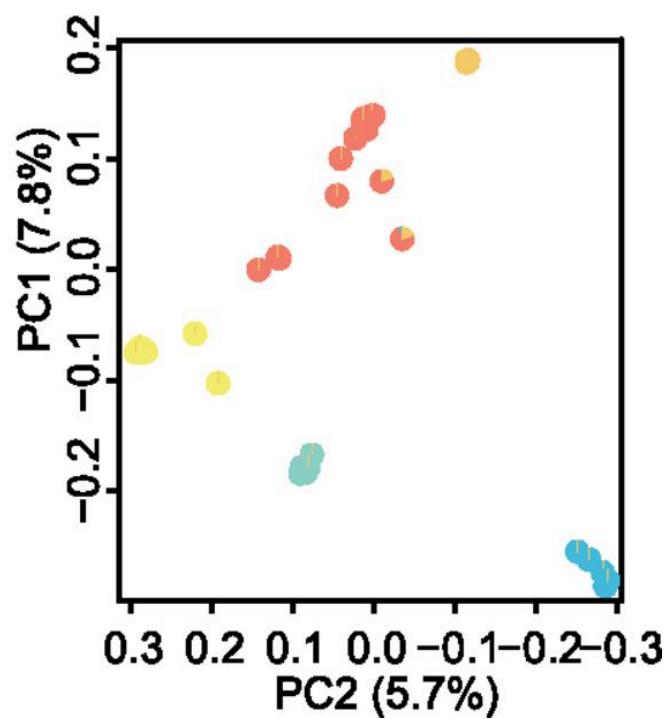
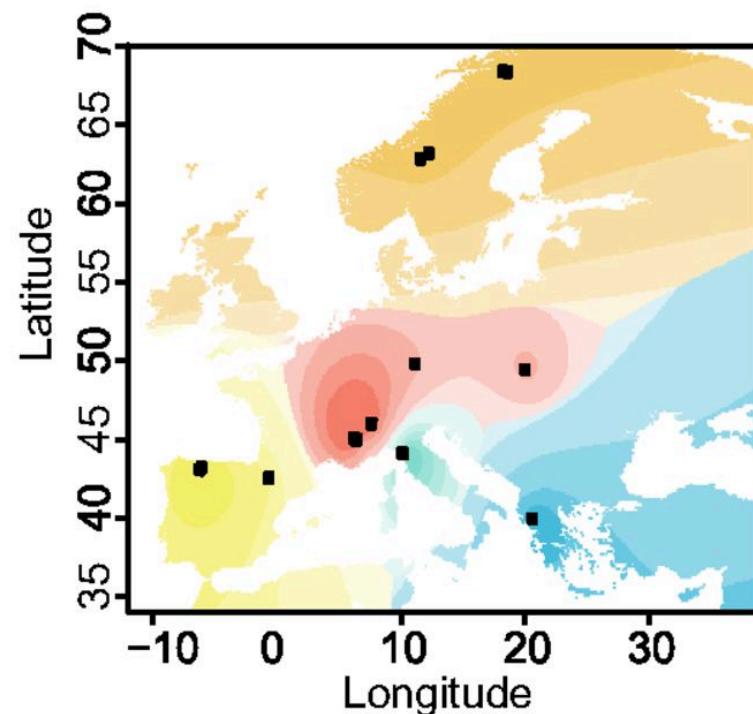
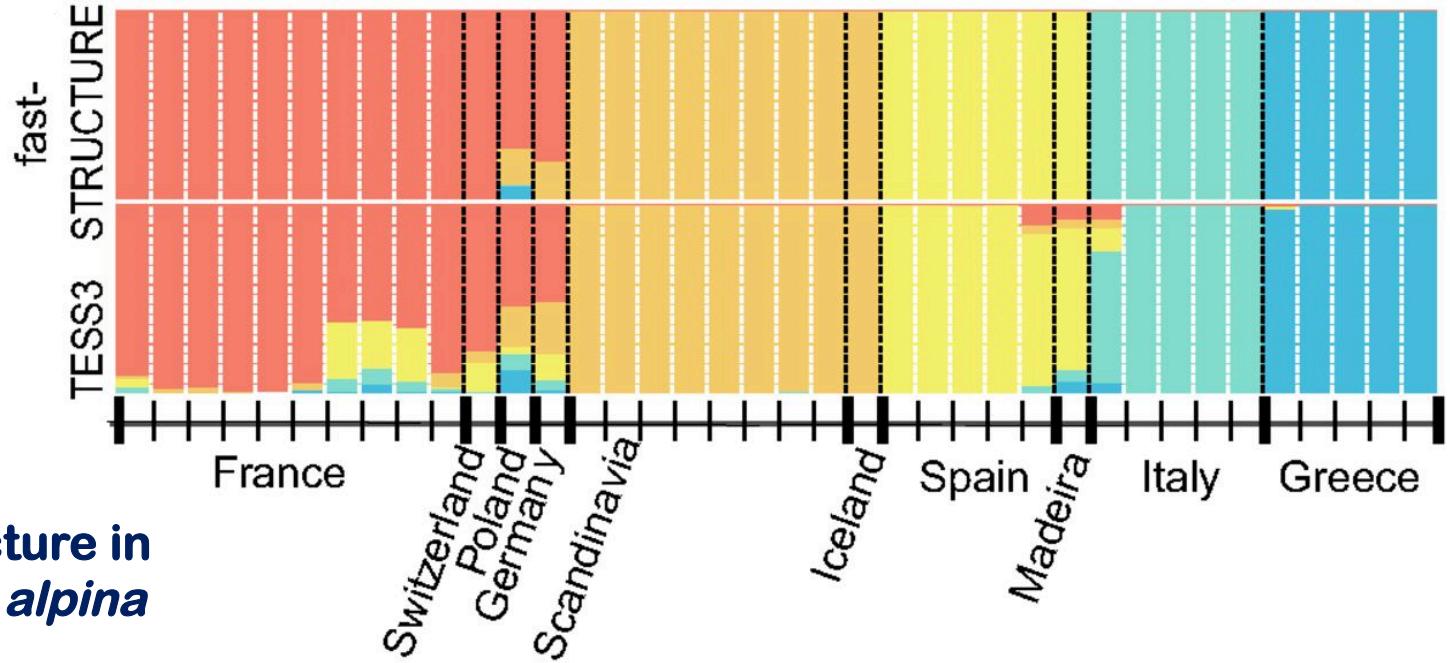


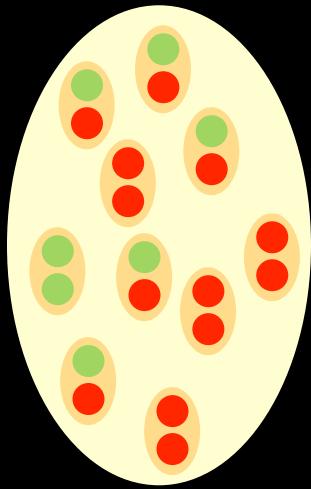
Population structure, admixture and changes in population size

Geographic structure in European *Arabis alpina*



Today

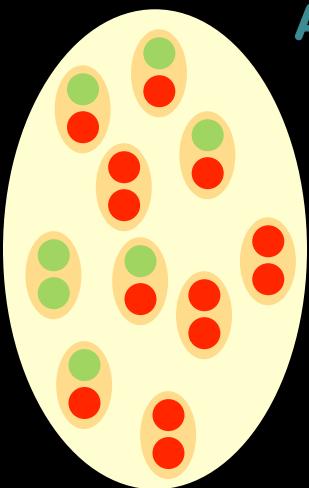
- **Genetic structure of populations**
- **Hardy-Weinberg Equilibrium**
- **Genetic variation in space and time**
- **Variation in natural populations, Wright's Fixation Index (F_{ST})**
- **Other methods to measure population structure/gene flow**



● A

● a

Allele frequencies



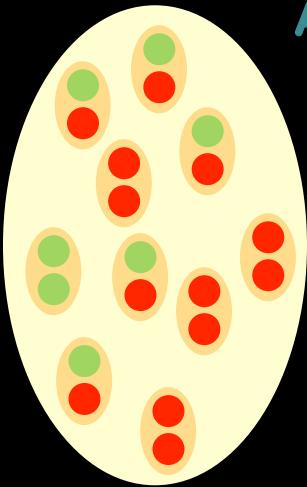
$$f_A = \frac{N_A}{2N} = \frac{7}{2*10}$$

or

$$f_a = \frac{N_a}{2N} = \frac{13}{2*10}$$

● A

● a



Allele frequencies

$$f_A = \frac{N_A}{2N} = \frac{7}{2*10}$$

or

$$f_a = \frac{N_a}{2N} = \frac{13}{2*10}$$

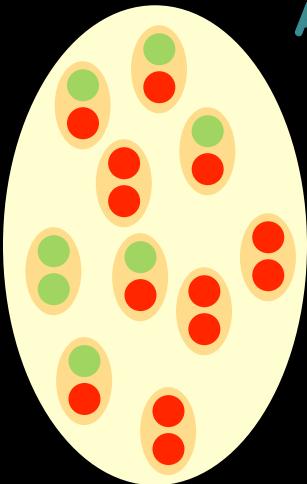
Genotype frequencies

$$f_{AA} = \frac{N_{AA}}{2N} = \frac{1}{10}$$

$$f_{Aa} = \frac{N_{Aa}}{2N} = \frac{5}{10}$$

$$f_{aa} = \frac{N_{aa}}{2N} = \frac{4}{10}$$

- A
- a



Allele frequencies

$$f_A = \frac{N_A}{2N} = \frac{7}{2*10}$$

or

$$f_a = \frac{N_a}{2N} = \frac{13}{2*10}$$

● A
● a

Genotype frequencies

$$f_{AA} = \frac{N_{AA}}{2N} = \frac{1}{10}$$

$$f_{Aa} = \frac{N_{Aa}}{2N} = \frac{5}{10}$$

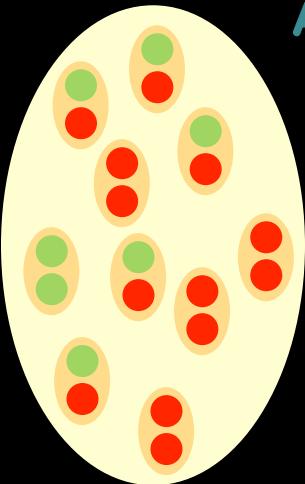
$$f_{aa} = \frac{N_{aa}}{2N} = \frac{4}{10}$$

K-allelic loci

$$f_i = f_{ii} + \sum f_{ij}/2$$

f_{Aa} = Heterozygosity

$1 - f_{Aa}$ = Homozygosity
 $= f_{AA} + f_{aa}$



Allele frequencies

$$f_A = \frac{N_A}{2N} = \frac{7}{2*10}$$

or

$$f_a = \frac{N_a}{2N} = \frac{13}{2*10}$$

● A
● a

Genotype frequencies

$$f_{AA} = \frac{N_{AA}}{2N} = \frac{1}{10}$$

$$f_{Aa} = \frac{N_{Aa}}{2N} = \frac{5}{10}$$

$$f_{aa} = \frac{N_{aa}}{2N} = \frac{4}{10}$$

K-allelic loci

$$f_i = f_{ii} + \sum f_{ij}/2$$

f_{Aa} = Heterozygosity

$1 - f_{Aa}$ = Homozygosity
 $= f_{AA} + f_{aa}$

You can also calculate allele frequencies by:

$$f_A = \frac{2N_{AA} + N_{Aa}}{2N}$$

Hardy-Weinberg:

- **Explains how Mendelian segregation influences allelic and genotypic frequencies in a population**

Hardy-Weinberg:

- Explains how Mendelian segregation influences allelic and genotypic frequencies in a population

Assumptions:

1. Population is infinitely large, to avoid effects of *genetic drift* (= change in genetic frequency due to chance)
1. Mating is random (with regard to traits under study)
2. No natural selection (for traits under study)
3. No mutation
4. No migration

Hardy-Weinberg law:

- If assumptions are met, population will be in genetic equilibrium

Two expected predictions:

1. Allele frequencies do not change over generations.
2. After one generation of random mating, genotypic frequencies will remain in the following proportions:

Hardy-Weinberg law:

- If assumptions are met, population will be in genetic equilibrium

Two expected predictions:

1. Allele frequencies do not change over generations.
2. After one generation of random mating, genotypic frequencies will remain in the following proportions:

p^2 (frequency of AA)

$2pq$ (frequency of Aa)

q^2 (frequency of aa)

p = allelic frequency of A

q = allelic frequency of a

$p^2 + 2pq + q^2 = 1$

$$f_A^2 + 2f_A f_a + f_a^2 = (f_A + f_a)^2 = 1$$

Basis of the Hardy-Weinberg

Hardy-Weinberg state: $p^2 + 2pq + q^2 = 1$ ~ at equilibrium

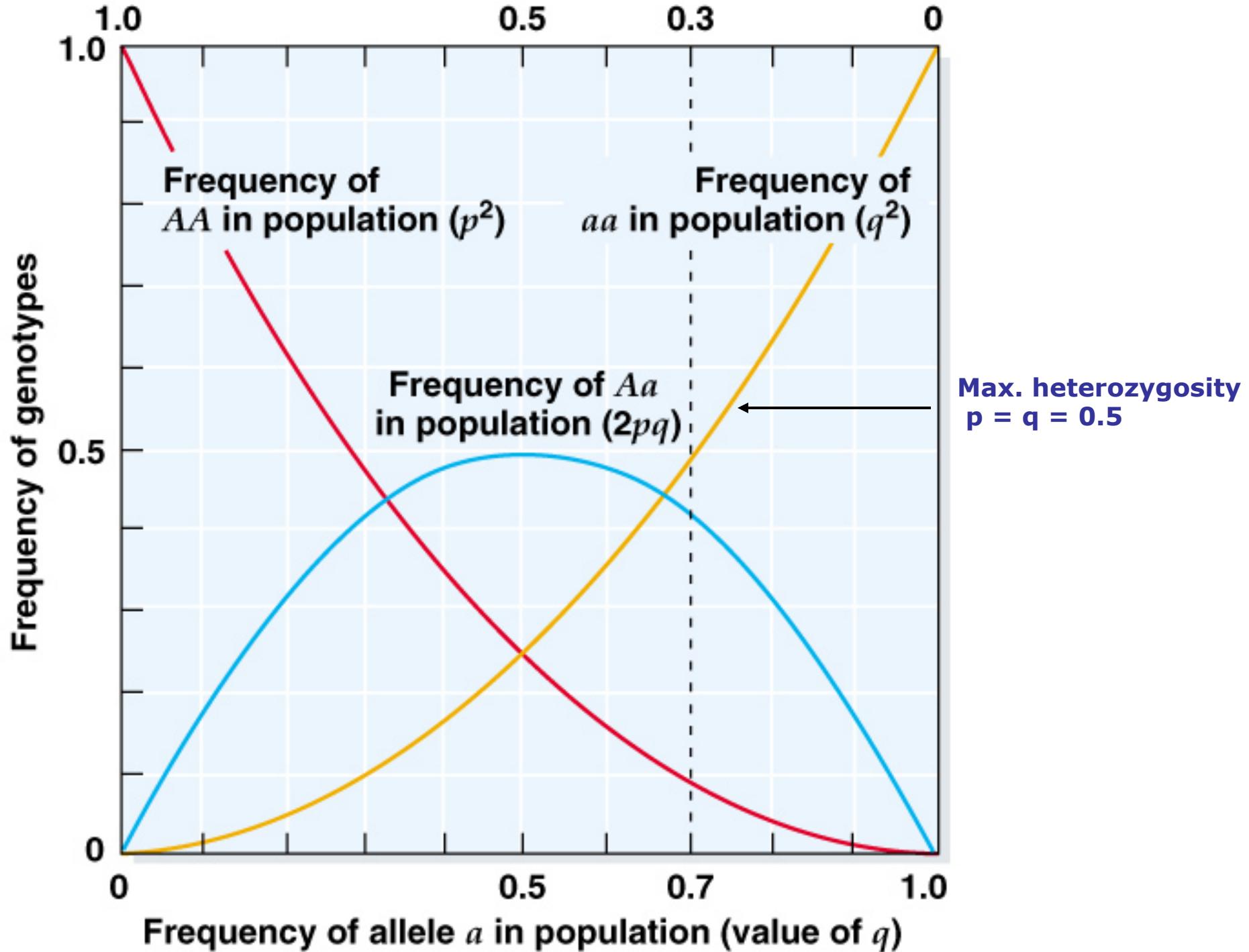
1. Zygotes form by random combinations of alleles, in proportion to the abundance of the alleles in the population
2. If $f(p) = 0.5$ and $f(q) = 0.5$, outcome is as follows:

	A(p)	a(q)
A(p)	$AA(p^2)$ $0.5 \times 0.5 = 0.25$	$Aa(pq)$ $0.5 \times 0.5 = 0.25$
a(q)	$Aa(pq)$ $0.5 \times 0.5 = 0.25$	$aa(q^2)$ $0.5 \times 0.5 = 0.25$

3. When population is at equilibrium:

$$p^2 + 2pq + q^2 = 1$$

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$



Assumptions of Hardy-Weinberg:

1. Population is infinitely large

- **Assumption is unrealistic**
- **Large populations are mathematically similar to infinitely large populations**
- **Finite populations with rare mutations, rare migrants, and weak selection generally fit Hardy-Weinberg proportions**

2. Mating is random

- Few organisms mate randomly for all traits or loci
- Hardy-Weinberg applies to any locus for which mating occurs randomly, even if mating is non-random for other loci
- This works because different loci assort independently due to recombination

- 3. No natural selection
- 4. No mutation
- 5. No migration

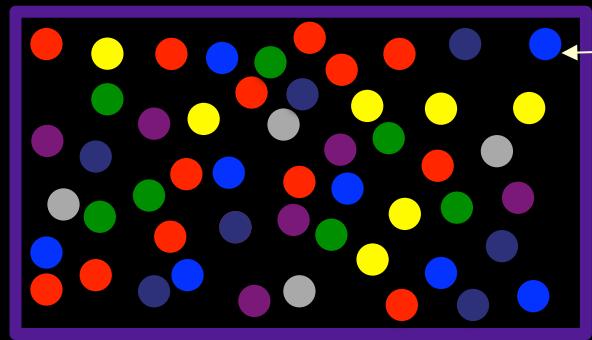
- Gene pool must be closed to the addition/subtraction of new alleles
- Selection subtract alleles or cause some alleles to increase in frequency
- Mutation adds variation (generates novel alleles)
- Effects of mutation can be accommodated with a model (e.g., infinite alleles model)
- Migration can add or subtract variation depending on which alleles migrants carry and whether they immigrate or emigrate

Testing Hardy-Weinberg assumptions:

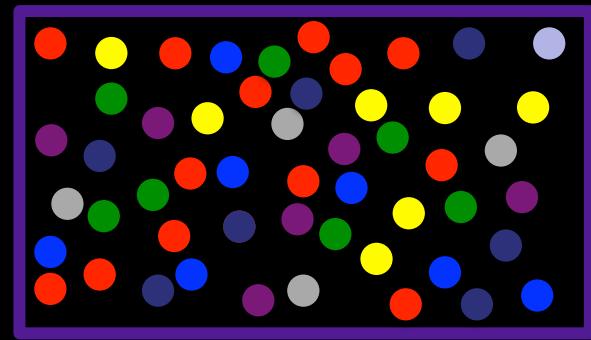
- Data from real populations rarely match Hardy-Weinberg proportions
- Test observed and expected proportions with a goodness of fit (GF) test such as a chi-square test.
- If deviation is larger than expected, begin to determine which assumptions are violated (this is where the real work of population genetics begins).
- Factors that contribute to non-equilibrium:
 - Population differentiation (through drift and mutation)
 - Migration
 - Fluctuations in population size
 - Selection

Mutations change allele frequencies

Pop 1

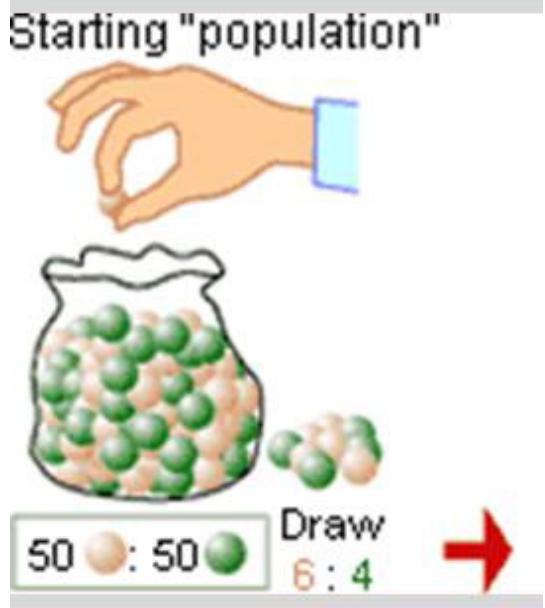


Pop 1 (next generation)



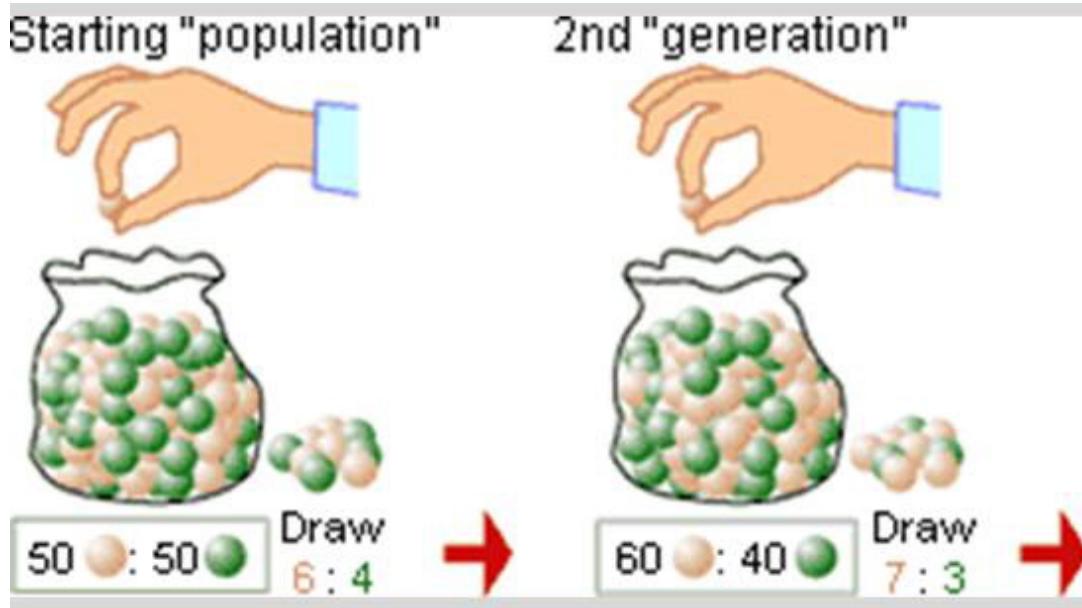
Mutation

Drift also changes allele frequencies (sampling error)

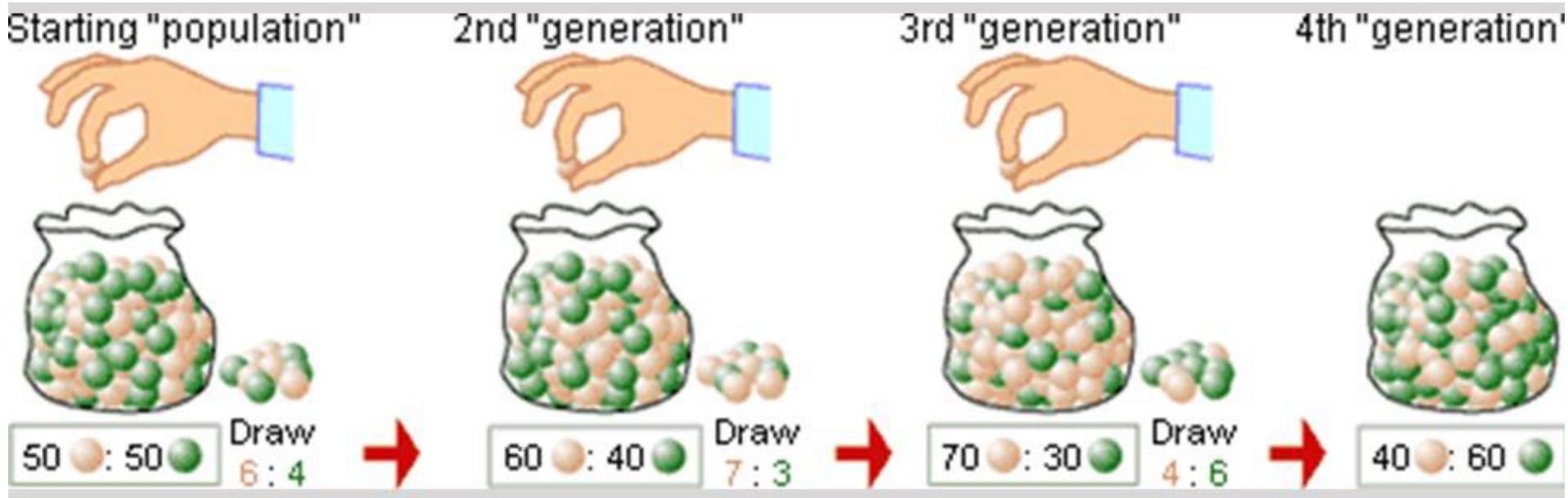


Drift occurs in finite populations, even with completely random mating, and leads to changes in *both* genotype and allele frequencies.

Drift changes allele frequencies (sampling error)



Drift changes allele frequencies (sampling error)



Mutation and Drift change allele frequencies



Mutation and Drift change allele frequencies



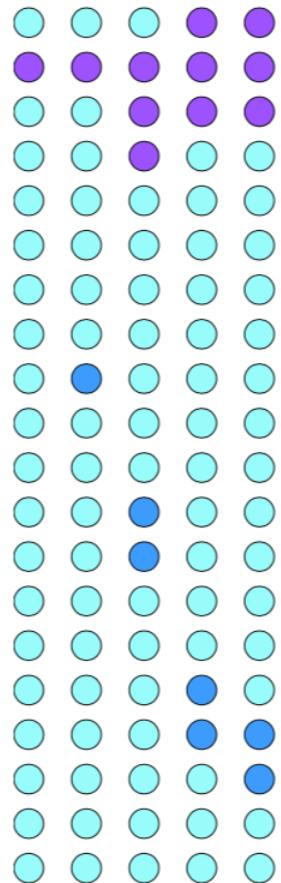
Mutation and Drift change allele frequencies



Mutation and Drift change allele frequencies



Mutation and Drift change allele frequencies



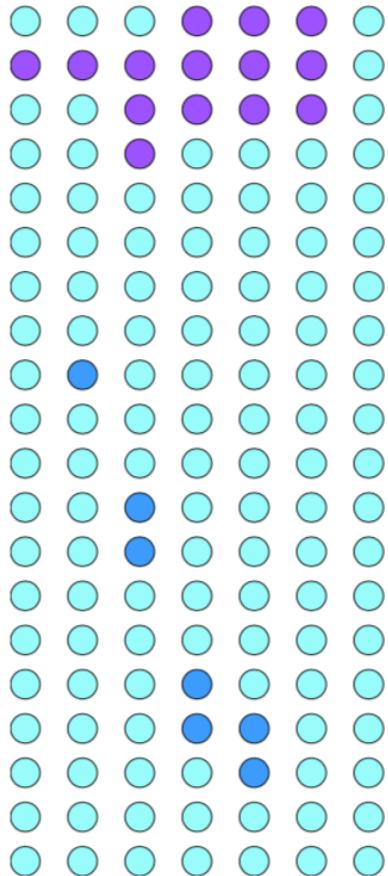
past

present

Mutation and Drift change allele frequencies



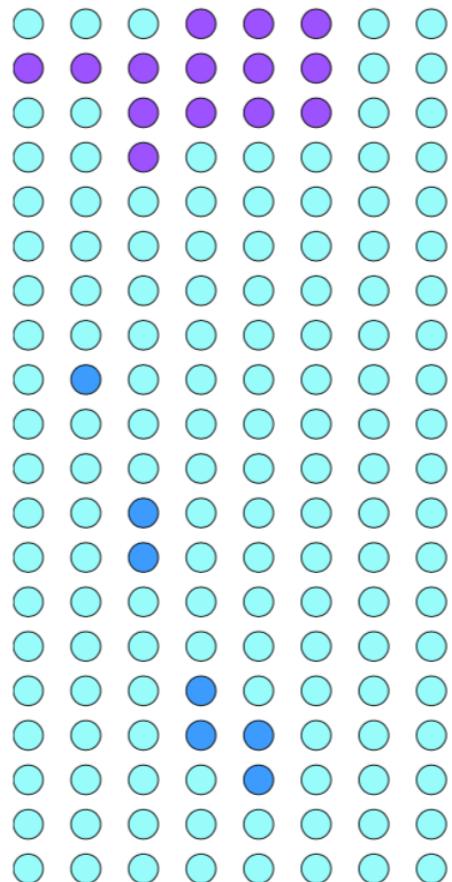
Mutation and Drift change allele frequencies



past

present

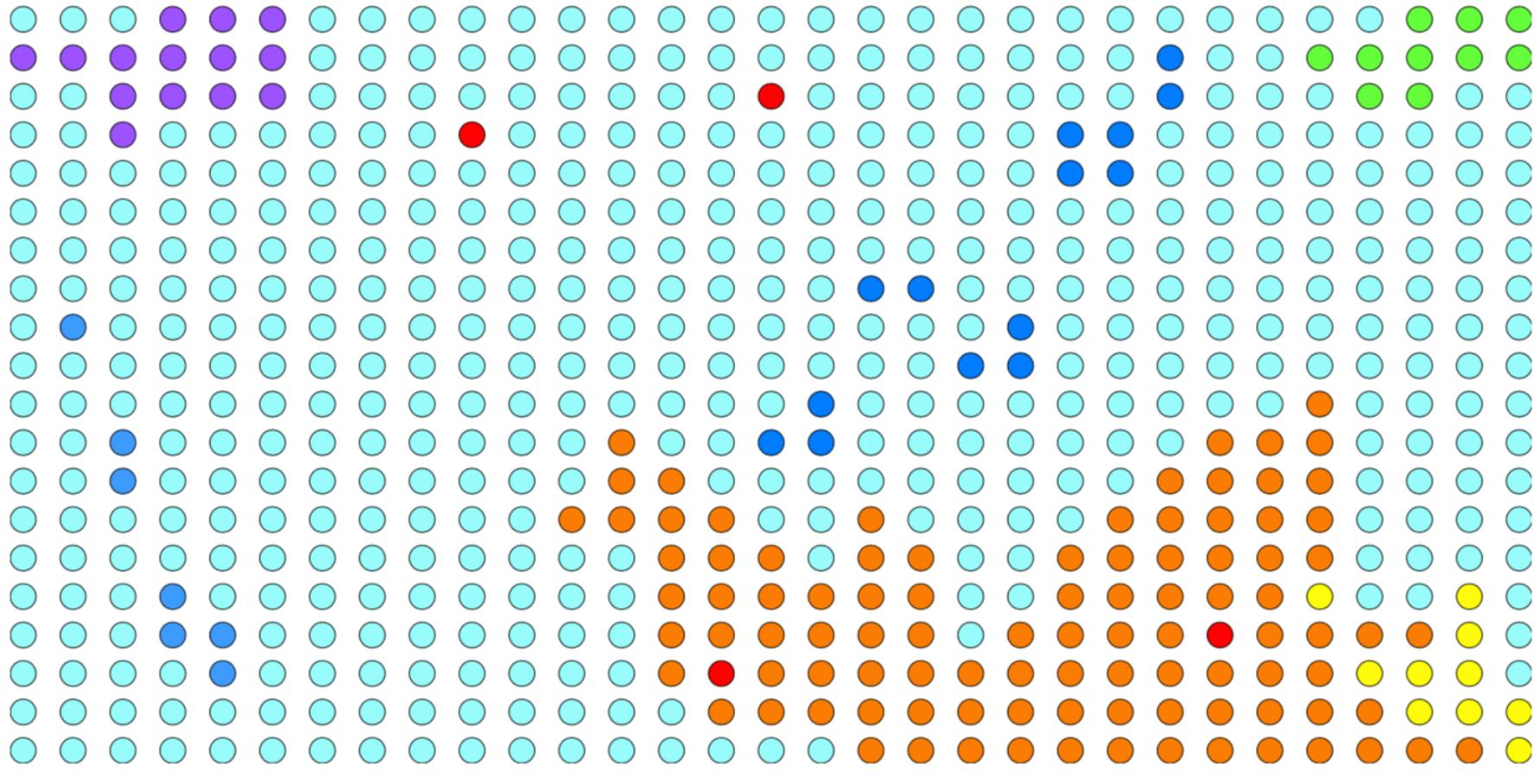
Mutation and Drift change allele frequencies



past

present

Mutation and Drift change allele frequencies



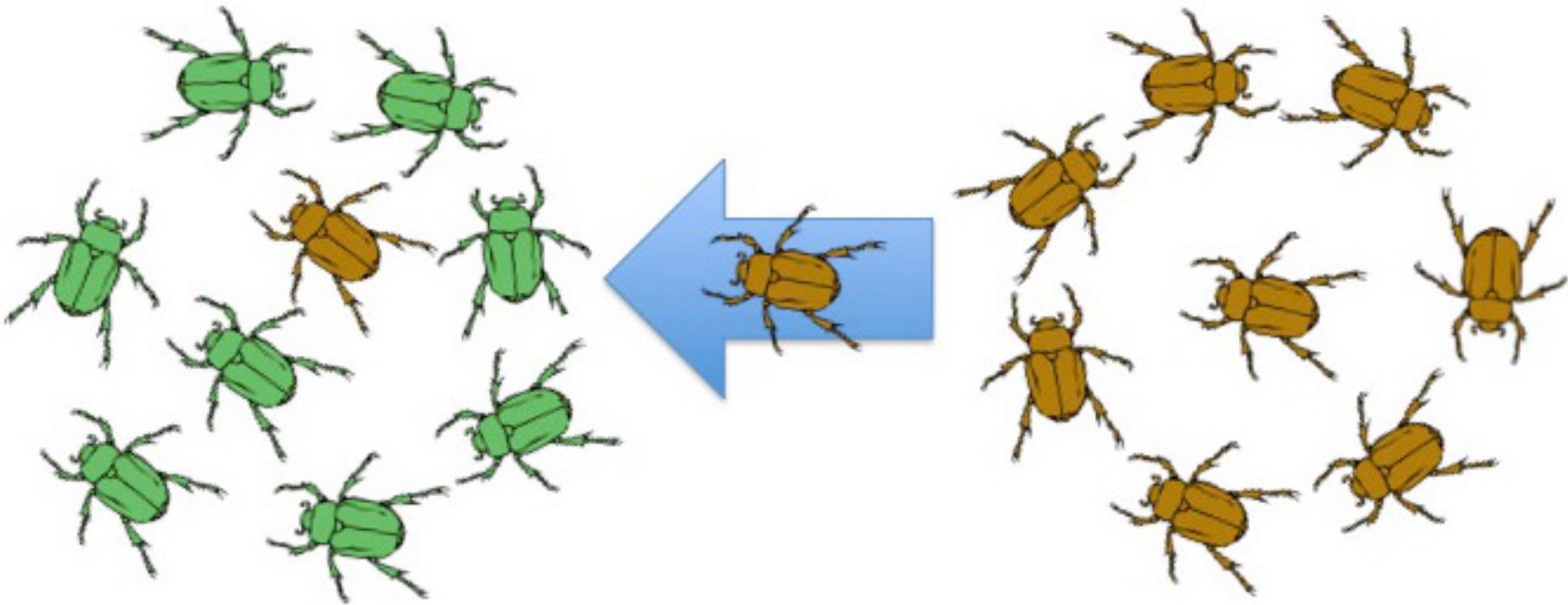
past

present

<https://phytools.shinyapps.io/drift-selection/>

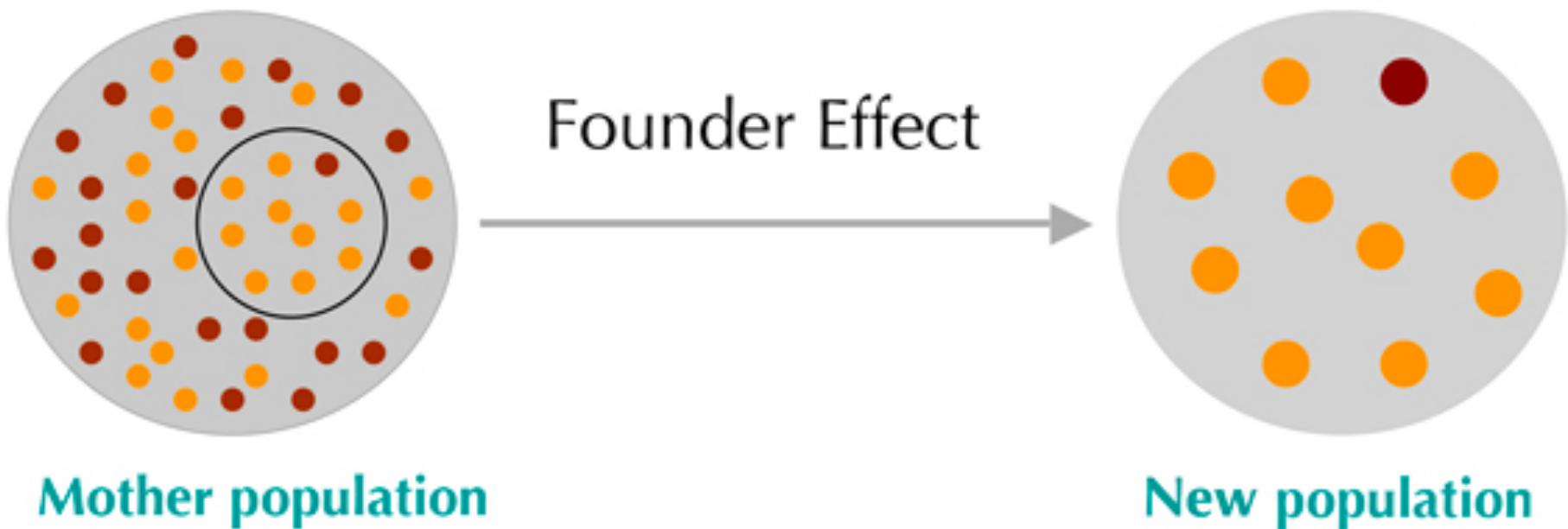
<https://phytools.shinyapps.io/drift-selection/>

Gene flow (migration) also changes allele frequencies

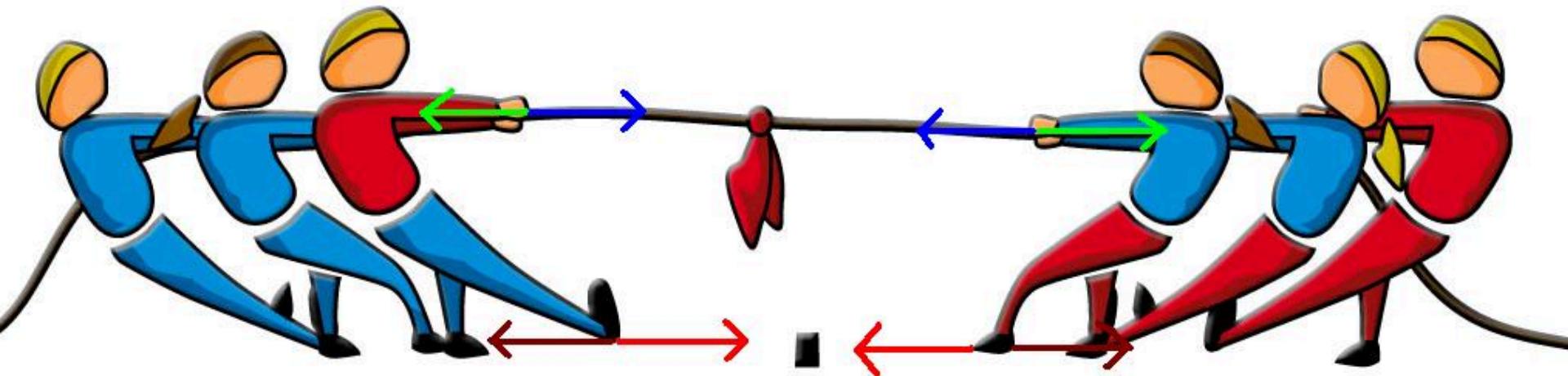


Migration refers to the movement of some organisms (or their gametes) among subpopulations and subsequent breeding, also referred to as gene flow, or admixture. Migration acts to homogenize populations, or reduce differences between populations

Variations in population size also changes allele frequencies



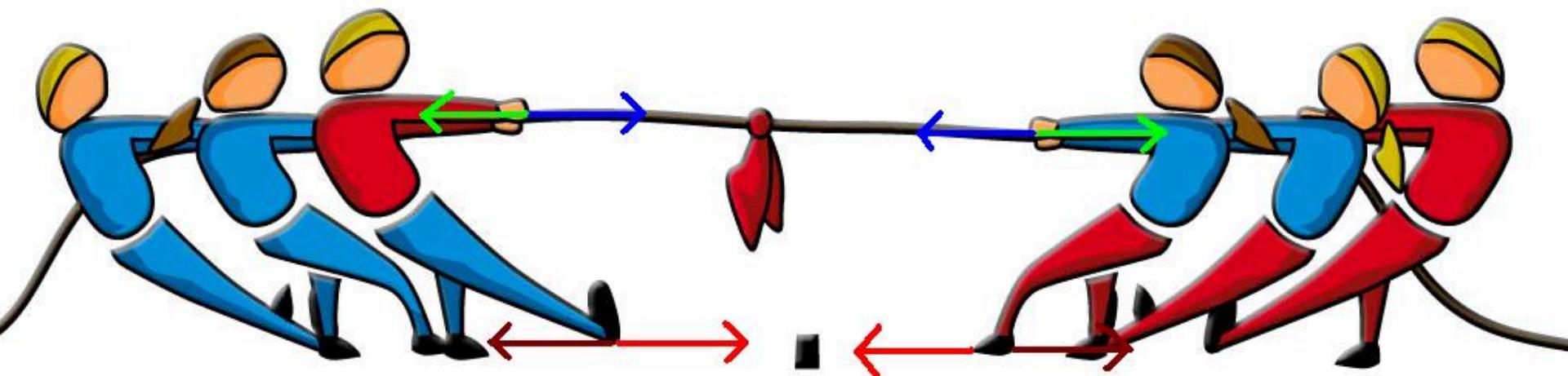
Balance between Drift and Migration



Balance between

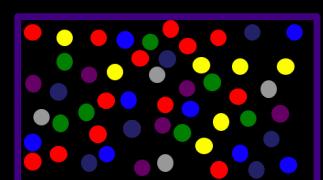
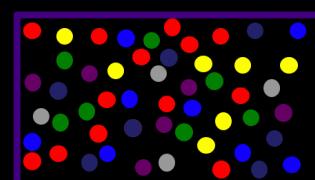
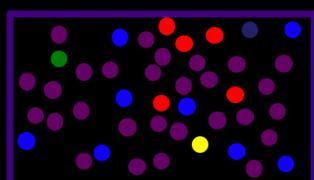
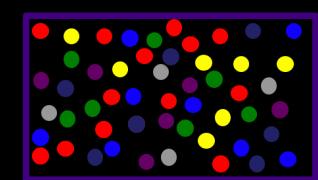
Drift

Migration



Genetic structure
(subpopulations)

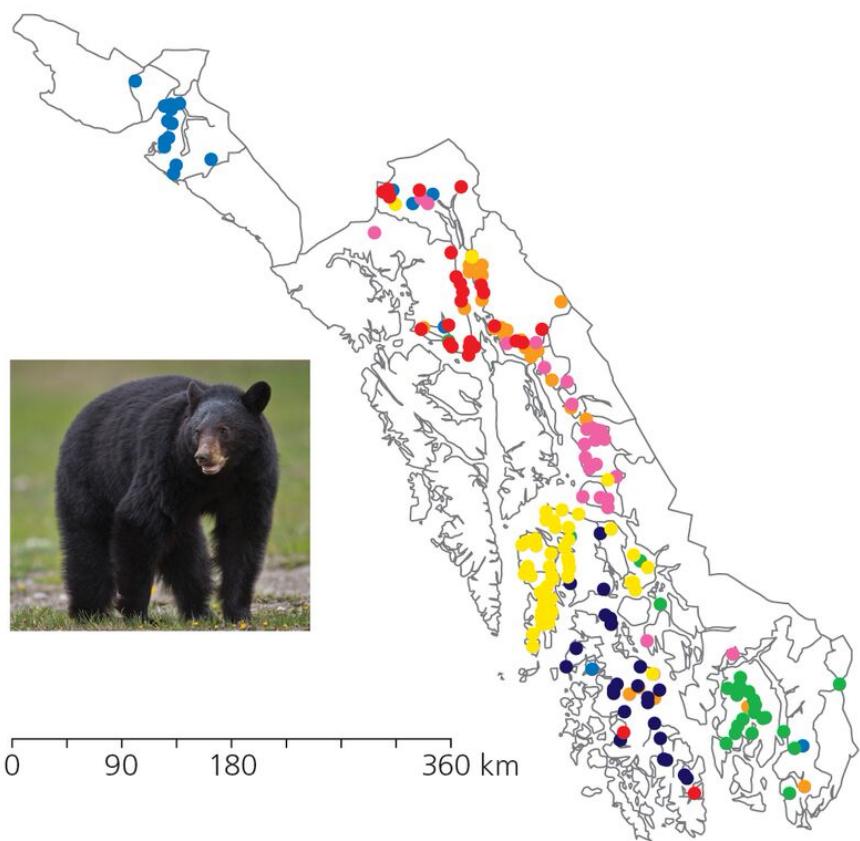
Genetic homogeneity
Panmixia (single population)



<https://phytools.shinyapps.io/drift-selection/>

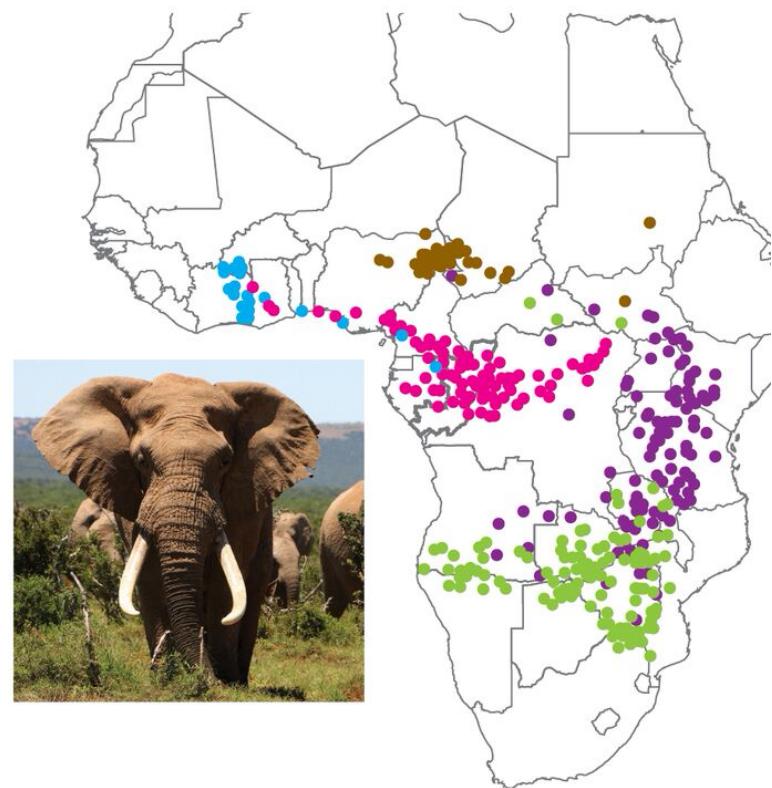
Subdivided populations show distinct genetic structure

A



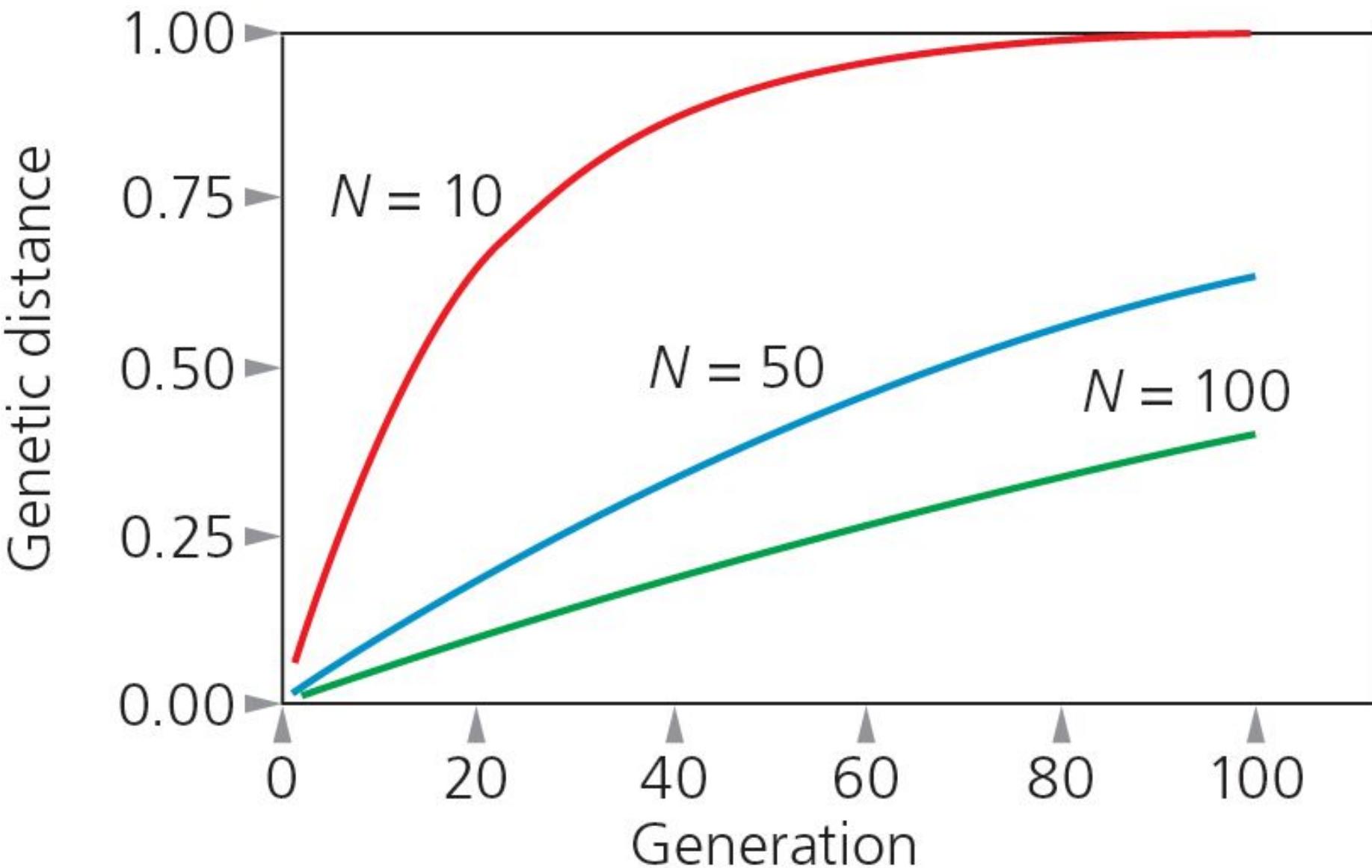
Black Bears in SE Alaska

B



African Elephants

Isolated populations become genetically distinct



Methods to detect population differentiation and gene flow

Measuring genetic differentiation among populations

- Useful to partition genetic variation into components:
 - within populations
 - between populations
 - among populations
- Sewall Wright's Fixation index (F_{ST}) is a useful index of genetic differentiation and comparison of overall effect of population substructure
 - Measures reduction in heterozygosity (H)

Measuring genetic differentiation among populations

- Useful to partition genetic variation into components:
 - within populations
 - between populations
 - among populations
- Sewall Wright's Fixation index (F_{ST}) is a useful index of genetic differentiation and comparison of overall effect of population substructure

- Measures reduction in heterozygosity (H)

$$F_{ST} = \frac{H_{\text{Total}} - H_{\text{subpopulation}}}{H_{\text{Total}}}$$

- F_{ST} ranges between minimum of 0 and maximum of 1:

= 0 ⇒ no genetic differentiation

<< 0.5 ⇒ little genetic differentiation

>> 0.5 ⇒ moderate to great genetic differentiation

= 1.0 ⇒ populations fixed for different alleles

- Consider three subpopulations with 2 alleles at frequencies p and q ,

$$p \quad q \quad H_S = 2pq$$

▪ Subpop 1:	0.7	0.3	0.42
▪ Subpop 2:	0.5	0.5	0.50
▪ Subpop 3:	0.3	0.7	0.42

Average $H_S = 0.446$

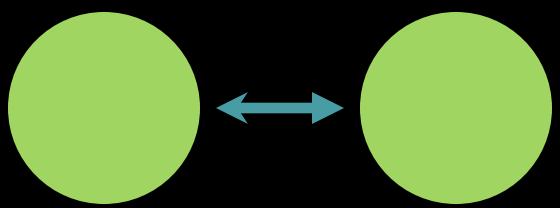
The *total expected heterozygosity* (H_T) across all subpopulations is calculated from the average allele frequency

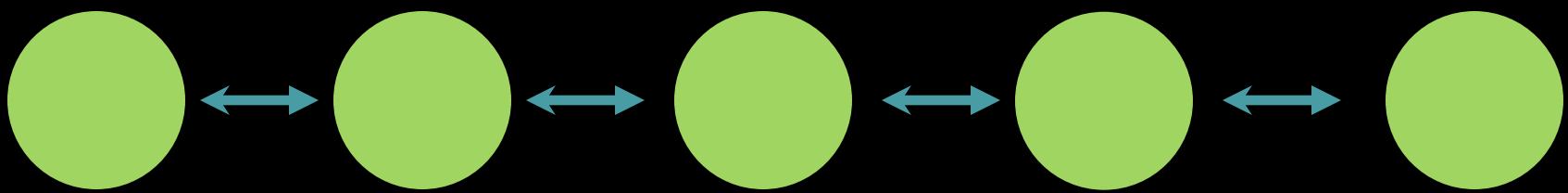
	p	q	
Subpop 1:	0.7	0.3	
Subpop 2:	0.5	0.5	
Subpop 3:	0.3	0.7	
$p = 0.5 \quad q = 0.5$			$H_T = 2pq = 0.5$

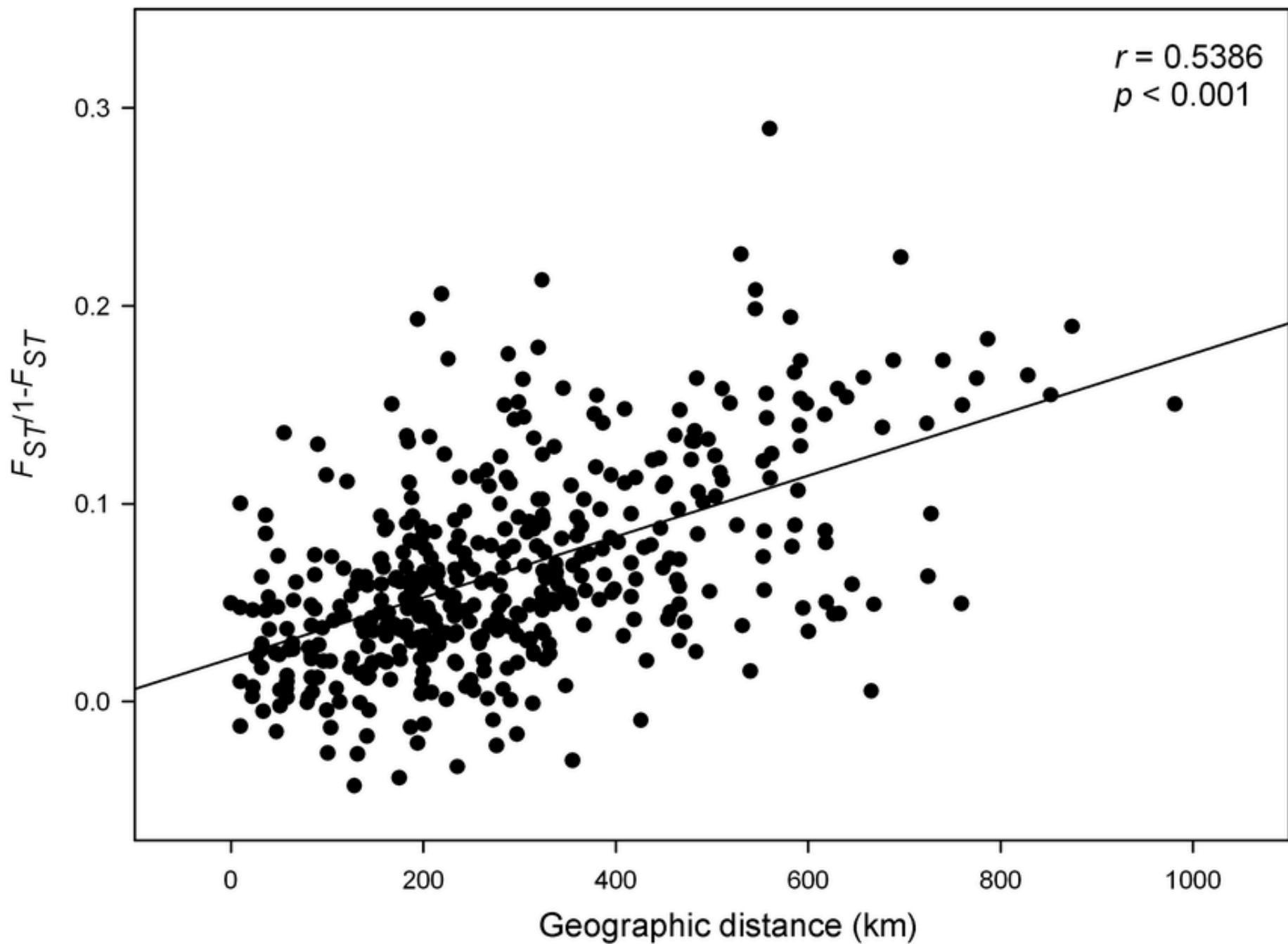
Remember that,

$$F_{ST} = \frac{H_{Total} - H_{subpopulation}}{H_{Total}}$$

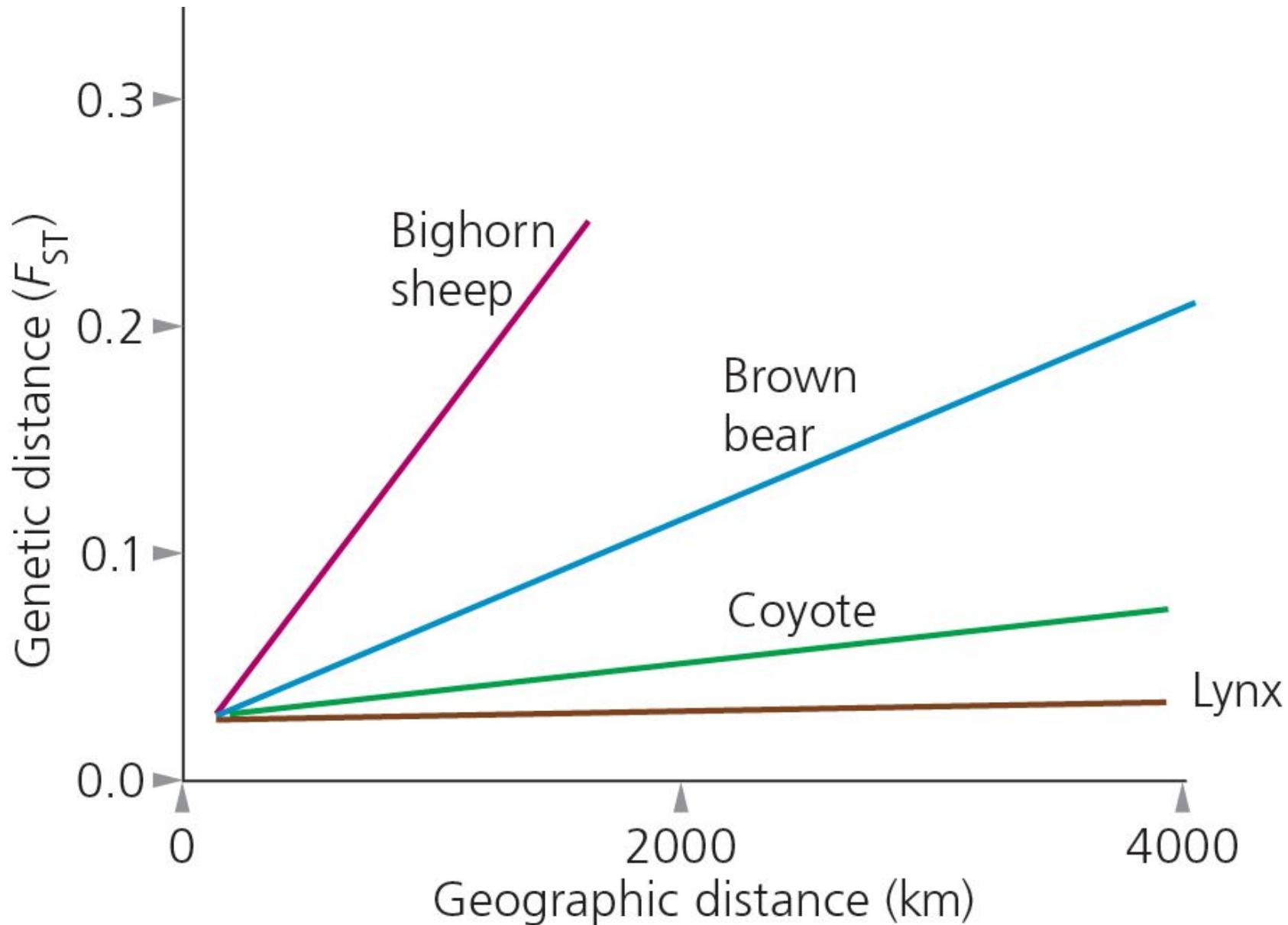
$$F_{ST} = \frac{0.5 - 0.466}{0.5} = 0.11$$







Amount of gene flow varies with the biology of the organism



Demographic Inference Methods

Many approaches!

My overview will be somewhat selective and historical

See Schraiber and Akey (2015)
Nat Rev Genet for a recent review

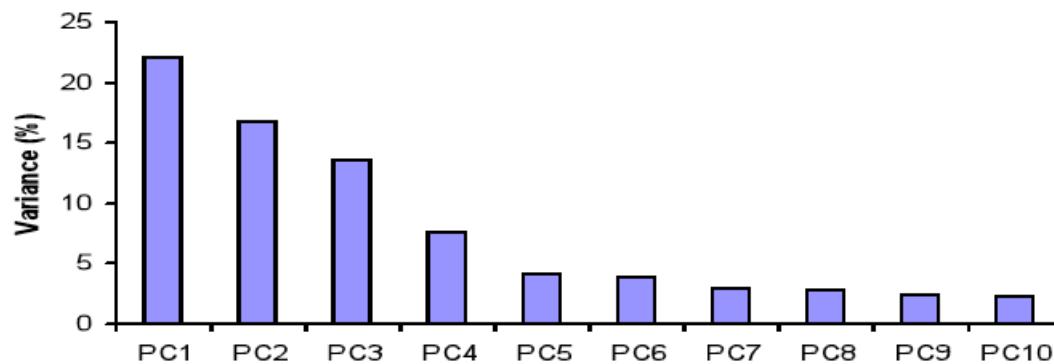
See also

[http://methodspopgen.com/
methods-to-infer-population-
structure/](http://methodspopgen.com/methods-to-infer-population-structure/)

Name	Data type	Inference	Notes	Refs
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User-friendly GUI; can be computationally demanding	32
FRAPPE	Unlinked bi-allelic SNVs	Population structure, admixture	Alexander et al. ⁴¹ argue that convergence is not guaranteed	40
ADMIXTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Estimate the number of populations via cross-validation error	41
fastSTRUCTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Obtains variational Bayesian estimates of posterior probability distribution	42
Structurama	Unlinked multi-allelic genotypes	Population structure, admixture	Uses a Dirichlet process to estimate the number of populations	43
HAPMIX	Phased haplotypes; reference panel	Chromosome painting	Requires populations to be specified <i>a priori</i>	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	49
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Extends the fineSTRUCTURE approach to estimate unsampled ancestral populations and admixture times	7
LAMP	Phased haplotypes; reference panel	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM, so is more discrete than other approaches	52
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by an HMM to estimate local ancestry	53
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62, 63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristics search; redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71, 72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in PSMC's PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rates should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89, 92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP ²¹	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97, 98

Dimensionality Reduction

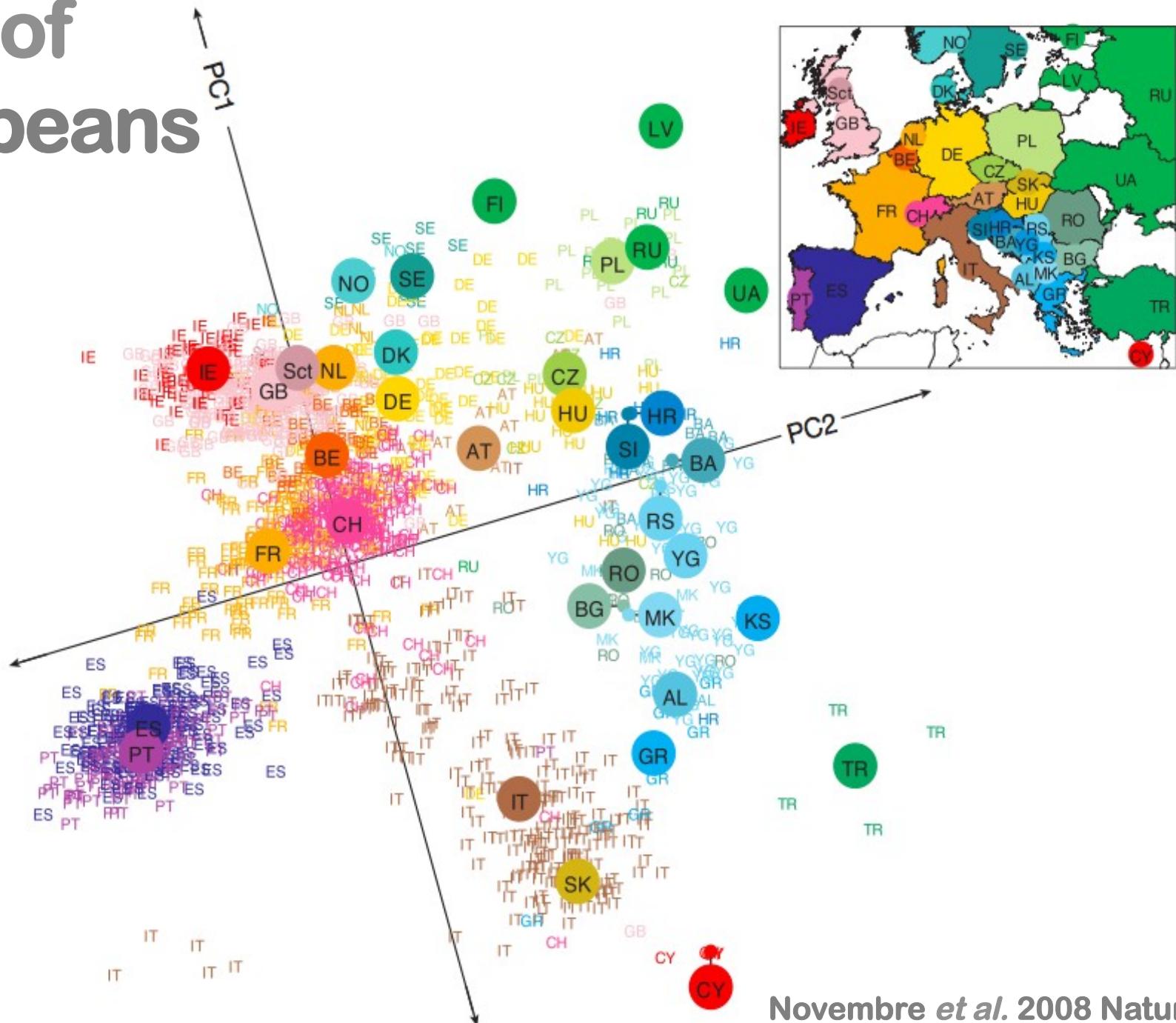
Can *ignore* the components of lesser significance.



You do *lose some information*, but if the eigenvalues are small, you don't lose much

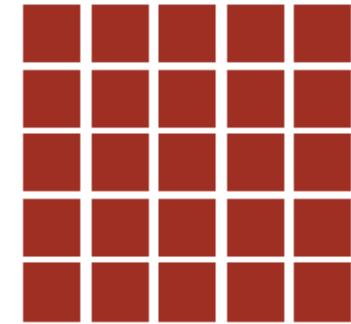
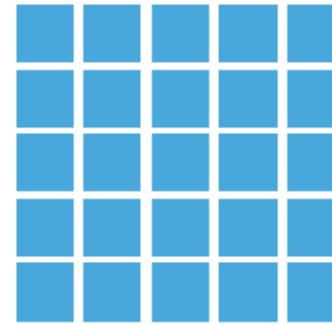
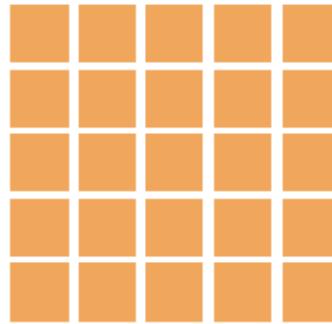
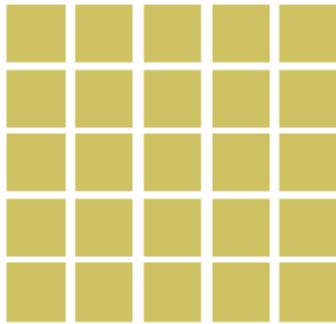
- n dimensions in original data
- calculate n eigenvectors and eigenvalues
- choose only the first p eigenvectors, based on their eigenvalues
- final data set has only p dimensions

PCA of Europeans



Novembre *et al.* 2008 Nature

Mixture Model



Identify **origins** of individuals
each with a **single** ancestry

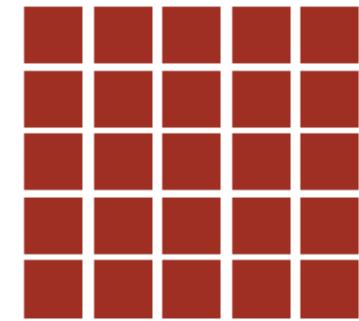
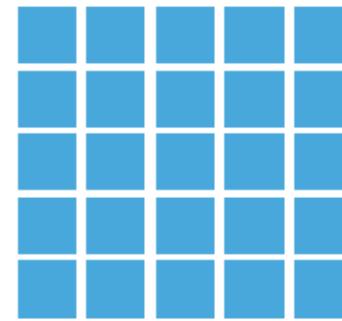
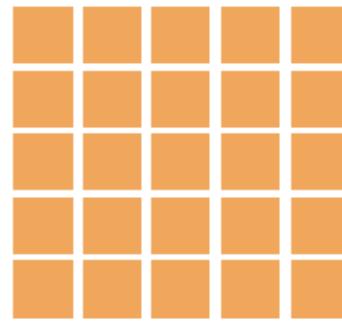
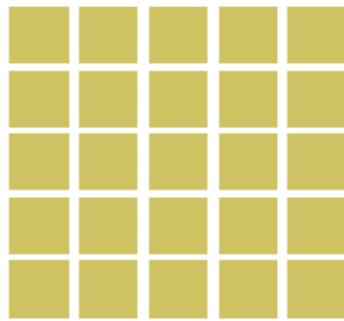
Approaches: Based on Hardy-Weinberg genotype proportions

Admixture Model

Several recent and ongoing genetic studies have focused on admixed populations: populations characterized by ancestry derived from two or more ancestral populations that were reproductively isolated

Admixed populations have arisen in the past several hundred years as a consequence of historical events such as the transatlantic slave trade, the colonization of the Americas and other long-distance migrations

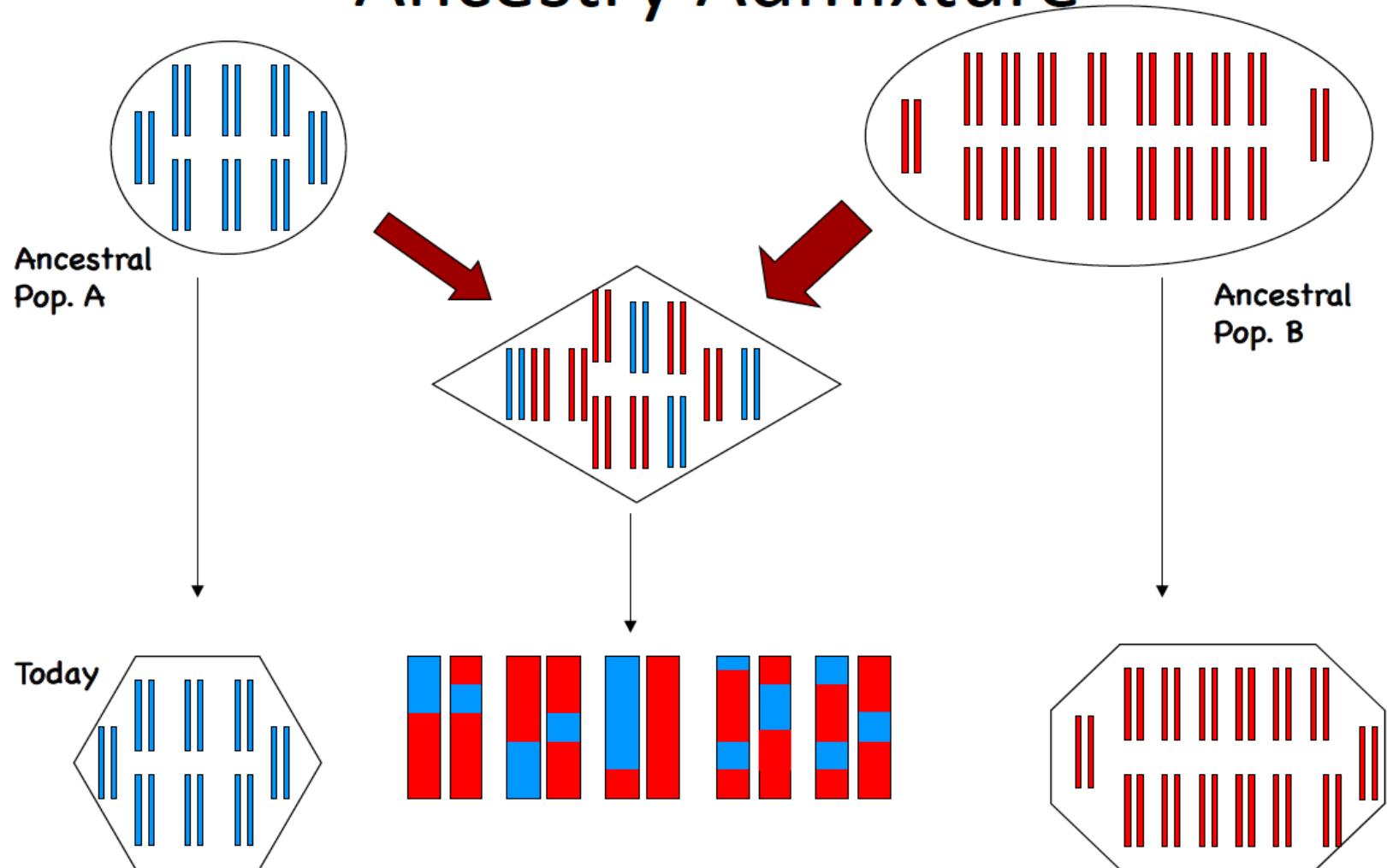
Admixture Model



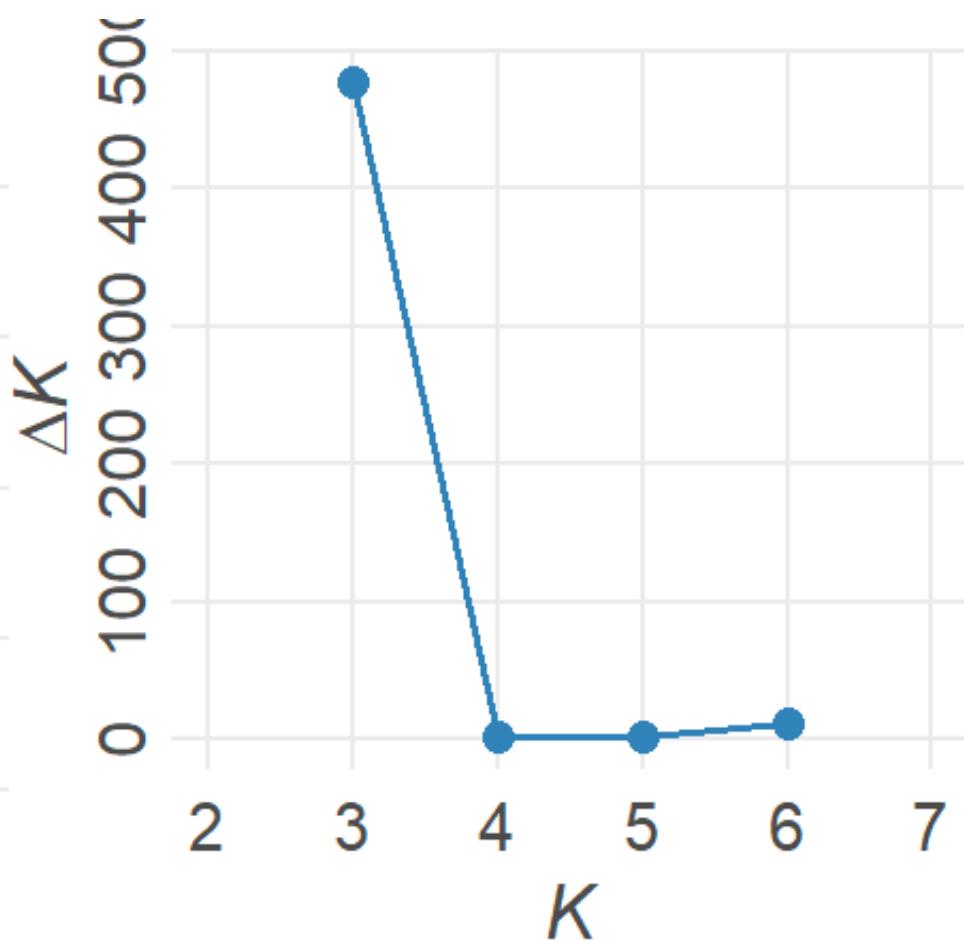
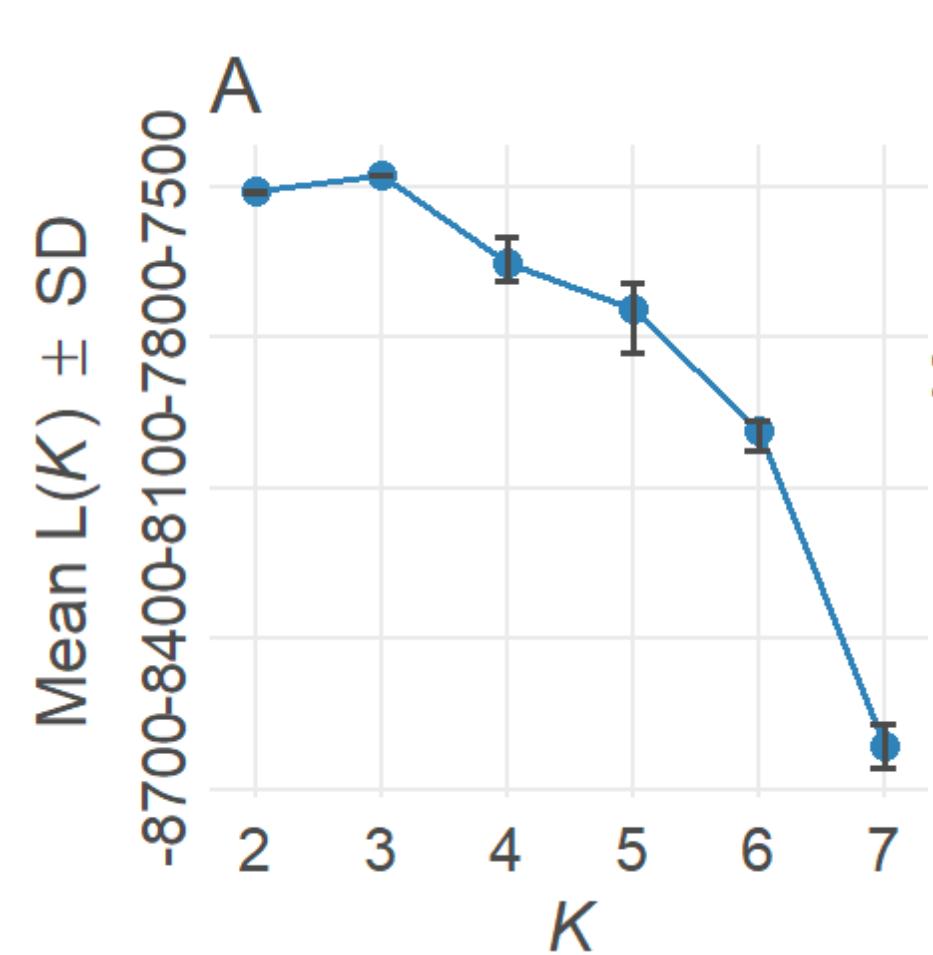
Identify **ancestry proportions** for individuals with **admixed** ancestry

Approaches: Structure (MCMC, Bayesian)
Or ADMIXTURE (quadratic programming)

Ancestry Admixture

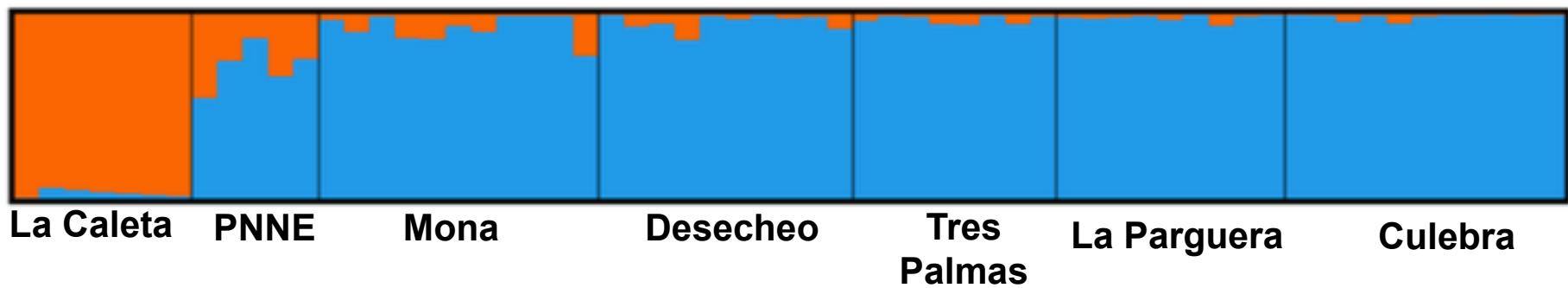


- ▶ The chromosomes of an admixed individual represent a mosaic of chromosomal blocks from the ancestral populations.



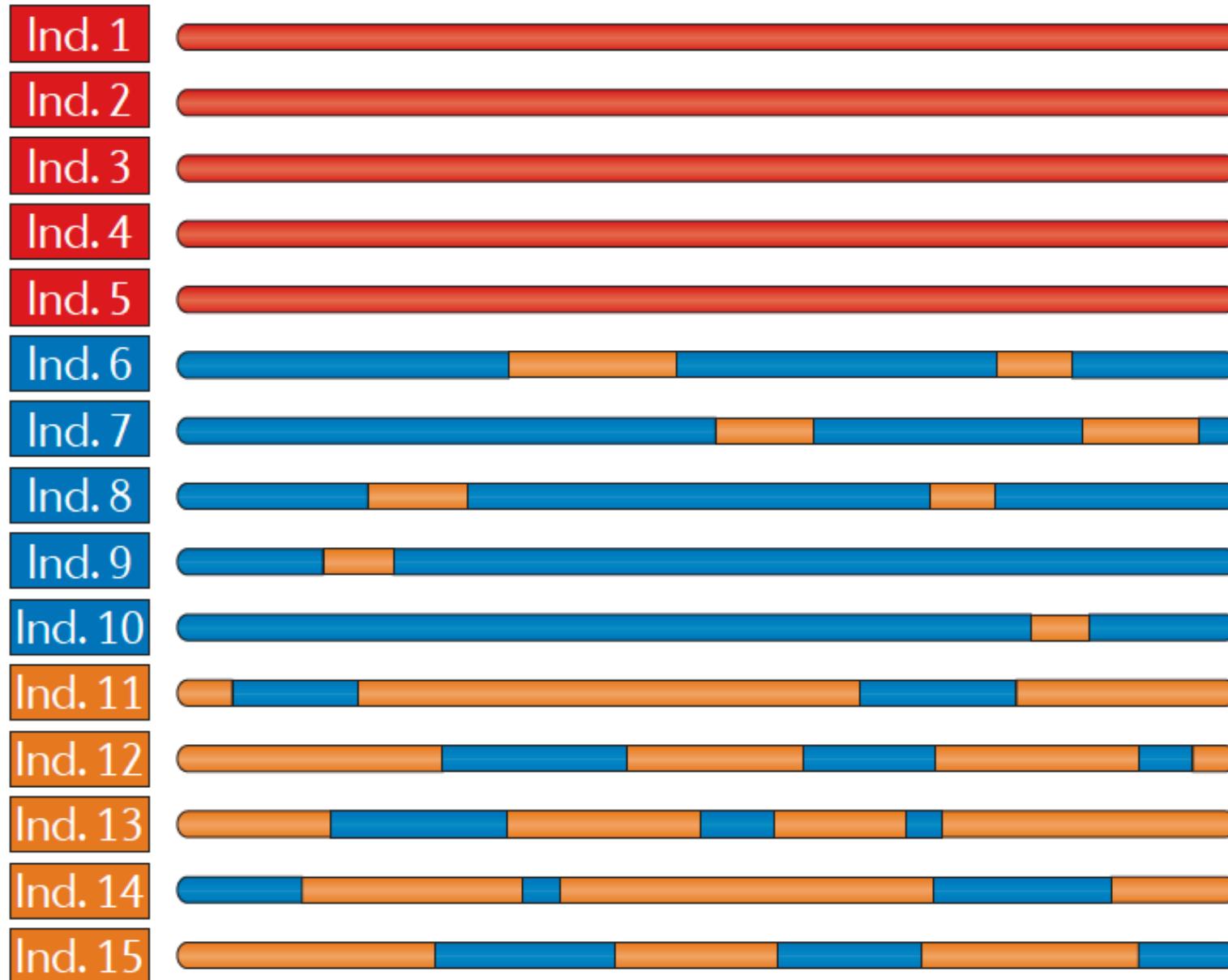
10,059 SNPs

Probability of Membership



Chromosome painting

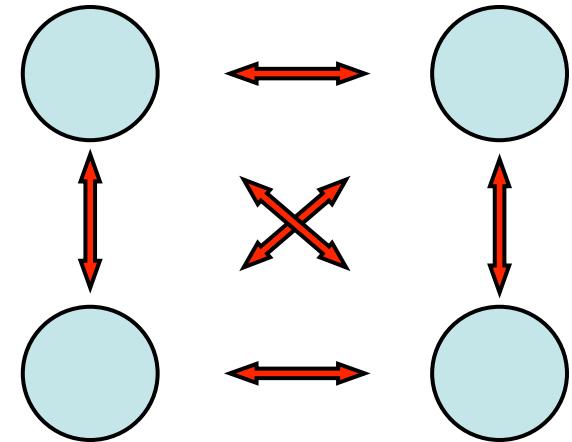
FineSTRUCTURE, HapMix, Globetrotter



Methods to estimate migration

RELATIONSHIP BETWEEN F_{ST} AND Nm IN THE ISLAND MODEL

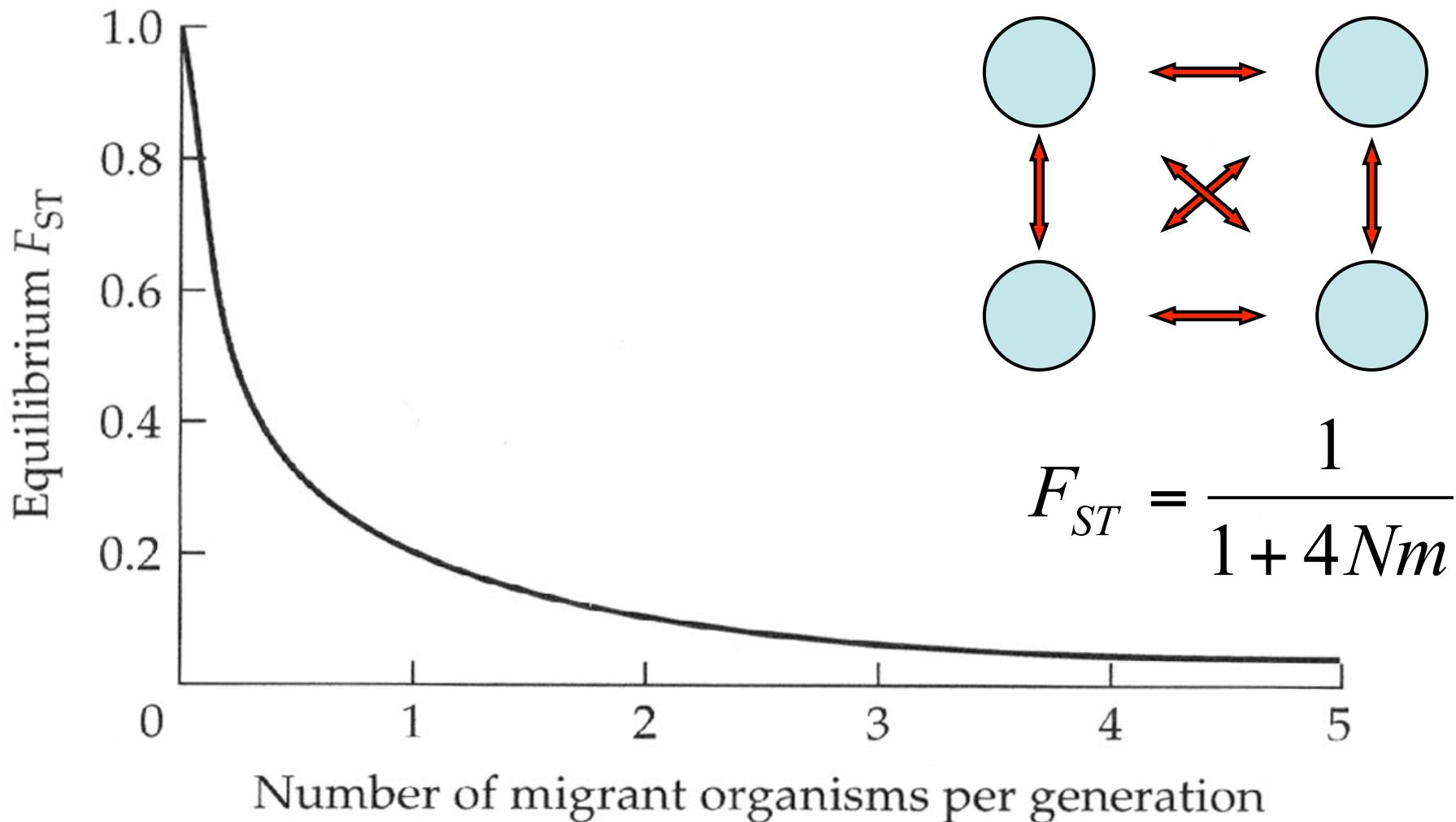
Nm is the absolute number of migrant organisms that enter each subpopulation per generation



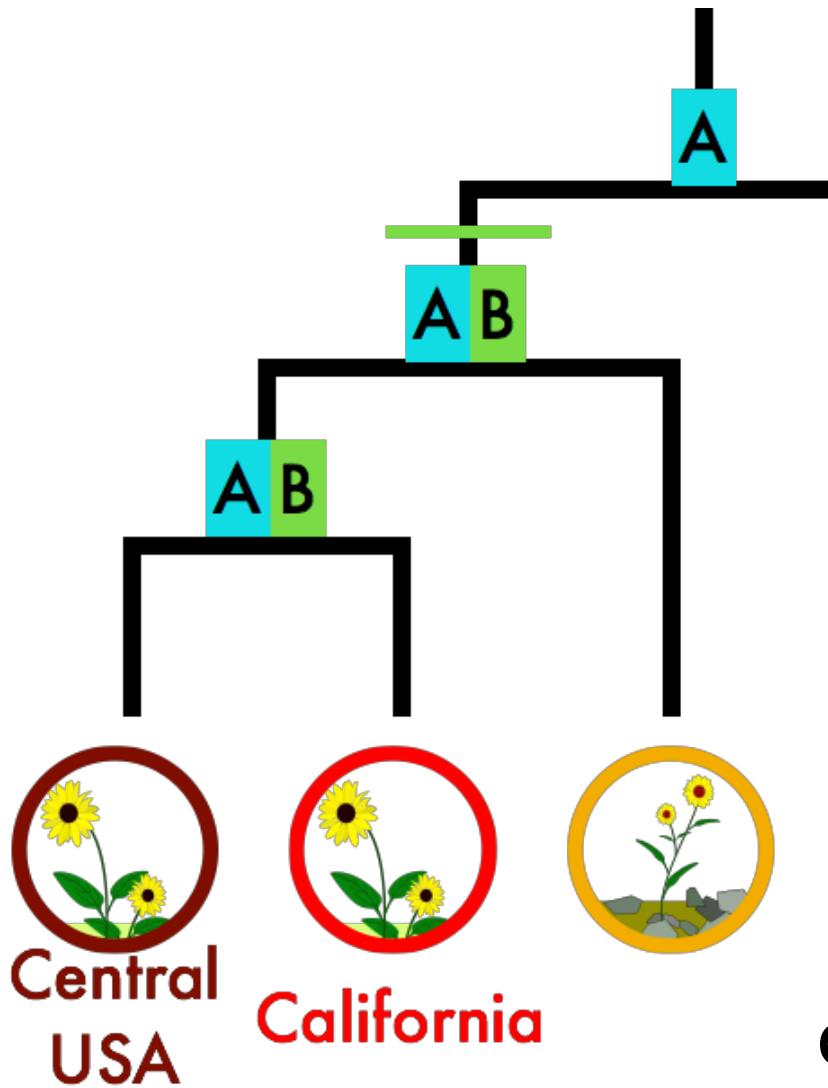
$$F_{ST} = \frac{1}{1 + 4Nm}$$

RELATIONSHIP BETWEEN F_{ST} AND Nm IN THE ISLAND MODEL

Nm is the absolute number of migrant organisms that enter each subpopulation per generation

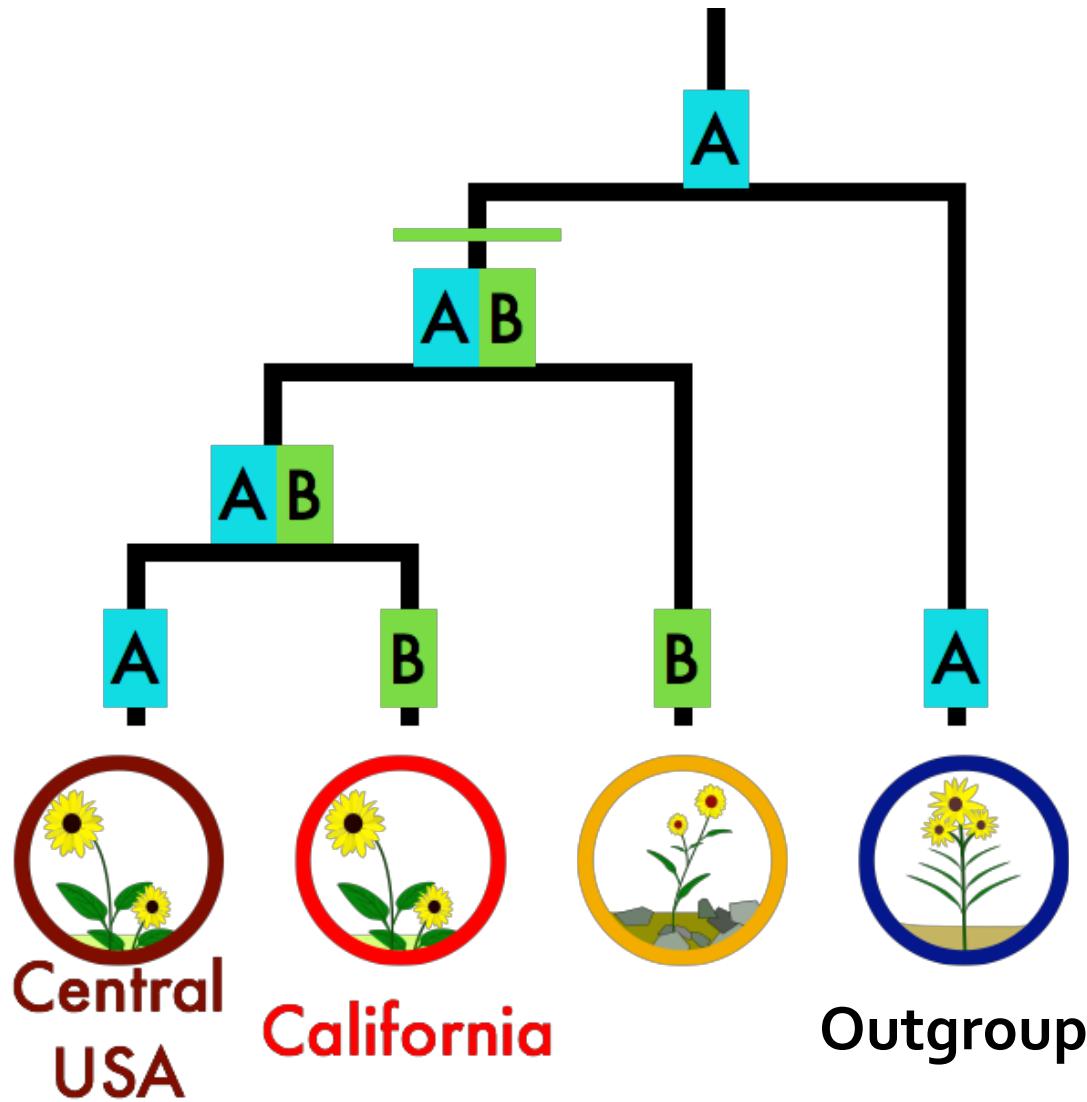


ABBA-BABA test



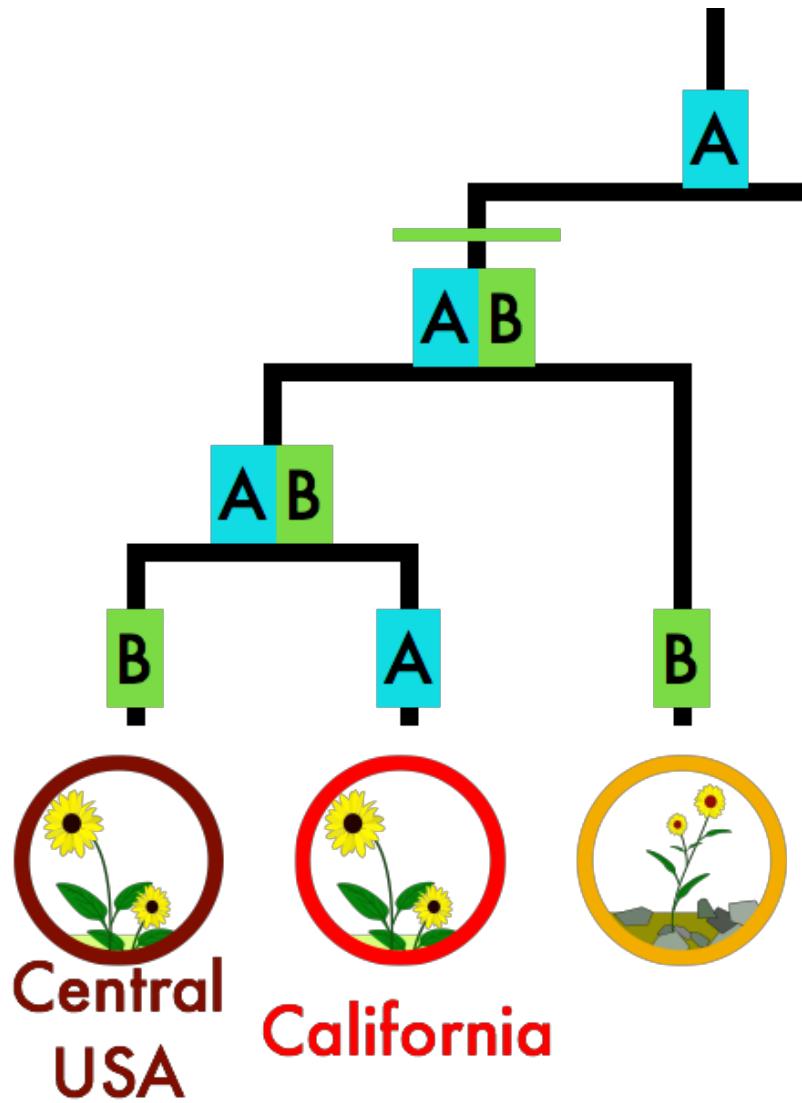
A = Ancestral
B = Derived

ABBA-BABA test



A = Ancestral
B = Derived

ABBA-BABA test



A = Ancestral
B = Derived

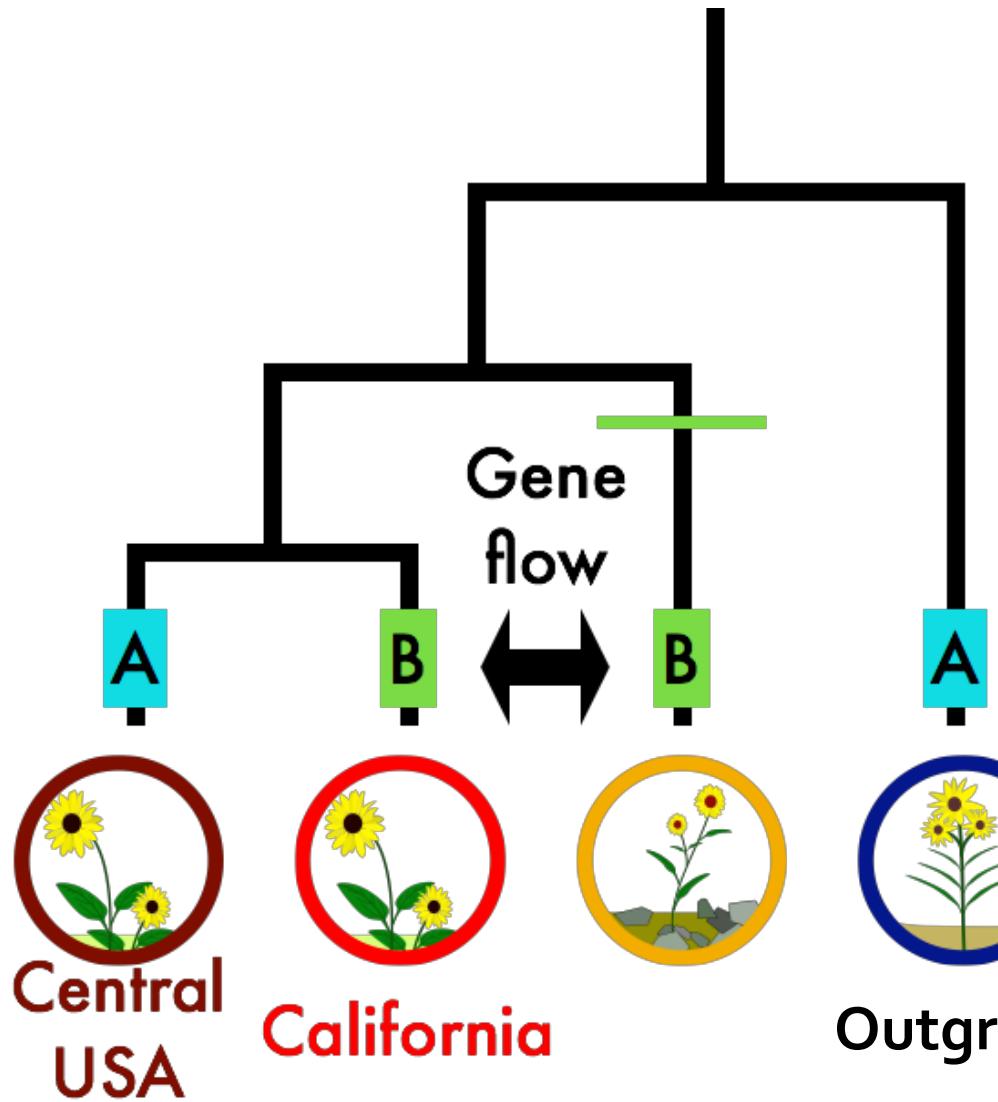
ABBA-BABA test

D-statistic = #ABBA - #BABA

Neutrality: D = 0

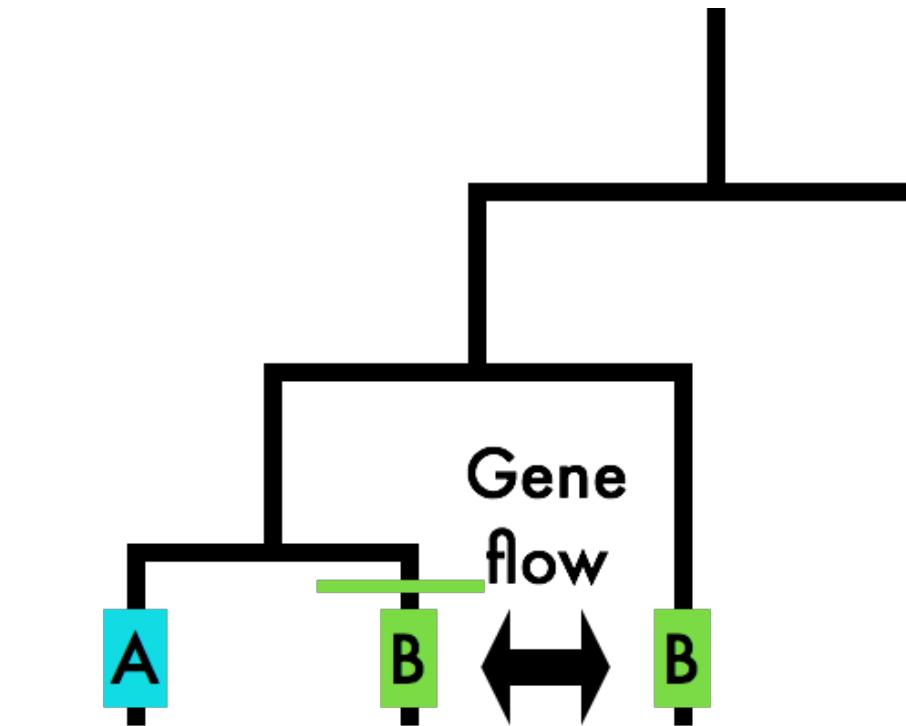
Gene flow: D = +

ABBA-BABA test



A = Ancestral
B = Derived

ABBA-BABA test



A = Ancestral
B = Derived



Central
USA



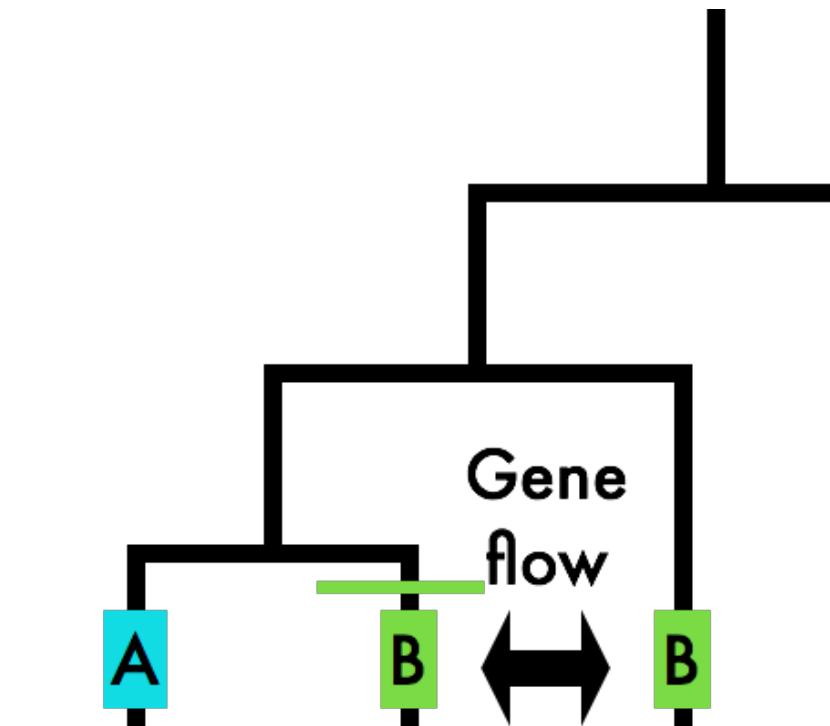
California



Outgroup



ABBA-BABA test



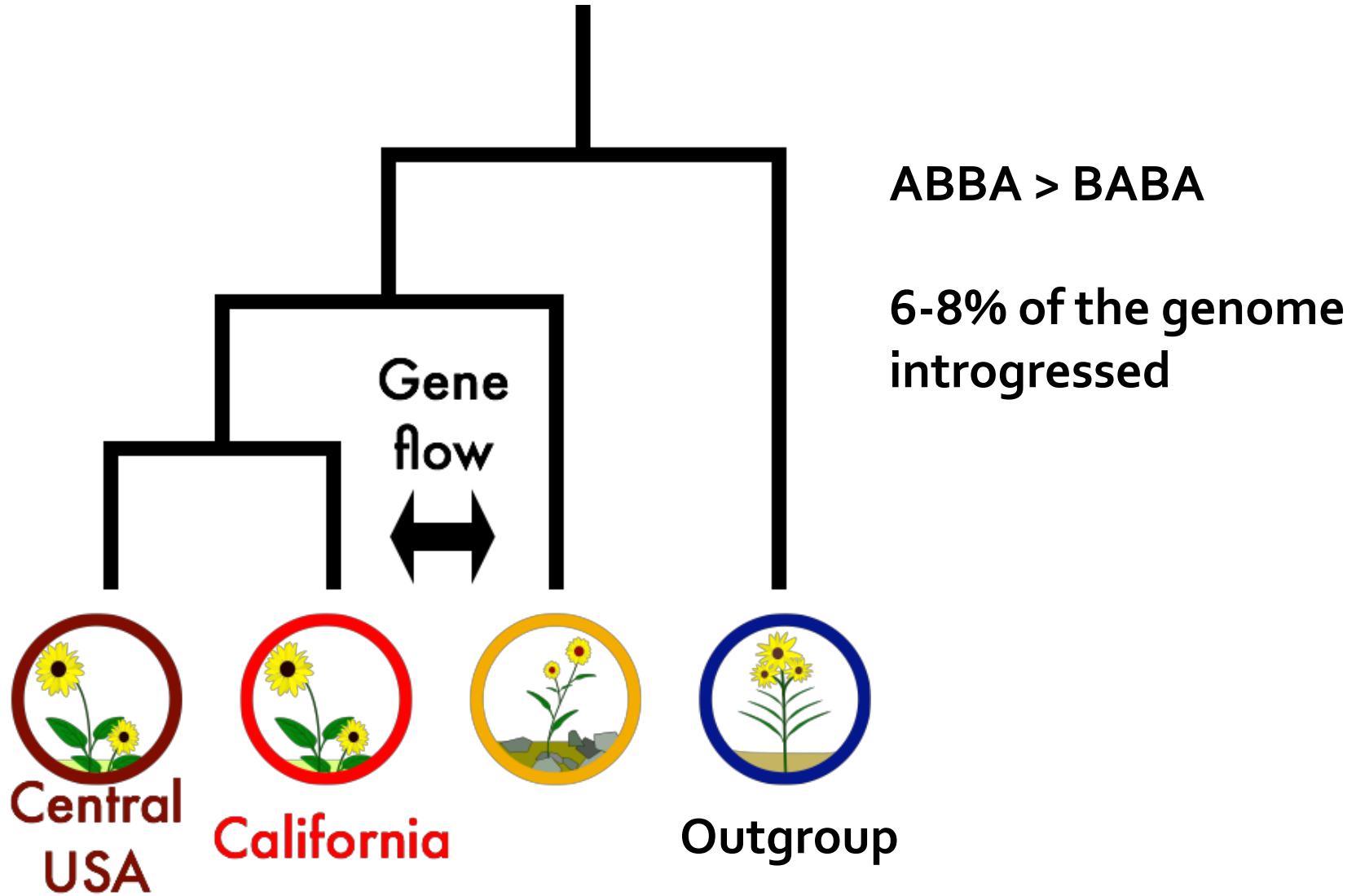
A = Ancestral
B = Derived

Central
USA

California

Outgroup

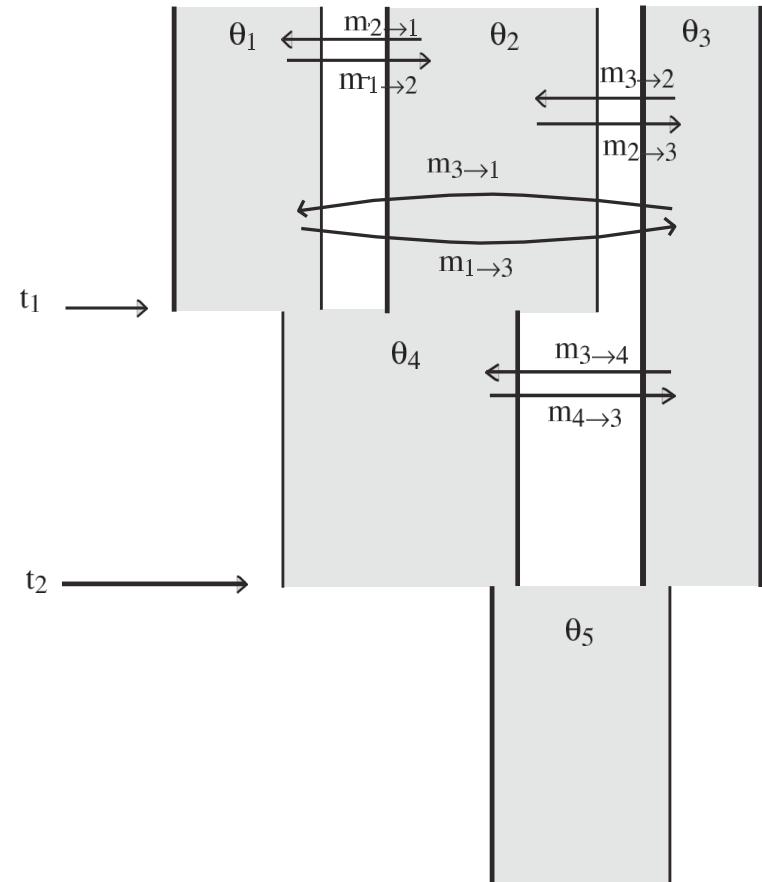
$$D \text{ statistic} = 0.123 \pm 0.033$$



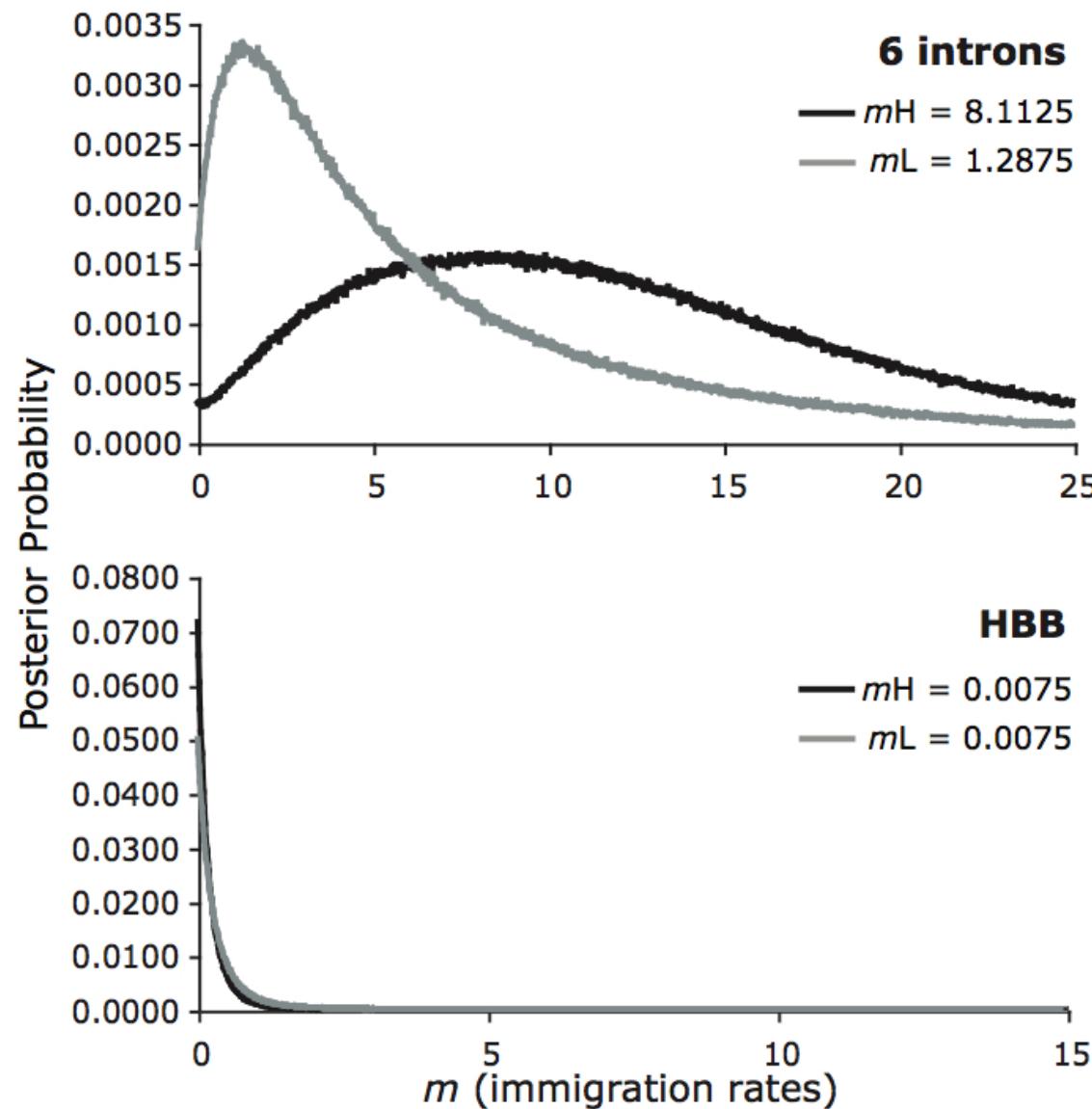
Methods to estimate both structure and migration

IM/IMa/IMa2

- **Uses coalescent simulation to calculate the full likelihood of the data given the model, for non-recombining regions (mitochondria, Y chromosome, small autosomal regions).**
- **Bayesian inference based on MCMC walk through parameter space, can be computationally expensive.**
- **Handles arbitrary number of populations.**

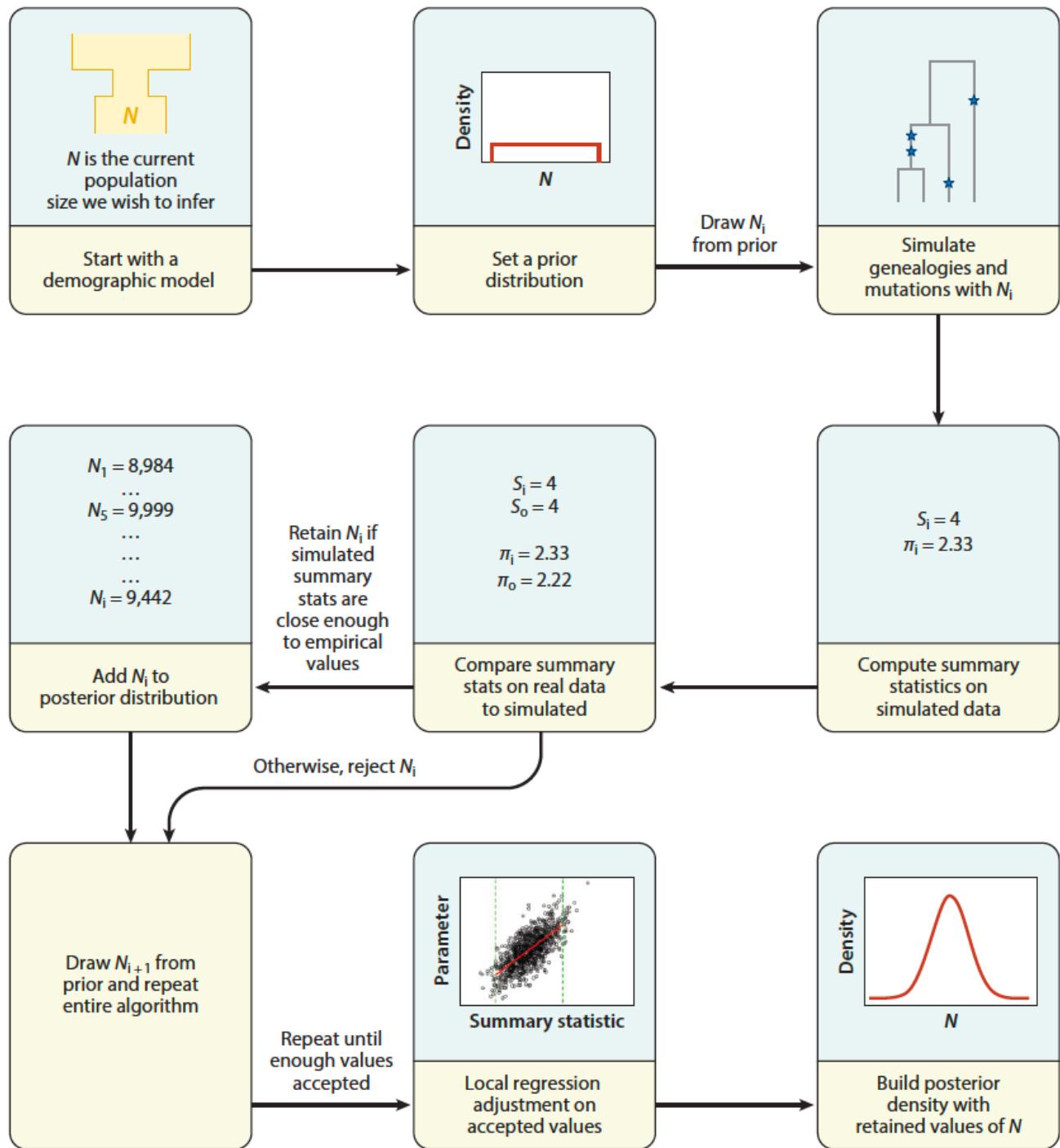


High F_{ST} = 1.00 results in zero migration rate for the β -globin locus but not other loci.

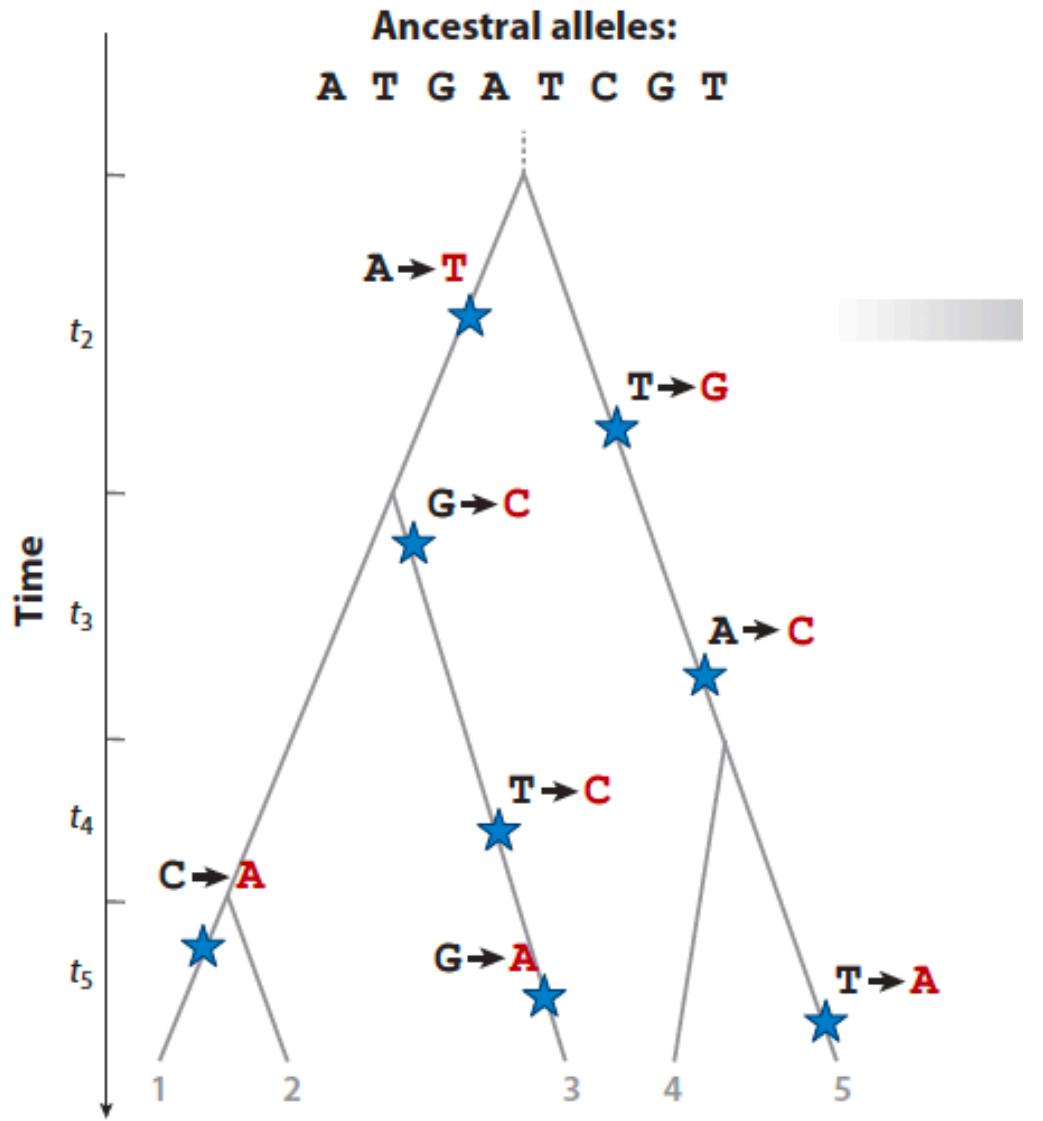


Bulgarella et al. (2012) showing inter-locus contrasts of migration rates in Crested Duck.

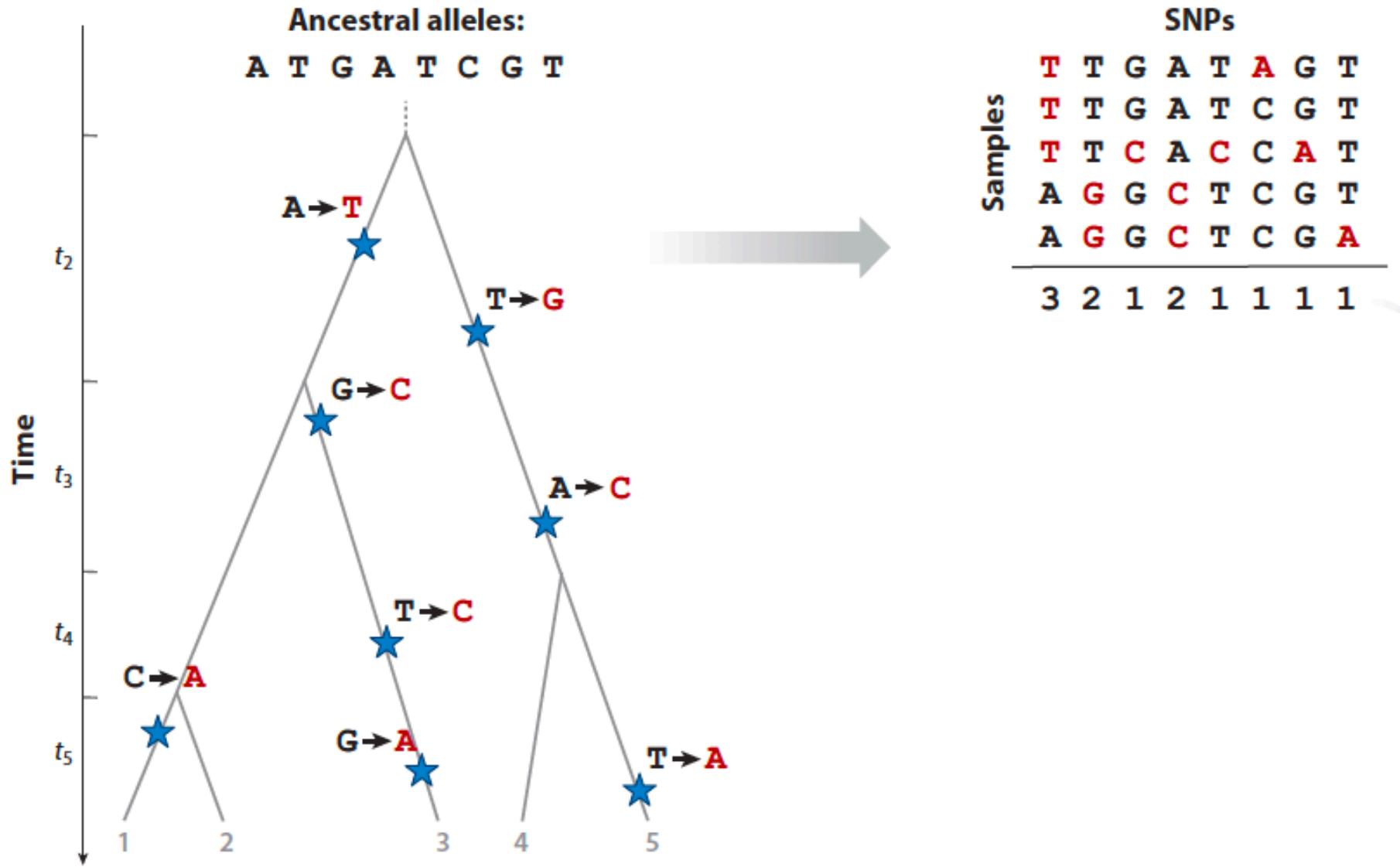
Approximate Bayesian computation



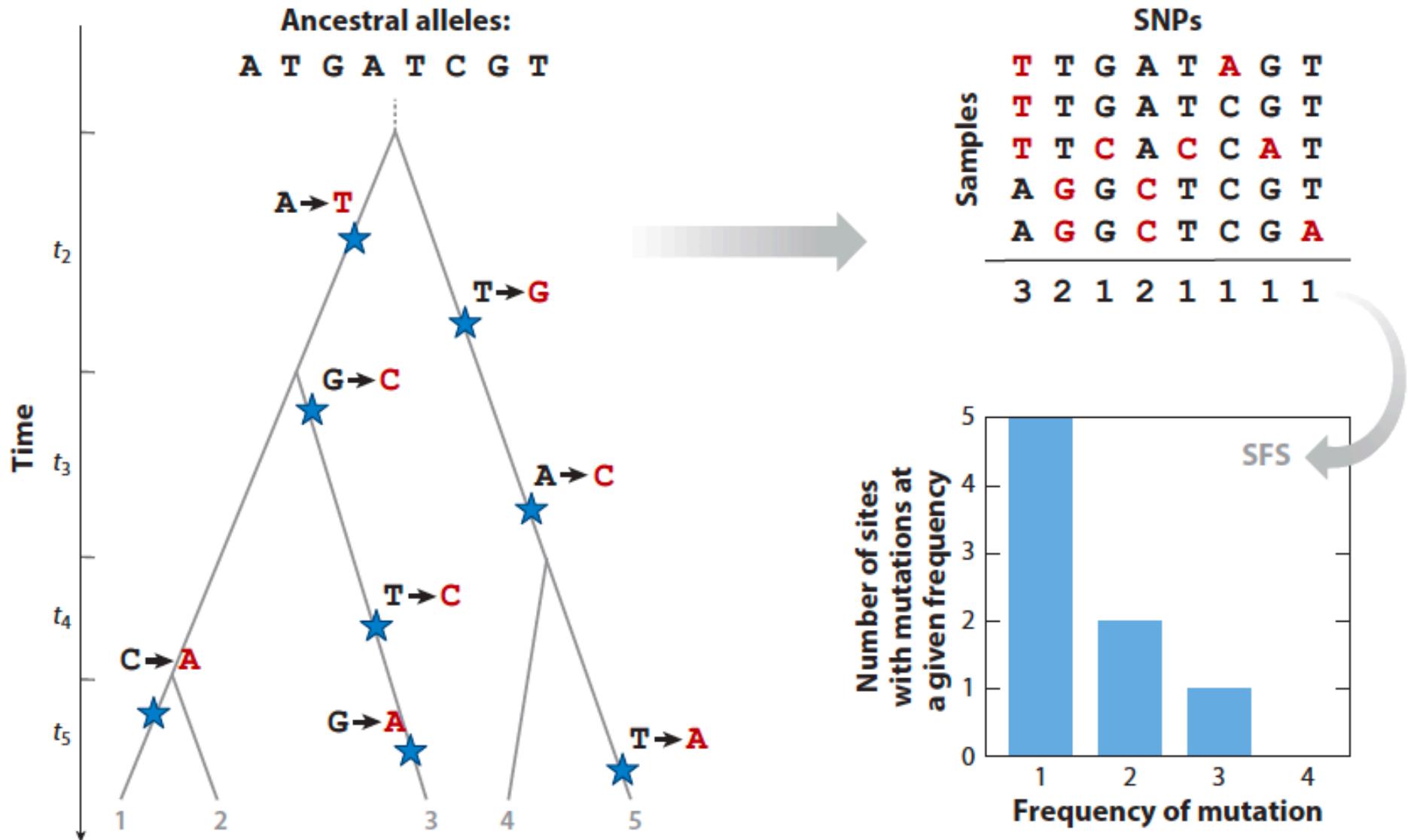
Allele Frequency Spectrum



Allele Frequency Spectrum

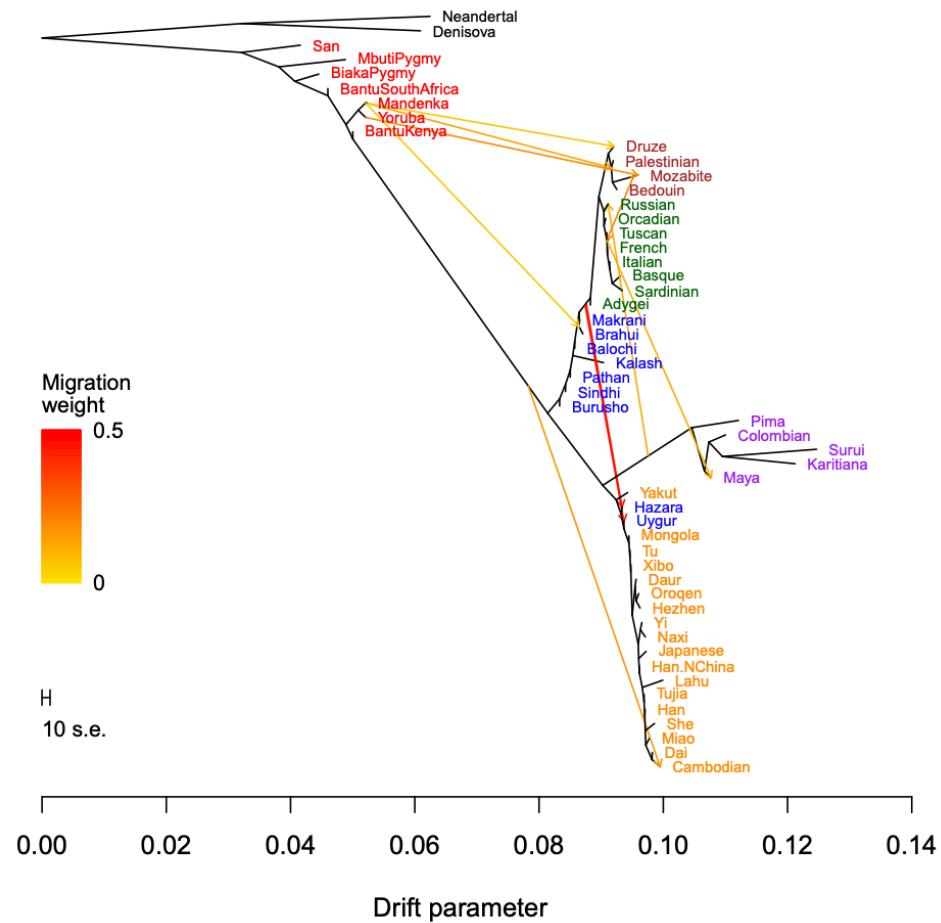
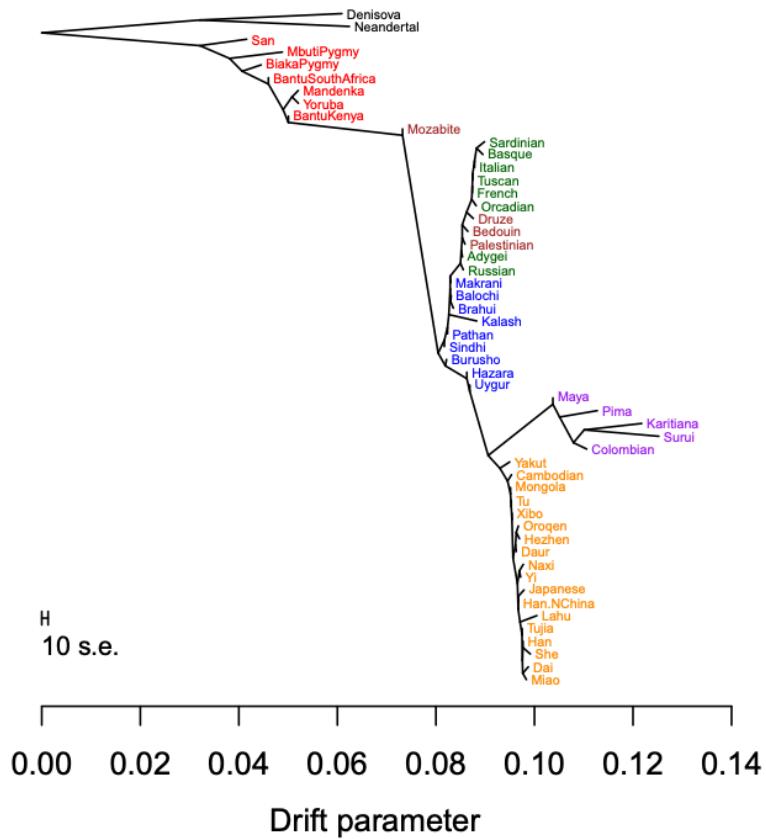


Allele Frequency Spectrum

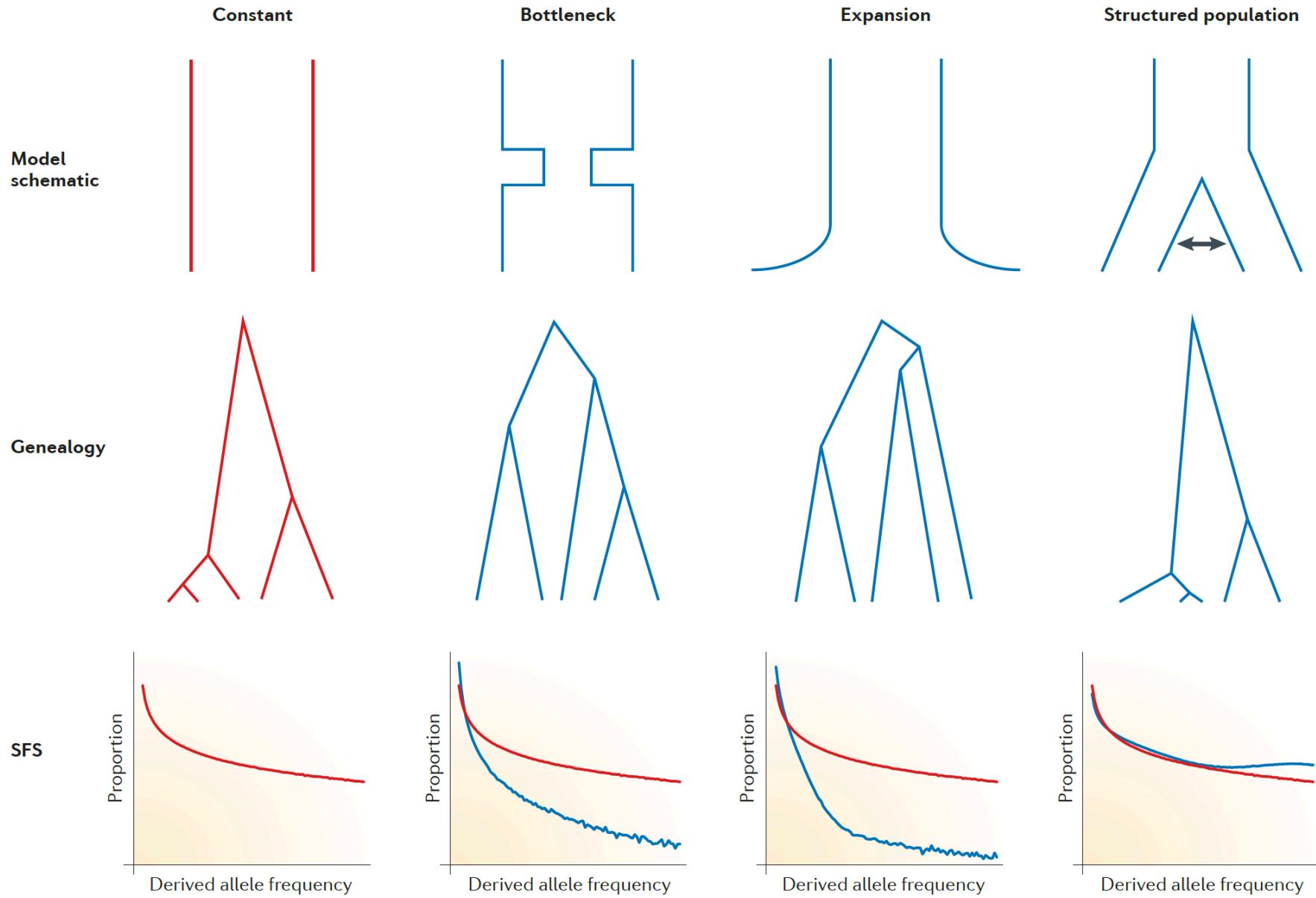


TreeMix: Structure with migration bands

A. Maximum likelihood human tree



Allele Frequency Spectrum



Model based methods

- 1. Collect data**
- 2. Develop models of evolution**
- 3. Calculate likelihood support for models based on your data**

Model based methods

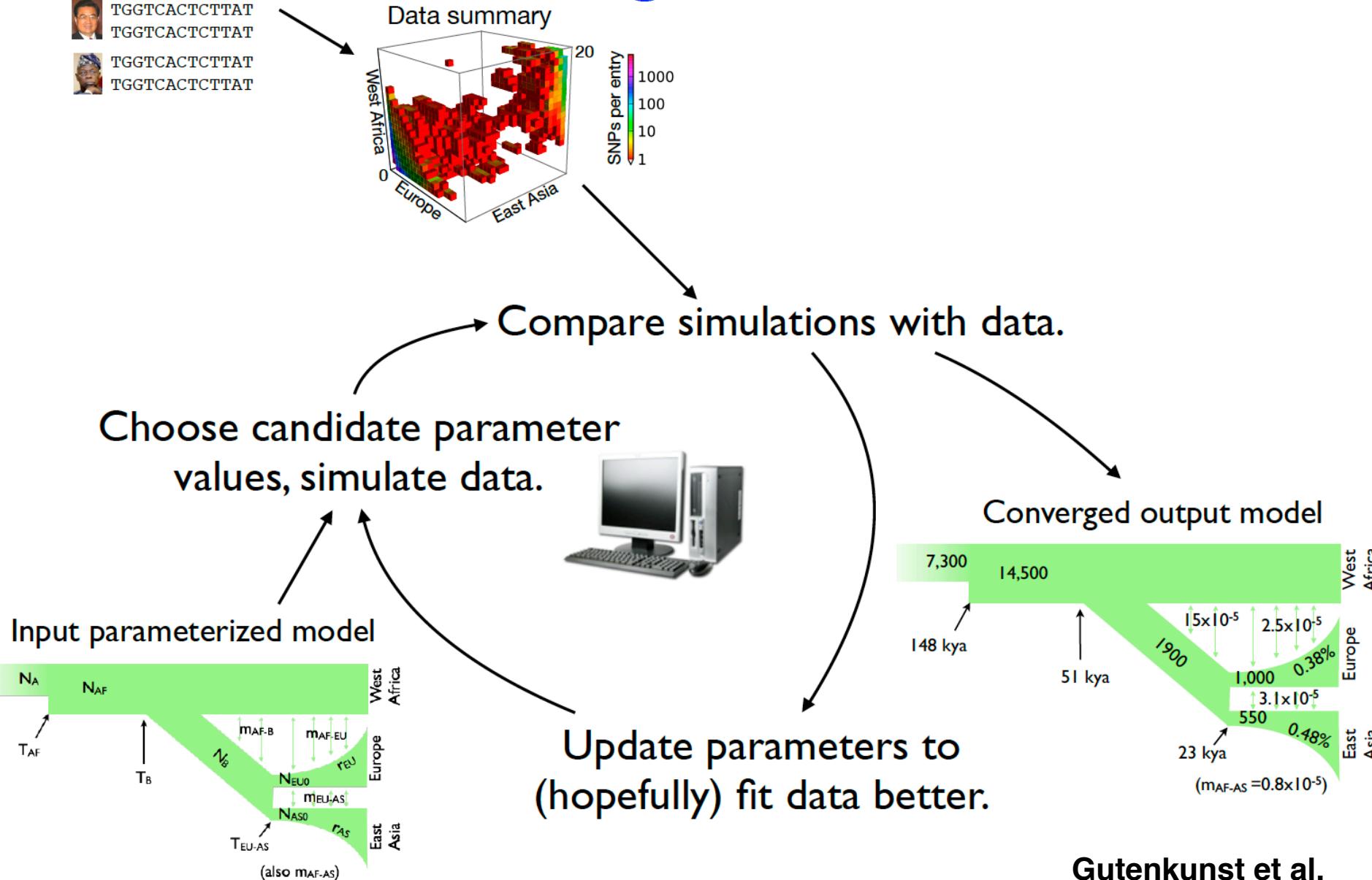
- 1. Collect data**
- 2. Develop models of evolution**
- 3. Calculate likelihood support for models based on your data**
- 4. Pick best supported model**

Data

TGGTCACTCTTAT
TGGTCACTCTTAT
TGGTCACTCTTAT
TGGTCACTCTTAT
TGGTCACTCTTAT
TGGTCACTCTTAT

Modeling workflow

∂a∂i
fastsimcoal2

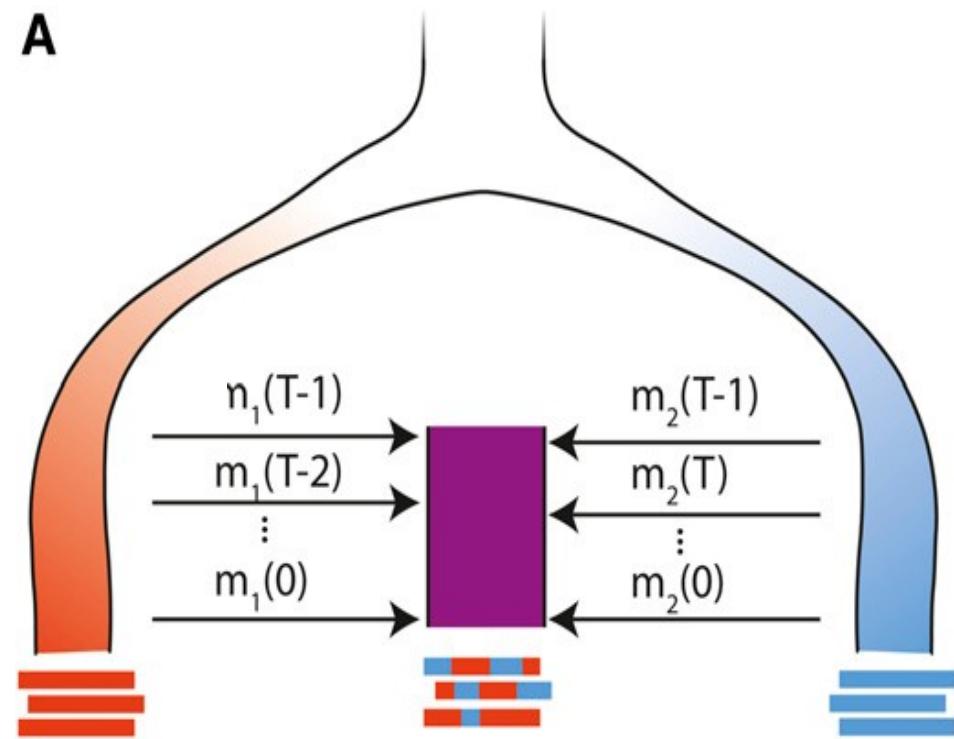


Haplotype lengths

The genomes of admixed individuals
will be mosaics of the source
populations

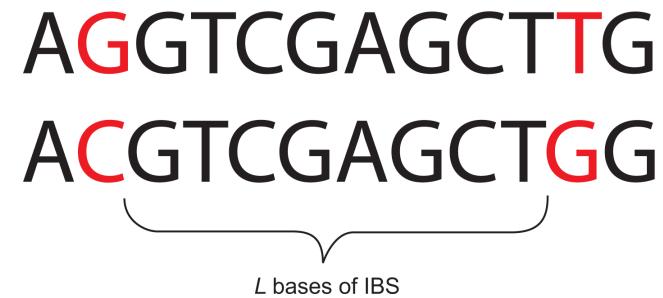
As time passes since admixture,
recombination breaks up admixture
tracts

TRACTS Infers admixture times
(potentially multiple pulses) and
proportions from the spectrum of
haplotype lengths



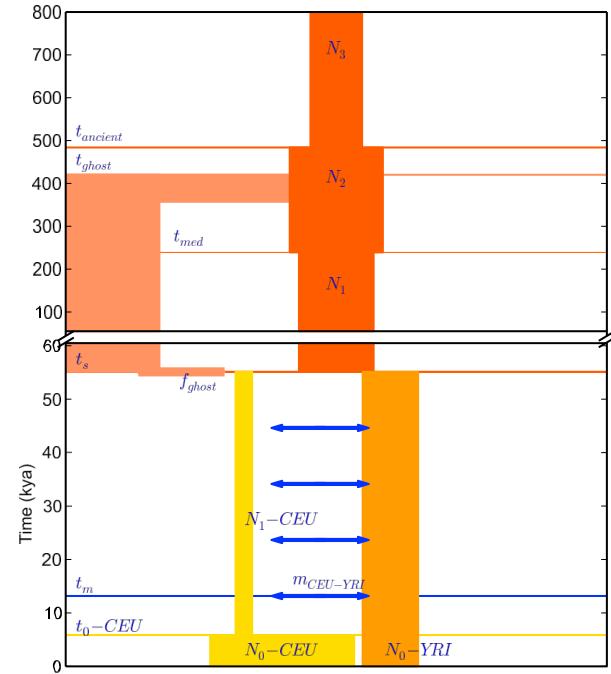
- Sequences that are Identical By State (IBS) with and between populations are informative about demographic history

AGGTCGAGCTTG
ACGTCGAGCTGG

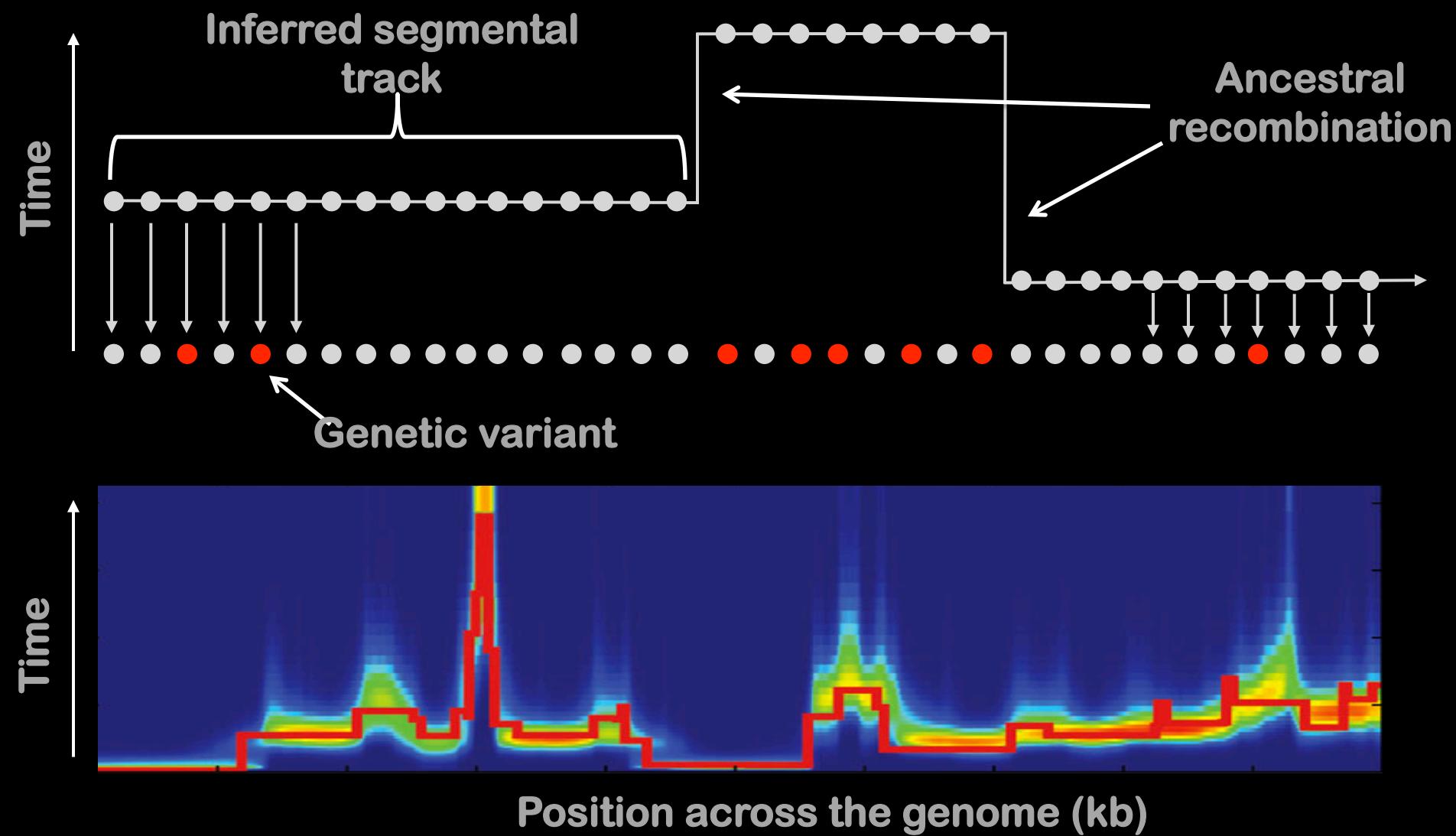


L bases of IBS

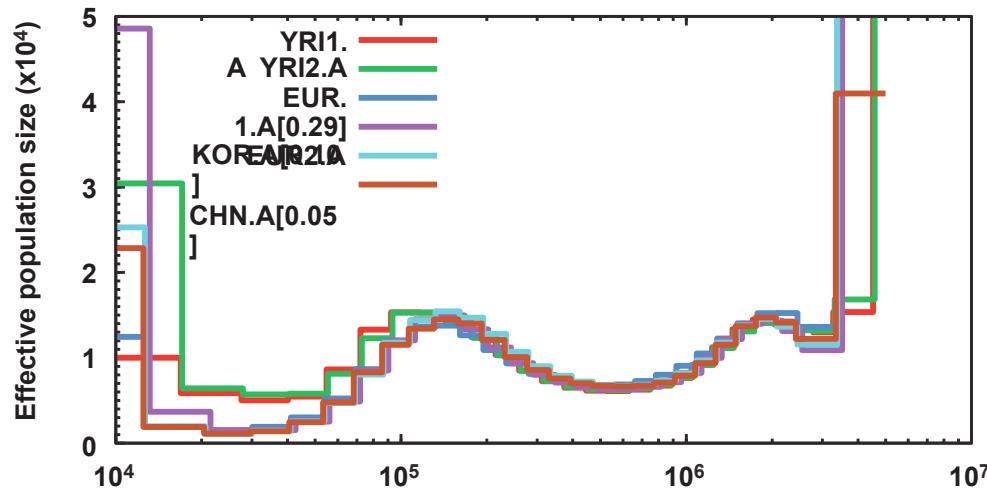
- Calculate expected spectrum of IBS tract lengths using coalescent theory
- Can fit very complex models



We can infer population sizes using genetic variation across the genome

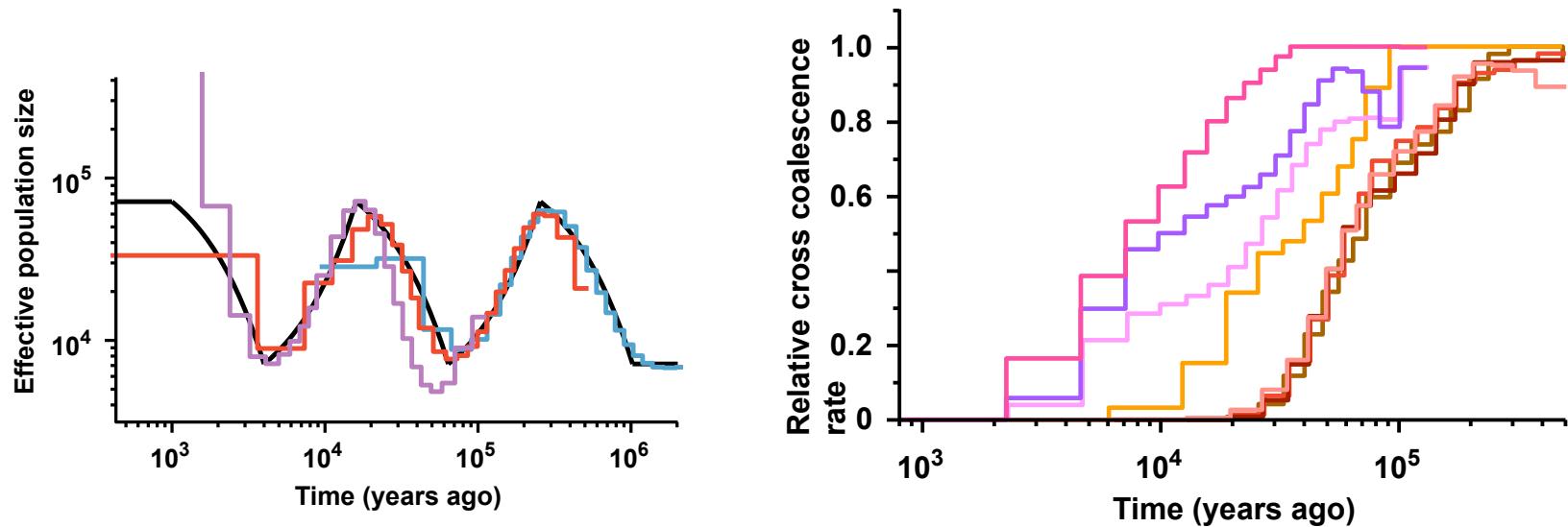


PSMC



- Estimate effective population size over time from a single unphased genome.
- No parametric model (e.g. exponential growth) assumed.

MSMC



- SMC model for multiple phased sequences
- Inferences of population sizes for more recent times than PSMC.
- Inferences of cross-coalescent rates between populations, which are indicative of population divergence and migration.

For Tuesday we will be
discussing:

Moreno-Mayar *et al.* 2018

and

Rougemont *et al.* 2018