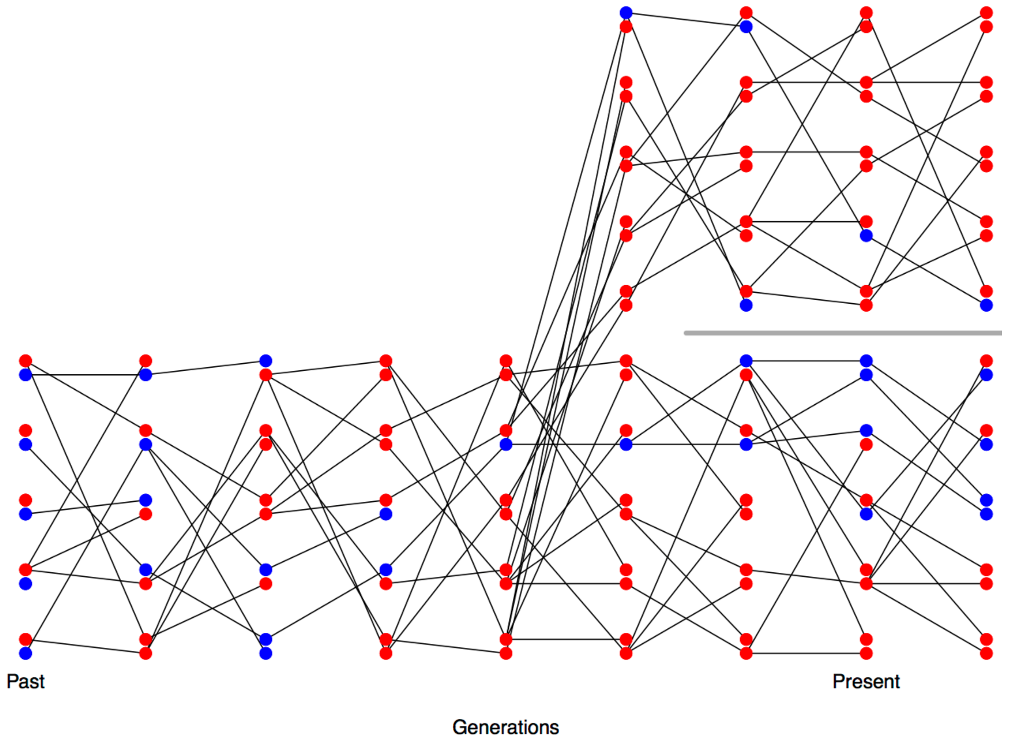
- Where H_S is the probability of randomly choosing different alleles from a subpopulation

- And where H_T is the probability of randomly choosing different alleles from the total population



6.1 A simple population split model

- We can imagine a population of constant size N_e that splits into two daughter populations, also of constant size N_e
- Our daughter populations do not exchange migrants and we sample an equal number of alleles from each of these at the present day
- If we consider a pair of alleles drawn from one subpopulation, their probability of differing (H_S) is simply $4N_e\mu$
- Things get trickier when we sample across subpopulations…



6.1 A simple population split model

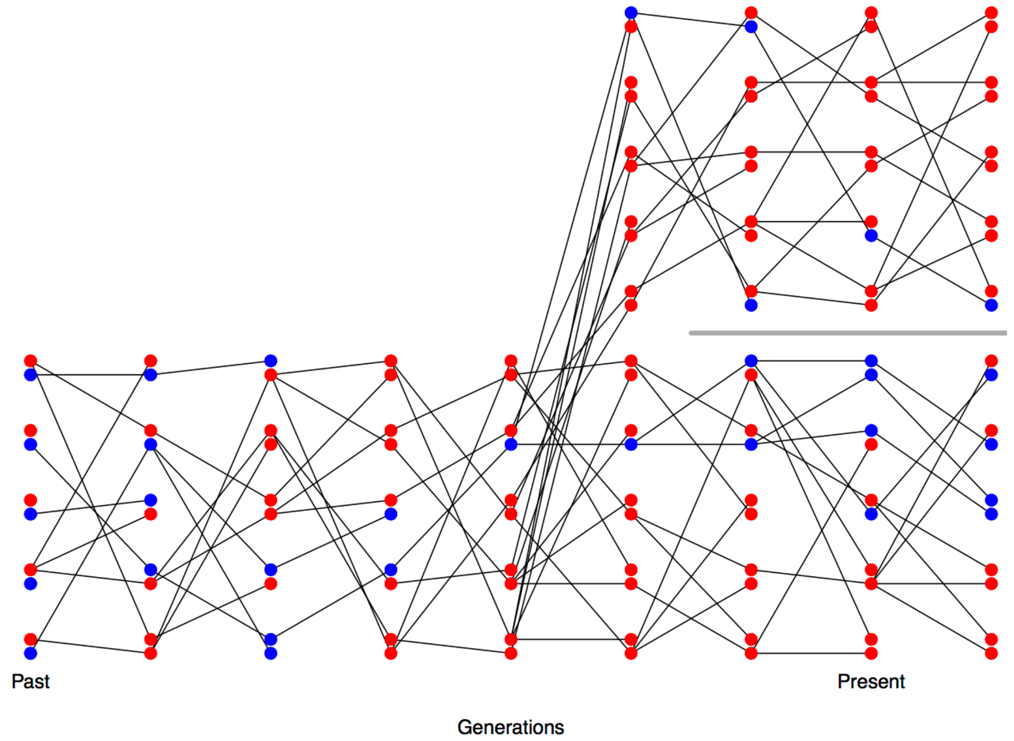
- Assume we sample equal numbers of alleles from both subpopulations and pool them

- When we randomly sample alleles from this pool, 50% of the time they are from the same subpopulation and 50% of the time they are from different subpopulations

- Therefore, the total heterozygosity (H_T) in the pool becomes:

$$H_T = \frac{1}{2}H_S + \frac{1}{2}H_B \qquad (6.2)$$

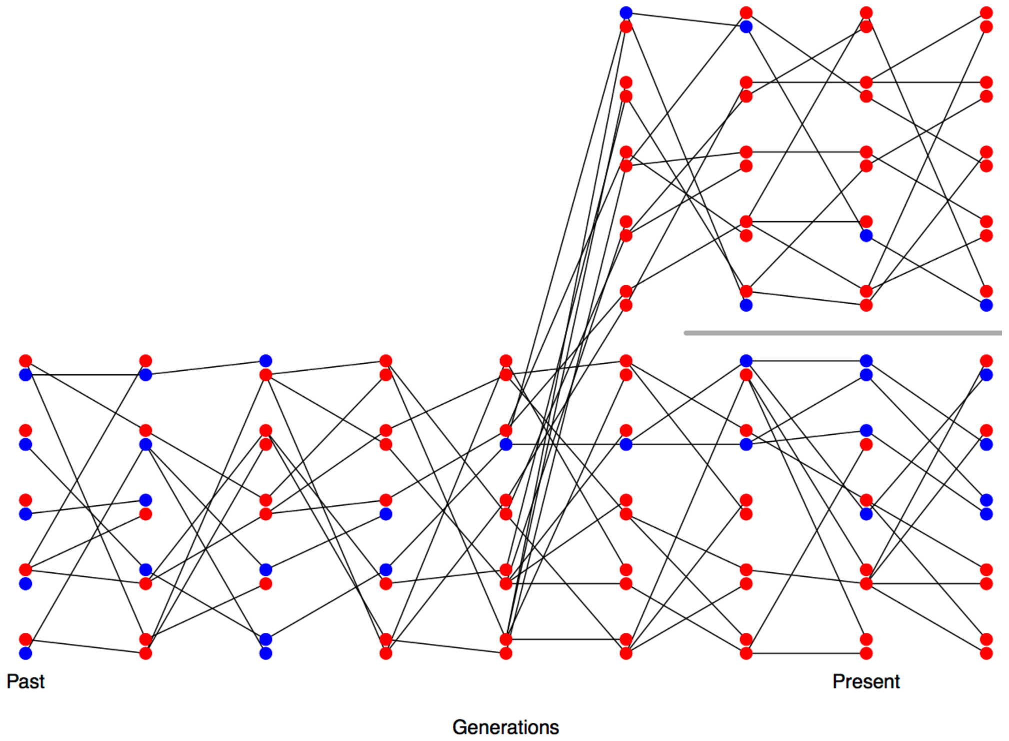
- H_B is the probability that alleles drawn from different subpopulations are different



6.1 A simple population split model

- In order for alleles drawn from different subpopulations to be the same (coalesce), we have to go back in time until when they were in the same population
- We will call the time until the subpopulations merge T, and once they are in the same population, the alleles will require a further $2N$ generations until they coalesce
- We can therefore say our alleles from different subpopulations are separated by $2(T+2N)$ meioses and, after accumulating mutations, the probability of different alleles from subpopulations becomes:

$$H_B \approx 2\mu(T+2N) \qquad (6.3)$$



6.1 A simple population split model

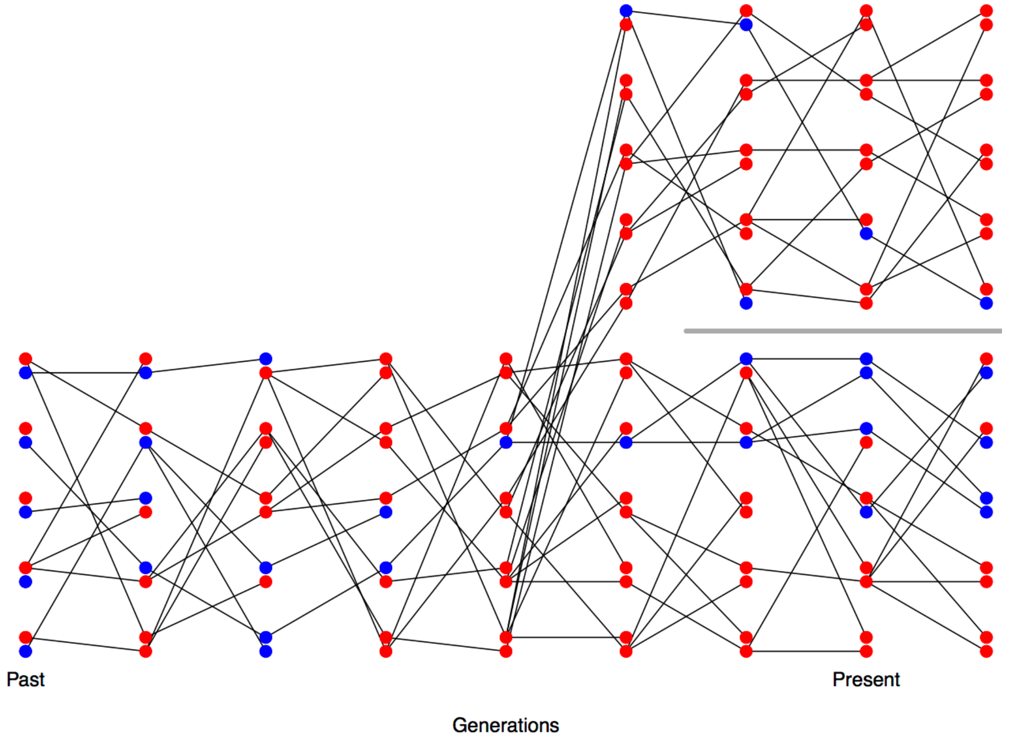
- We can now substitute our values of H_B (subpopulations) and H_T (total population) into our F_{ST} equation and find that:

$$F_{\mathrm{ST}} \approx \frac{\mu T}{\mu T + 4N_e\mu} = \frac{T}{T + 4N_e} \qquad (6.4)$$

- By looking at this equation we can see that that F_{ST} will increase as:

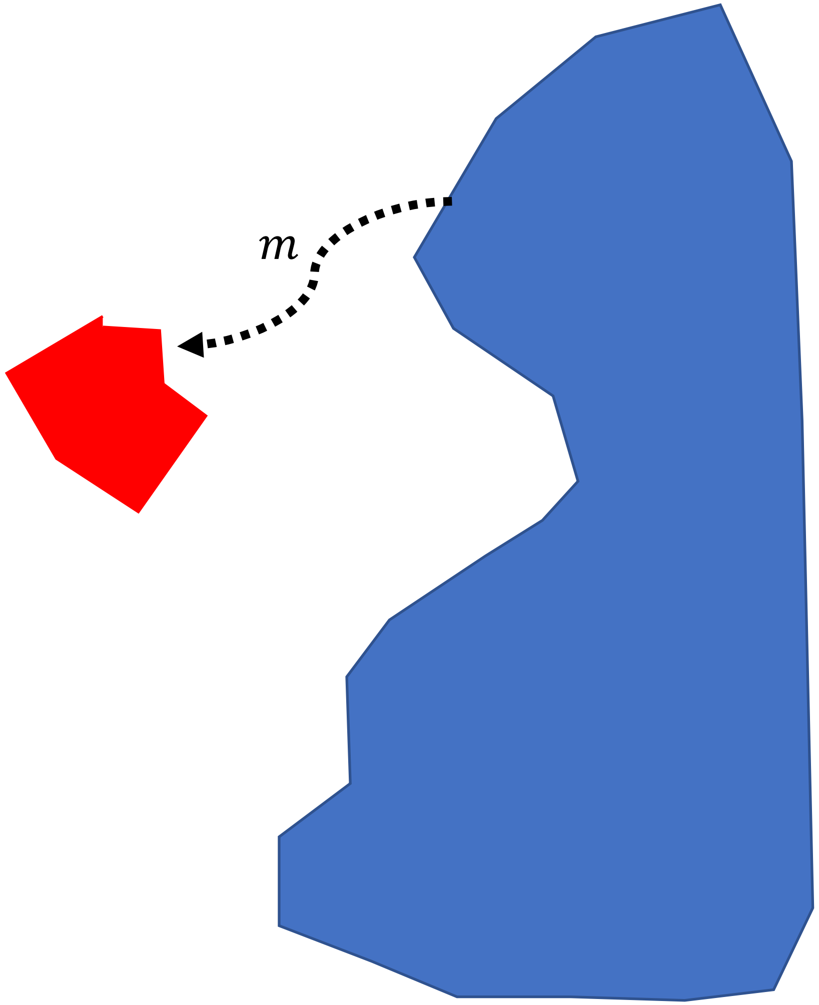
  1. Time in generations is longer

  2. N_e is smaller



6.2 A simple model of migration between an island and the mainland

- Now we'll consider a migration-drift equilibrium model between the mainland and an island
- Our expected heterozygosity on the mainland is H_M
- The island will have a very small number of individuals (N_I) relative to the mainland
- Each generation, some small fraction m of individuals on the island have mainland parents due to migration; migration in the other direction is negligible



6.2 A simple model of migration between an island and the mainland

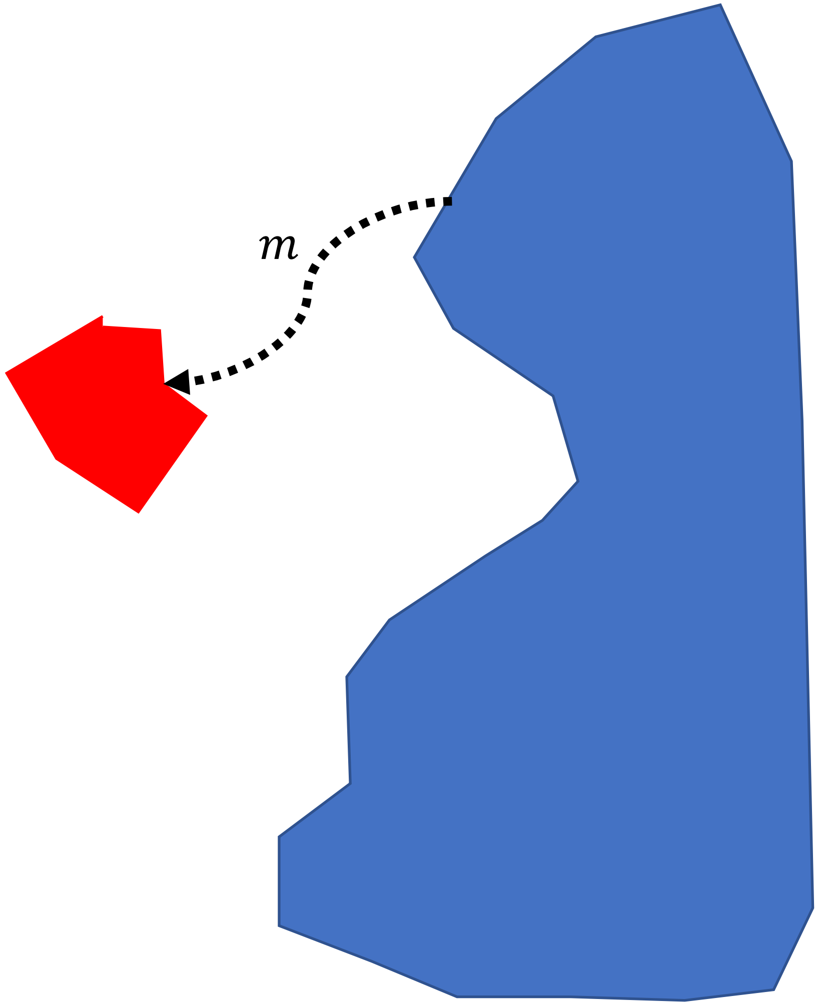
- If we sample an island allele and try to trace its lineage back in time, each generation it has a small probability m of being descended from the mainland

- The probability that an island allele coalesces with another island allele is the probability of coalescence without migration having occurred in either of these alleles from the mainland

$$\frac{1}{2N_I}(1-m)^{2(t+1)}\left(1-\frac{1}{2N_I}\right)^t \approx \frac{1}{2N_I}\exp\left(-t\left(\frac{1}{2N_I}+2m\right)\right), \quad (6.5)$$

Coalesces without migration

Doesn't coalesce



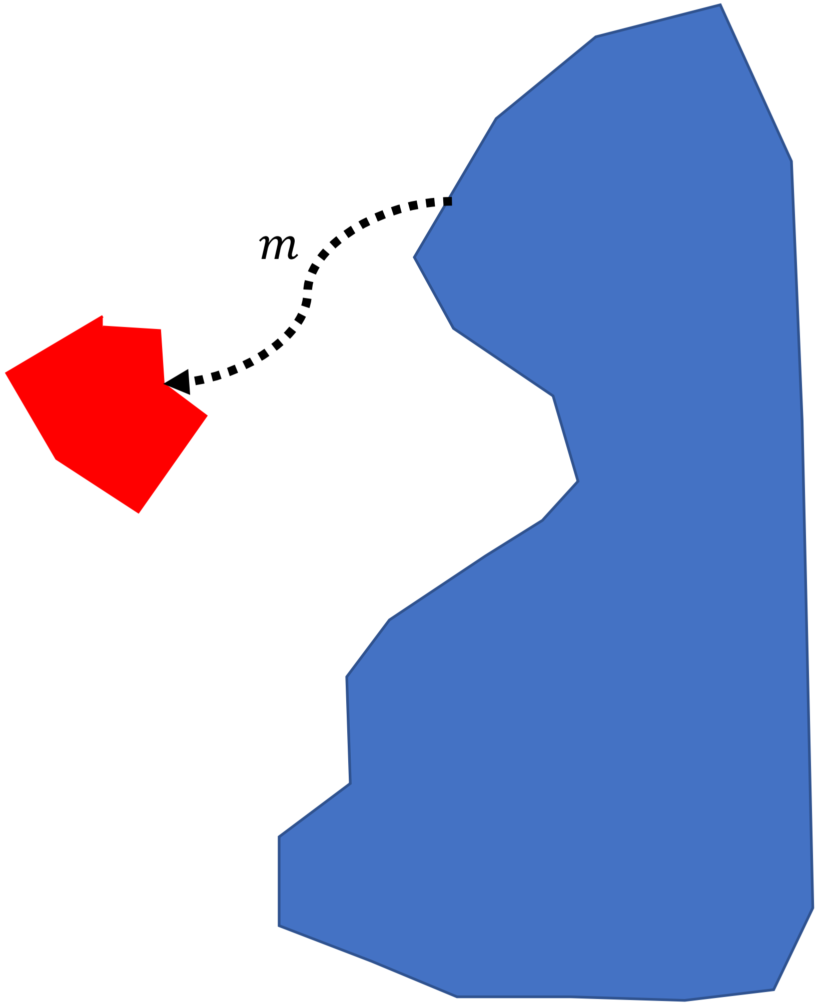
6.2 A simple model of migration between an island and the mainland

- We can also adjust our probability of coalescence on the island before either allele has migrated from the mainland to be irrespective of time:

$$\int_0^\infty \frac{1}{2N_I} \exp\left(-t\left(\frac{1}{2N_I} + 2m\right)\right) dt = \frac{1/(2N_I)}{1/(2N_I) + 2m}. \qquad (6.6)$$

- If we assume the probability of mutation is very low, the only way two alleles can be different is if migration from the mainland has occurred. Therefore:

$$1 - \frac{1/(2N_I)}{1/(2N_I) + 2m} \qquad (6.7)$$



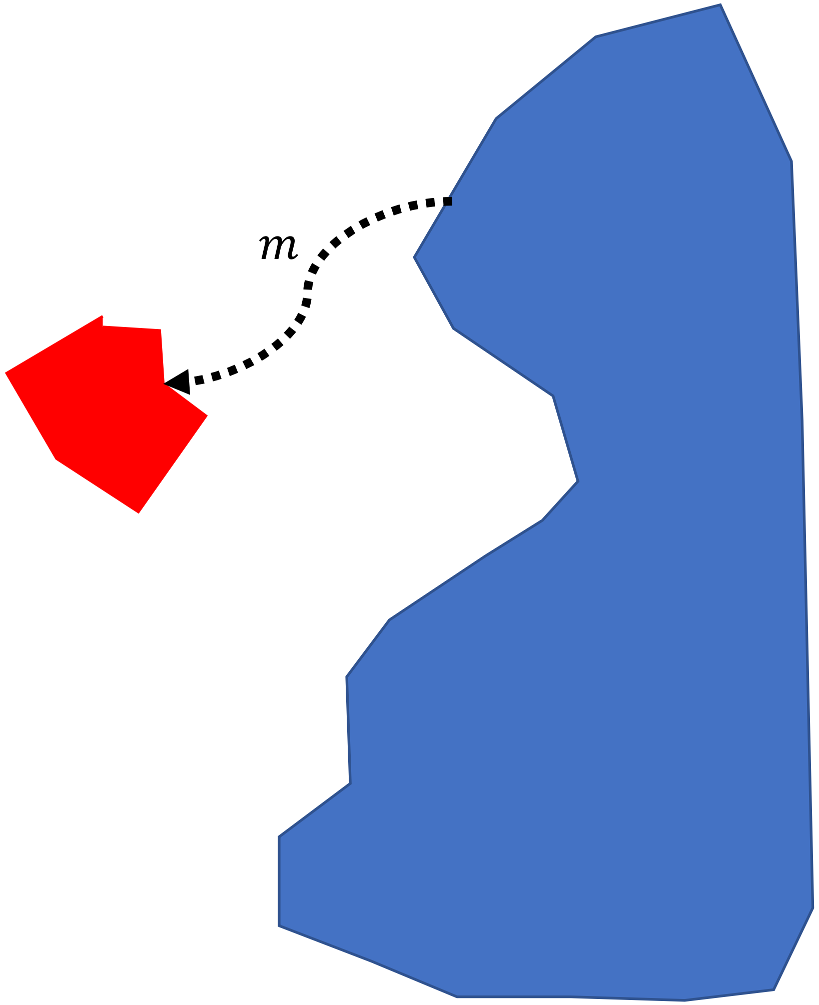
6.2 A simple model of migration between an island and the mainland

- To have heterozygosity on our island, then, we first have to see that an allele has migrated from the mainland and that these represent different draws from the mainland:

$$H_I = \left(1 - \frac{1/(2N_I)}{1/(2N_I) + 2m}\right) H_M \qquad (6.8)$$

- We can now also consider the level of inbreeding on our island relative to the mainland by calculating F_{IM}

$$F_{IM} = 1 - \frac{H_I}{H_M} = \frac{1/(2N_I)}{1/(2N_I) + 2m} = \frac{1}{1 + 4N_I m}. \qquad (6.9)$$



6.2 A simple model of migration between an island and the mainland

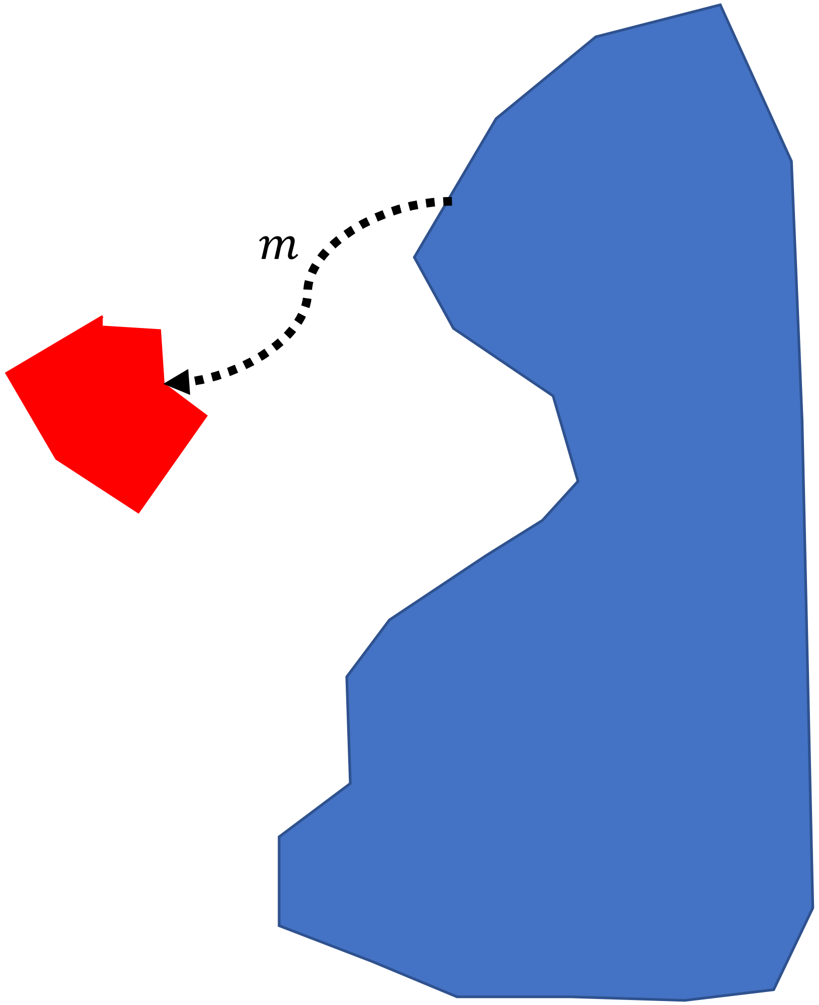
$$F_{IM} = 1 - \frac{H_I}{H_M} = \frac{1/(2N_I)}{1/(2N_I) + 2m} = \frac{1}{1 + 4N_I m}. \qquad (6.9)$$

- Our F_{IM} equation shows that inbreeding will be high on our island if:

  1. The migration rate is low

  2. The population size on the island is low

- This is because we're losing diversity due to strong drift on the island, and this diversity is not being replenished through gene flow from the mainland

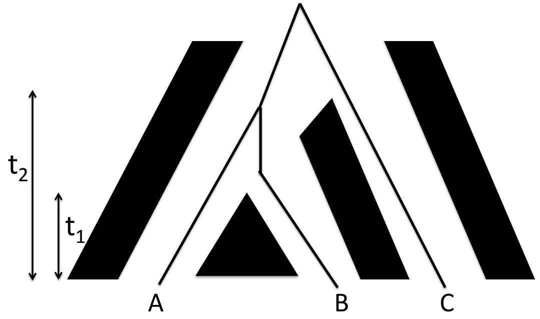


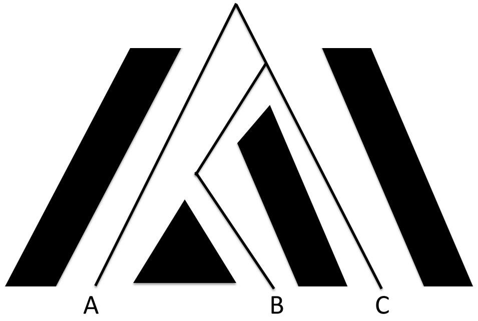
6.3 Incomplete lineage sorting

- Two alleles at a particular locus can be segregating in a population for a very long time, during which population splits may occur

- This means that coalescence occurs in the ancestral population and not within subpopulations or at the time of their split

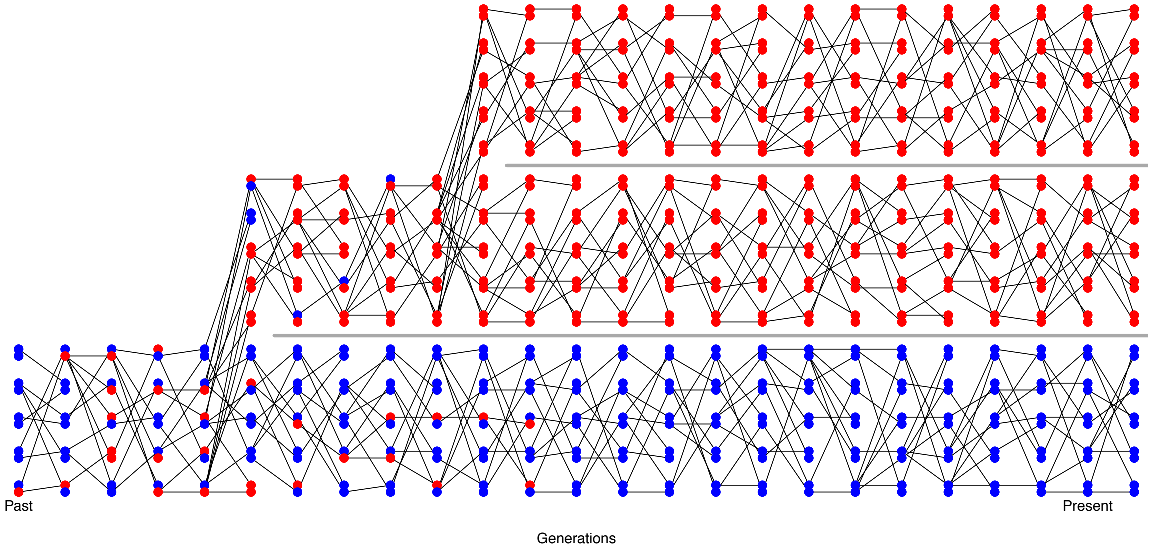
- In this situation, we can have different relationships observed at individual loci (in the gene tree) than we see when we look at relationships of populations (at the population tree) based on genome-wide data

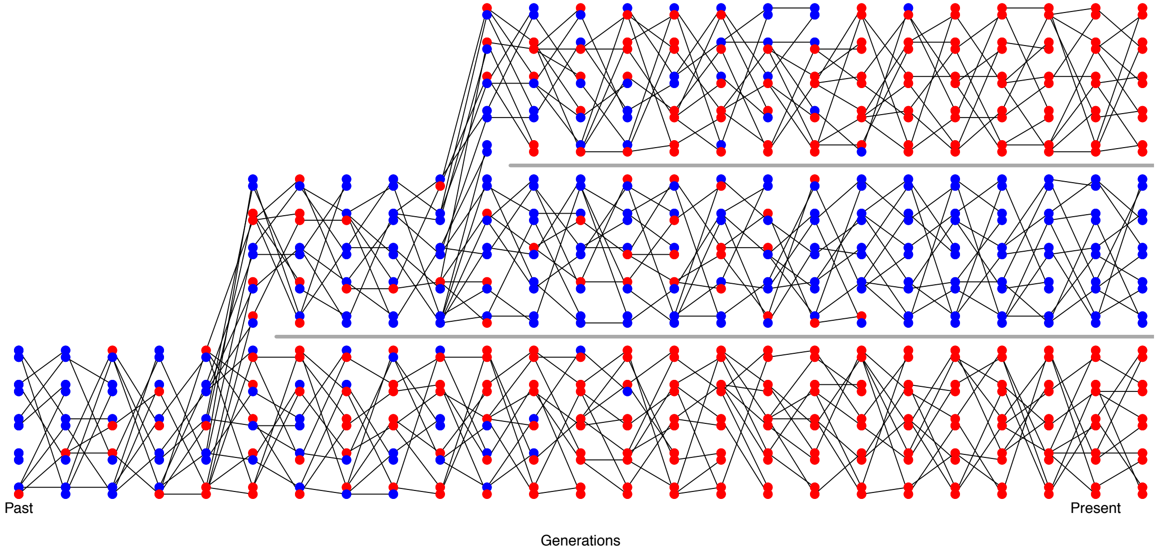


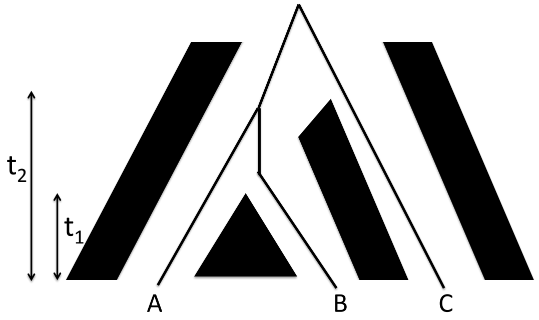


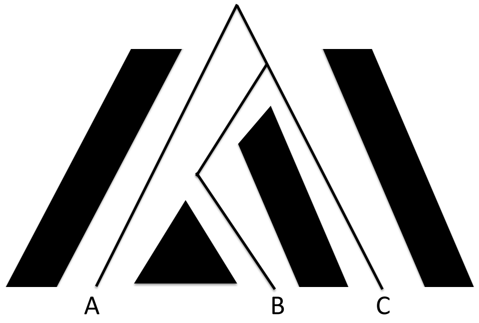
6.3 Incomplete lineage sorting

For example:







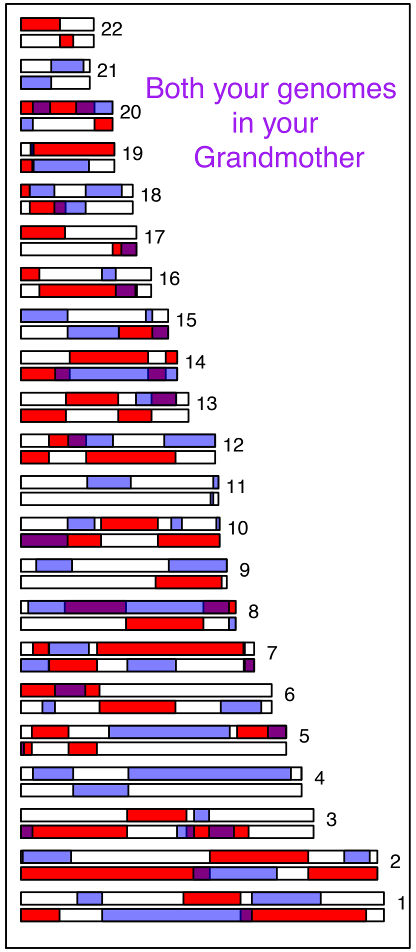


6.3 Incomplete lineage sorting

- A pedigree analogy that helps clarify is the patterns of inheritance we might see between two full siblings and their cousin

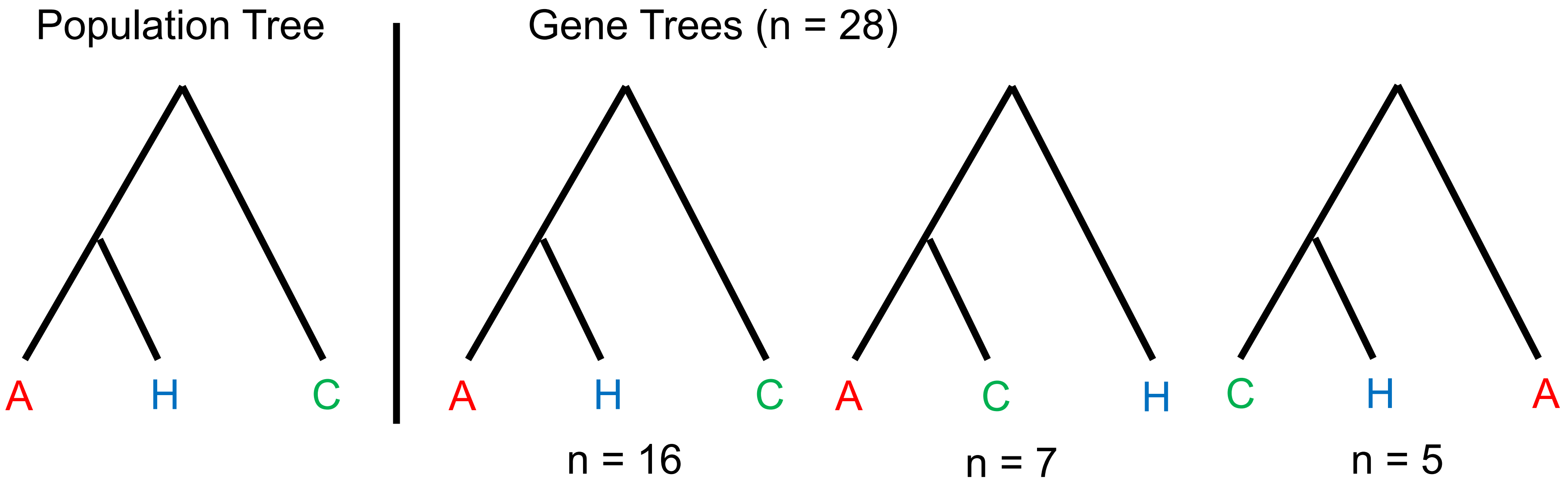
- At some loci across the genome, cousins will share alleles from their common ancestors that full siblings do not share, despite the closer relationship between full siblings

- In this instance, the average relatedness of the full siblings relative to their cousin is not reflected at a specific locus, but genome-wide, this would be clear



6.3 Incomplete lineage sorting

- As an empirical example, Jennings and Edwards (2005) were able to obtain sequence from 28 genes in three closely related species of the Australian grass finches: two long-tailed finches that are sister species (*Poephila acuticauda* (A) and *P. hecki* (H)) and the outgroup, a black-throated finch (*Poephila cincta* (C))



Population Tree

Gene Trees (n = 28)

n = 16

n = 7

n = 5

6.3 Incomplete lineage sorting

- In the context of coalescence, we can also develop an equation that allows us to predict the extent of Incomplete Lineage Sorting (ILS) given population genetic parameters
- We will assume that two sister populations, A and B, split t_1 generations in the past
- A deeper split from an outgroup population C occurred t_2 generations in the past
- No gene flow occurs between populations after they split
- Therefore, the first opportunity for alleles from the A and B populations to coalesce is t_1 generations ago
- As long as alleles from A and B coalesce before t_2, their gene trees will match the population tree

6.3 Incomplete lineage sorting

- Deviations from the population tree can occur when A and B alleles fail to coalesce between t_1 and t_2
- We can write this probability in our coalescent notation as $\left(1 - \frac{1}{2N}\right)^{t_2 - t_1}$
- Additionally, once alleles from A and B are in their ancestral population with population C without coalescing, if lineages from one of these species first coalesces with C we will see ILS where population and gene trees are discordant
- A discordant tree will happen at this stage with probability $\frac{2}{3}$ because 2 of the three pairs possible (A & C and B & C) are discordant

6.3 Incomplete lineage sorting

- With this information we can write the full probability for discordant trees:

$$\frac{2}{3}\left(1 - {}^{1}\!/_{2N}\right)^{t_2 - t_1} . \qquad (6.10)$$

- This equation tells us that discordant trees (ILS) will be more likely when:

1. Population sizes are large

2. The time between splits is small

6.3 Incomplete lineage sorting

- Many empirical tests have been developed to assess the prevalence of gene flow between two populations

- For example, think back to how much was learned from the fact that humans and Neanderthals experienced gene flow in their histories

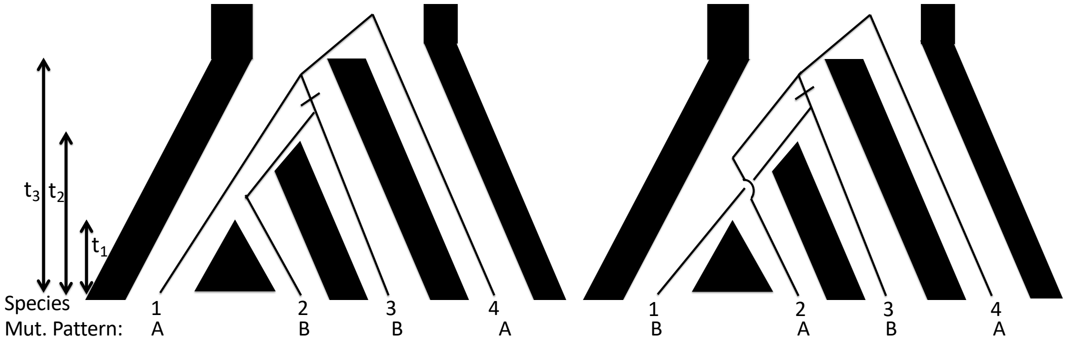
- Coop describes one of these methods that infers gene flow between populations based on the level of gene tree discordance with the population tree; this is different than the discordance due to ILS which is cause by rapid splitting of populations in their history

6.3 Incomplete lineage sorting

- In the example to the right, Species 1 and 2 are more closely related to each other than they are to species 3

- We see in this figure two discordant gene trees from the population tree in which species 2 first coalesces with species 3 (left) and species 1 first coalesces with species 3 (right)

- Under ILS, both of these discordant trees will happen with equal probability

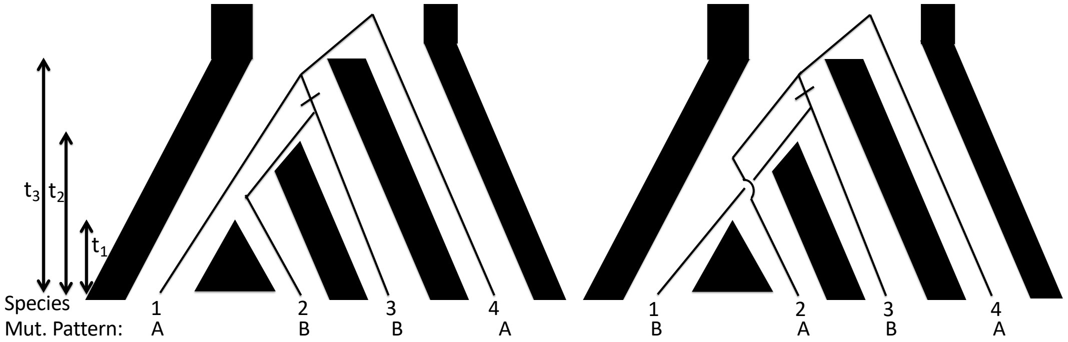


6.3 Incomplete lineage sorting

- Gene flow from species 3 into species 1 or species 2 will also result in these discordant patterns

- However, if gene flow occurs primarily from species 3 into species 1, we will observe more of the “BABA” pattern (right) than the “ABBA” pattern (left)

- When the BABA and ABBA patterns are uneven, this suggests gene flow rather than ILS

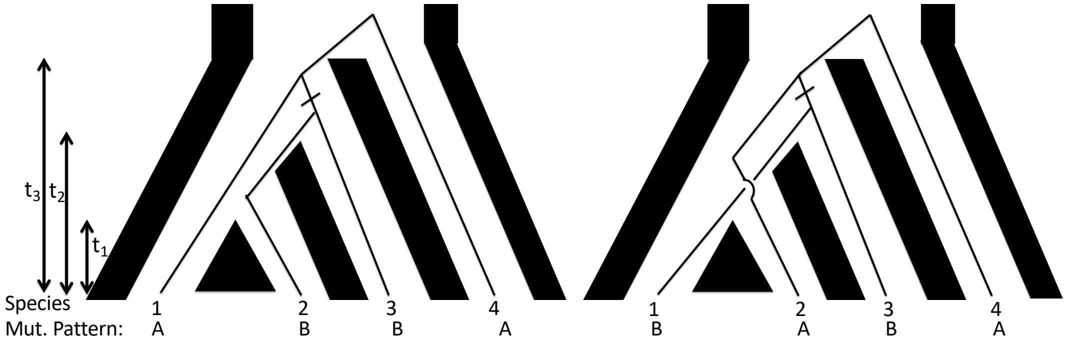


6.3 Incomplete lineage sorting

- To test for gene flow between species 3 and either of species 1 or 2, we can generate sequence data from across these genomes and determine whether ABBA and BABA patterns are even or skewed:

$$\frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}} \qquad (6.11)$$

- If discordance is due to ILS, the statistic will be zero, when gene flow is from species 3 to species 1 it will be negative, when gene flow is from species 3 to species 2, it will be positive


t_3
t_2
t_1
Species
Mut. Pattern:
1
A
2
B
3
B
4
A
1
B
2
A
3
B
4
A