Lecture 8: Population Structure and Correlations Among Loci (Ch3&6)

Population genetic PCB4553/6685

- Within a species, there's often geographically-restricted mating among individuals.
- Individuals tend to mate with individuals from the same, or closely related sets of populations.
- This form of non-random mating is called **population structure** and can have profound effects on the distribution of genetic variation within and among natural populations.

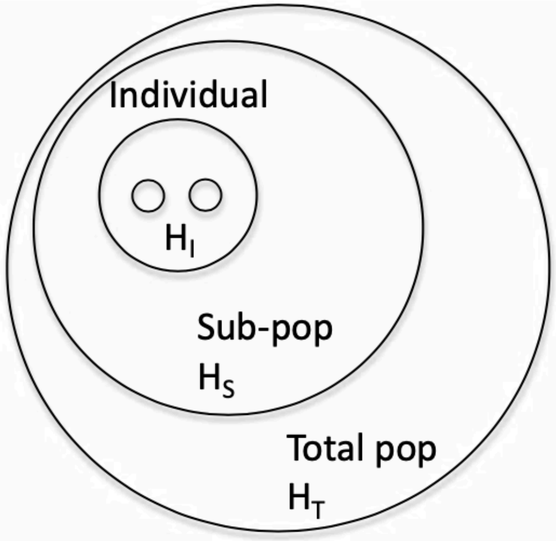
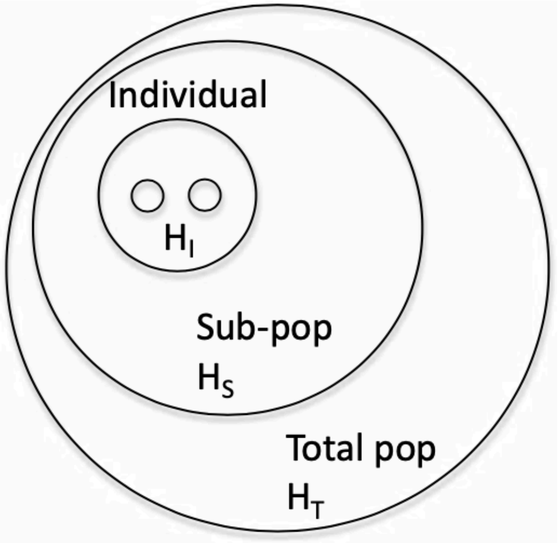


Figure 3.1: The hierarchical nature of F-statistics. The two dots within an individual represent the two alleles at a locus for an individual I. We can compare the heterozygosity in individuals (H_I), to that found by randomly drawing alleles from the sub-population (S), to that found in the total population (T).

Review: inbreeding coefficient, structure within a population

- F: probability two randomly chosen alleles (gametes) are IBD
- only way the offspring can be heterozygous (A1A2) is if their two alleles at a locus are not IBD
- Fraction heterozygous individuals:
  - $P(A_1A_2) = 2pq(1-F) = f_{12} = H_I$
- Expected heterozygosity under random mating is
- $H_S = 2p_S(1-p_S) = 2p_Sq_S$
- $$F_{\mathrm{IS}} = 1 - \frac{H_I}{H_S} = 1 - \frac{f_{12}}{2p_Sq_S}$$
- Interpretations of F_{IS} :
  - relative difference between observed and expected heterozygosity due to a deviation from random mating within the subpopulation.
  - Correlation between gametes within a population relative to a random mating model



Wright's F-statistics

- Wright's definition: *correlations* between random gametes, drawn from the same level X, relative to level Y

- Comparing observed heterozygosity in individuals (H_I) to that expected in the total population, H_T

- $F_{IT} = 1 - \frac{H_I}{H_T} = \frac{H_T - H_I}{H_T} = 1 - \frac{f_{12}}{2p_T q_T}$

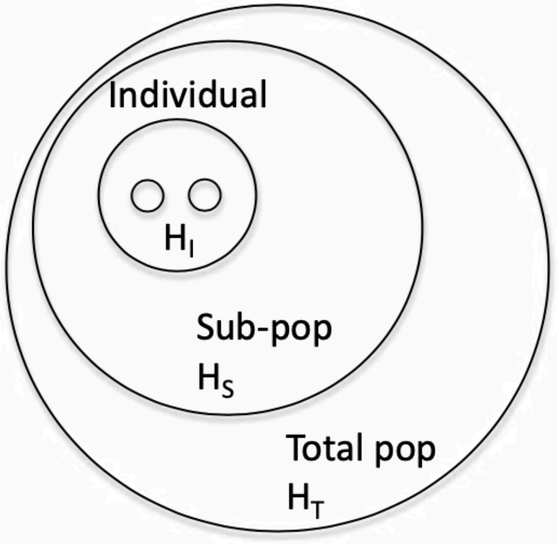
- Comparing the expected heterozygosity in the subpopulation (H_S) to that expected in the total population H_T

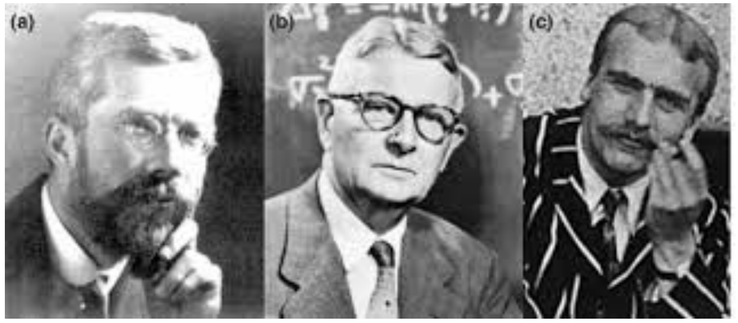
- $F_{ST} = 1 - \frac{H_S}{H_T} = \frac{H_T - H_S}{H_T} = 1 - \frac{2p_S q_S}{2p_T q_T}$

- Relating all three F-statistics

- $(1 - F_{\mathrm{IT}}) = \frac{H_I}{H_S}\frac{H_S}{H_T} = (1 - F_{\mathrm{IS}})(1 - F_{\mathrm{ST}})$

- $H_I = H_T \times (1 - F_{IS}) \times (1 - F_{ST})$





Ronald Fisher **Sewall Wright** JBS Haldane

Multiple subpopulations and multiple sites

- For multiple subpopulations,
- $\overline{F}_{ST} = 1 - \frac{\overline{H}_S}{H_T}$
- $\overline{H}_S = 1/k \sum_{i=1}^{k} H_S^{(i)}$
- $H_S^{(i)} = 2p_i q_i$
- For multiple sites, replace H_I, H_S, and H_T with their averages across loci

Wahlund effect

- When there are multiple sub-pops, average expected heterozygosity always smaller than total expected heterozygosity, i.e.

- $\overline{H}_S \leq H_T$

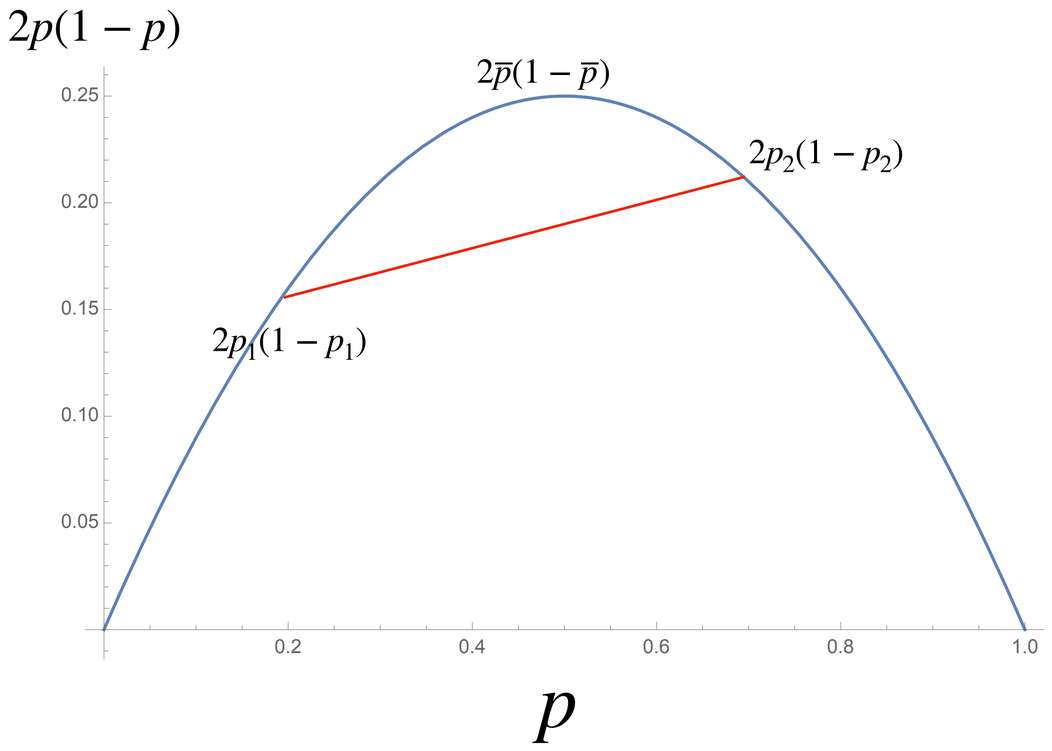
- So there is always a deficiency of heterozygotes in sub-pops

- Can be shown using *concavity* of the quadratic function

- $\overline{F}_{ST} \geq 0$

  - $\overline{F}_{ST} = 1 - \frac{\overline{H}_S}{H_T}$

- $\overline{F}_{IS} \leq \overline{F}_{IT}$



Question 1.

In a species of lemurs, you estimate the allele frequency to be 20%. In a particular population, you estimate that the allele frequency is 10%. In this population, only 9% of individuals are heterozygote. What is F_{IT}, F_{ST}, and F_{IS} for this population?

Interpretations of *F*-statistics

- *F*-statistics according to Wright: **correlations** between random gametes, drawn from the same level X, relative to level Y

- F_{IS}: X is individuals and Y is subpop

- $$F_{\mathrm{IS}} = \frac{2p_S q_S - f_{12}}{2p_S q_S} = \frac{f_{11} + f_{22} - p_S^2 - q_S^2}{2p_S q_S}$$

- Correlation between alleles drawn from a population (or an individual) above that expected by chance

- Same results can be derived for F_{ST}

Interpretations of *F*-statistics

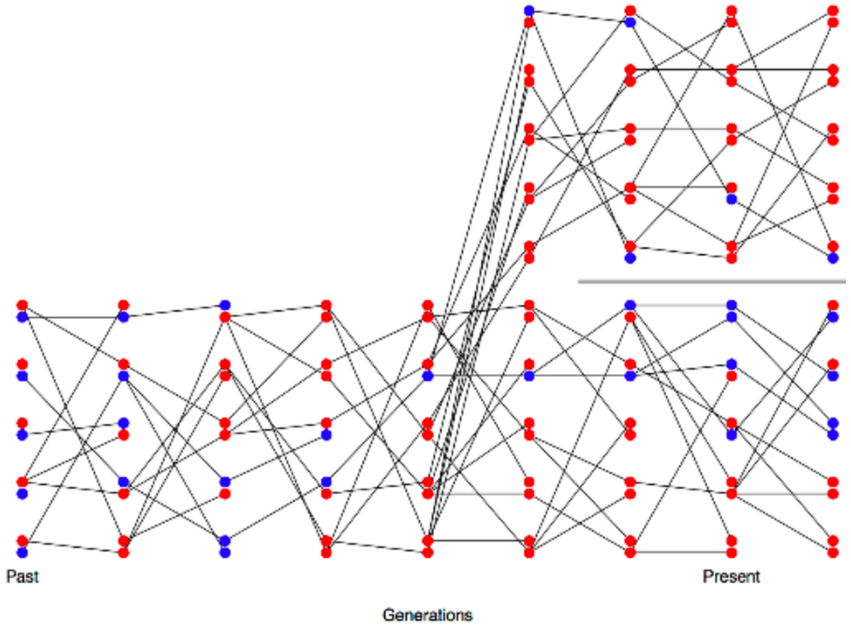
- We can also interpret *F*-statistics as **proportions of variance** explained by different levels of population structure. To see this

$$F_{\mathrm{ST}} = \frac{2\bar{p}\bar{q} - \frac{1}{K}\sum_{i=1}^{K} 2p_i q_i}{2\bar{p}\bar{q}} = \frac{\left(\frac{1}{K}\sum_{i=1}^{K} p_i^2 + \frac{1}{K}\sum_{i=1}^{K} q_i^2\right) - \bar{p}^2 - \bar{q}^2}{2\bar{p}\bar{q}}$$
$$= \frac{\mathrm{Var}(p_1, \ldots, p_K)}{\mathrm{Var}(\bar{p})}, \qquad (3.7)$$

- This means we can test the significance level F_{ST} using ANOVA!

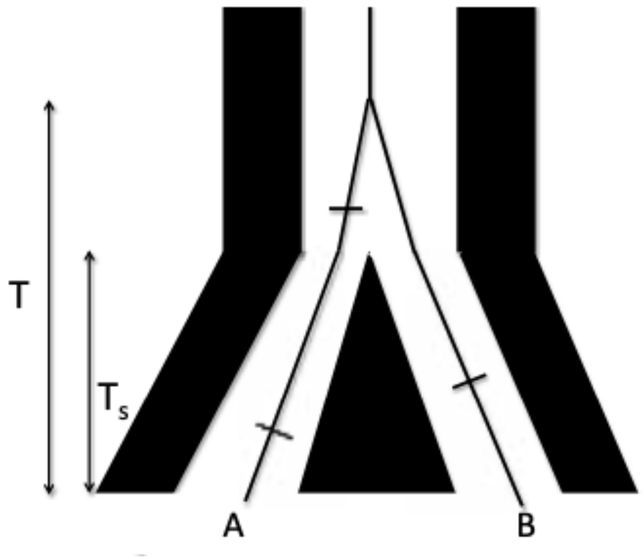
Relating F_{ST} to population history - population split

- Ancestral population size N_e
- at time T, split into to populations of size N_e
- Expected Heterozygosity within a descendent sub-pop
- $$H_S = \frac{4N_e\mu}{1 + 4N_e\mu} \approx 4N_e\mu$$
- assuming $N_e\mu \ll 1$
- Total heterozygosity
- $H_T = 1/2H_S + 1/2H_B$
- H_B: probability that a pair of alleles drawn from our two different sub-pops differ from each other.



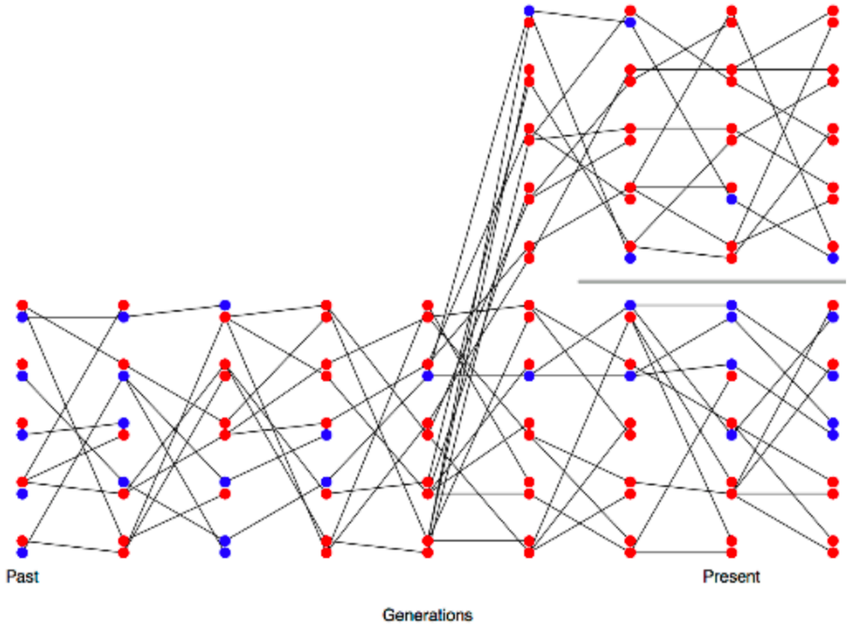
What is H_B?

- T_s: time since the two species split from the common ancestor
- Total time to coalescent: $T = T_s + 2N_A$



Relating F_{ST} to population history - population split

- $H_B \approx 2\mu(T + 2N)$
- $H_S \approx 4N\mu$
- $F_{ST} = 1 - \frac{H_S}{H_T} = 1 - \frac{4N\mu}{1/2H_S + 1/2H_B}$
- $F_{ST} = 1 - \frac{4N\mu}{2N\mu + \mu(T + 2N)} = 1 - \frac{4N}{4N + T} = \frac{T}{4N + T}$



Question 1.

The genome-wide F_{ST} between Bornean and Sumatran orangutan species samples (*Pongo pygmaeus* and *Pongo abelii*) is ≈ 0.37 (Locke *et al.*, 2011), representing a deep population split between the species (potentially with little subsequent gene flow). Within the populations the genome-wide average Watterson's θ is $\theta_W = 1.4\text{kb}^{-1}$, estimated from the number of segregating sites. Assume a generation time of 20 years, and a mutation rate of 2×10^{-8} per base per generation. How far in the past did the two populations diverge?

Relating F_{ST} to population history - island model

- Mainland pop: expected heterozygosity = H_M
- Island pop size = N_I , very small compared to mainland pop.
- Each generation some low fraction m of island inds have migrant parents from the mainland the generation before.
- Immigrants back to the mainland negligible

Relating F_{ST} to population history - island model

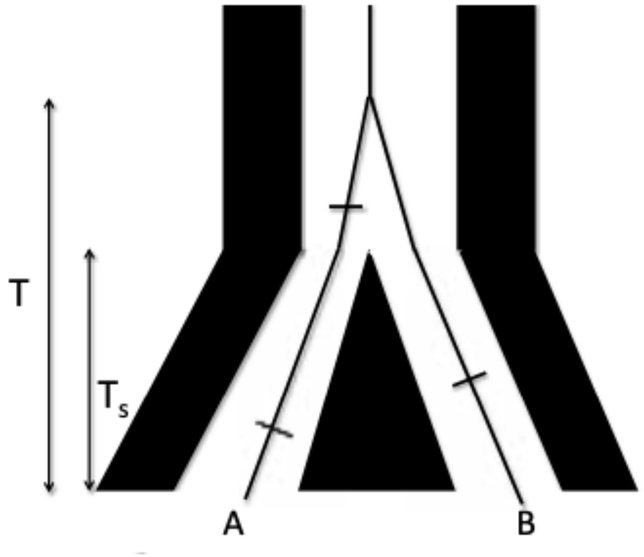
- probability that our lineages coalesce before either allele migrates back to mainland:
  - $\frac{1/2N_I}{1/2N_I + 2m}$
- Level of heterozygosity on the island is:
  - $$H_I = \left(1 - \frac{1/(2N_I)}{1/(2N_I) + 2m}\right) H_M$$
- Reduction of heterozygosity on the island
  - $$F_{IM} = 1 - \frac{H_I}{H_M} = \frac{1/(2N_I)}{1/(2N_I) + 2m} = \frac{1}{1 + 4N_I m}.$$
- considering the island as our sub-pop, $F_{IM} = F_{ST}$

Question 2.

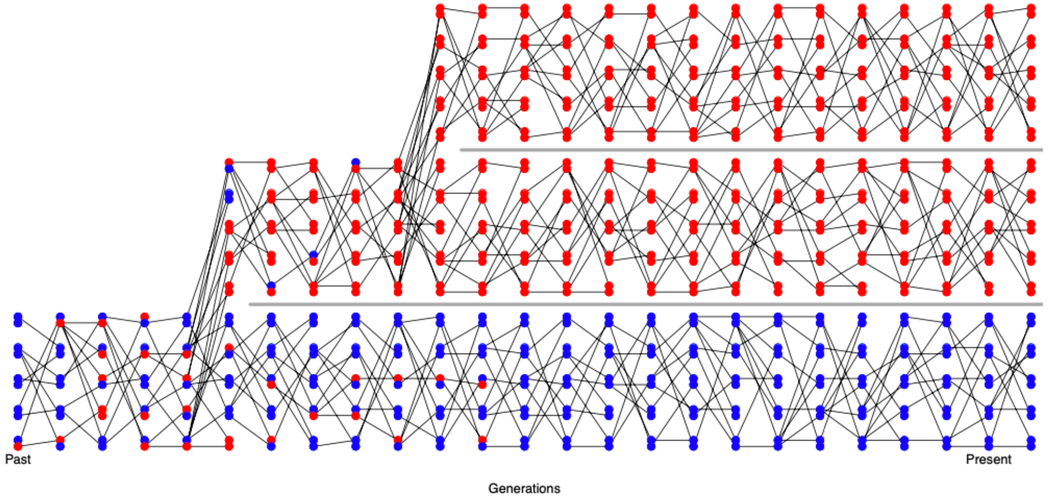
You are investigating a small river population of sticklebacks, which receives infrequent migrants from a very large marine population. At a set of putatively neutral biallelic markers the freshwater population has frequencies:
0.2, 0.7, 0.8
at the same markers the marine population has frequencies:
0.4, 0.5 and 0.7.
From studying patterns of heterozygosity at a large collection of markers, you have estimated the long term effective size of your freshwater population is 2000 individuals.
What is your estimate of the migration rate from the marine populations into the river?

Review: contribution of ancestral polymorphism to divergence

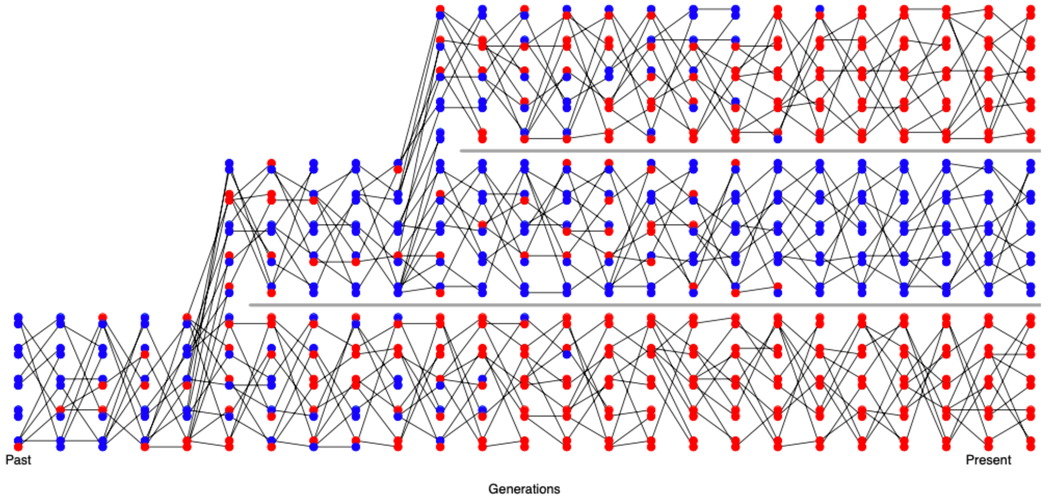
- T_s: time since the two species split from the common ancestor
- $T = T_s + 2N_A$
- If recent split, then ancestral polymorphisms contribute to divergence



Incomplete lineage sorting

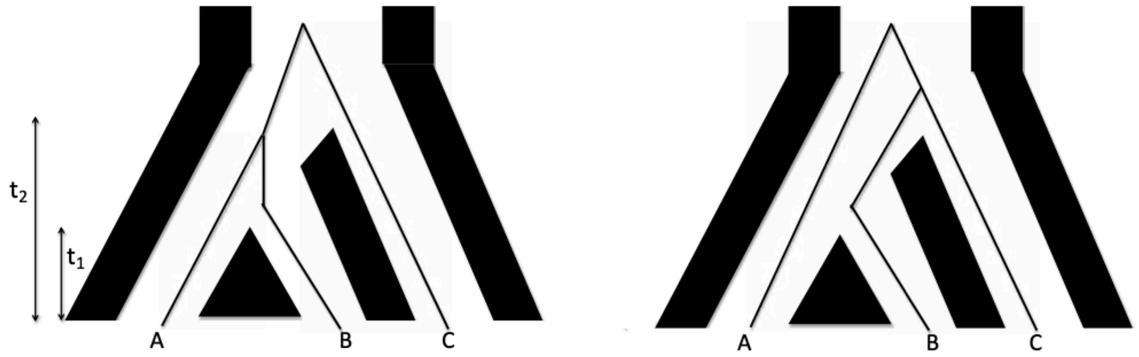


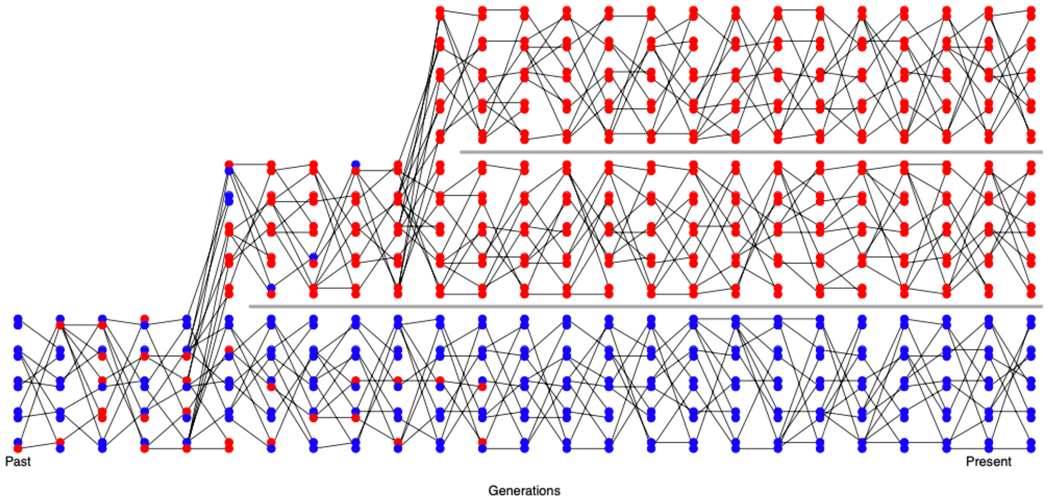
- Incomplete lineage sorting (ILS) occurs when ancestral polymorphisms are preserved after species split

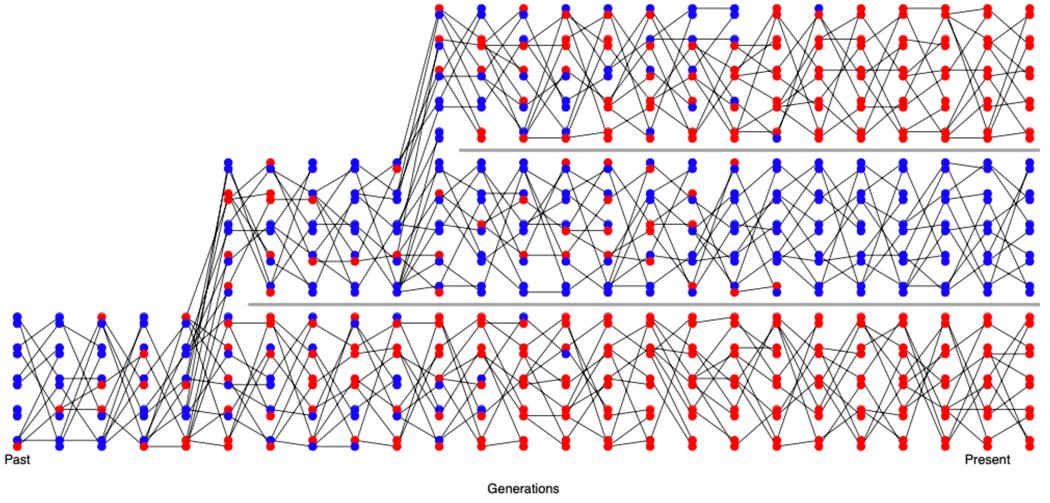


Incomplete lineage sorting

- Ancestral polymorphism could lead to discordance between species tree and gene tree

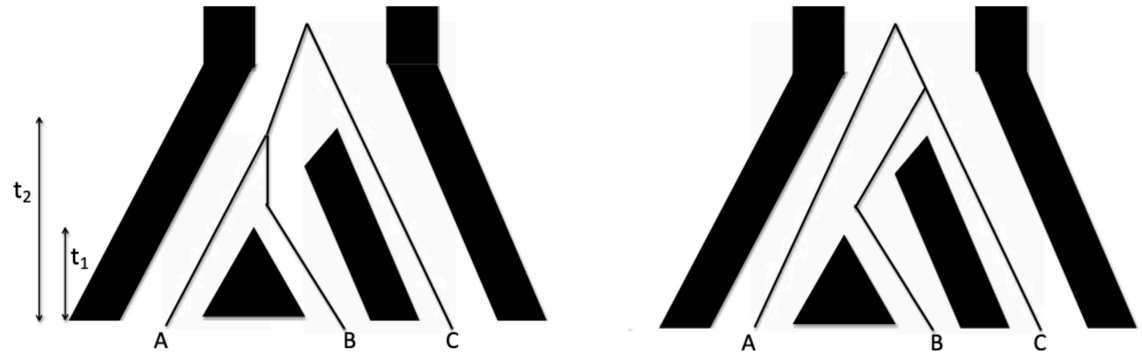






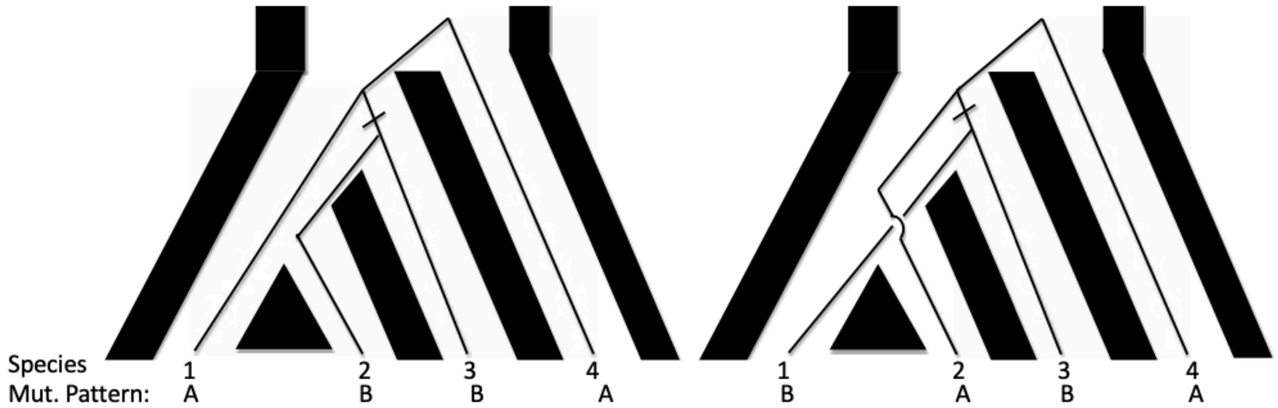
How often does species tree and gene tree disagree?

- Probability species A and B not coalescing before the split of the MRCA of A&B and C
  - $(1 - 1/2N)^{t_2 - t_1}$
- Probability A&C or B&C coalesce before A&B coalesce (discordance)
  - $2/3$
- Answer: $\frac{2}{3}(1 - 1/2N)^{t_2 - t_1}$



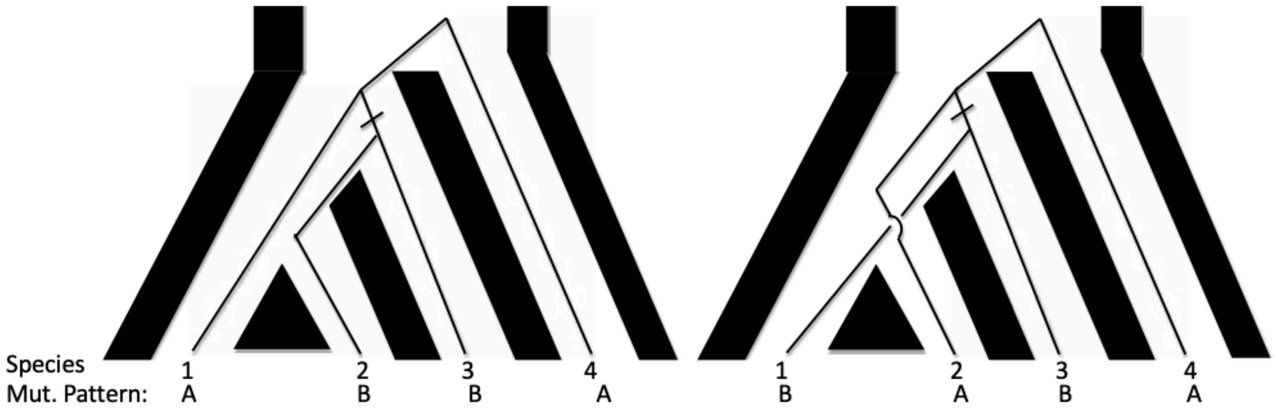
Using incomplete lineage sorting to test for gene flow

- Set up:
  - Four species (namely, 1,2,3,4)
  - Two sister species (1, 2)
  - One outgroup (4)
- Ancestral state: A
- Derived state: B, *present in two spps*
- Two possible patterns:
  - ABBA
  - BABA



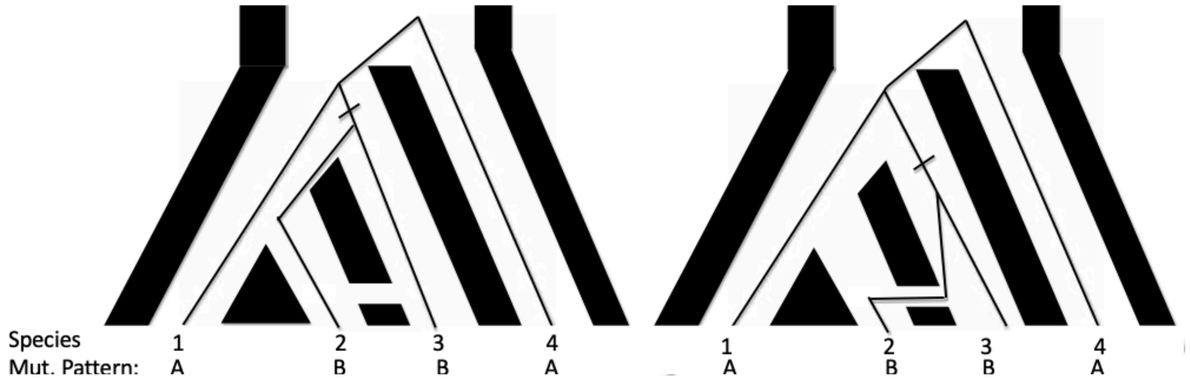
Using incomplete lineage sorting to test for gene flow

- Null hypothesis: there is no gene flow between species 1&3 or 2&3
- Discordance between species tree and gene tree should be symmetric if there is only ILS
- $n_{ABBA} = n_{BABA}$



Using incomplete lineage sorting to test for gene flow

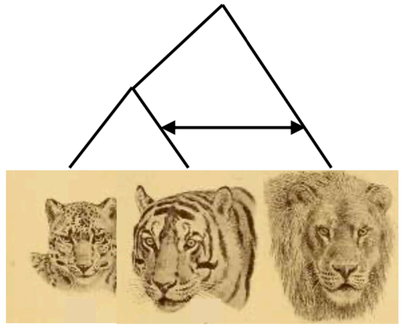
- Symmetry is broken if there is gene flow
- $n_{ABBA} > n_{BABA}$, if gene flow between spp 2&3
- $n_{ABBA} < n_{BABA}$, if gene flow between spp 1&3



Example

- ABBA-BABA statistic = 0.07±0.0026
- Indicating gene flow between lions and tigers
- Standard error can be calculated by resampling

Snow leopard	Tiger	Lion	Domestic cat	Counts
A	B	B	A	1,434,106
B	A	B	A	1,250,134



Figueiró et al. (2017)

Other approaches to population structure - Assignment Methods

- Aim: Finding the probability that an individual of unknown population comes from one of K **predefined** populations.

- Example:

  - There are three broad pops of common chimps in Africa: western, central, and eastern.

  - Imagine that we have a chimps whose origin is unknown. If we have genotyped a set of unlinked markers from a panel of inds representative of these populations,

  - Calculate the probability that our chimp comes from each of these populations.

Basic principles

- New individual's genotype at locus l is g_l
- g_l : number of copies of allele A_1 this individual carries at this locus (g_l = 0, 1, 2).
- P of genotype at locus l conditional on coming from population k,

$$\mathbb{P}(g_l|\text{pop k}) = \begin{cases} (1-p_{k,l})^2 & g_l = 0 \\ 2p_{k,l}(1-p_{k,l}) & g_l = 1 \\ p_{k,l}^2 & g_l = 2 \end{cases}$$

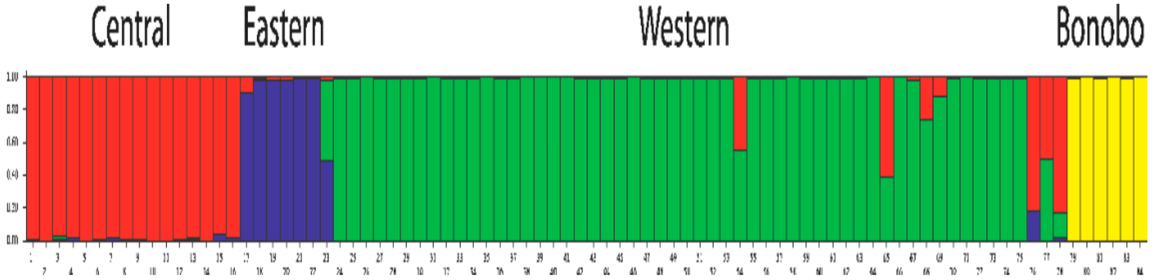
-

Basic principles

- Assuming loci are independent
- Prob of the individual's genotype across all S loci, conditional on the individual coming from population k, is
- $$\mathbb{P}(\text{ind.}|\text{pop k}) = \prod_{l=1}^{S} \mathbb{P}(g_l|\text{pop k})$$
- Probability ind is from population k
  - $\mathbb{P}(\text{pop k}|\text{ind.})$
- Apply *Baye's Rule*
  - $$\mathbb{P}(\text{pop k}|\text{ind.}) = \frac{\mathbb{P}(\text{ind.}|\text{pop k})\mathbb{P}(\text{pop k})}{\mathbb{P}(\text{ind.})}$$
  - $$\mathbb{P}(\text{ind.}) = \sum_{k=1}^{K} \mathbb{P}(\text{ind.}|\text{pop k})\mathbb{P}(\text{pop k})$$

Basic principles

- $$\mathbb{P}(\text{pop k}|\text{ind.}) = \frac{\mathbb{P}(\text{ind.}|\text{pop k})\mathbb{P}(\text{pop k})}{\mathbb{P}(\text{ind.})}$$
  - posterior probability that our new individual comes from each of our 1,⋯ ,K populations.
- We need to know the population definitions and allele frequencies *a priori*
- Full Bayesian approach:
  - Jointly infer allele frequencies and assignments
- Softwares: STRUCTURE

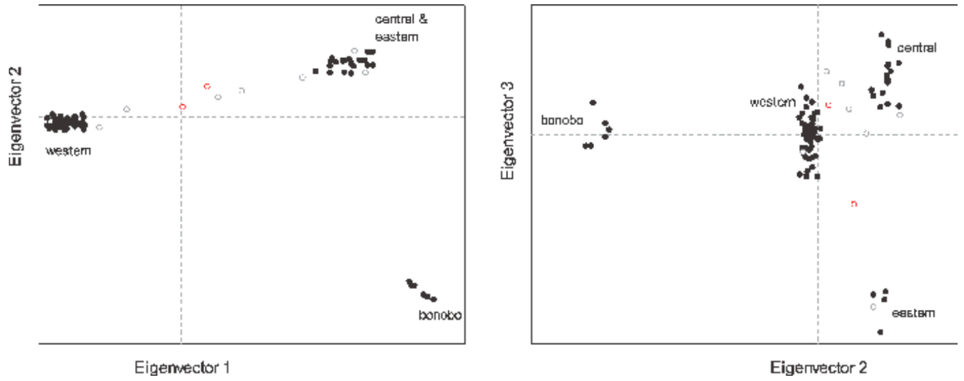


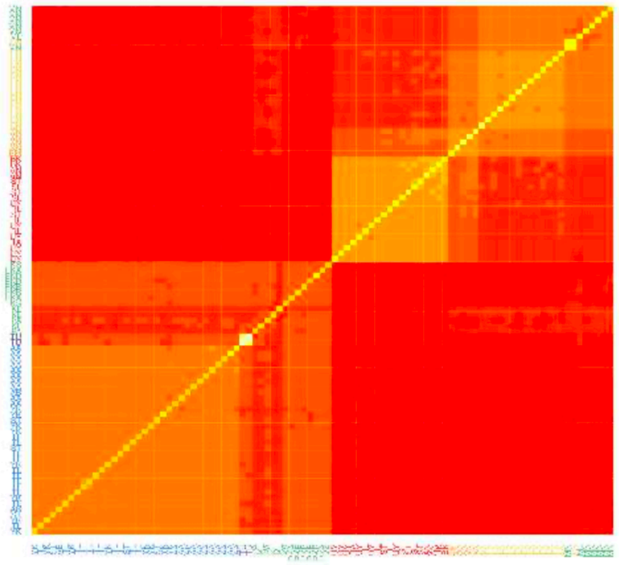
STRUCTURE plot for 78 common chimpanzee and 6 bonobo at over 300 polymorphic microsatellites markers. K = 4. Becquet et al. (2007)

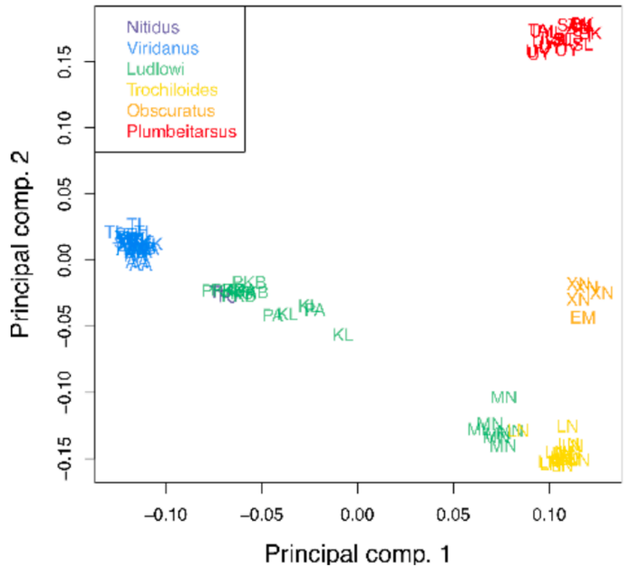
Pitfalls

- How to choose the K
  - taking the results of STRUCTURE-like approaches for some particular value of K and taking this to represent the best way to describe population-genetic variation.
- What do the “clusters” represent?
  - They are not ancestral populations

PCA


central &
eastern
western
bonobo
Eigenvector 2
Eigenvector 1
central
western
bonobo
eastern
Eigenvector 3
Eigenvector 2




Nitidus
Viridanus
Ludlowi
Trochiloides
Obscuratus
Plumbeitarsus
Principal comp. 2
Principal comp. 1
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.10
−0.05
0.00
0.05
0.10

Correlations between alleles at *different* loci

- F-statistics are about the correlation between alleles on the same locus

- Correlations between alleles at *different loci* can also occur under forces including drift, selection, population structure, and *recombination*

Recombination

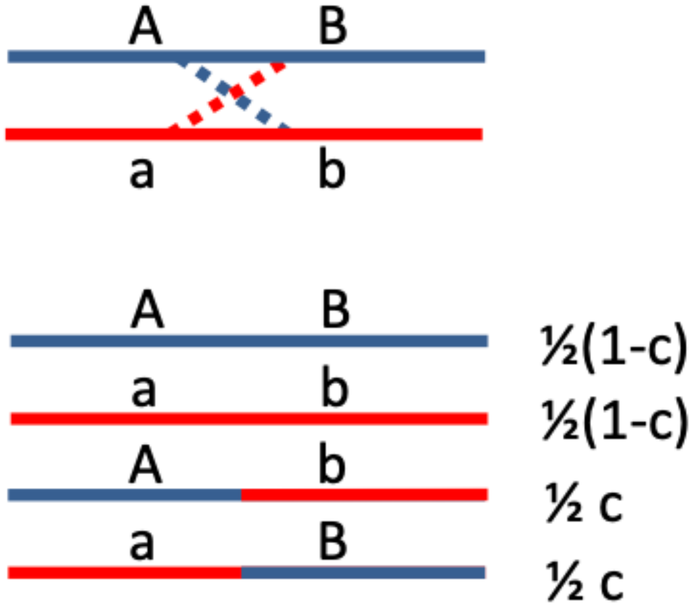
- Breaking the Law of Independent Assortment
- Expected to have 9:3:3:1 ratio for two dominant phenotypes
- Exceptions to this were quickly discovered since the rediscovery of Mendel's work
- Revised by Morgan

Bateson, Saunders, and Punnett experiment

Phenotype and genotype	Observed	Expected from 9:3:3:1 ratio
Purple, long (*P_L_*)	284	216
Purple, round (*P_ll*)	21	72
Red, long (*ppL_*)	21	72
Red, round (*ppll*)	55	24

Recombination

- Consider a heterozygous ind with haplotypes AB and ab
- If no recombination, gametes will be AB and ab
- c: probability of an *odd* number of crossing over events between our loci in a single meiosis
- $0 \leq c \leq 1/2$
- $c = 1/2$ for loci on different chromosomes



Linkage disequilibrium (LD)

- LD: *statistical* non-independence (correlation) of alleles in a *population* at different loci

- Genetic linkage refers to the linkage of multiple loci due to the fact that they are transmitted through meiosis together.

- LD refers to the covariance between the alleles at different loci;

  - May be due to the genetic linkage

  - but does not necessarily imply this (e.g. genetically unlinked loci can be in LD due to population structure).

Measuring LD

- $D_{AB} = p_{AB} - p_A p_B$
- For loci with two alleles, only need one parameter to describe LD (i.e. only 1 degrees of freedom to change the genotype frequencies)
- $p_{AB} = p_A p_B + D$

LD can rise due to population structure

Question 4.

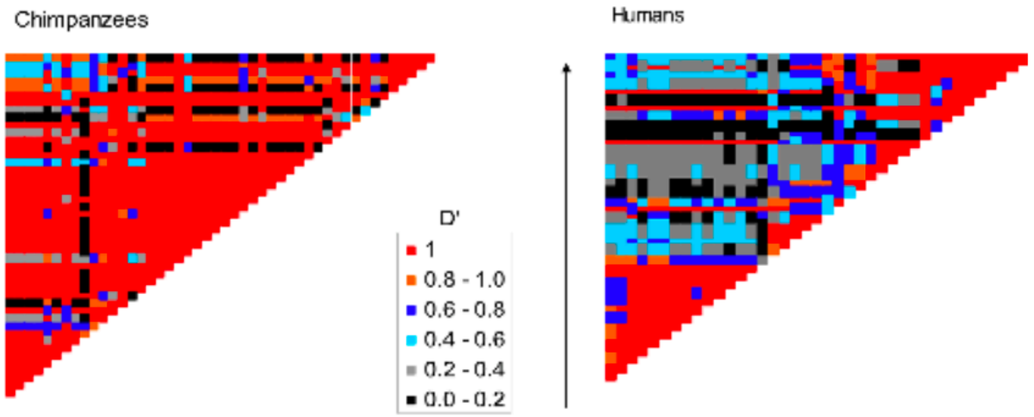
You genotype 2 bi-allelic loci (A & B) segregating in two mouse subspecies (1 & 2) which mate randomly among themselves, but have not historically interbreed since they speciated. The frequencies of haplotypes in each population are:

Pop	p_{AB}	p_{Ab}	p_{aB}	p_{ab}
1	.02	.18	.08	.72
2	.72	.18	.08	.02

A) How much LD is there within species? (i.e. estimate D)
B) If we mixed individuals from the two species together in equal proportions, we could form a new population with p_{AB} equal to the average frequency of p_{AB} across species 1 and 2. What value would D take in this new population before any mating has had the chance to occur?

Measuring LD

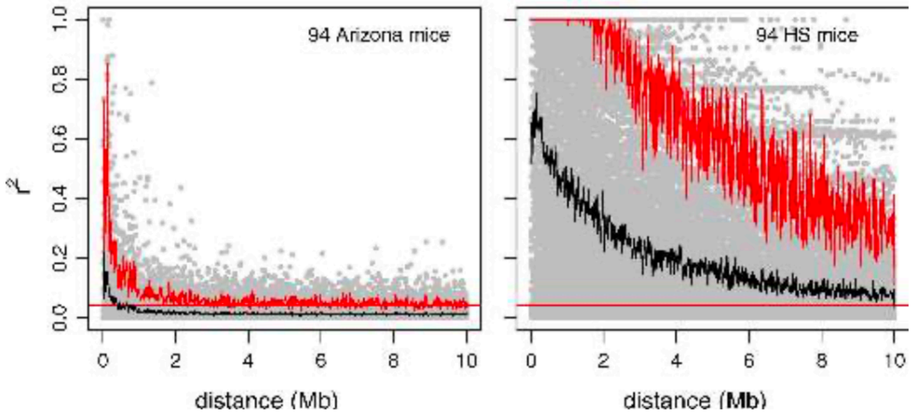
- We can remove the dependence of D only allele frequency by normalizing D with the max possible value
- $$D' = \frac{D}{D_{\max}}$$
- $$D_{\max} = \begin{cases} \max\{-p_A p_B, -(1-p_A)(1-p_B)\} & \text{when } D < 0 \\ \min\{p_A(1-p_B), (1-p_A)p_B\} & \text{when } D > 0 \end{cases}$$



LD across the TAP2 gene region in a sample of Humans and Chimps. The rows and columns are consecutive SNPs, with each cell giving the absolute D′ value between a pair of SNPs.
Ptak et al. (2004)

Measuring LD

- LD is the correlation between alleles at different loci
- $$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} = \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}$$
- $$= \frac{(\text{cov}(A,B))^2}{\text{var}(A)\text{var}(B)}$$
- r^2 is the squared correlation coefficient



The decay of LD for autosomal SNP in *Mus musculus domesticus*, as measured by r^2, in a wild-caught mouse population from Arizona and a set of advanced- generation crosses between inbred lines of lab mice. Each dot gives the r2 for a pair of SNPs a given physical distance apart, for a total of ~ 3000 SNPs. The solid black line gives the mean, the jagged red line the 95th percentile, and the flat red line a cutoff for significant LD. Laurie et al. (2007)

Decay of LD due to recombination

- Frequency of our AB haplotype in the next generation p'_{AB}

$$p'_{AB} = (1-c)p_{AB} + cp_A p_B$$

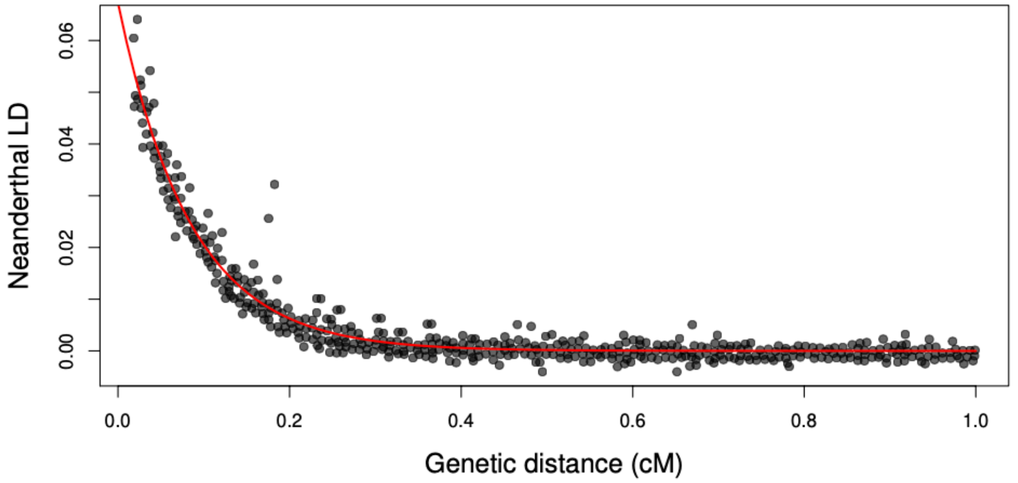
$$\Delta p_{AB} = p'_{AB} - p_{AB} = -cp_{AB} + cp_A p_B = -cD$$

$$\begin{aligned} D' &= p'_{AB} - p'_A p'_B \\ &= (p_{AB} + \Delta p_{AB}) - (p_A + \Delta p_A)(p_B + \Delta p_B) \\ &= p_{AB} + \Delta p_{AB} - p_A p_B \\ &= (1-c)D \end{aligned}$$

$$D_t = (1-c)^t D_0 \approx e^{-ct} D_0$$

Estimating interbreeding time between human and neantherthals

- Neanderthals and modern humans diverged from each other ~.5 MYA, allowing time for allele frequency differences to accumulate between them
- Interbreeding occurred sometime after human ancestors migrated out of Africa
- How to date this time when the two human species met and interbred?



Sankararaman et al..

Estimating interbreeding time between human and neantherthals

- $D_t = (1-c)^t D_0 \approx e^{-ct} D_0$
- Use c estimated in contemporary humans and solve for t
- What is D_0?
- $t \approx 1200$ generations or $35{,}000$ years!