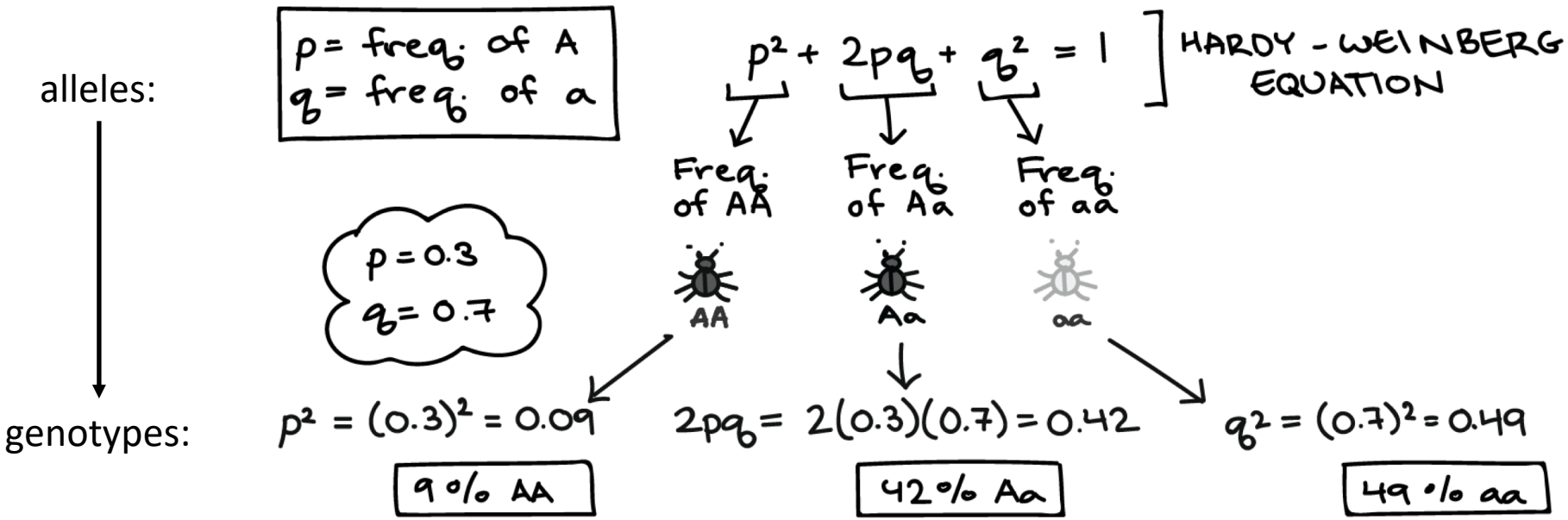
Population Genomics: Structure, Gene Flow, and Selection Part 2

Where does Evolution Begin?



Khan Academy

When a population is in Hardy Weinberg ***equilibrium***, allele frequencies will not change between generations.

Evolution is when the Hardy Weinberg equilibrium is violated and there is a ***change in allele frequencies***.

Where does Evolution Begin?

There are five mechanisms of evolution that can change allele frequencies.

1. Mutation (the generation of new alleles)

2. Non-random mating

3. Gene flow

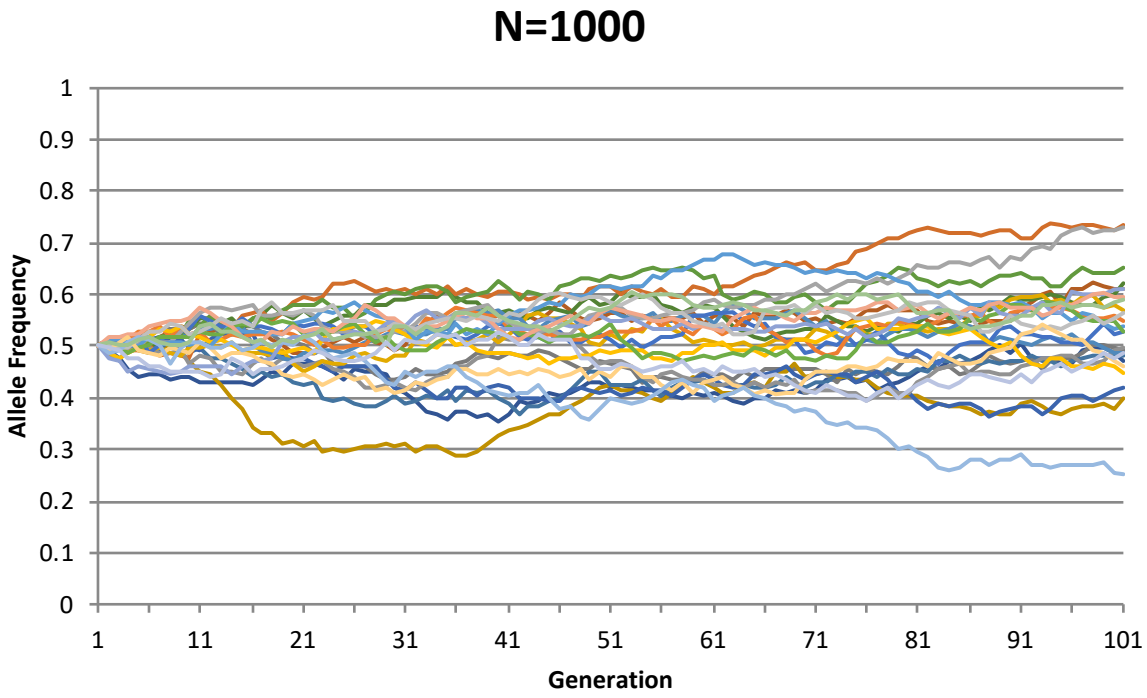
4. Genetic drift

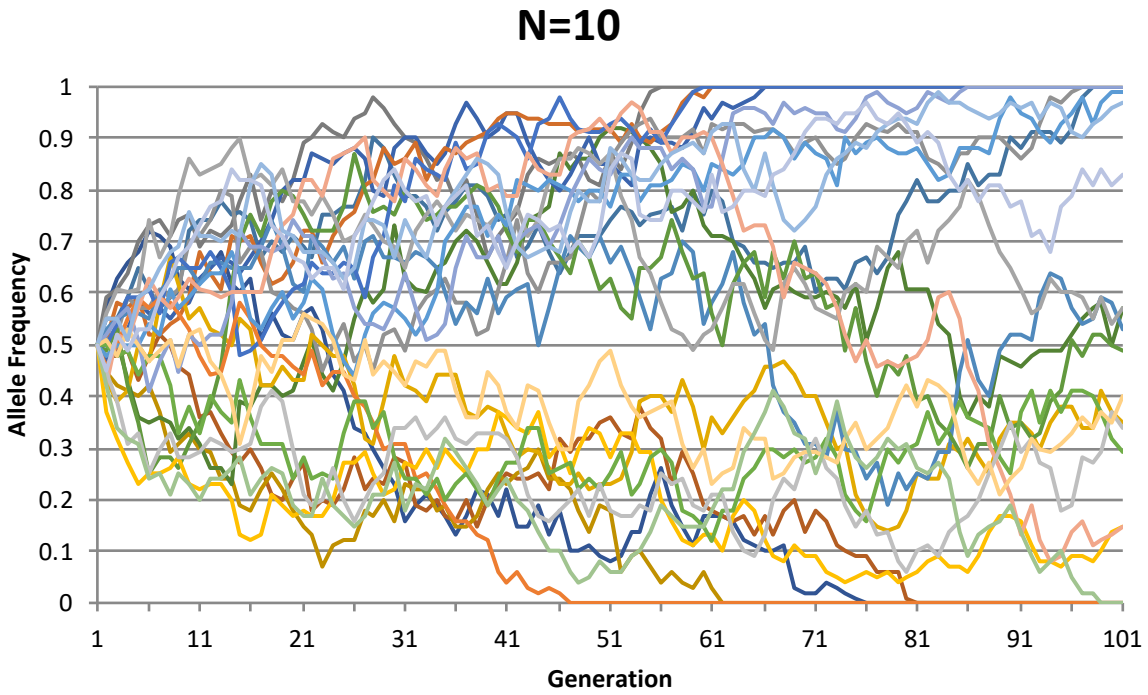
5. Natural selection

"Evolution begins as one mutation, on one chromosome, in one individual."

Matthew Hahn, 2019.
First line of Chapter 1, *Molecular Population Genetics*

Drift (and Selection) Causes Allele Frequencies to Differ Between Populations



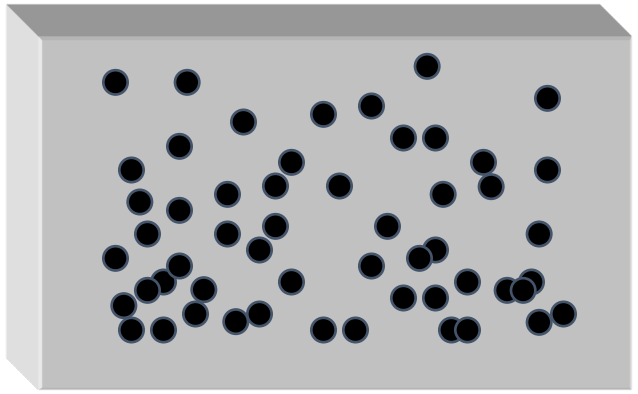


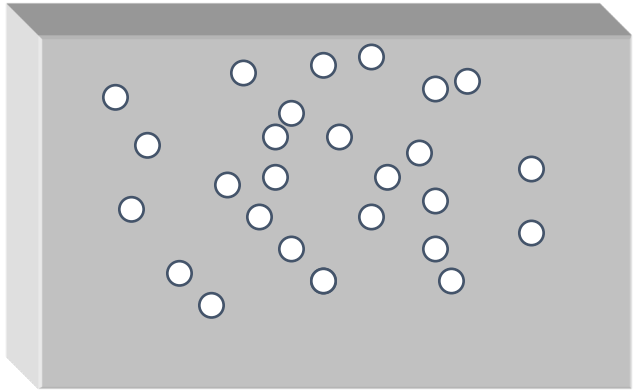
- Allele frequencies fluctuate in populations of small N_e
- In this model, the allele is selectively neutral

- The relationship between population size and allele frequency is known as the ***mutation-drift balance***.

Migration is a homogenizing force

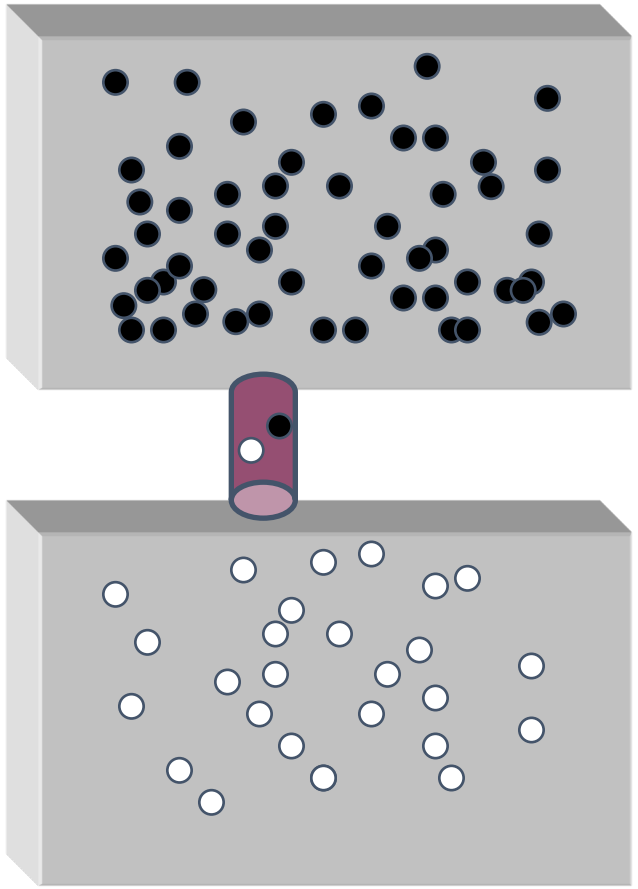
Differentiation is inversely proportional to gene flow





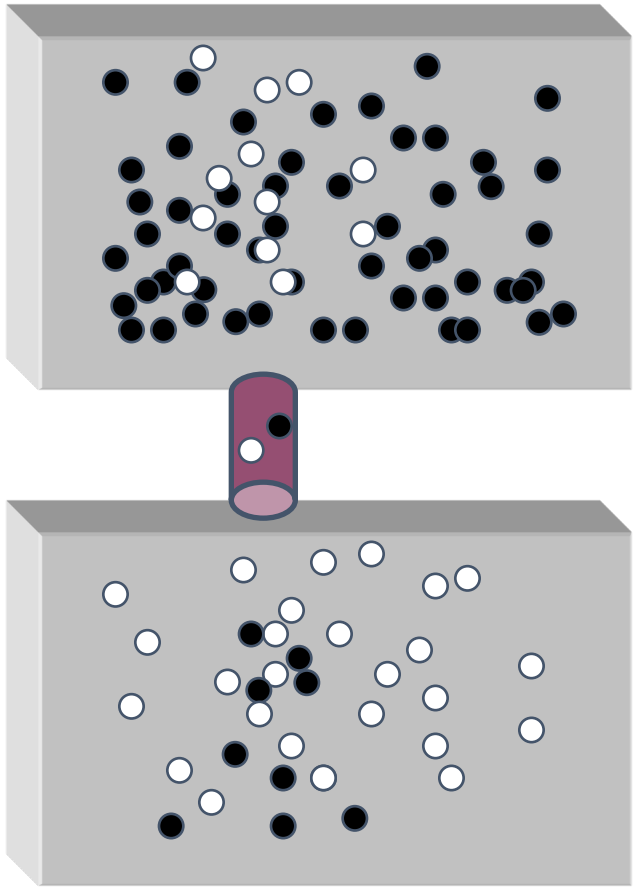
Migration is a homogenizing force

Differentiation is inversely proportional to gene flow



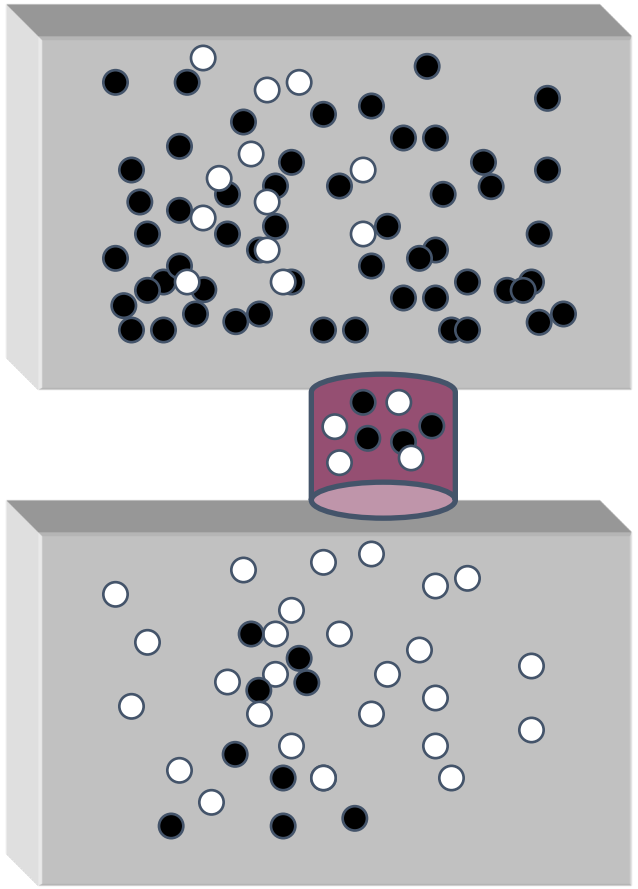
Migration is a homogenizing force

Differentiation is inversely proportional to gene flow



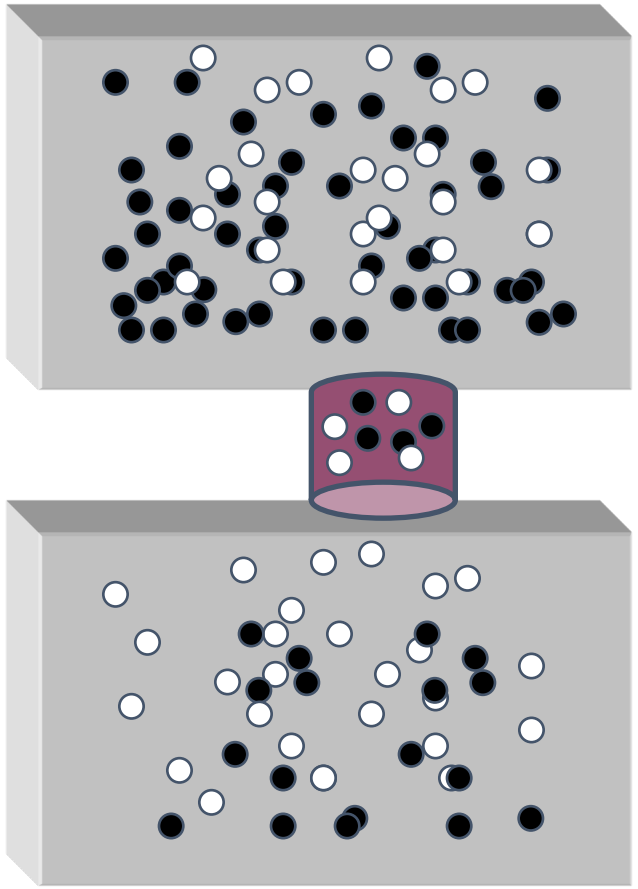
Migration is a homogenizing force

Differentiation is inversely proportional to gene flow



Migration is a homogenizing force

Differentiation is inversely proportional to gene flow



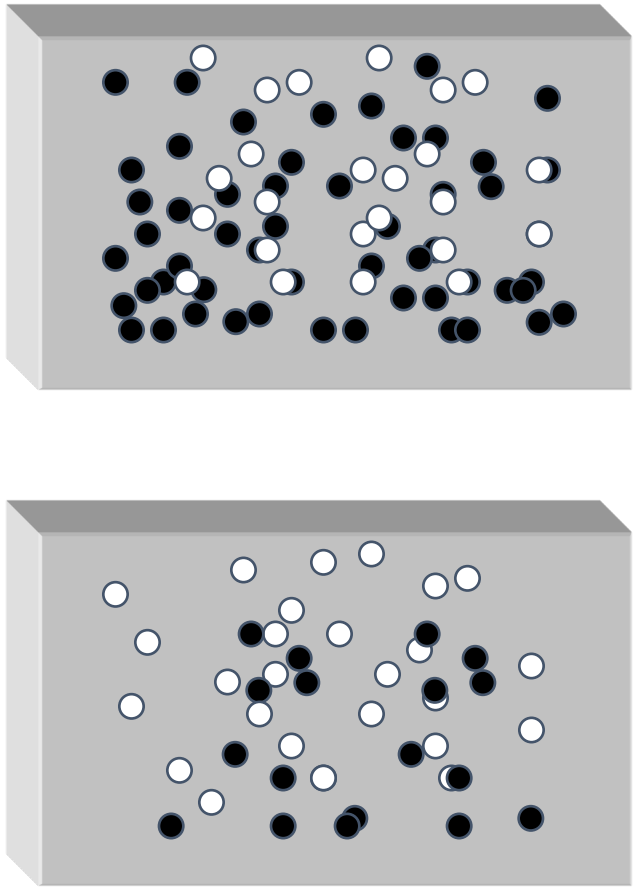
Migration is a homogenizing force

Differentiation is inversely proportional to gene flow

We can use the differentiation of the populations to estimate historical gene flow

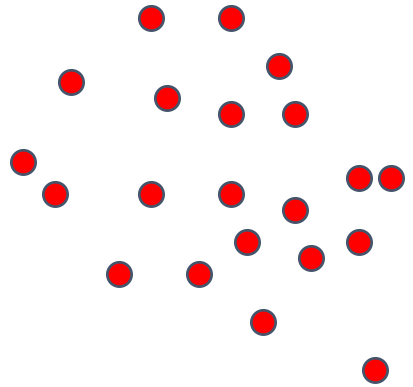
Gene flow is an important determinant of effective population size

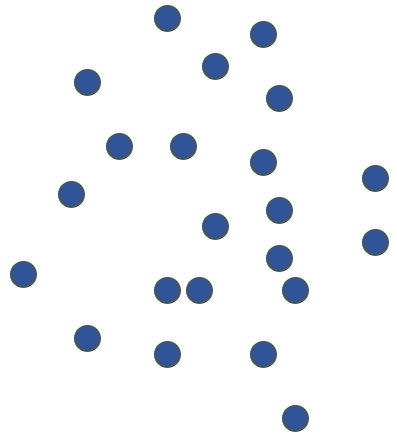
Estimation of gene flow is important for ecology, evolution, conservation, and forensics



Walhund Effect

Expected heterozygosity is *2pq*

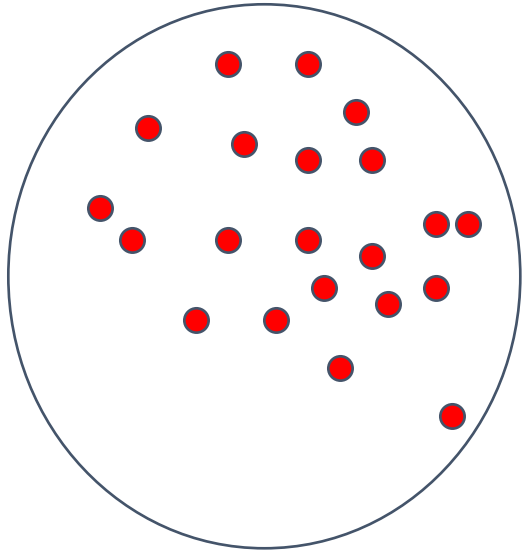


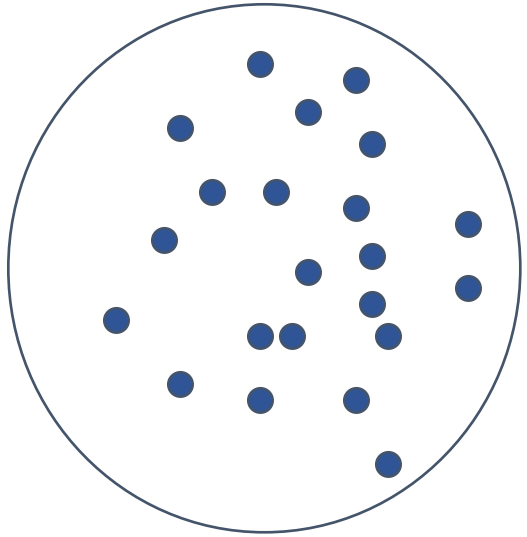


Walhund Effect

Expected heterozygosity is *2pq*

- *Separate populations:*
  - $H_E = 2pq = 2(1)(0) = 0$

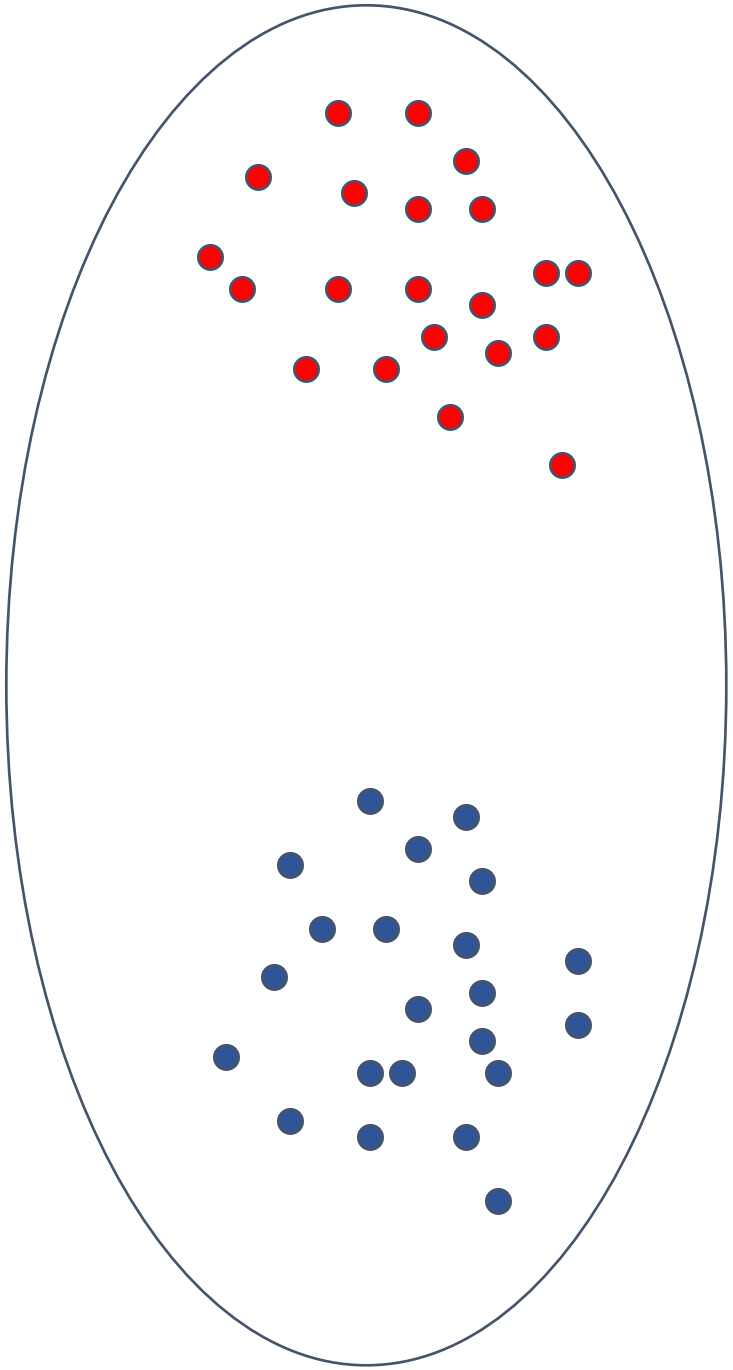




Walhund Effect

Expected heterozygosity is *2pq*

- *Separate populations:*
  - $H_E = 2pq = 2(1)(0) = 0$
- *Merged populations:*
  - $H_E = 2pq = 2(0.5)(0.5) = 0.5$

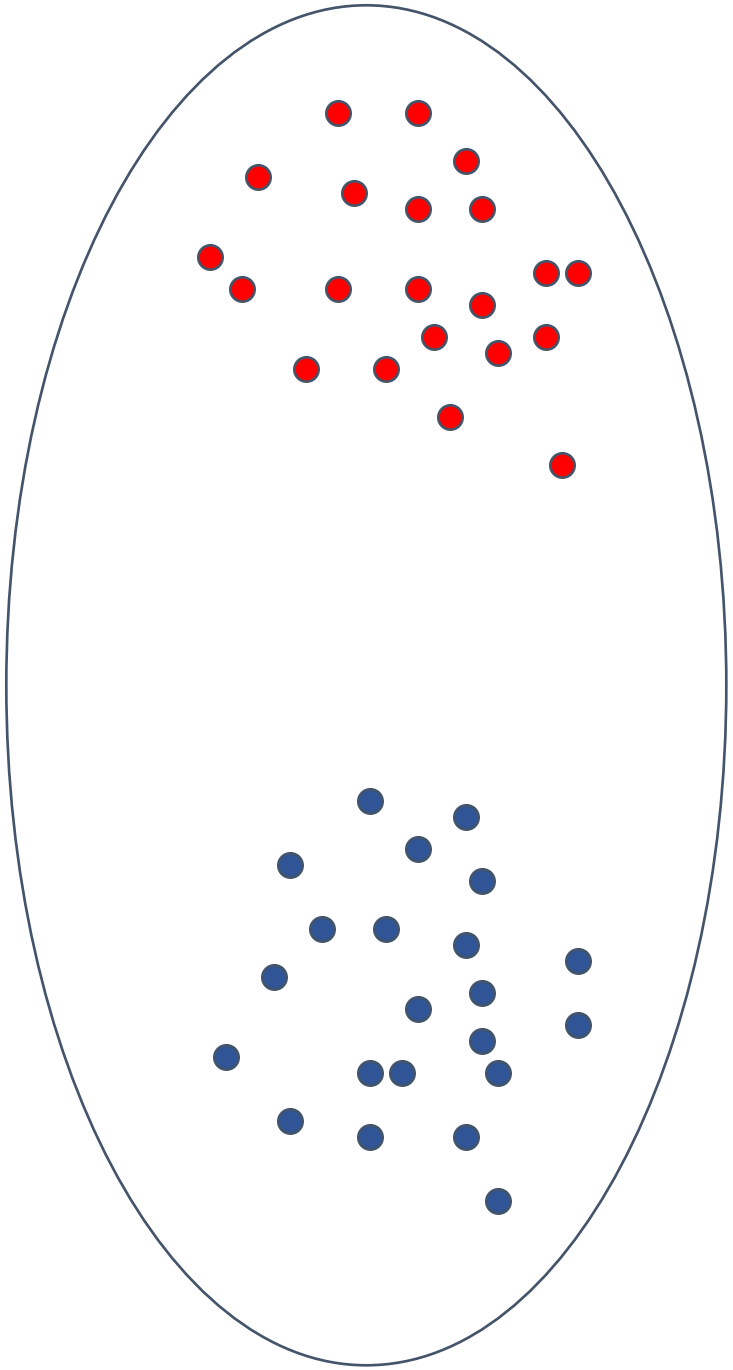


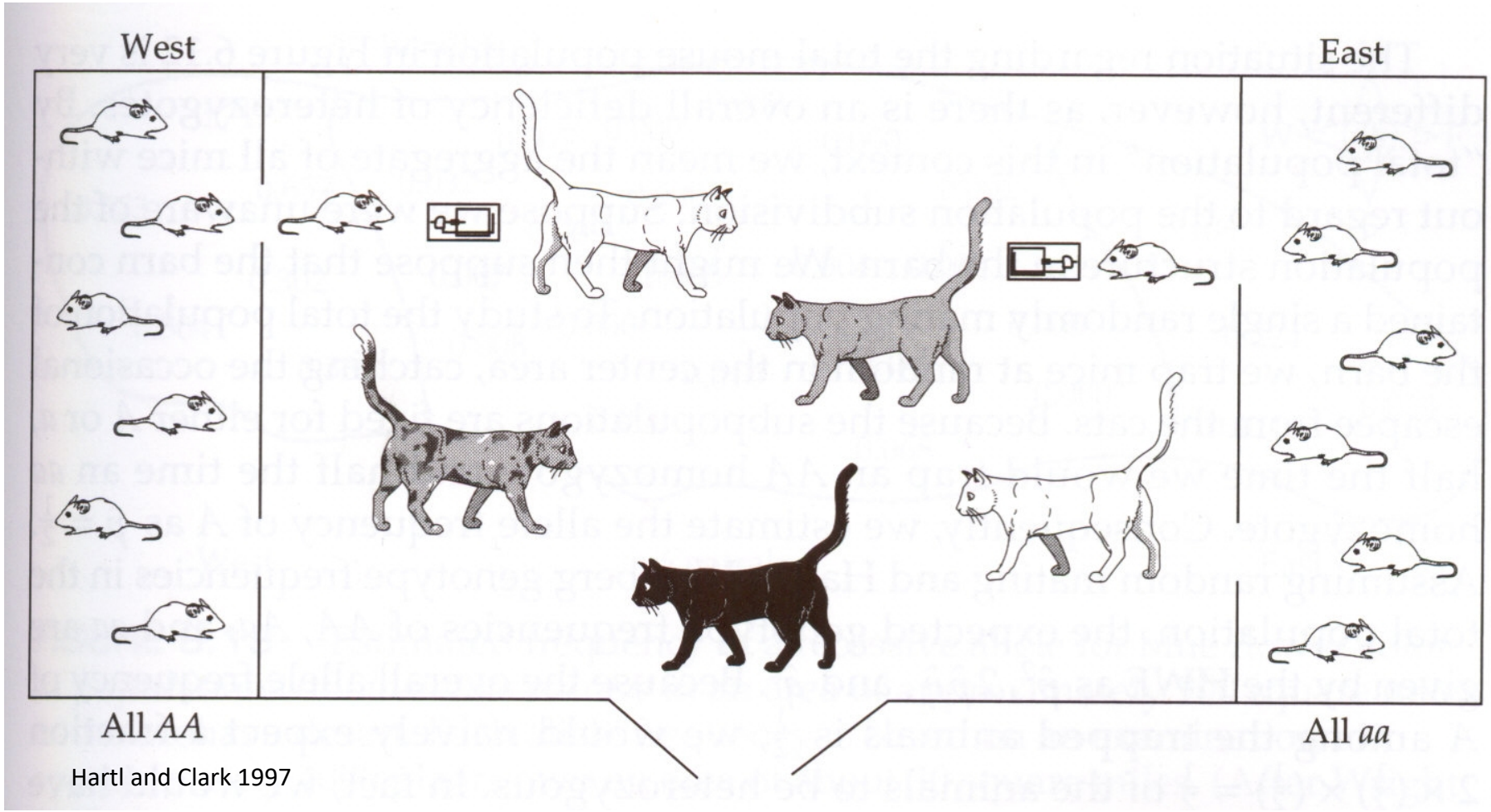
Walhund Effect

Expected heterozygosity is *2pq*

- *Separate populations:*
  - $H_E = 2pq = 2(1)(0) = 0$
- *Merged populations:*
  - $H_E = 2pq = 2(0.5)(0.5) = 0.5$

H_E always exceeds H_O when randomly-mating subpopulations are merged





Hartl and Clark 1997

Trapped mice will always be homozygous even though $H_E = 0.5$

Heterozygosity is the condition of having two different alleles at a locus.

The proportion of **heterozygous** individuals in a population is a good indicator of the effective population size.

Because populations are often **structured** in nature (thus restricting gene flow), heterozygosity is often less than expected.

How do we measure population structure?

F-Statistics

Based on the inbreeding coefficient, $\hat{F}$.

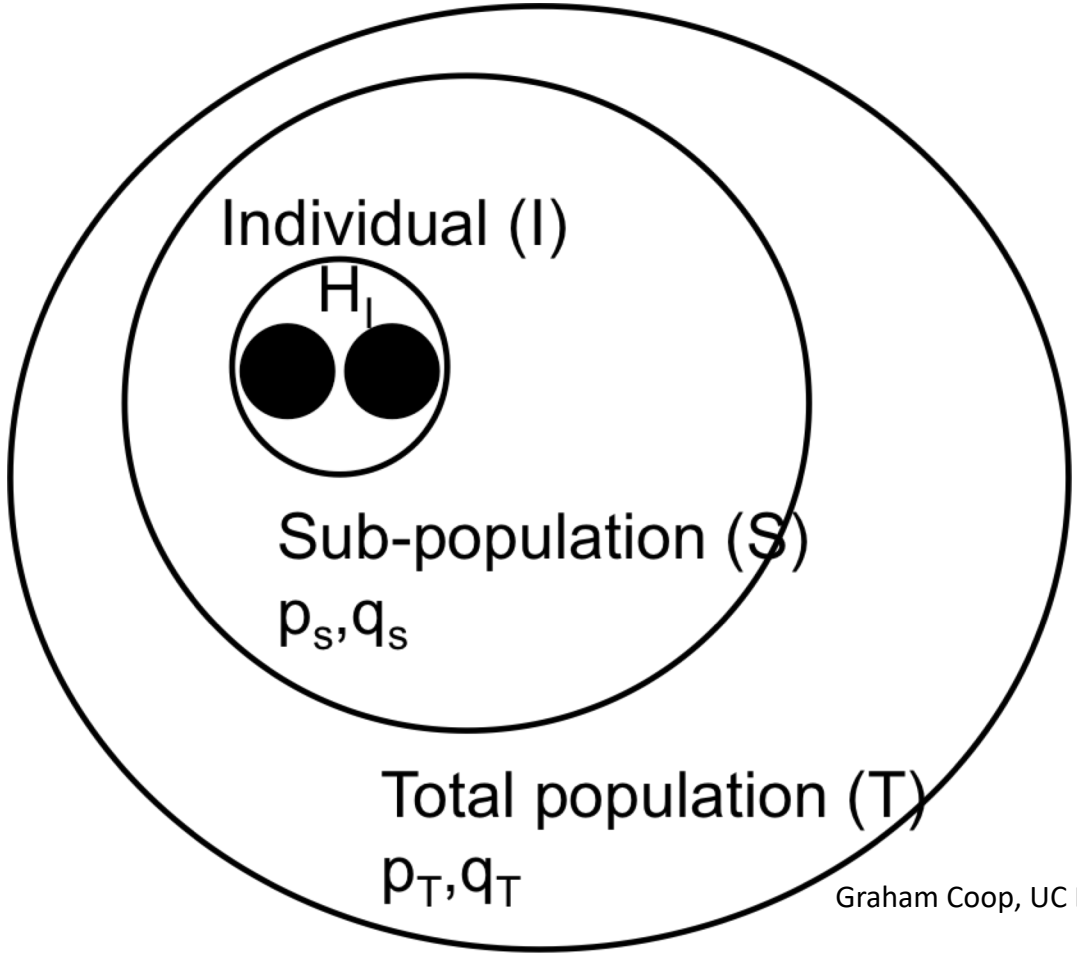
- $\hat{F} = 1 - \frac{H_o}{H_E}$, where H_E is the expected heterozygosity under Hardy-Weinberg equilibrium.

$$F_{IS} = 1 - \frac{H_I}{H_S}$$

$$F_{IT} = 1 - \frac{H_I}{H_T}$$

$$F_{ST} = 1 - \frac{H_S}{H_T}$$

Arelquin
DnaSP
vcftools
PopGenome



Graham Coop, UC Davis

A diagram showing the hierarchical nature of F-statistics. The two solid dots within an individual show the two alleles at a locus for an individual I. We can compare the heterozygosity on individuals to that found by randomly drawing alleles from the sub-population, to that found in the total population.

F-Statistics

Based on the inbreeding coefficient, $\hat{F}$.

- $\hat{F} = 1 - \frac{H_o}{H_E}$, where H_E is the expected heterozygosity under Hardy-Weinberg equilibrium.

$$F_{IS} = 1 - \frac{H_I}{H_S}$$

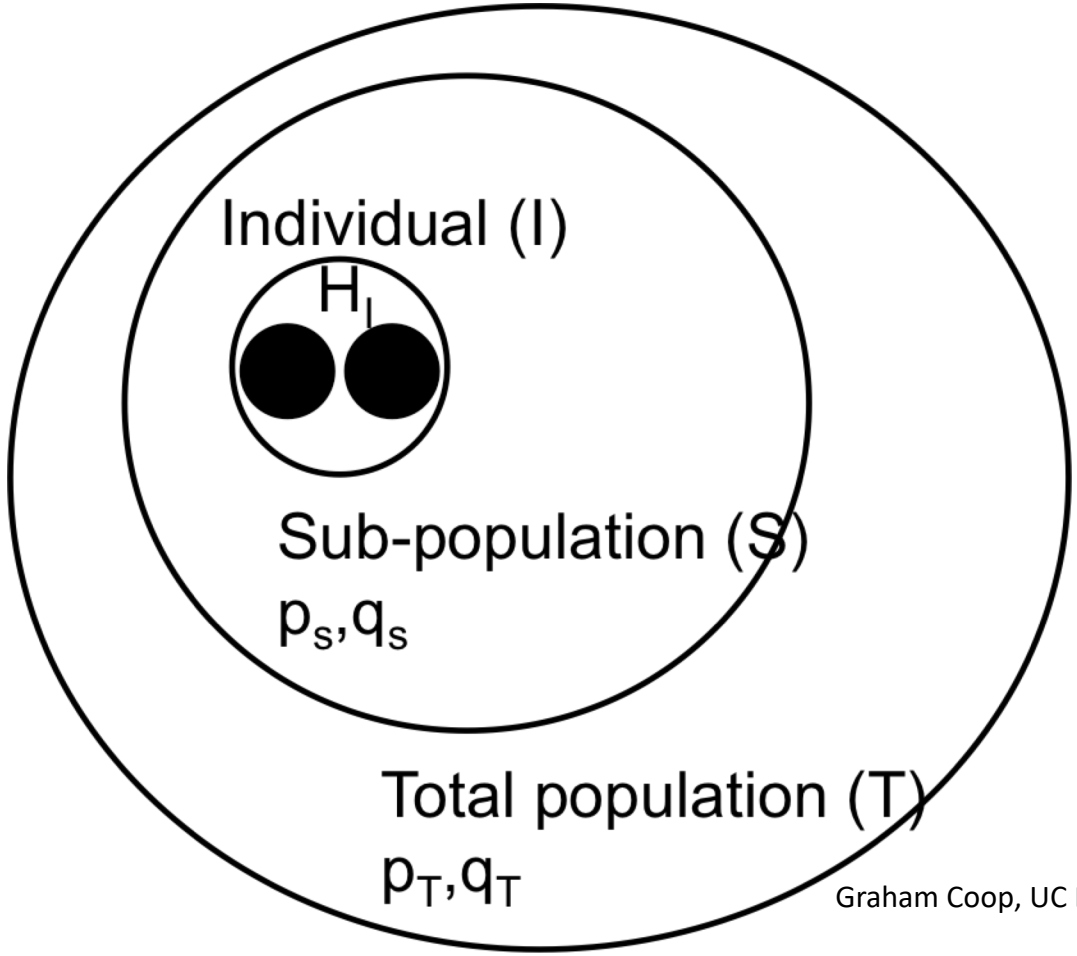
$$F_{IT} = 1 - \frac{H_I}{H_T}$$

$$F_{ST} = 1 - \frac{H_S}{H_T}$$

This is the FIXATION INDEX.

High Fst indicates more variation between subpopulations than within

Arelquin
DnaSP
vcftools
PopGenome



Graham Coop, UC Davis

A diagram showing the hierarchical nature of F-statistics. The two solid dots within an individual show the two alleles at a locus for an individual I. We can compare the heterozygosity on individuals to that found by randomly drawing alleles from the sub-population, to that found in the total population.

Approaches for Estimating Population Structure

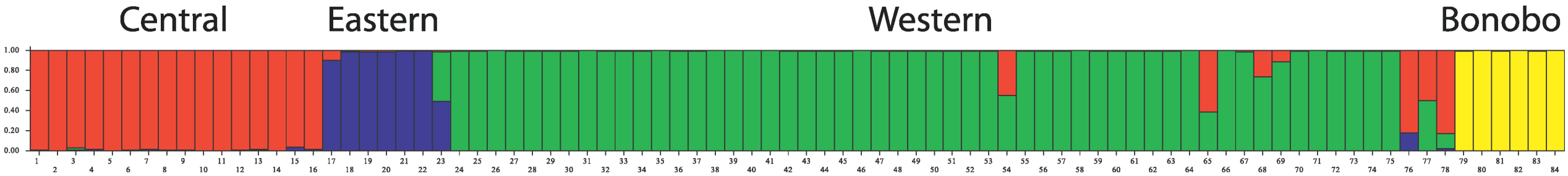
Cluster-based assignment methods

- *Find the probability that an individual of unknown population origin comes from one of the K predefined populations.*
- *Uses Bayesian statistics (with priors)*
    1. *Estimate allele frequencies from the data at all loci in each K population.*
    2. *Reassign each individual to a population k given the allele frequencies.*

Pritchard et al. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics.*

Often called STRUCTURE analysis.

With a K = 4, each individual is represented as a vertical bar divided into four colors depicting the estimate of the fraction of ancestry coming from each of the K = 4 populations.



78 *Pan troglodytes* and 6 *P. paniscus*, 300 polymorphic loci

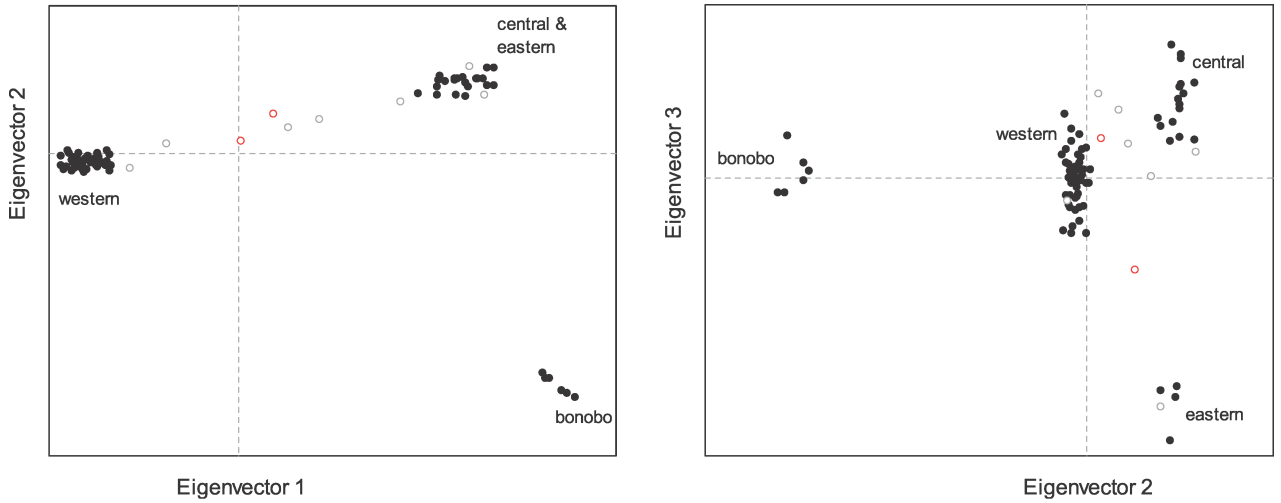
Becquet et al. 2007. PLoS Genetics

Approaches for Estimating Population Structure

Principal Components Analysis (PCA)

- *Consider a dataset of N individuals at S SNPs. At any SNP, an individual's genotype can be a 0, 1, or 2 based on the number of copies of one of the alleles.*
- *This can be expanded into an $N \times S$ covariance matrix upon which we can perform PCA analysis.*
- *PCA axes represent major axes of variation in the data.*

PC1 separates Western chimps from the others,
PC2 separates bonobos from all the three groups of common chimpanzee.
PC3 separates the Eastern chimp samples from the others.
Note also how the hybrid individuals (open black and red circles) tend to fall intermediate between groups.



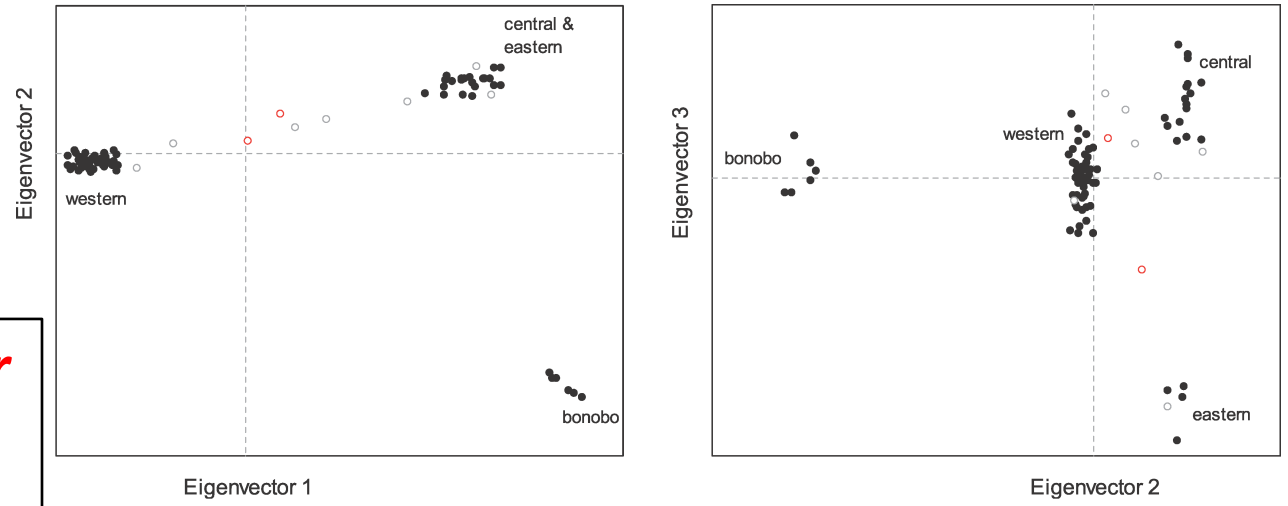
Becquet et al. 2007. PLoS Genetics

Approaches for Estimating Population Structure

Principal Components Analysis (PCA)

- *Consider a dataset of N individuals at S SNPs. At any SNP, an individual's genotype can be a 0, 1, or 2 based on the number of copies of one of the alleles.*
- *This can be expanded into an $N \times S$ covariance matrix upon which we can perform PCA analysis.*
- *PCA axes represent major axes of variation in the data.*

PC1 separates Western chimps from the others,
PC2 separates bonobos from all the three groups of common chimpanzee.
PC3 separates the Eastern chimp samples from the others.
Note also how the hybrid individuals (open black and red circles) tend to fall intermediate between groups.

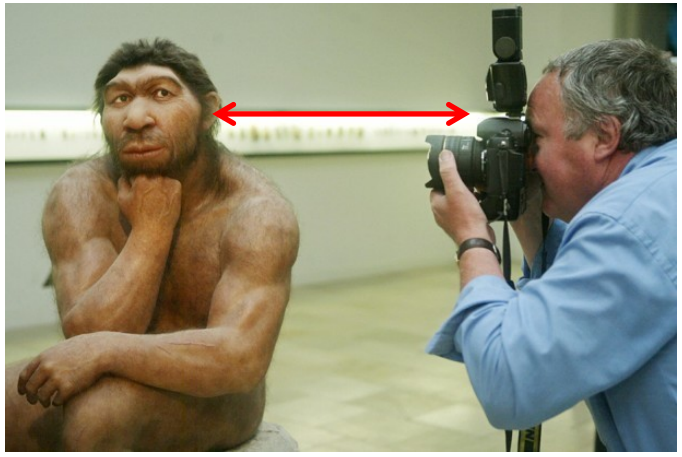


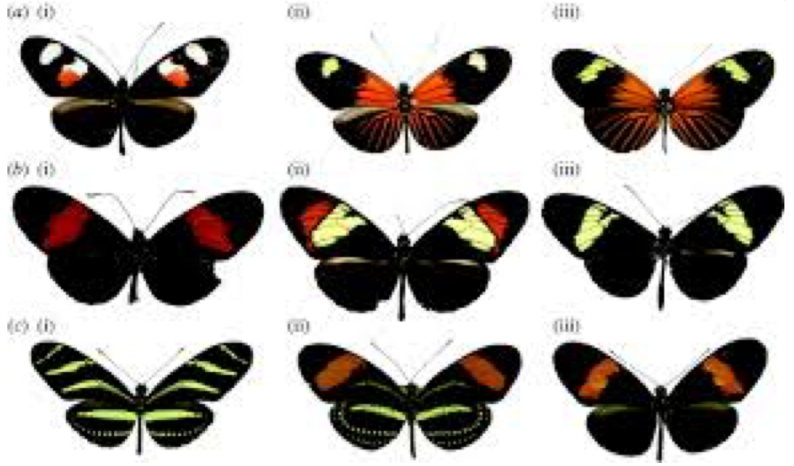
STRUCTURE and PCA are widely used together to assess population structure. PCA has an added benefit of being non-parametric.

Becquet et al. 2007. PLoS Genetics

Introgression

- the transfer of genetic material (gene flow) from one species to another as a result of hybridization
- (Surprisingly?) common
- Hybridization between species has the potential to contribute to phenotypic and genetic variation





Introgression can be adaptive

	Adaptation	Reference
	Darwin's finches: beak shape variation	Lamichhaney *et al.*, 2015
	*Anopheles:* insecticide resistance, dessication resistance	Fontaine *et al.*, 2015 Norris *et al.*, 2015 Fouet *et al.*, 2012 Gray *et al.*, 2009
	*Mus:* olfactory receptors, rodenticide resistance	Liu *et al.*, 2015 Song *et al.*, 2011
	*Heliconius:* wing pattern mimicry	The *Heliconius* Genome Consortium, 2012 Pardo-Diaz *et al.*, 2012
	Humans: high altitude adaptation	Huerta-Sánchez *et al.*, 2014

Figure 4. Summary of recent studies of adaptive introgression in animals. Examples highlight the exchange of distinct adaptive phenotypic traits between species.

Warfarin resistance in *Mus domesticus*

Mus musculus domesticus

Mus spretus

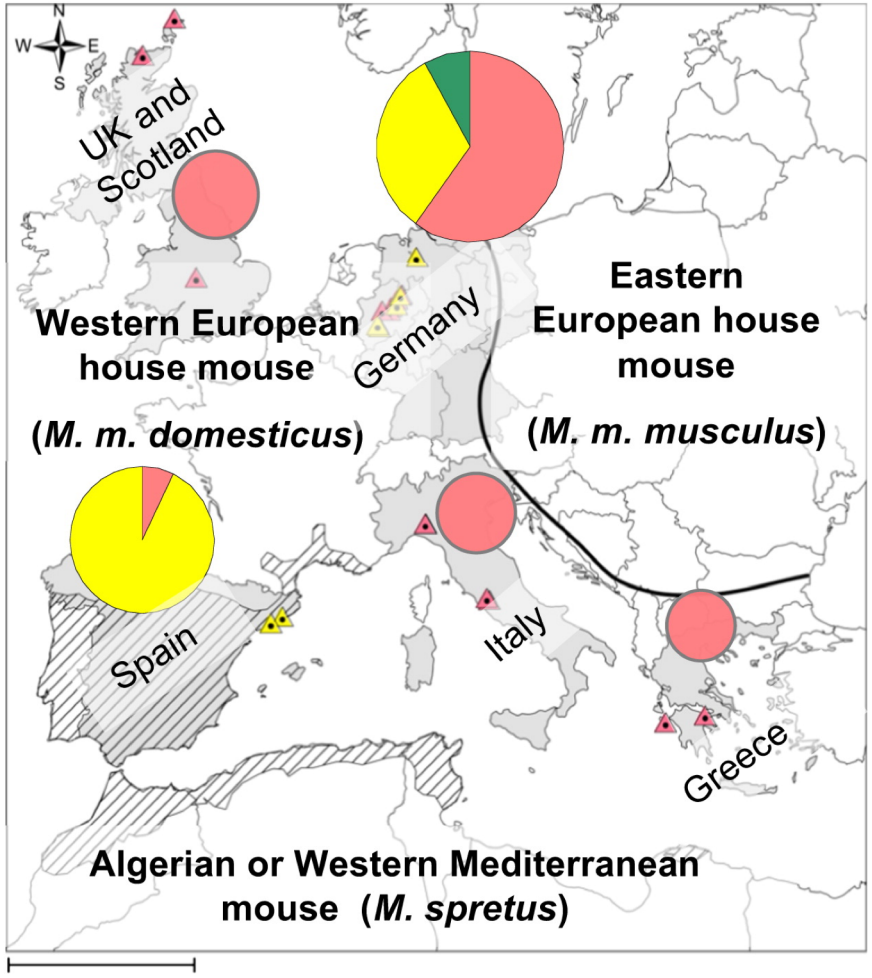
Adapted from Katya Mack

Adaptive introgression of warfarin resistance

Mus spretus carries variation in the gene *vkorc1* that confers resistance to anticoagulant rodenticides such as warfarin

Adaptive evolution at *vkorc1* in *M. spretus* may be an adaptation to a granivorous vitamin K-deficient diet

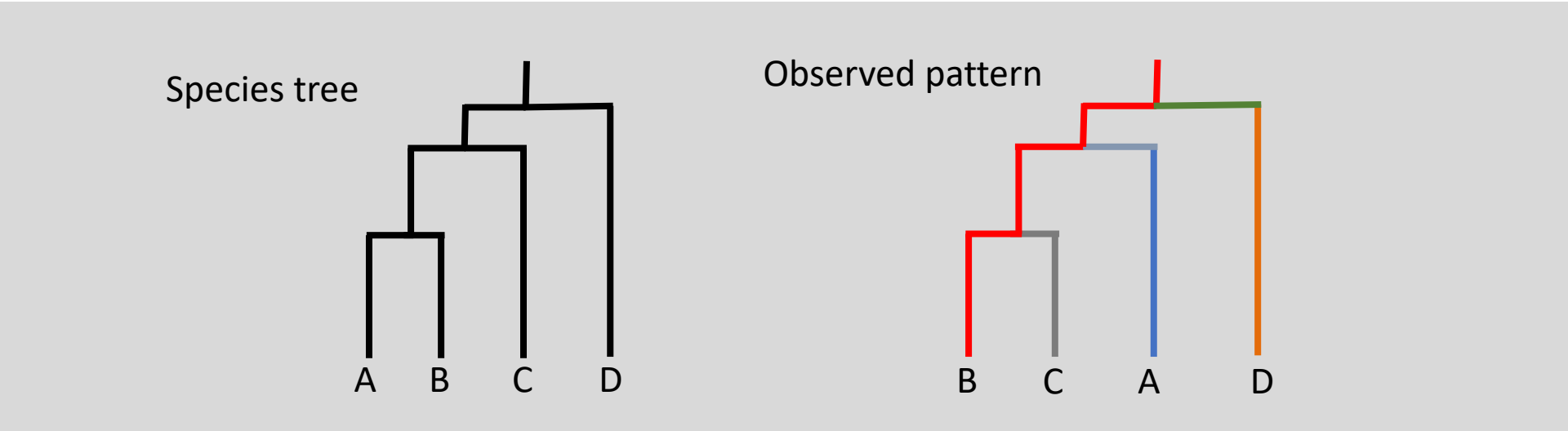
Adaptive introgression of *M. spretus vkorc1* allele into *M. domesticus* populations



Song et al., 2011, *Curr Biology*

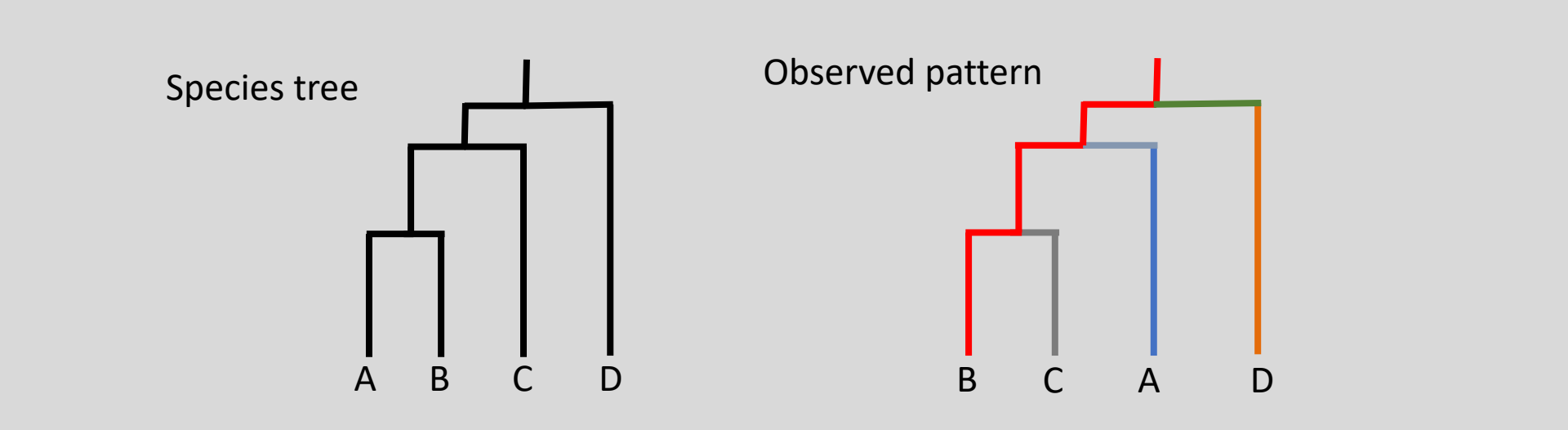
Adapted from Katya Mack

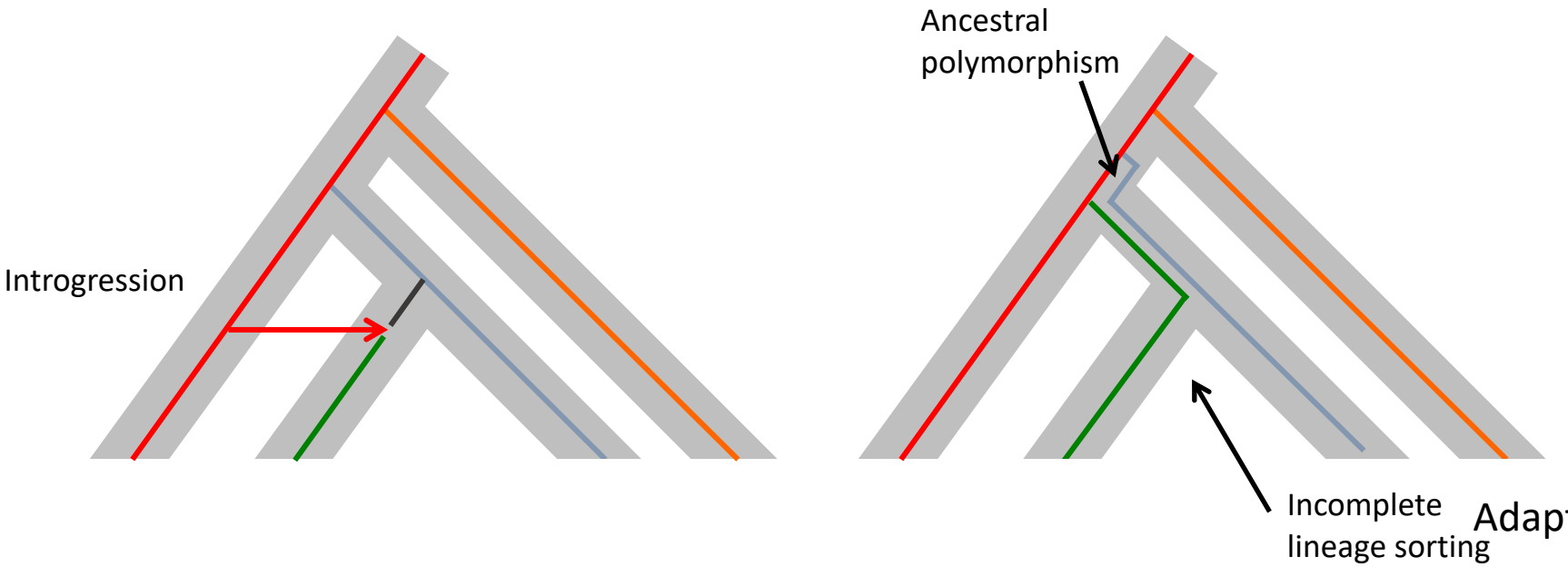
Incomplete lineage sorting



Adapted from Katya Mack

Incomplete lineage sorting

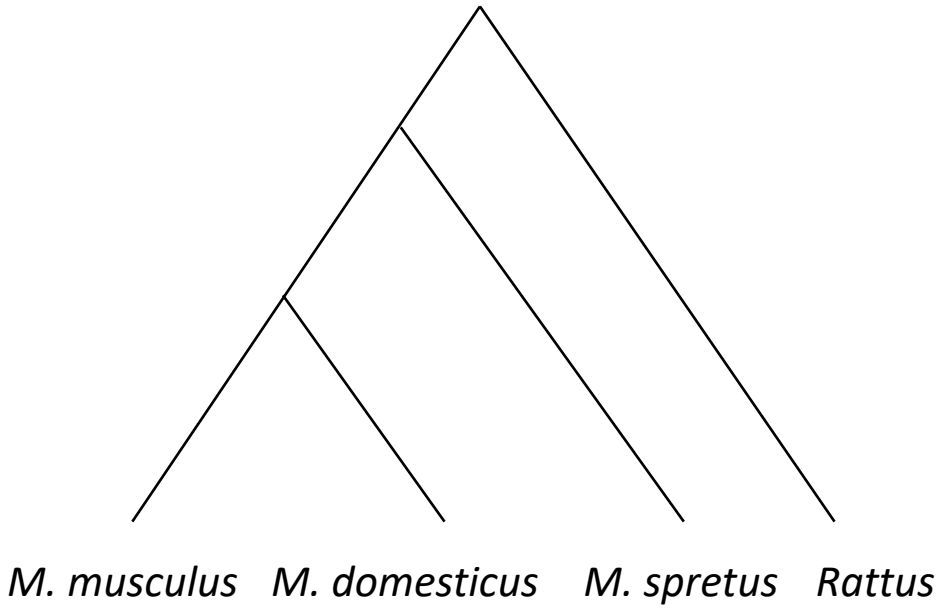




Adapted from Katya Mack

Identifying introgression

Phylogenetic tree

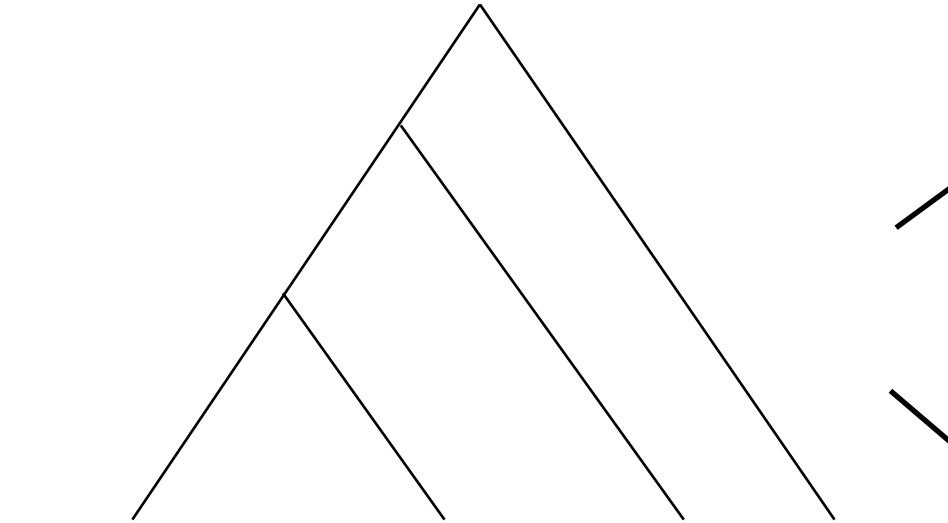


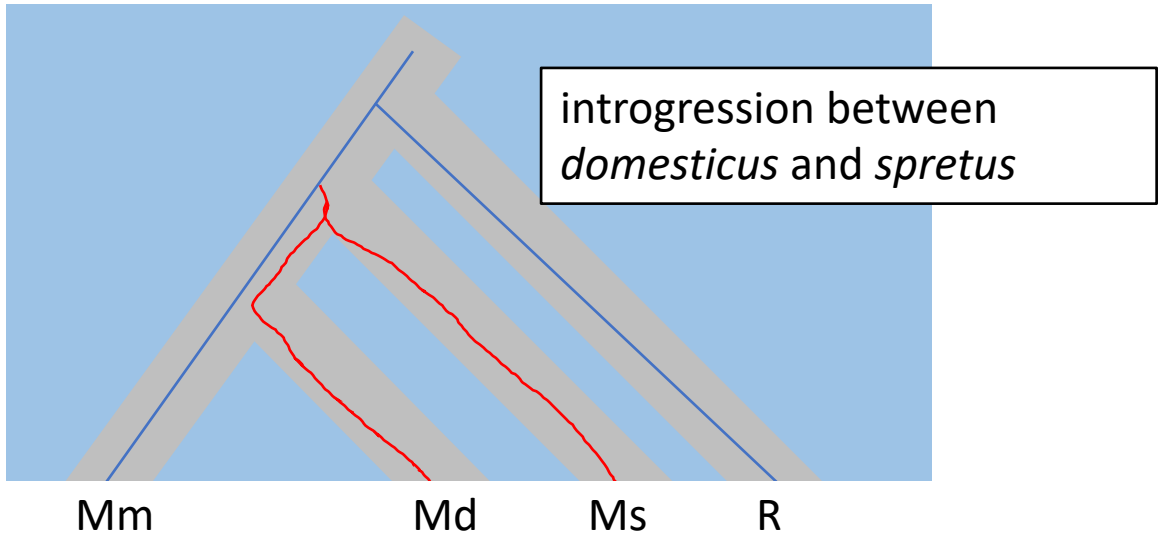
This genealogical arrangement (BBAA) represents the phylogeny and should be most frequent

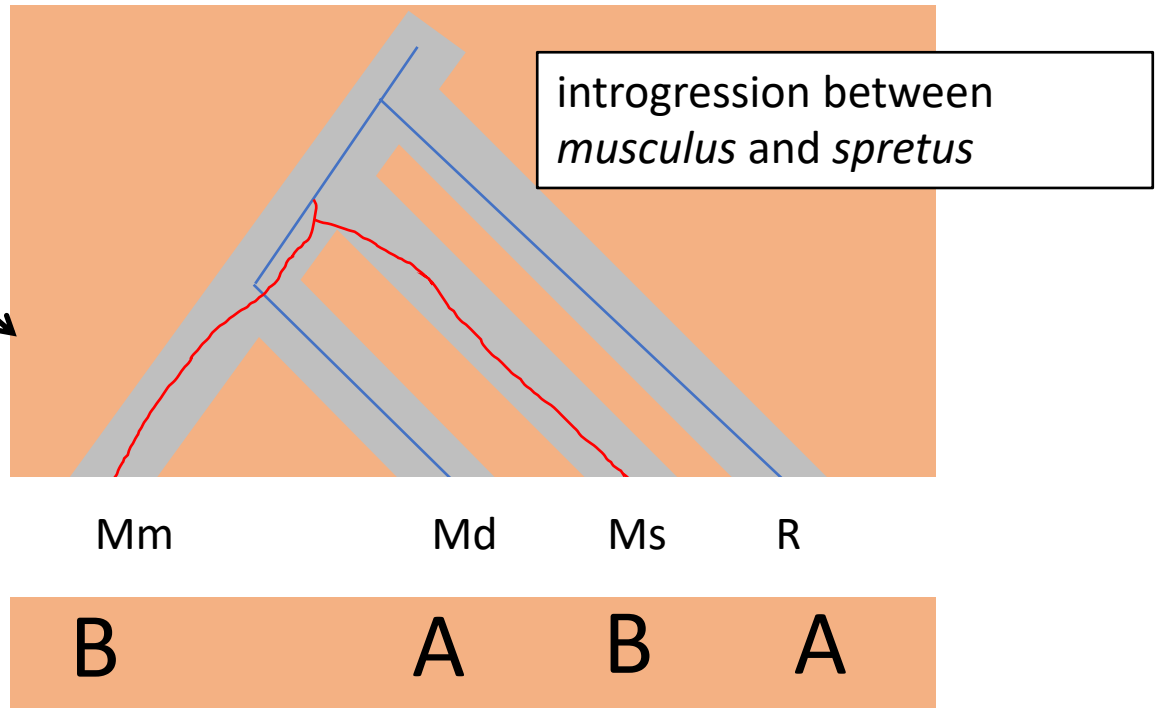
Adapted from Katya Mack

Identifying introgression

Phylogenetic tree







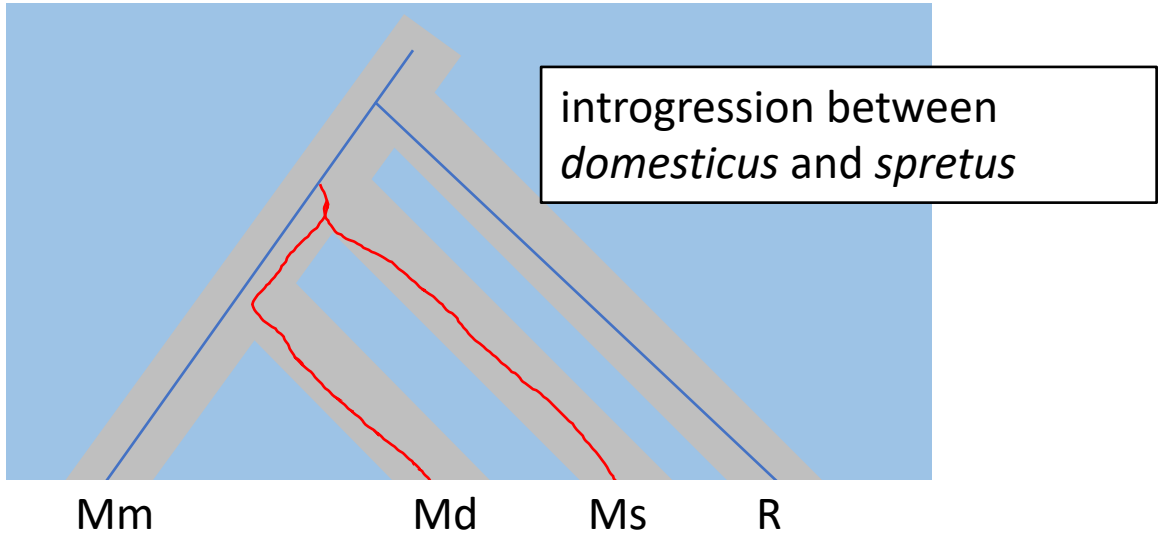
Adapted from Katya Mack

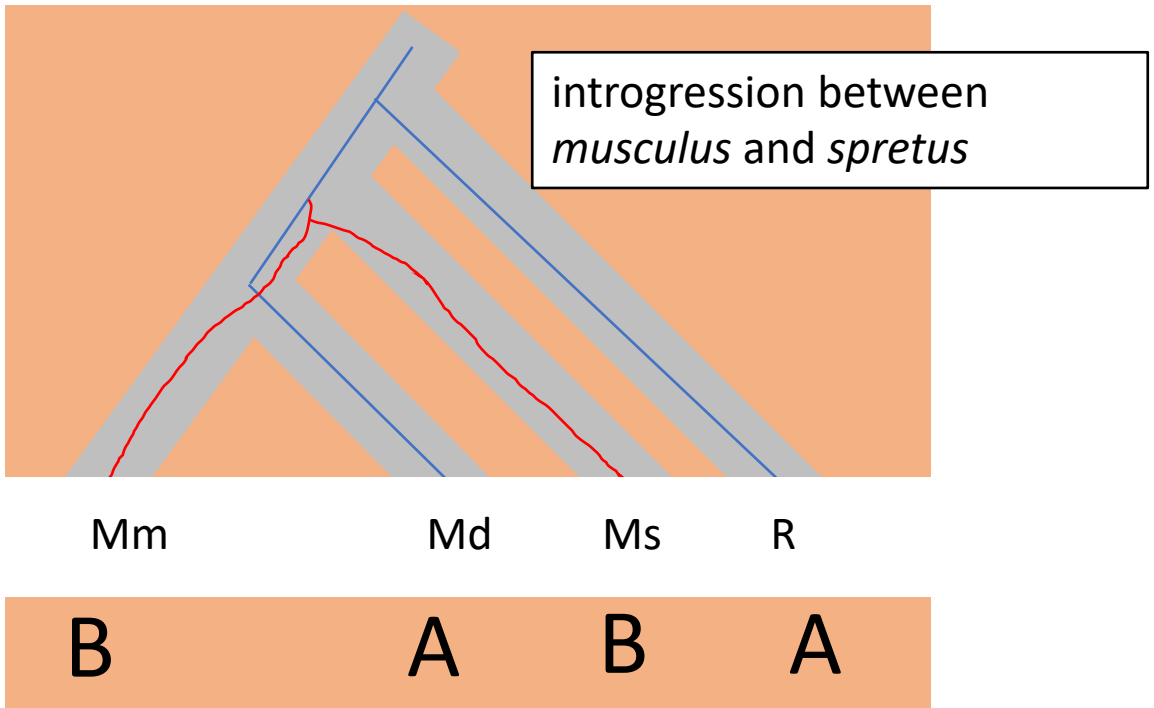
Identifying introgression

Patterson's D:

Expectation is that #(ABBA) = #(BABA)

$$D = \frac{\sum ABBA - BABA}{\sum ABBA + BABA}$$





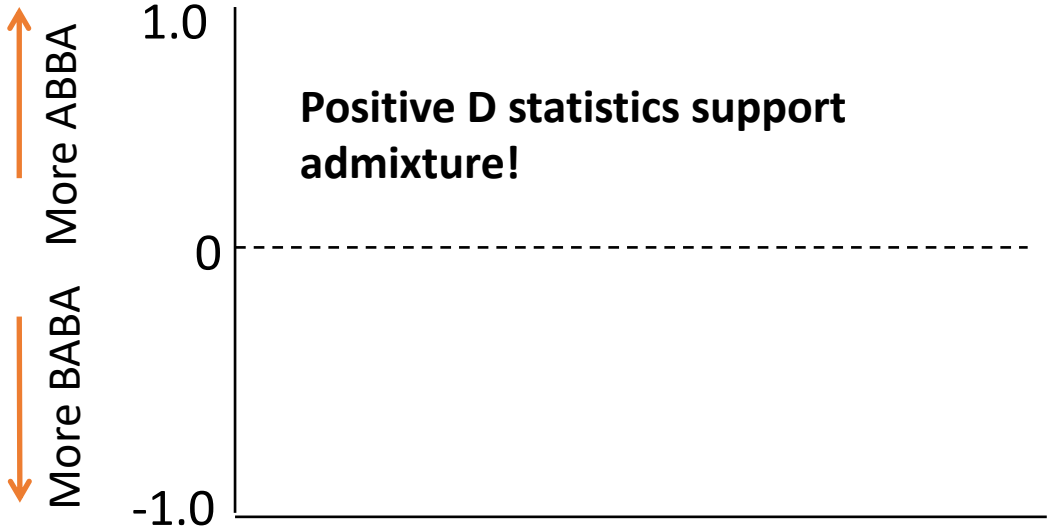
Adapted from Katya Mack

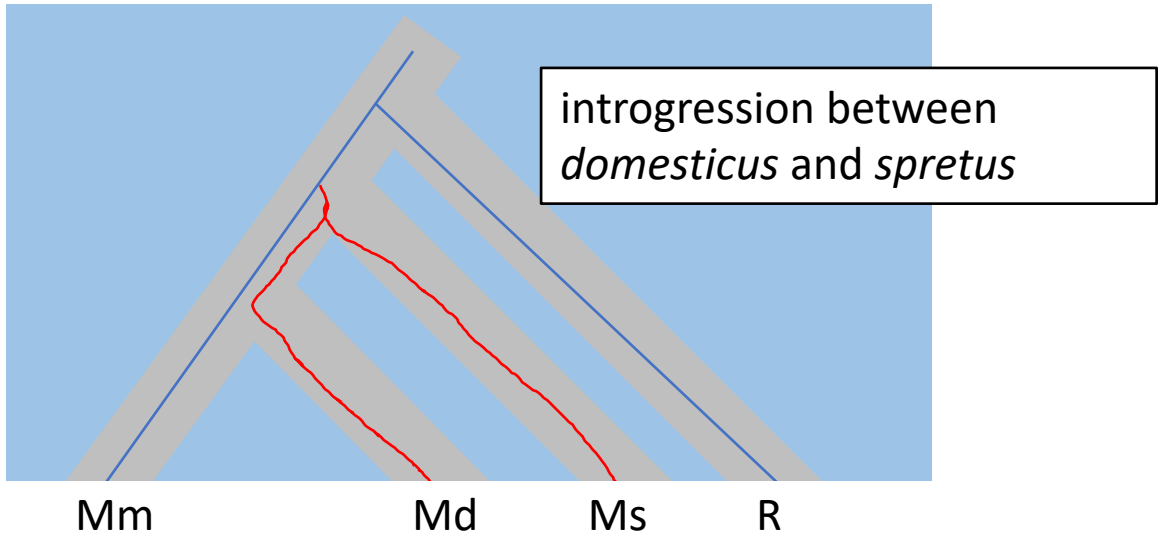
Identifying introgression

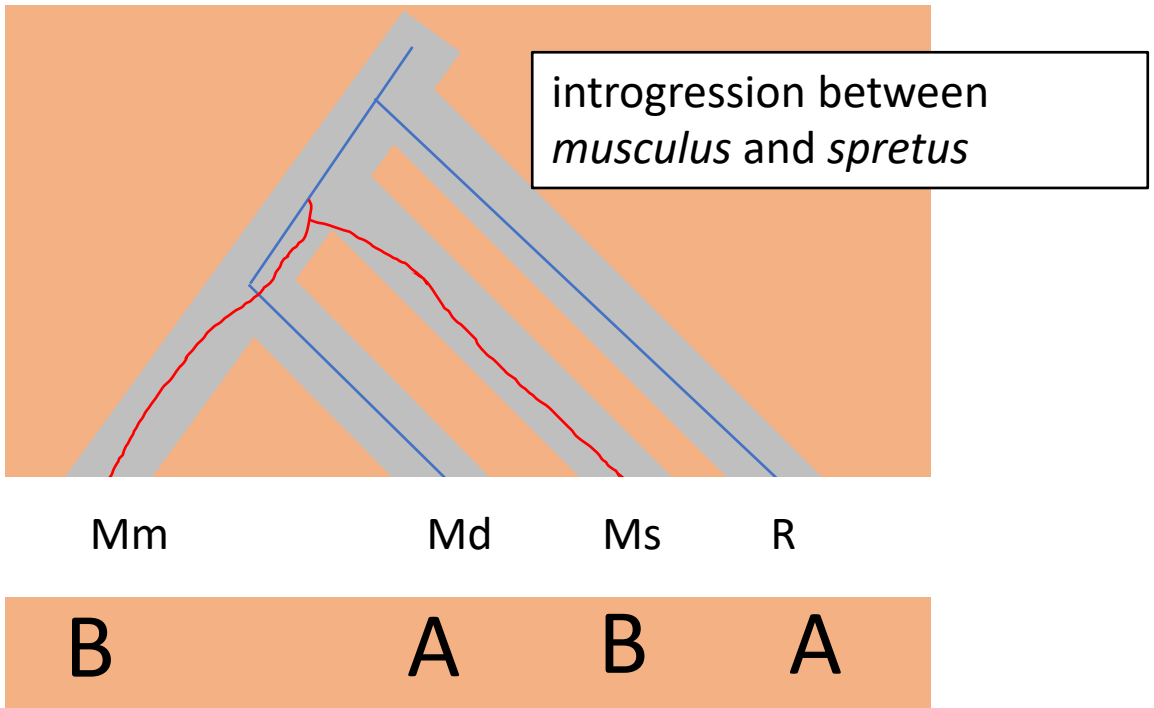
Patterson's D:

Expectation is that #(ABBA) = #(BABA)

$$D = \frac{\sum ABBA - BABA}{\sum ABBA + BABA}$$







Adapted from Katya Mack

OK, Back to selection…

Genetic Hitchhiking

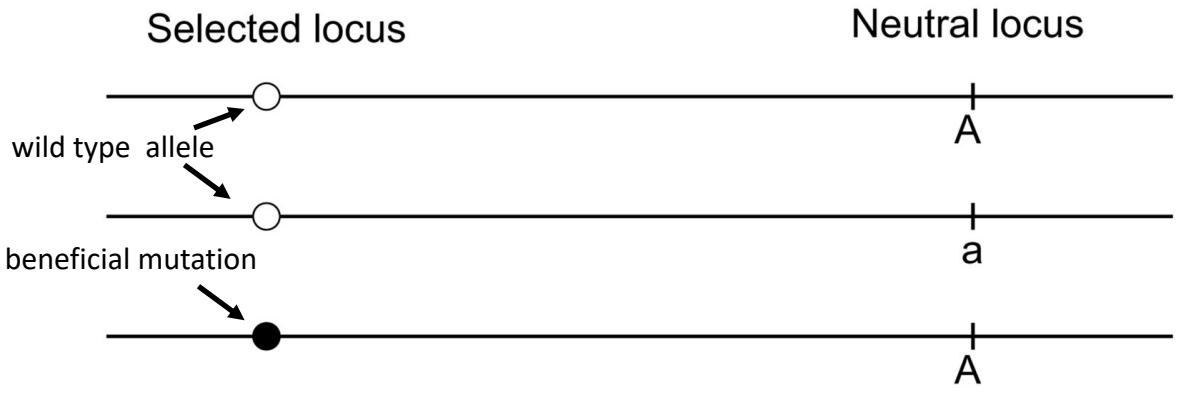
Consider three haplotypes present in a population.

- ***haplotype: a set of genes inherited from a parent.***

*There is a **selected** locus and a **neutral** locus.*

- *Note that a beneficial mutation has occurred at the selected locus.*
- *there are now two alleles at the selected locus.*

The neutral locus contains two alleles (A and a)



Stephan W. 2019. Selective Sweeps. *Genetics.*

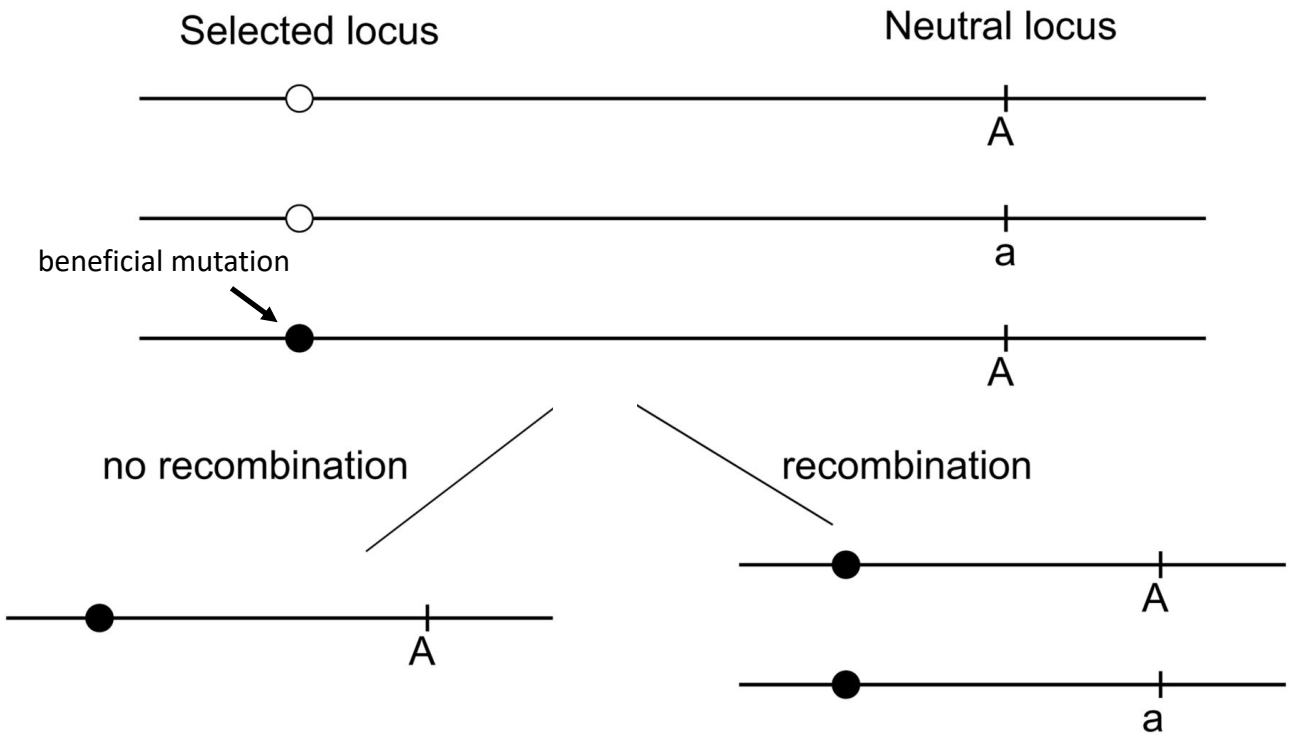
Genetic Hitchhiking

If no recombination occurs, one haplotype is present.

With recombination, the neutral locus stays ***polymorphic.***

With recombination, two haplotypes remain.

Recombination rate is positively correlated with polymorphism.

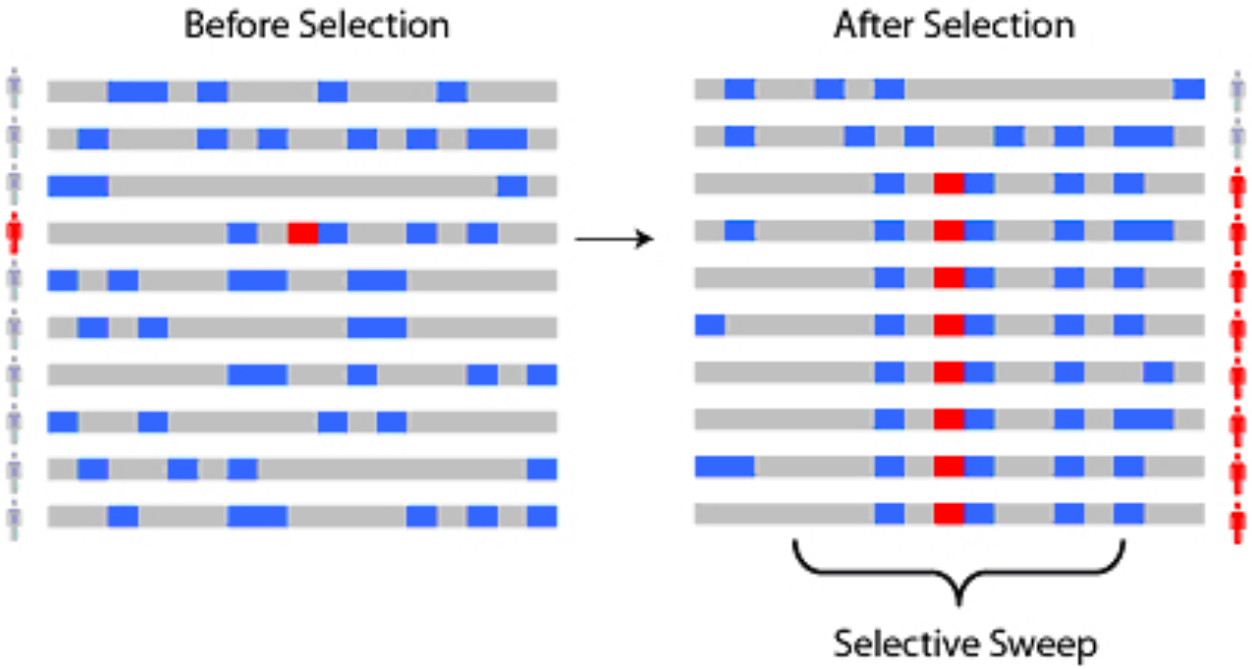


Stephan W. 2019. Selective Sweeps. *Genetics.*

Selective Sweeps

A new beneficial mutation will rise in frequency in the population.

Nearby linked alleles will also rise in frequency.

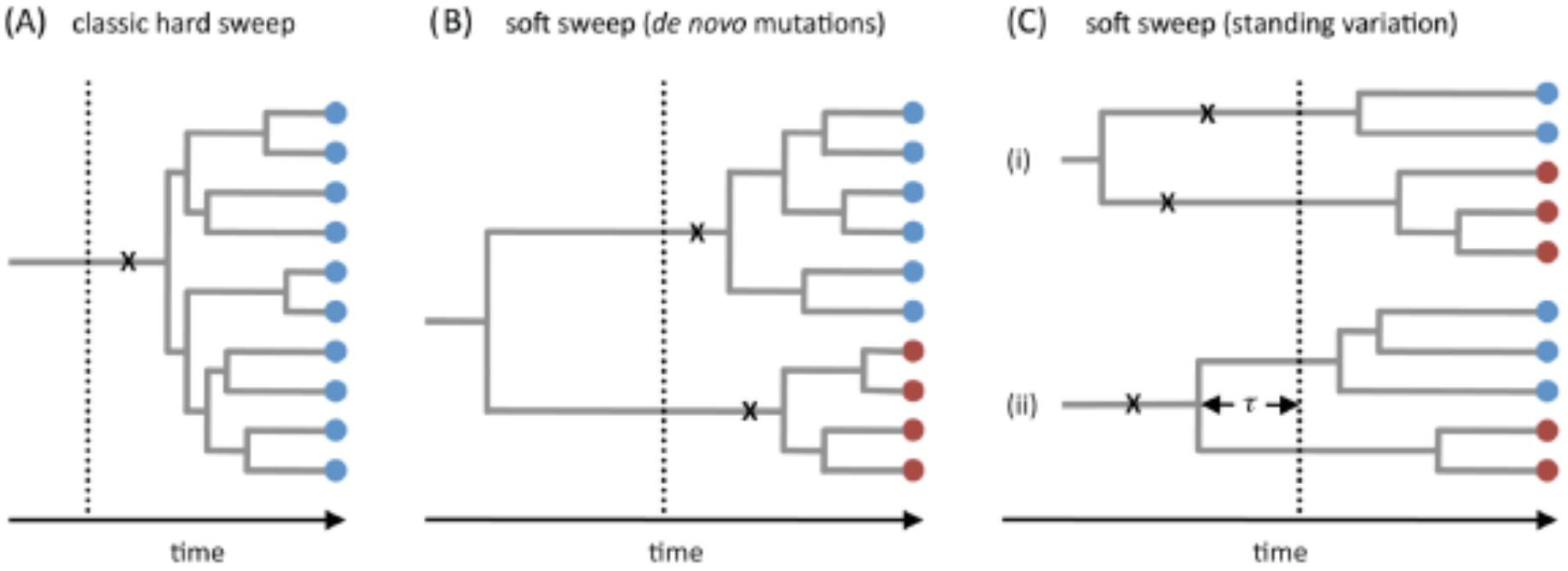


Schaffner SF. 2008. Evolutionary Adaptation in the Human Lineage. *Nature Education.*

Hard versus Soft Selective Sweeps

In a ‘hard’ sweep, a new mutation sweeps through the population.

This would mean that a genealogy of all samples would coalesce more recently than the the onset of positive selection.

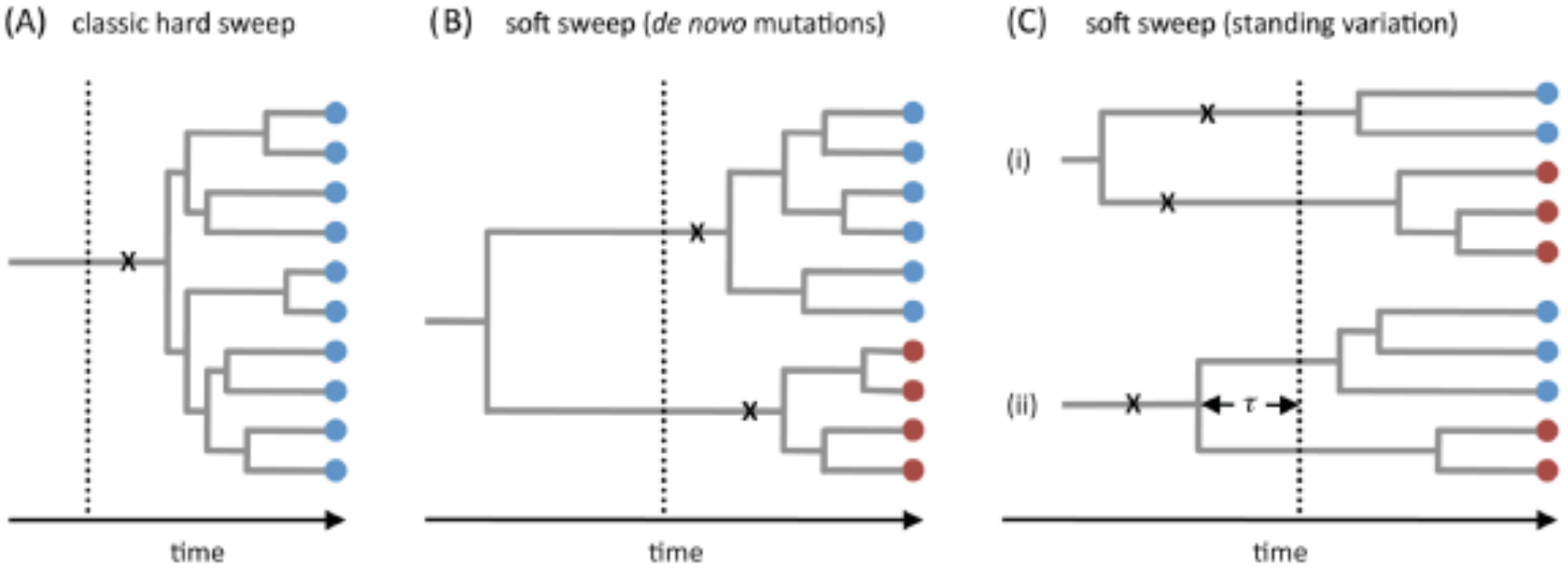


Messer and Petrov. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*.

Hard versus Soft Selective Sweeps

In contrast, a ‘soft’ sweep is when multiple adaptive alleles at the same locus sweep through the population to high frequency.

In this case, a genealogy would reveal coalescence ***before*** the onset of positive selection.

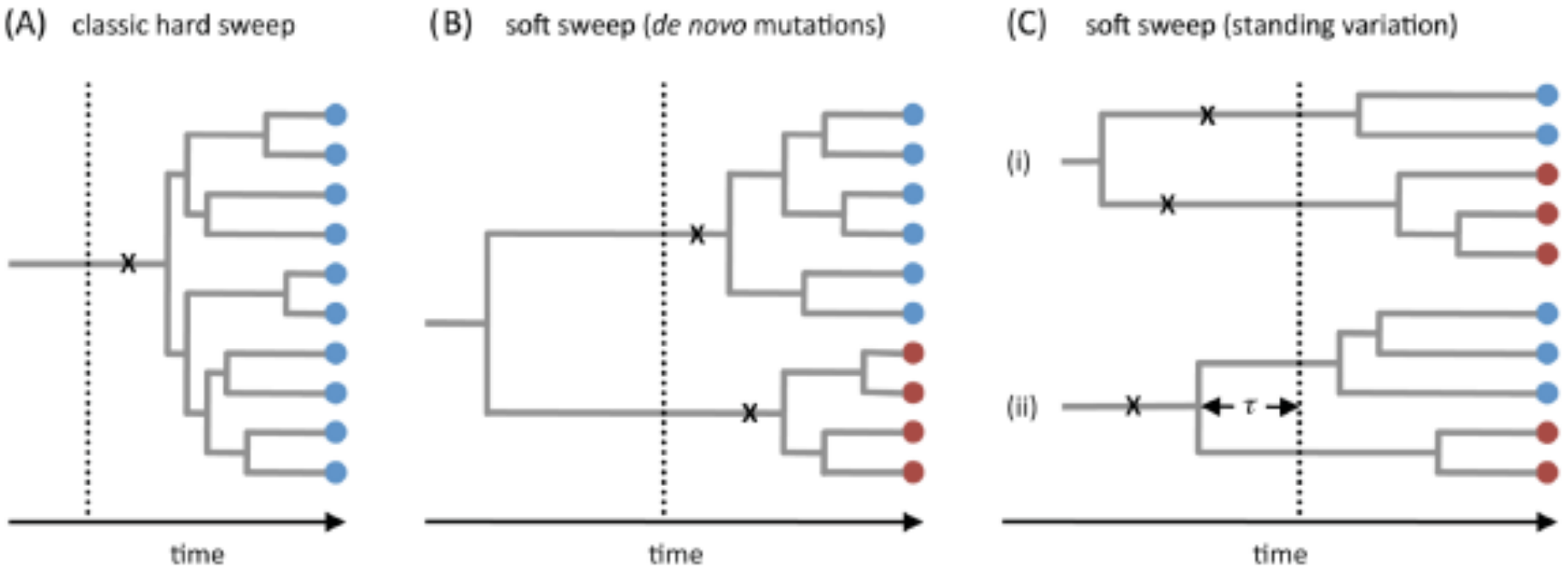


Messer and Petrov. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*.

Hard versus Soft Selective Sweeps

In B, the mutations arose *de novo* after the onset of positive selection.

In C, they may have arose *de novo* after positive selection (top) or were already present as ***standing variation***.

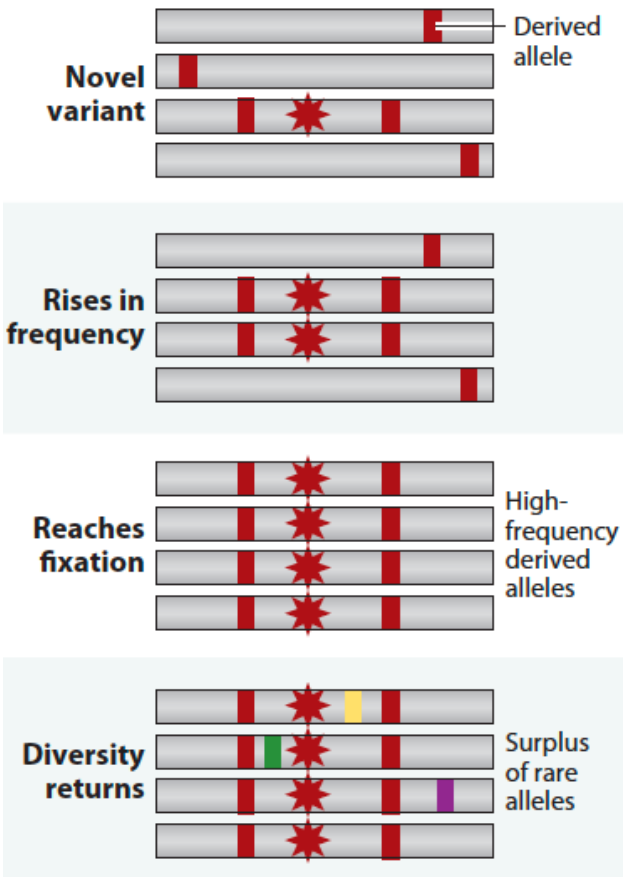


Messer and Petrov. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*.

Methods for Detecting Selection in Population Samples

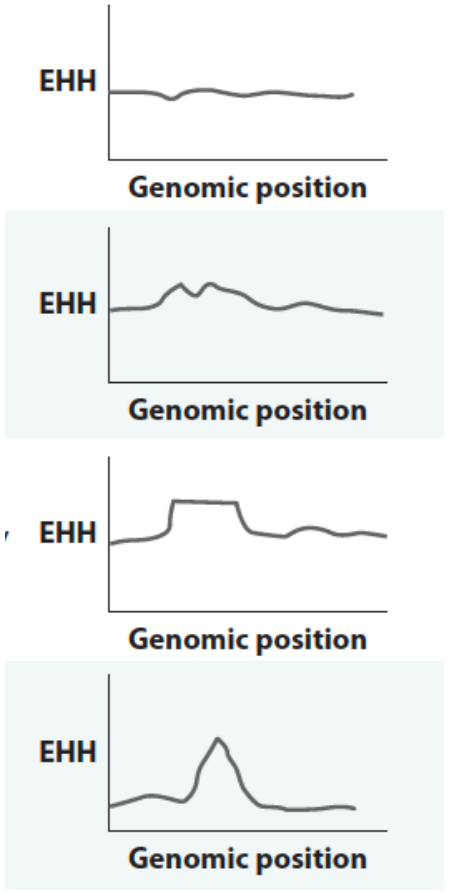
TIME

Frequency-based



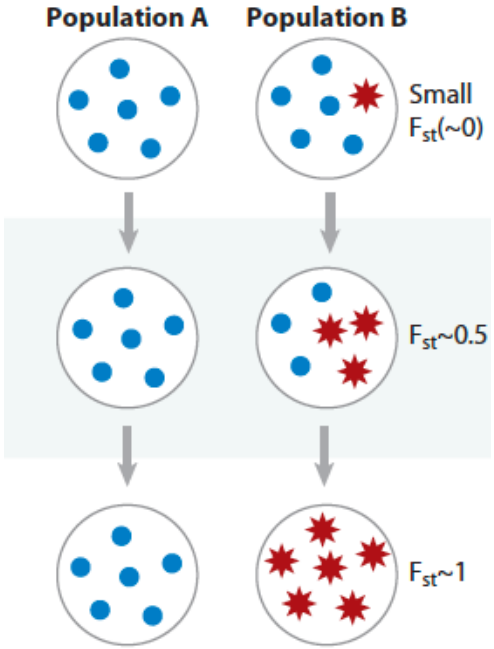
i.e., Tajima's D

Linkage-based



EHH = extended haplotype homozygosity

Population differentiation-based



Large values of Fst at a particular locus

Vitti et al. 2013. Detecting Natural Selection in Genomics Data. *Annual Reviews Genetics.*