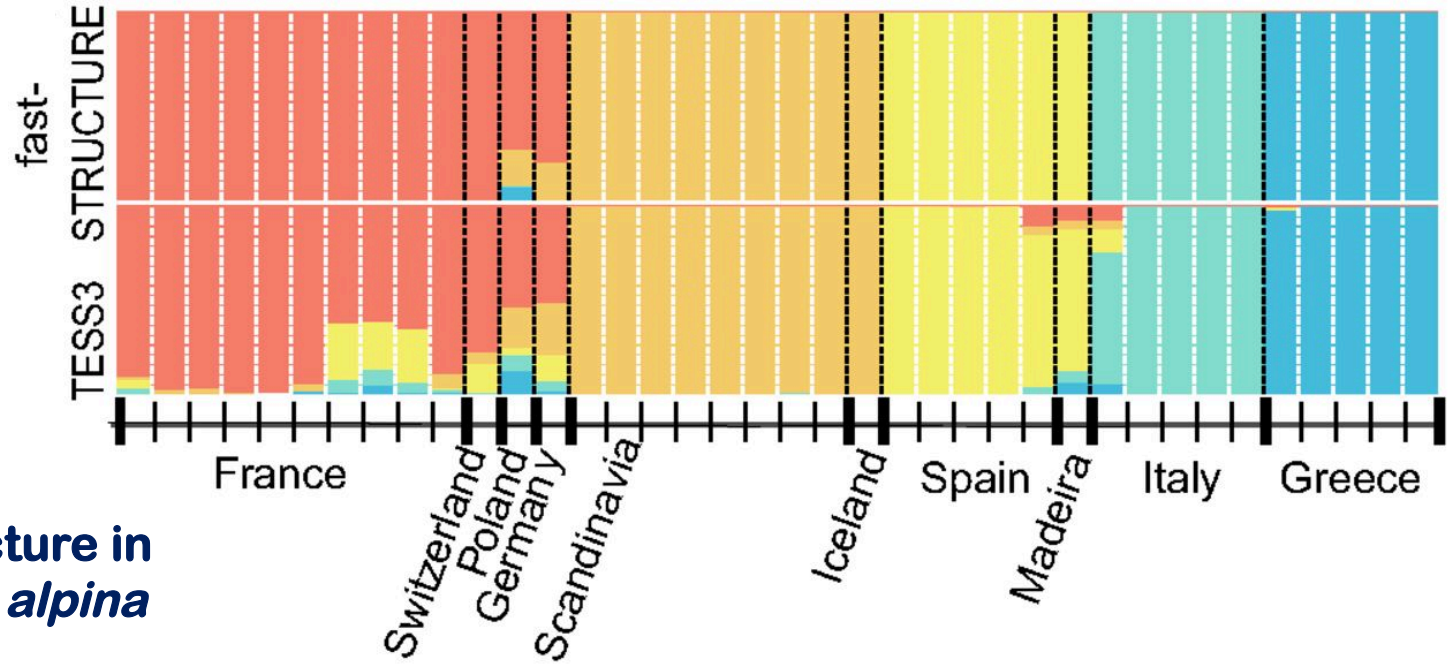
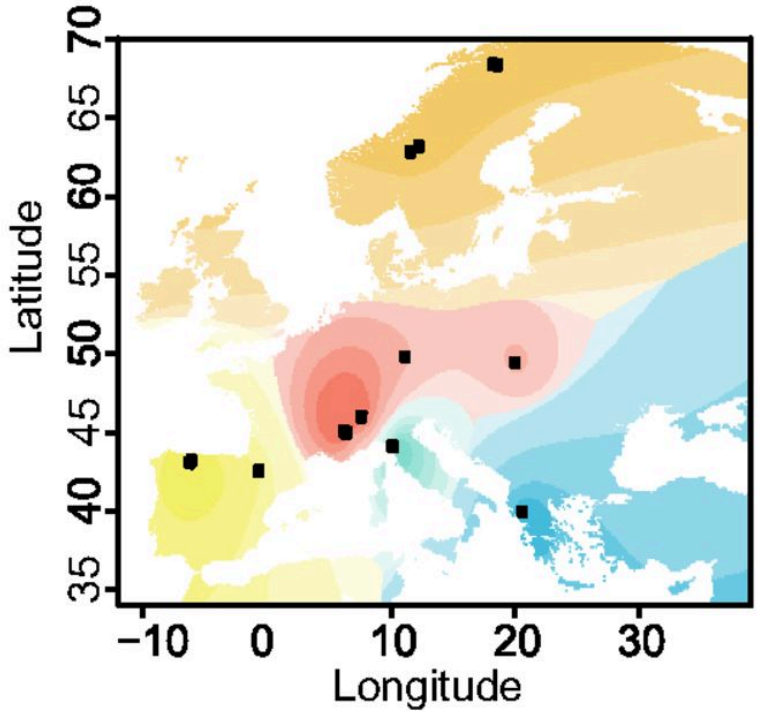
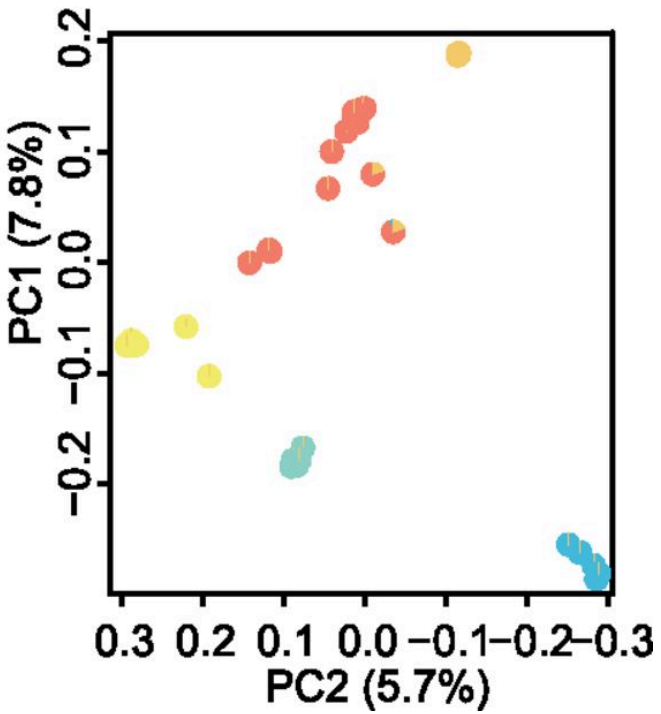
Population structure, admixture and changes in population size

Geographic structure in European *Arabis alpina*



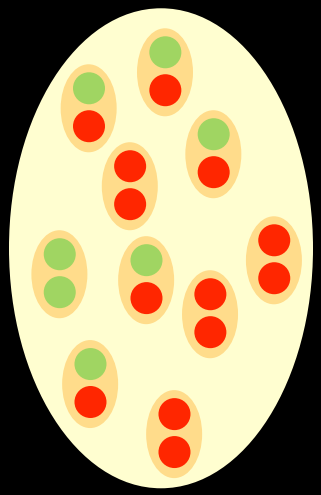




Benjamin Laenen et al.
PNAS 2018;115:4:816-821

Today

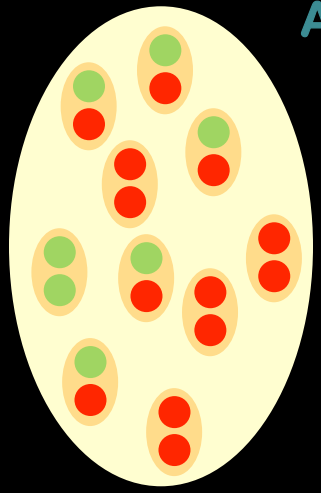
- Genetic structure of populations
- Hardy-Weinberg Equilibrium
- Genetic variation in space and time
- Variation in natural populations, Wright's Fixation Index (F_{ST})
- Other methods to measure population structure/gene flow



A

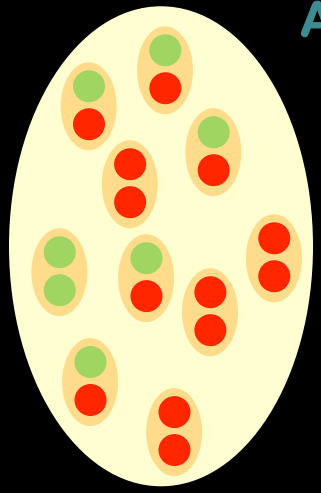
a

Allele frequencies



A
a

$$f_A = \frac{N_A}{2N} = \frac{7}{2*10} \quad \text{or} \quad f_a = \frac{N_a}{2N} = \frac{13}{2*10}$$



Allele frequencies

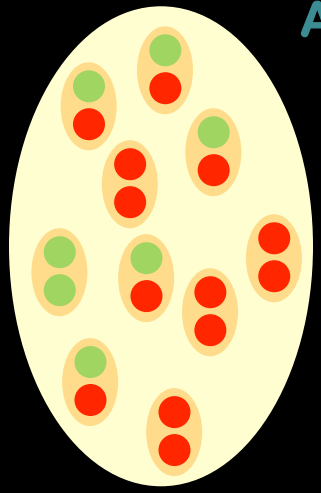
$$f_A = \frac{N_A}{2N} = \frac{7}{2*10} \quad \text{or} \quad f_a = \frac{N_a}{2N} = \frac{13}{2*10}$$

Genotype frequencies

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

$$fA_a = \frac{NA_a}{2N} = \frac{5}{10}$$

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



● A
● a

Allele frequencies

$$f_A = \frac{N_A}{2N} = \frac{7}{2*10} \quad \text{or} \quad f_a = \frac{N_a}{2N} = \frac{13}{2*10}$$

Genotype frequencies

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

$$fA_a = \frac{NA_a}{2N} = \frac{5}{10}$$

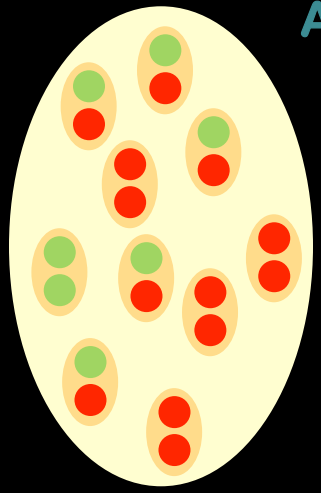
$$f_{aa} = \frac{N_{aa}}{2N} = \frac{4}{10}$$

K-allelic loci

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

fA_a = Heterozygosity

1 - fA_a = Homozygosity
= $fAA + f_{aa}$



● A
● a

Allele frequencies

$$f_A = \frac{N_A}{2N} = \frac{7}{2*10} \quad \text{or} \quad f_a = \frac{N_a}{2N} = \frac{13}{2*10}$$

Genotype frequencies

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

$$fA_a = \frac{NA_a}{2N} = \frac{5}{10}$$

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

K-allelic loci

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

fA_a = Heterozygosity

1 - fA_a = Homozygosity
= $fAA + f_{aa}$

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

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

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

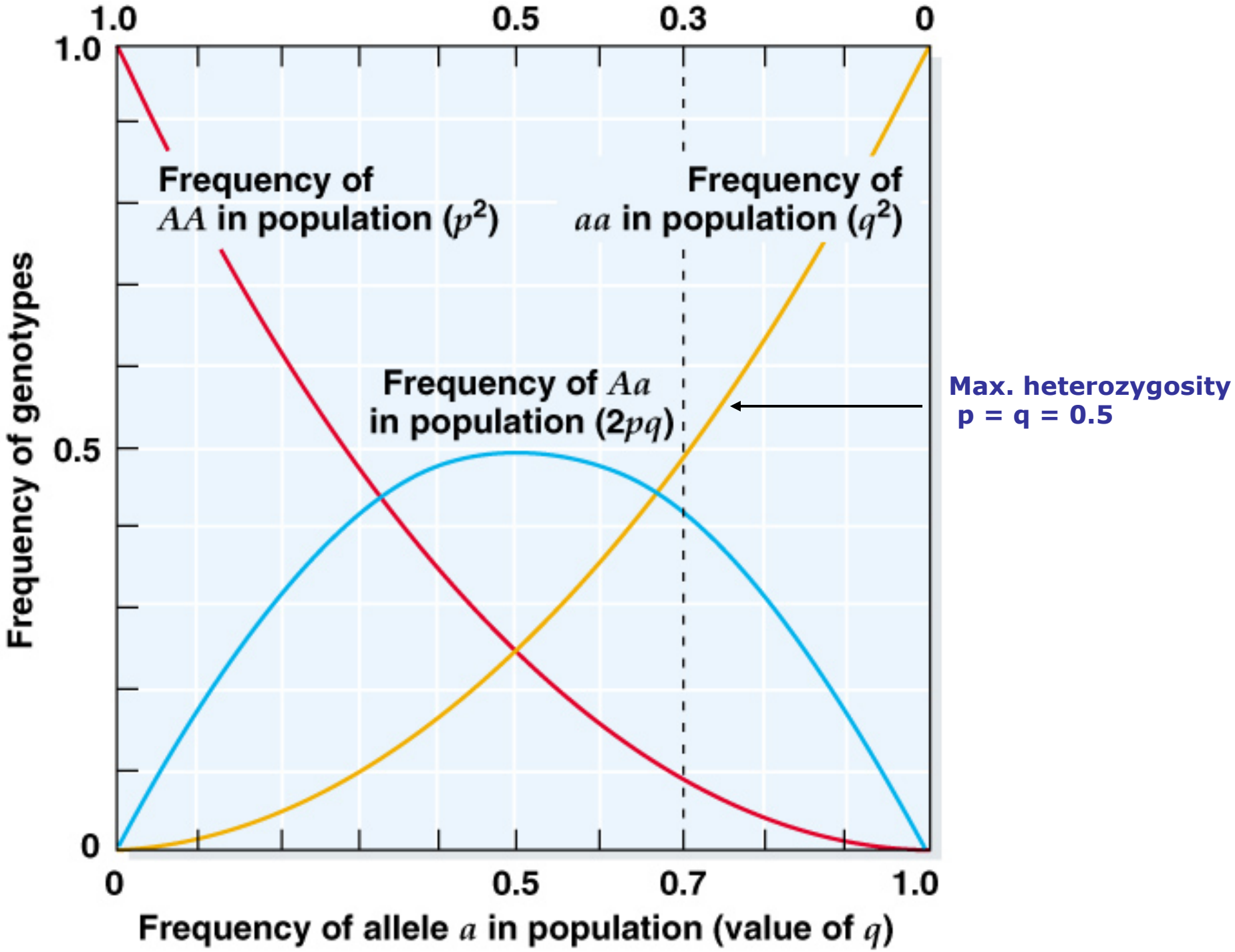
2. If *f(p)* = 0.5 and *f(q)* = 0.5, outcome is as follows:

	A(p)	a(q)
A(p)	AA(p^2) 0.5 x 0.5 = 0.25	Aa(pq) 0.5 x 0.5 = 0.25
a(q)	Aa(pq) 0.5 x 0.5 = 0.25	aa(q^2) 0.5 x 0.5 = 0.25

3. When population is at equilibrium:

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

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


1.0
0.5
0.3
0
1.0
0.5
0
Frequency of genotypes
Frequency of AA in population (p^2)
Frequency of aa in population (q^2)
Frequency of Aa in population ($2pq$)
Max. heterozygosity
p = q = 0.5
0
0.5
0.7
1.0
Frequency of allele a in population (value of q)

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)

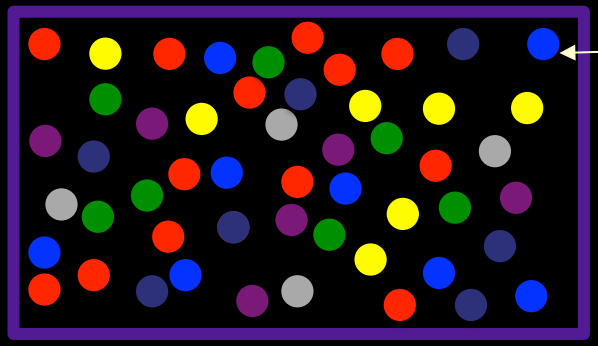
- Migration can add or subtract variation depending on which alleles migrants carry and whether they immigrate or emigrate

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

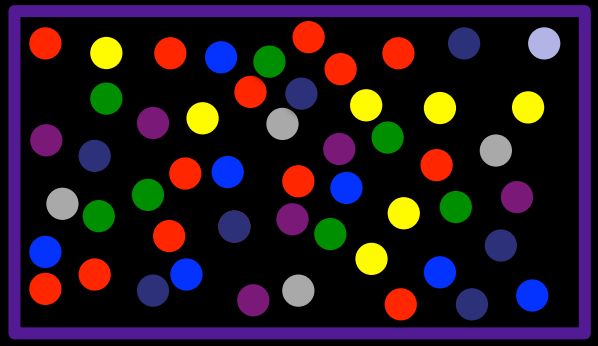
Mutations change allele frequencies

Pop 1

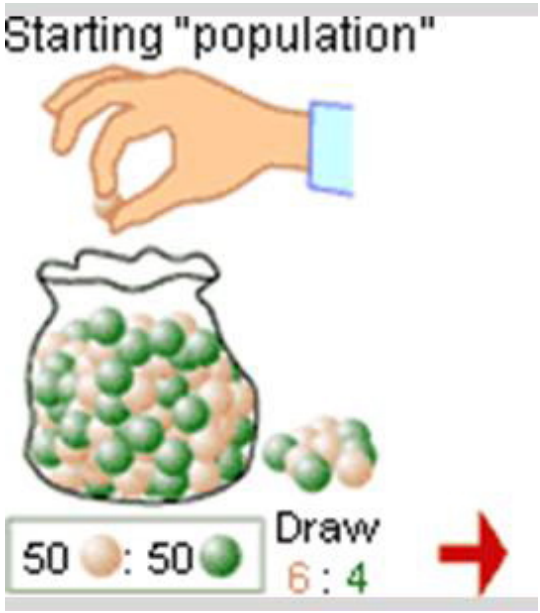


Mutation

Pop 1 (next generation)



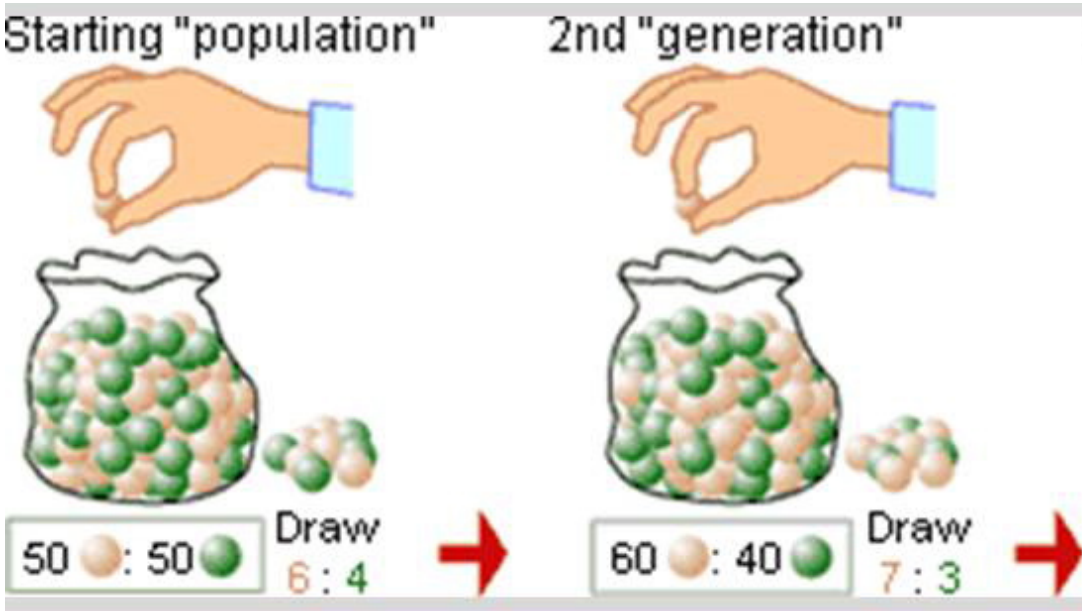
Drift also changes allele frequencies (sampling error)



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

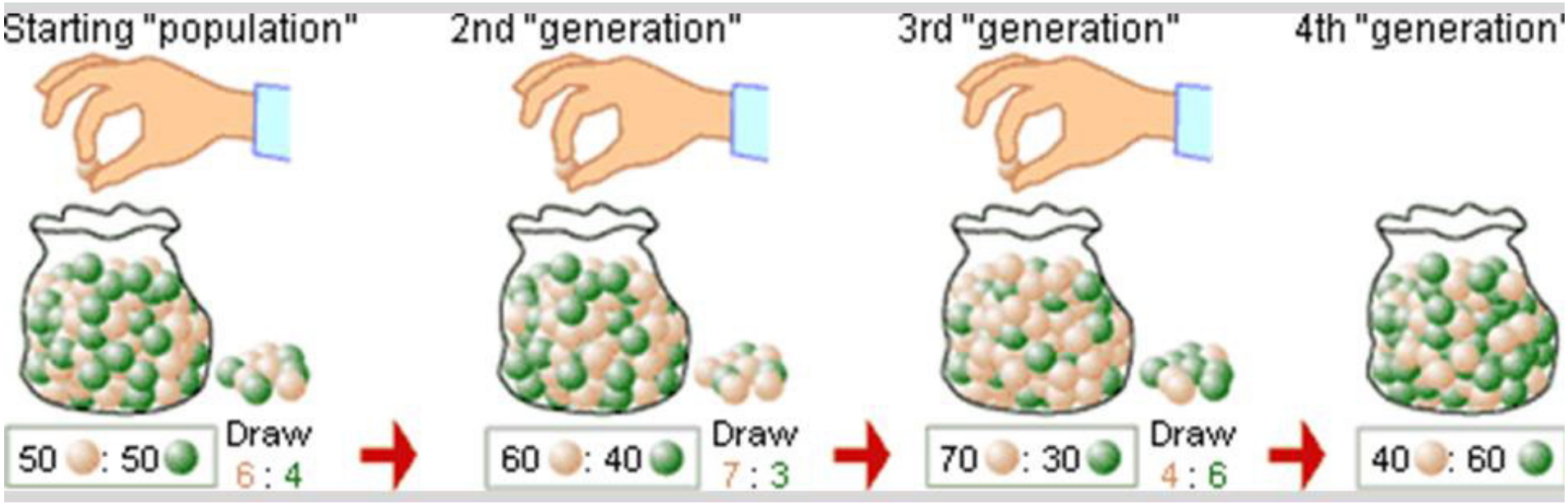
http://rosetta.ahmedmoustafa.io/drift/

Drift changes allele frequencies (sampling error)



http://rosetta.ahmedmoustafa.io/drift/

Drift changes allele frequencies (sampling error)



http://rosetta.ahmedmoustafa.io/drift/

Mutation and Drift change allele frequencies

©2008 Peter Beerli

Mutation and Drift change allele frequencies

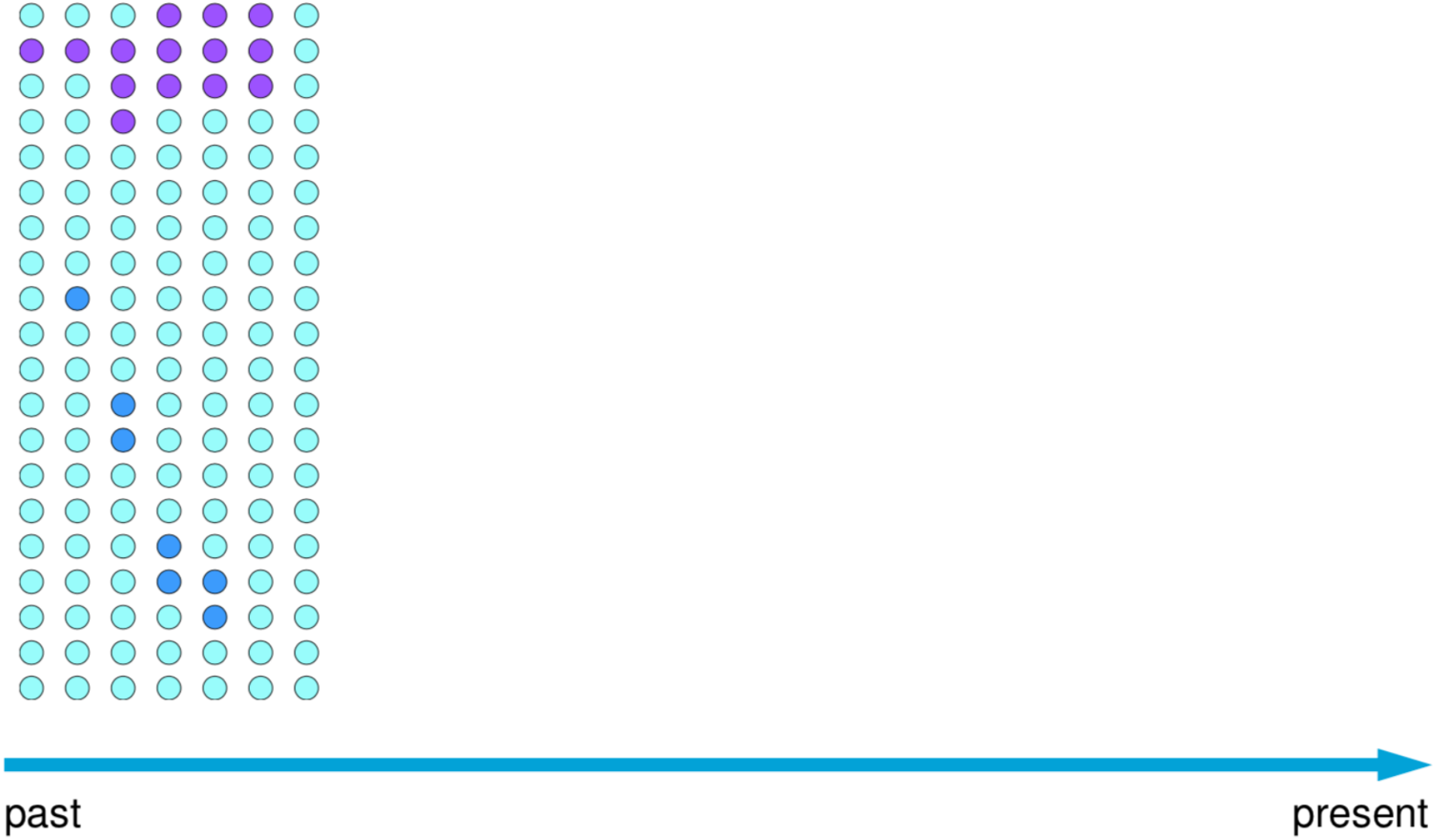
Mutation and Drift change allele frequencies

Mutation and Drift change allele frequencies

Mutation and Drift change allele frequencies

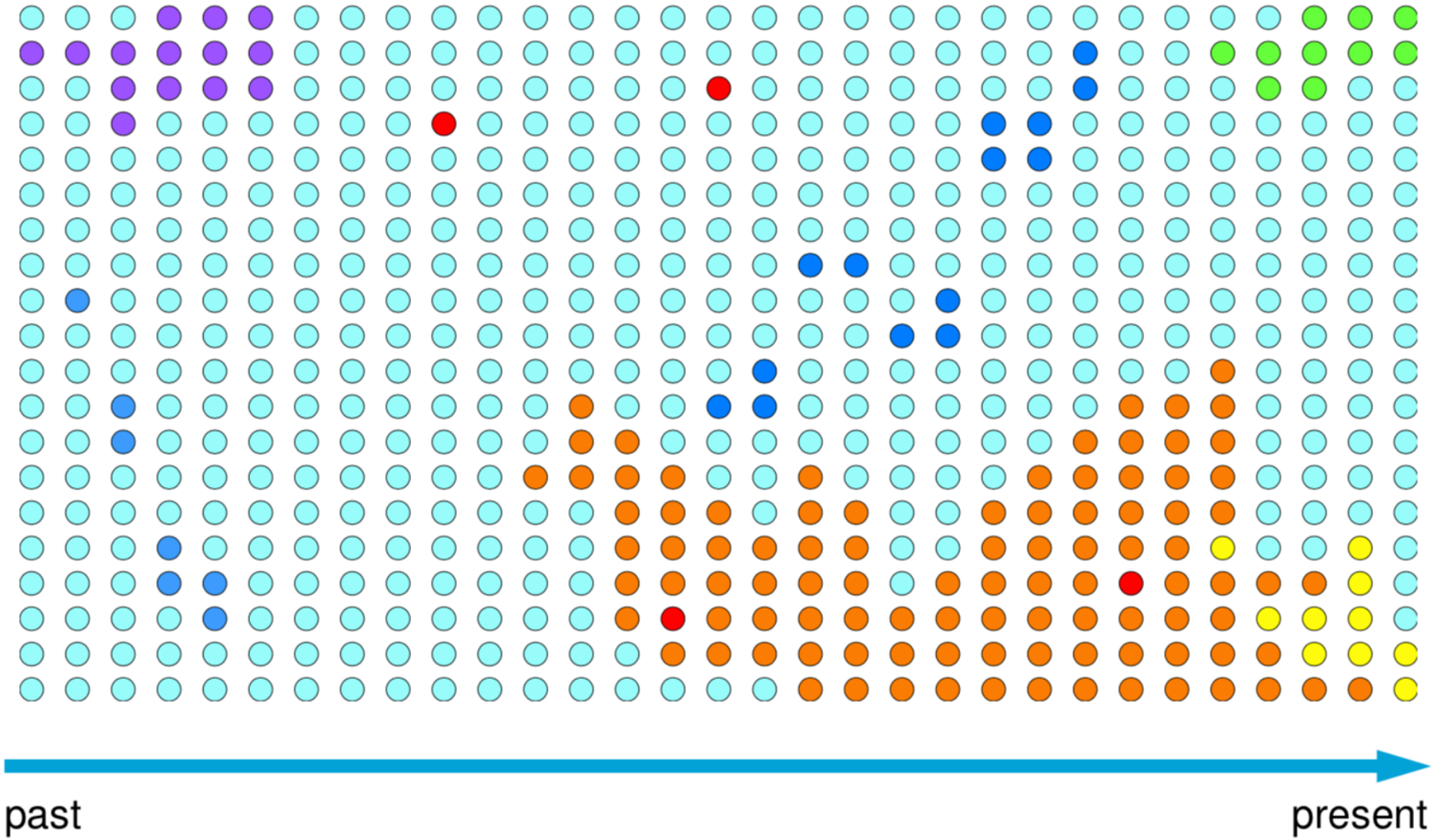
Mutation and Drift change allele frequencies

Mutation and Drift change allele frequencies



Mutation and Drift change allele frequencies

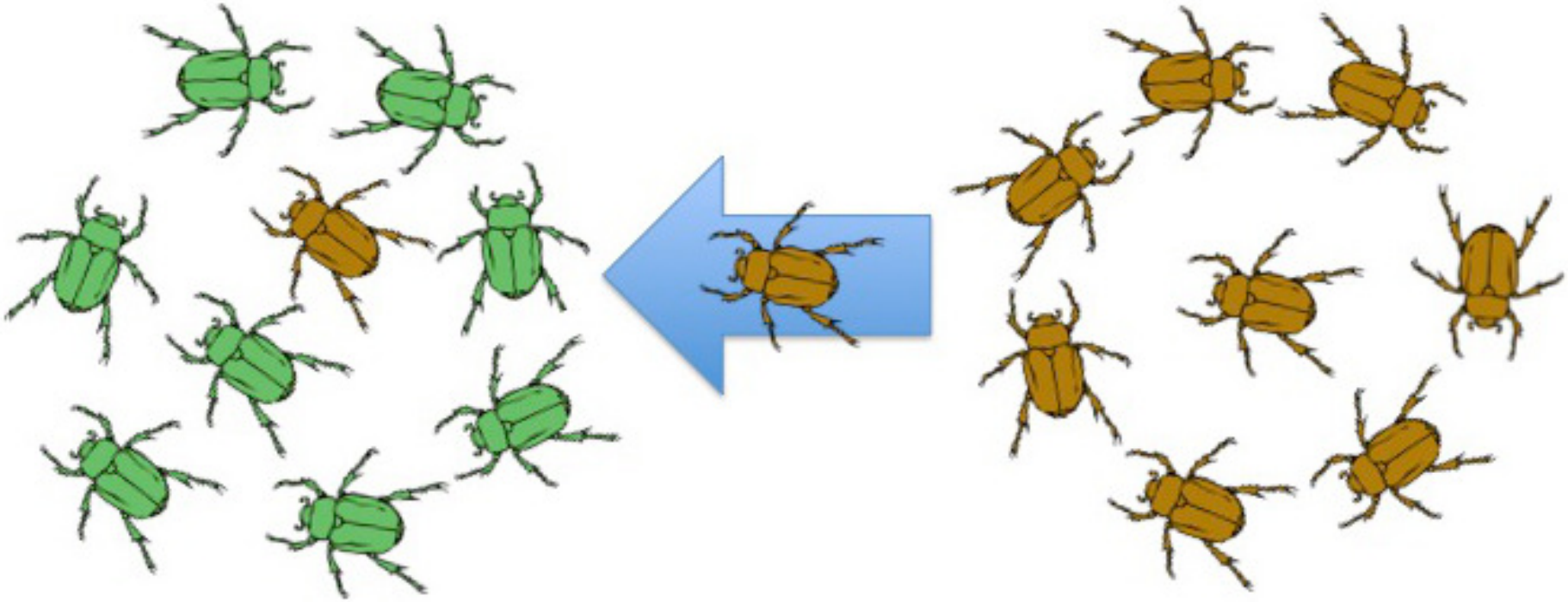
Mutation and Drift change allele frequencies



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

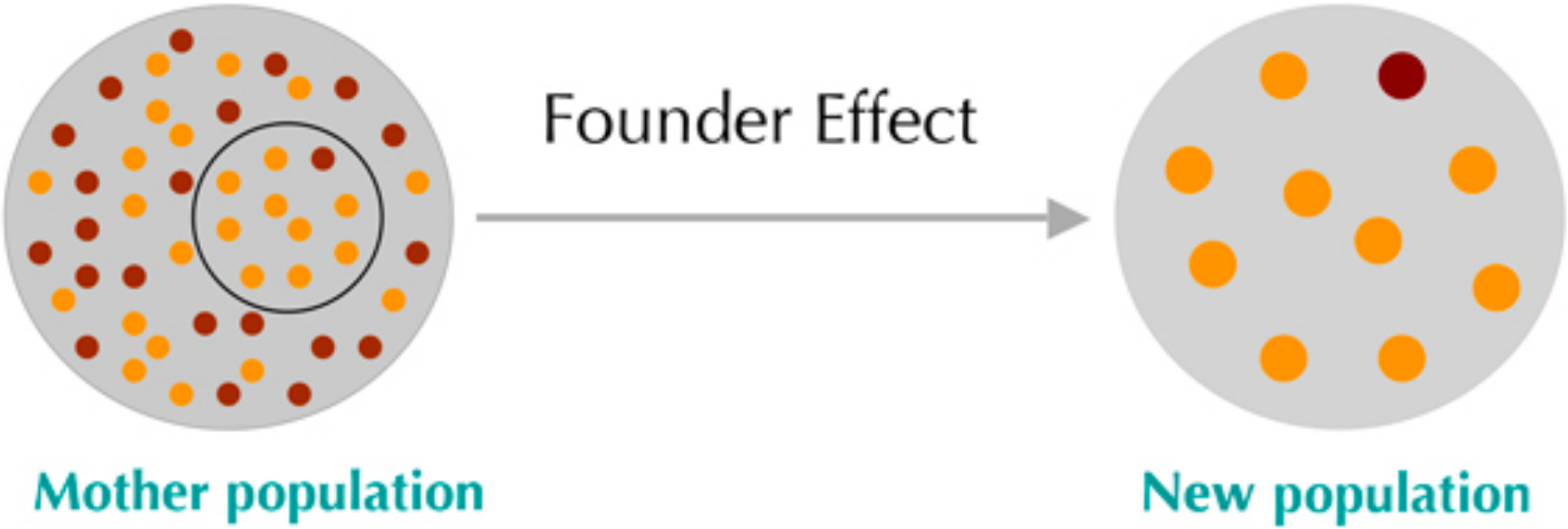
https://phytools.shinyapps.io/drift-selection/

Gene flow (migration) also changes allele frequencies



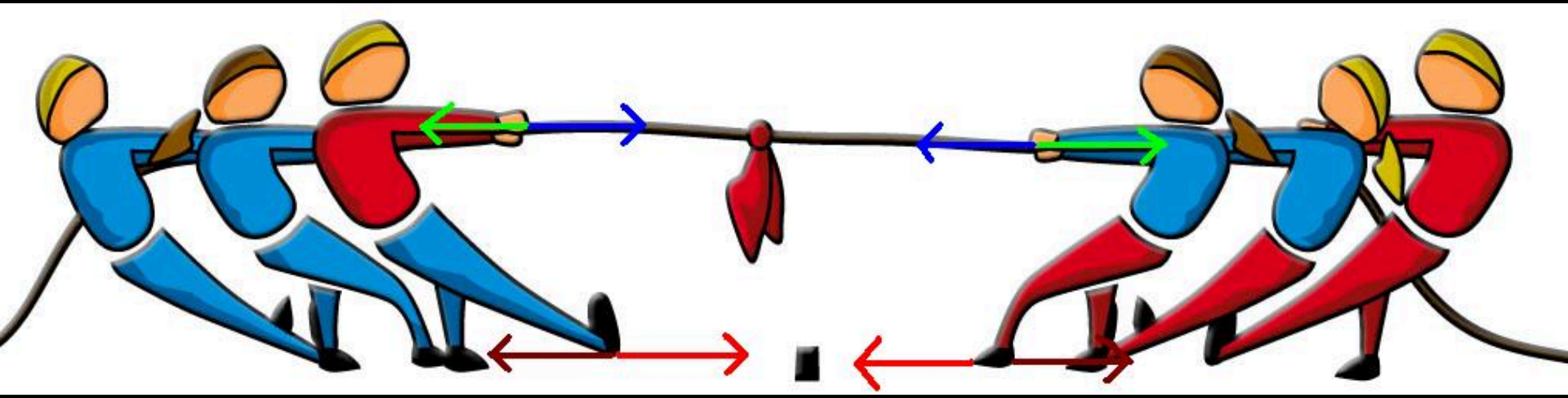
Migration refers to the movement of some organisms (or their gametes) among subpopulations and subsequent breeding, also referred to as gene flow, or admixture. Migration acts to homogenize populations, or reduce differences between populations

Variations in population size also changes allele frequencies



Balance between

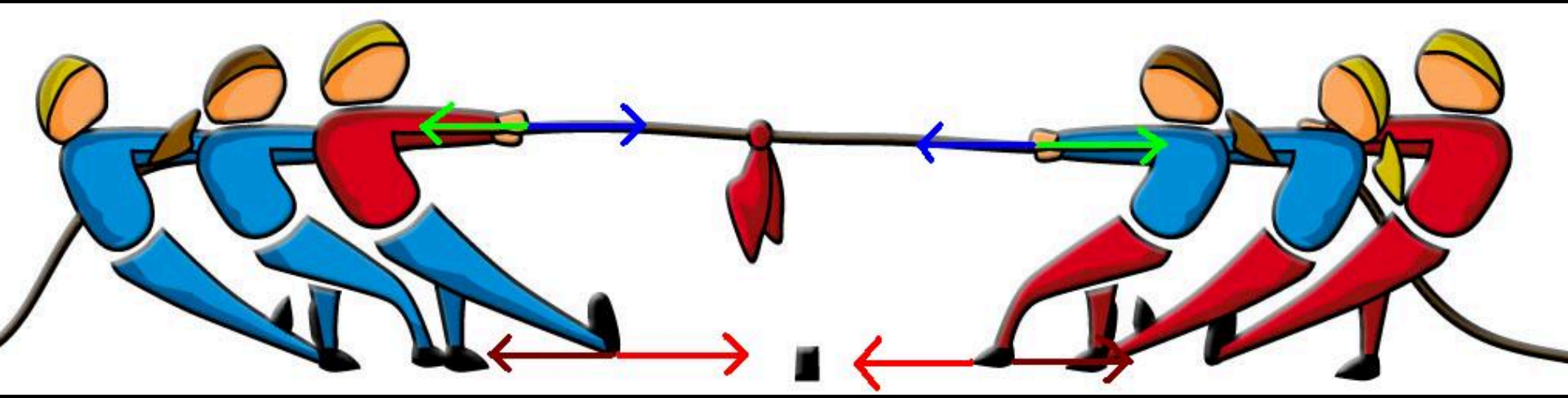
Drift ⟷ Migration



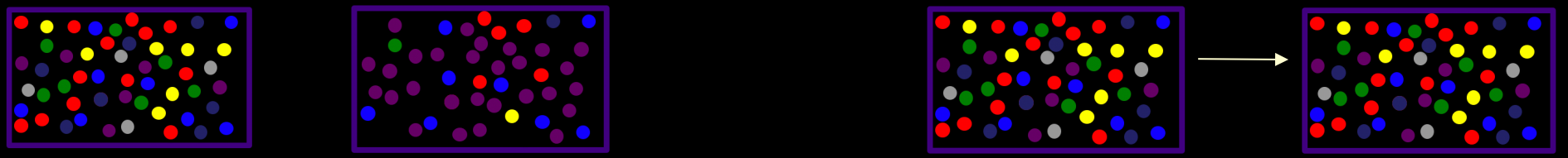
http://www.cienciaonline.com/2007/07/23/accion-reaccion-%C2%BFse-entiende/

Balance between

Drift ⟷ Migration



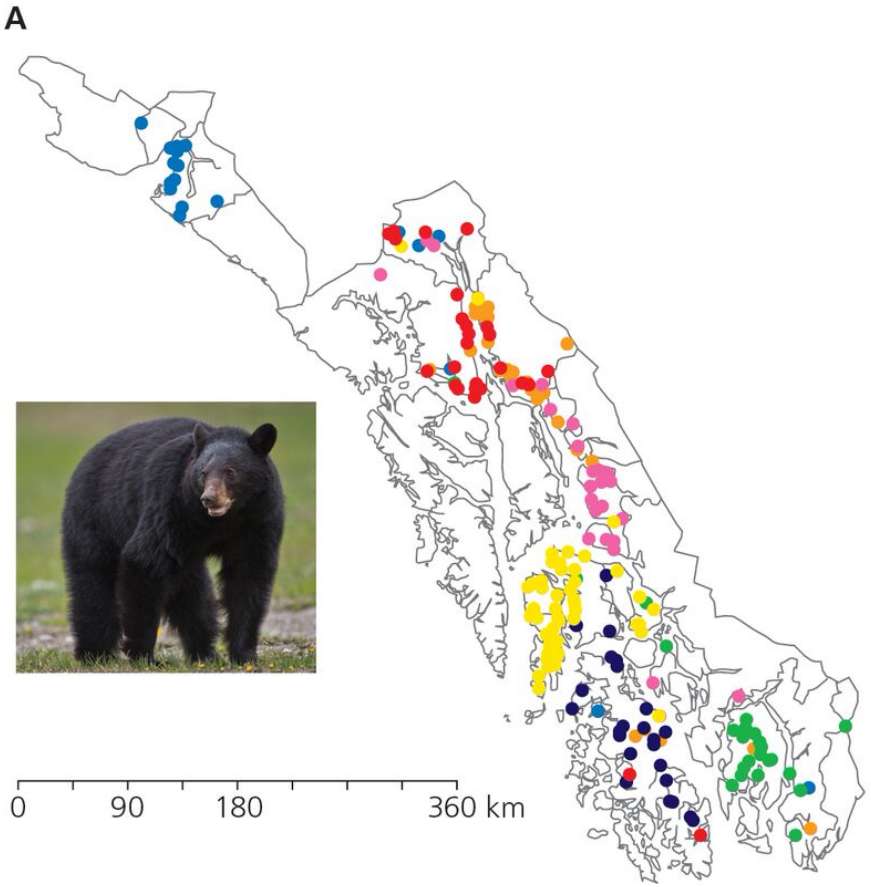
Genetic structure (subpopulations) ⟷ Genetic homogeneity Panmixia (single population



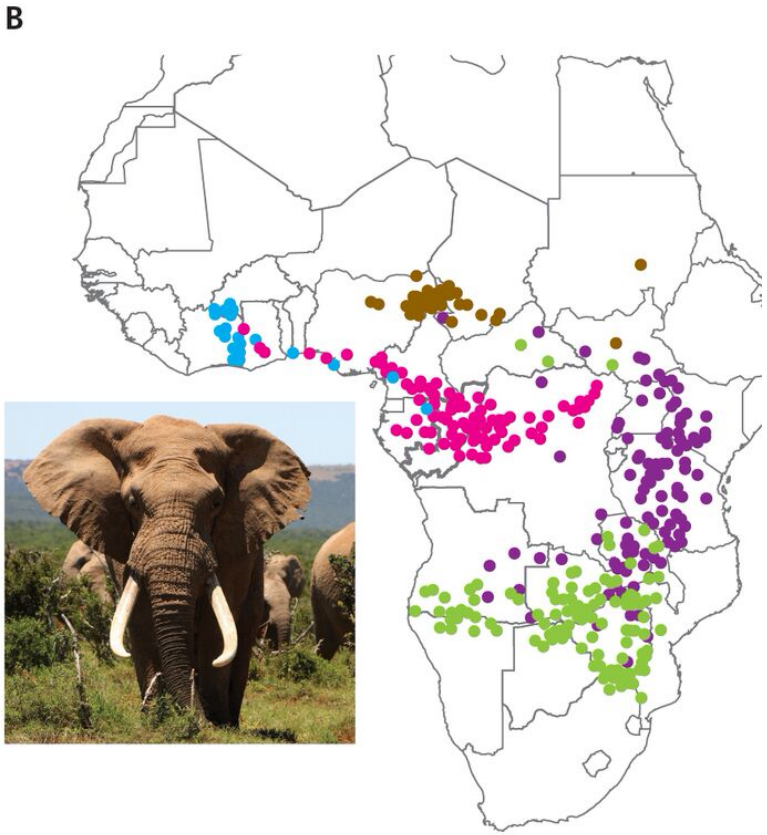
http://www.cienciaonline.com/2007/07/23/accion-reaccion-%C2%BFse-entiende/

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

Subdivided populations show distinct genetic structure

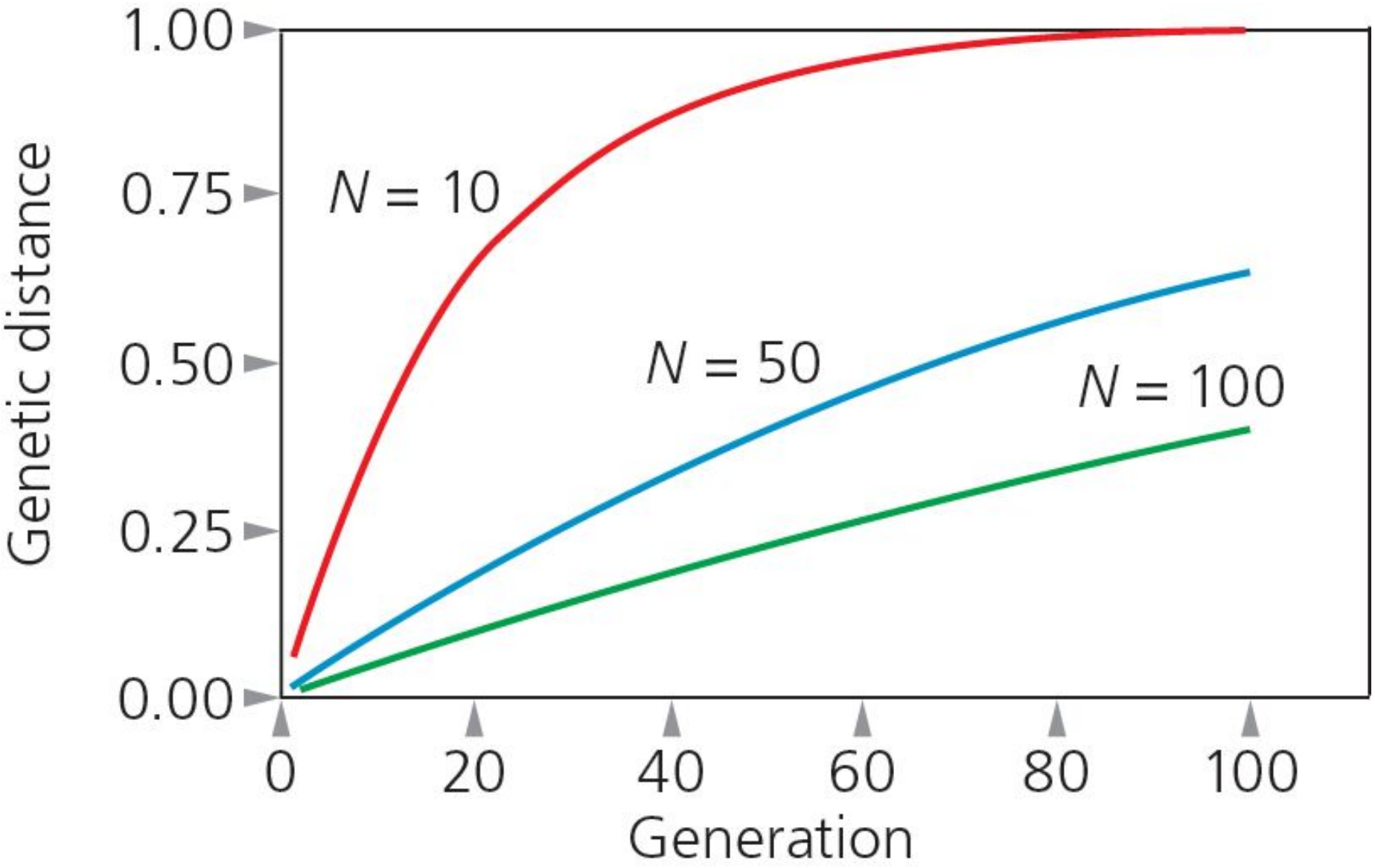


Black Bears in SE Alaska



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)

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

  - 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	q	$H_S=2pq$
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

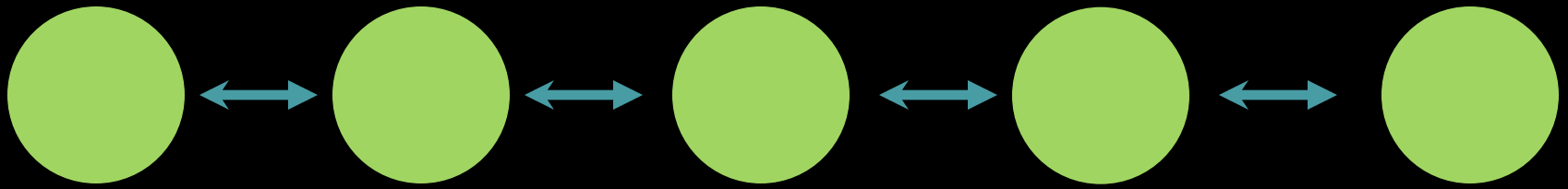
	p	q
Subpop 1:	0.7	0.3
Subpop 2:	0.5	0.5
Subpop 3:	0.3	0.7
	p = 0.5	q = 0.5

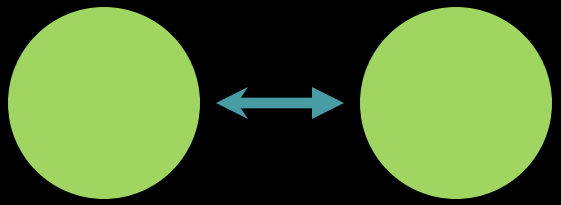
$$H_T = 2pq = 0.5$$

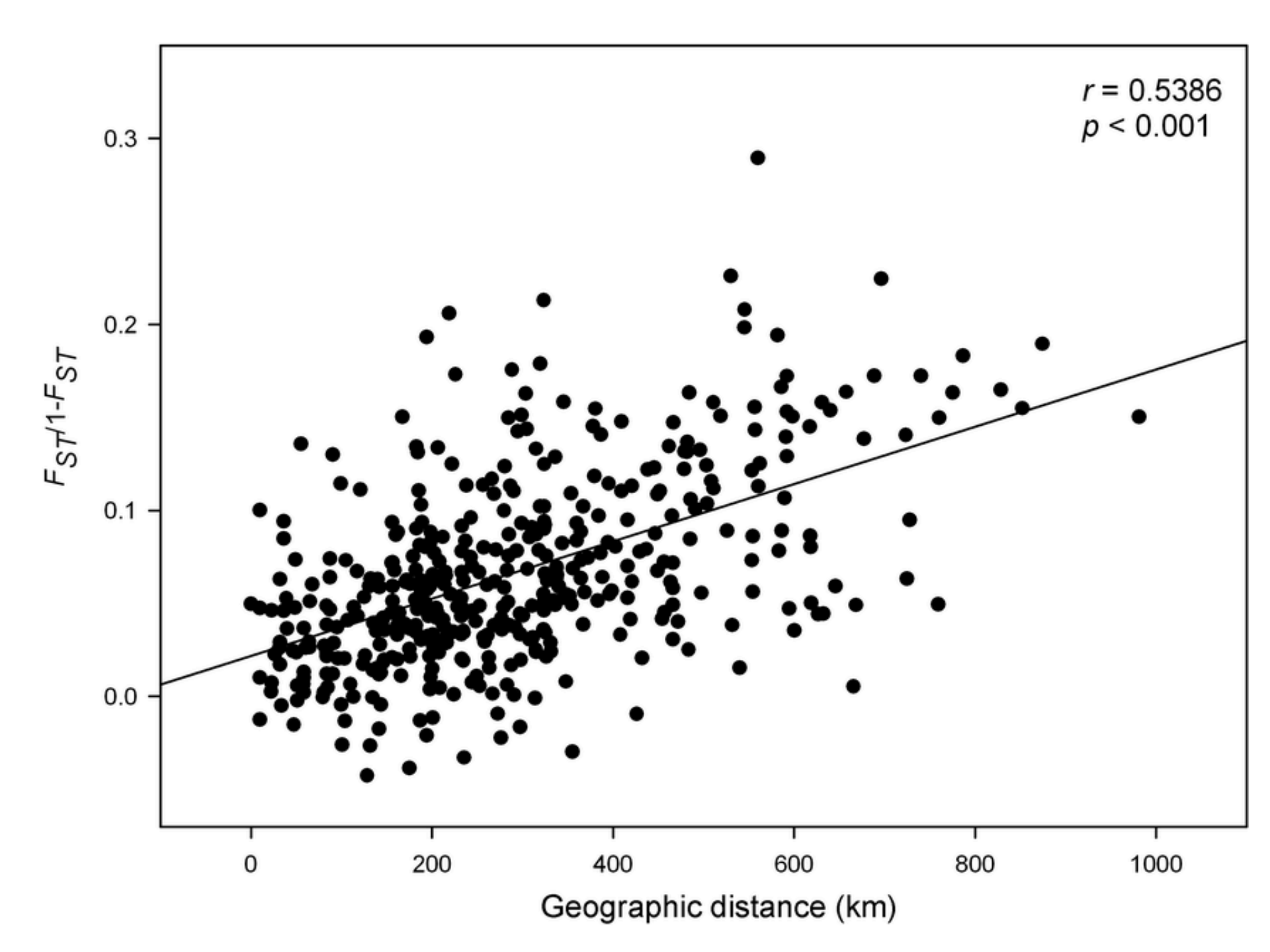
Remember that,

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

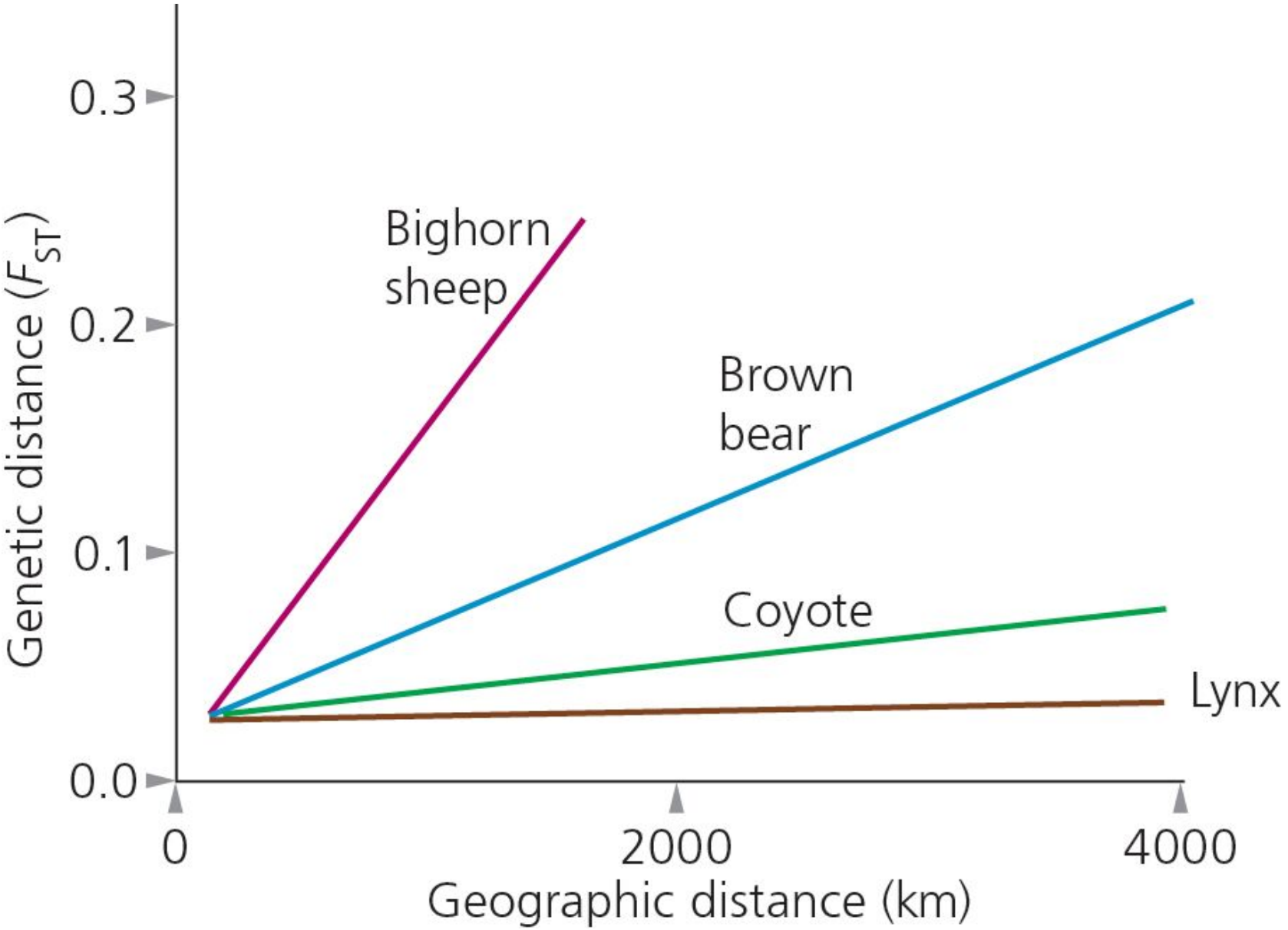
$$F_{ST} = \frac{0.5 - 0.466}{0.5} = 0.11$$






r = 0.5386
p < 0.001
0.3
0.2
0.1
0.0
$F_{ST}/1\text{-}F_{ST}$
0
200
400
600
800
1000
Geographic distance (km)

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

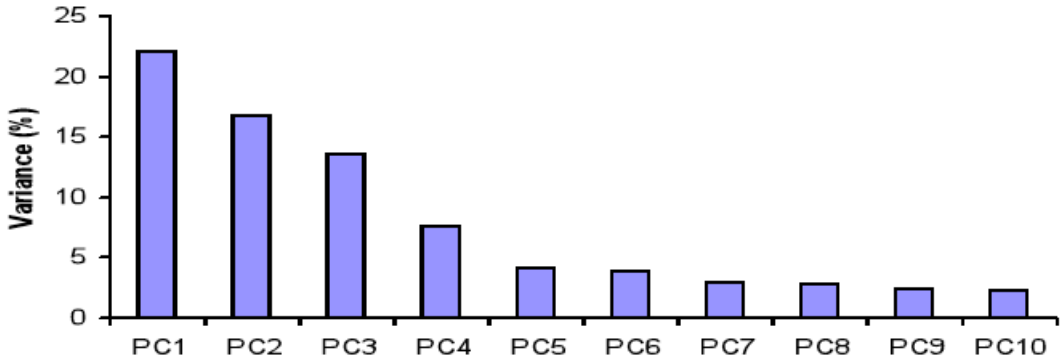
See also

http://methodspopgen.com/methods-to-infer-population-structure/

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	Estimates 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 a priori	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome ainting	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 heuristic 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 infer ence	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 MSMC'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 rate 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
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[131]	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

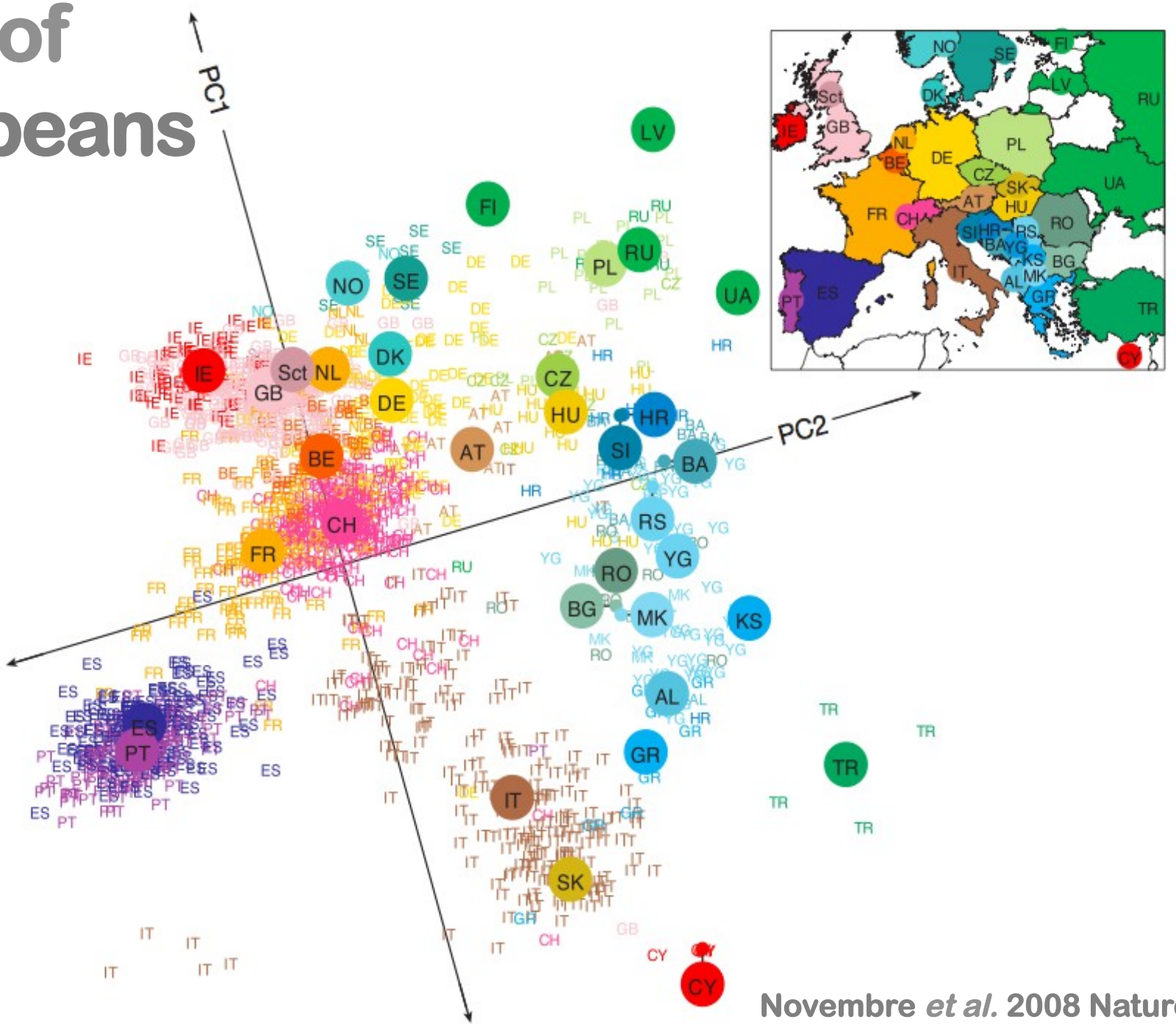
Can *ignore* the components of lesser significance.



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

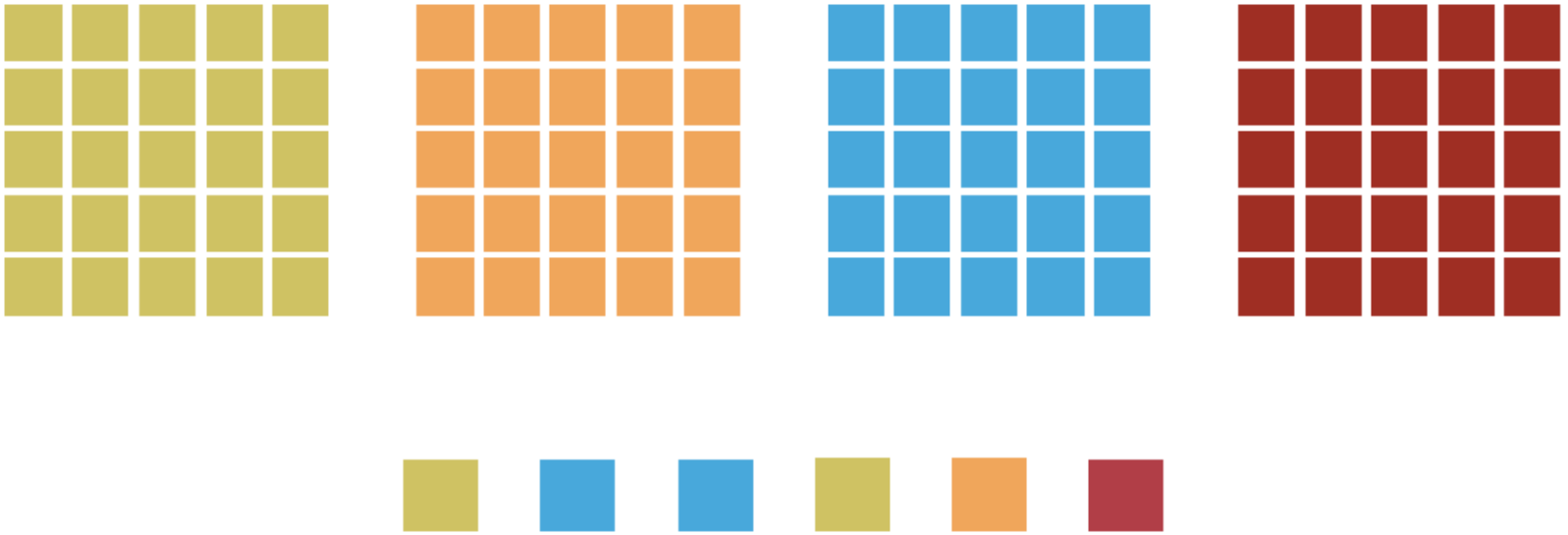
- n dimensions in original data
- calculate n eigenvectors and eigenvalues
- choose only the first p eigenvectors, based on their eigenvalues
- final data set has only p dimensions

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

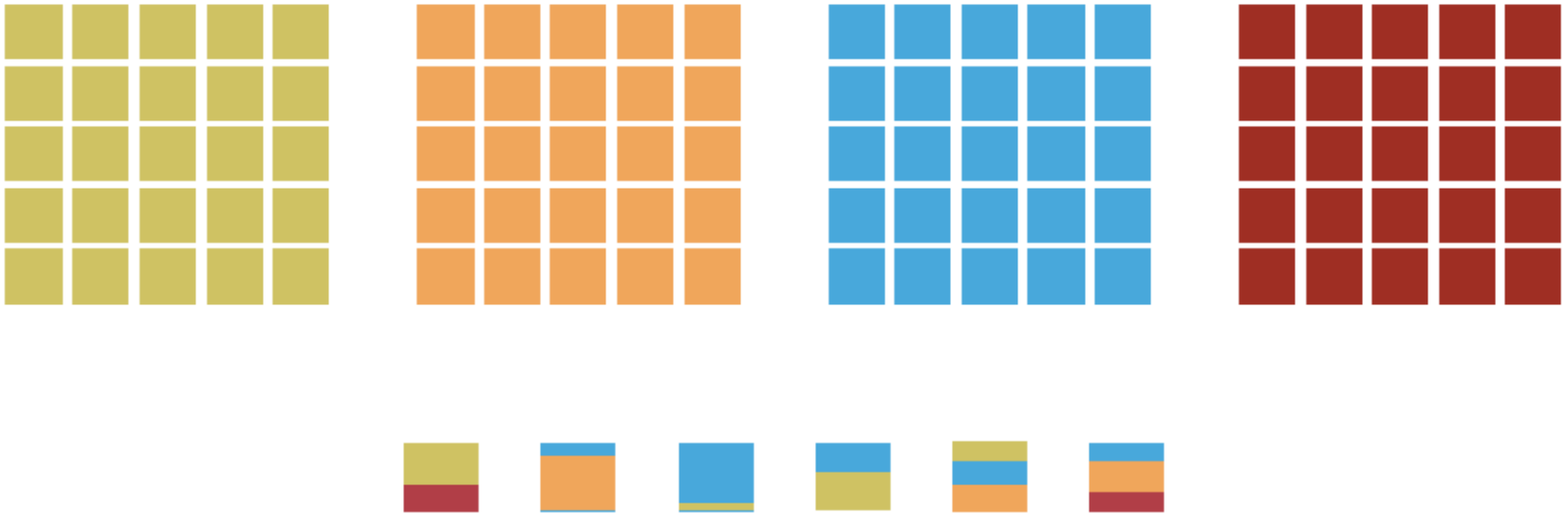
Adapted from John Novembre slide

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

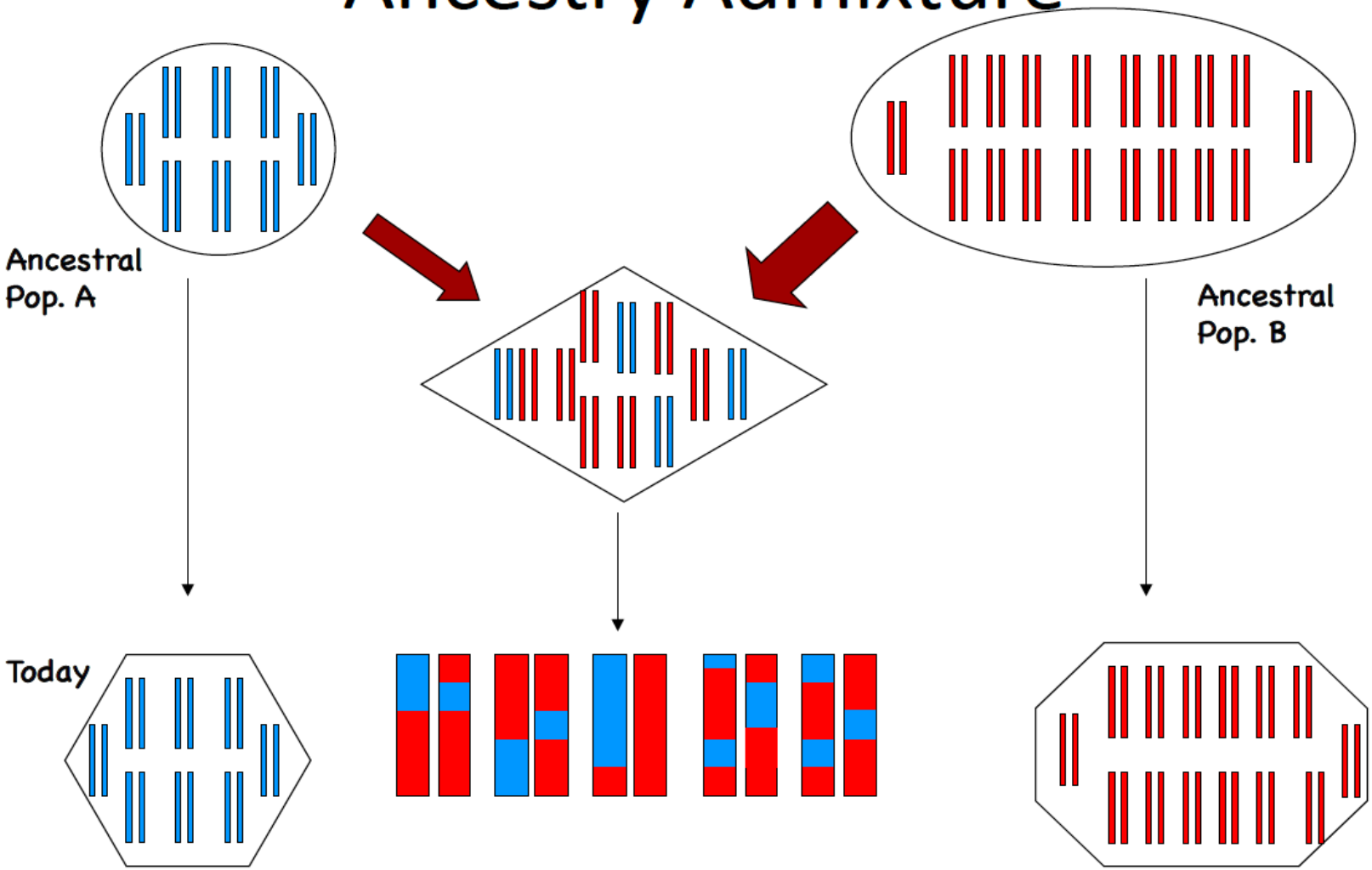


Identify **ancestry proportions** for individuals with **admixed** ancestry

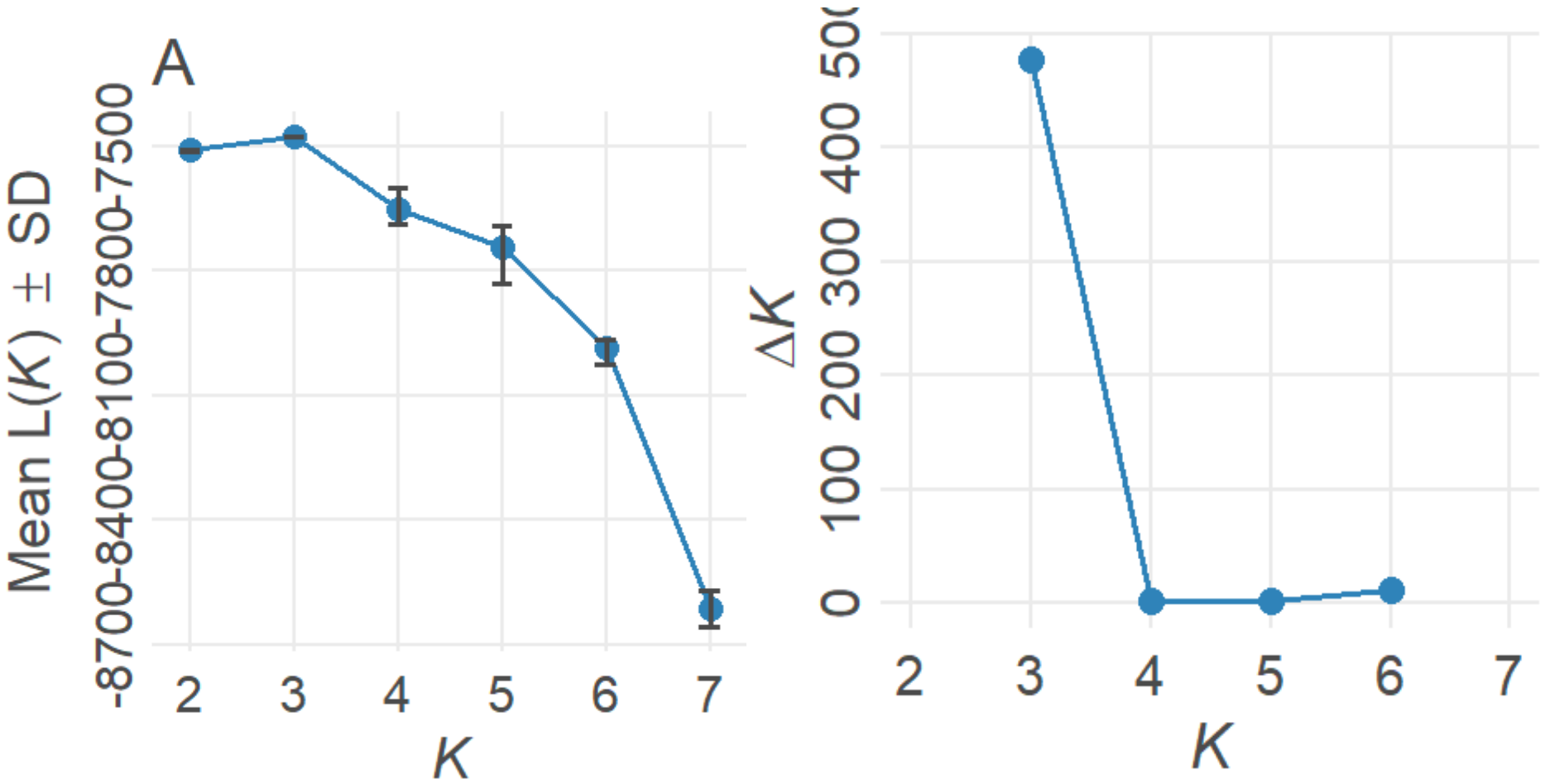
Approaches: Structure (MCMC, Bayesian)
Or ADMIXTURE (quadratic programming)

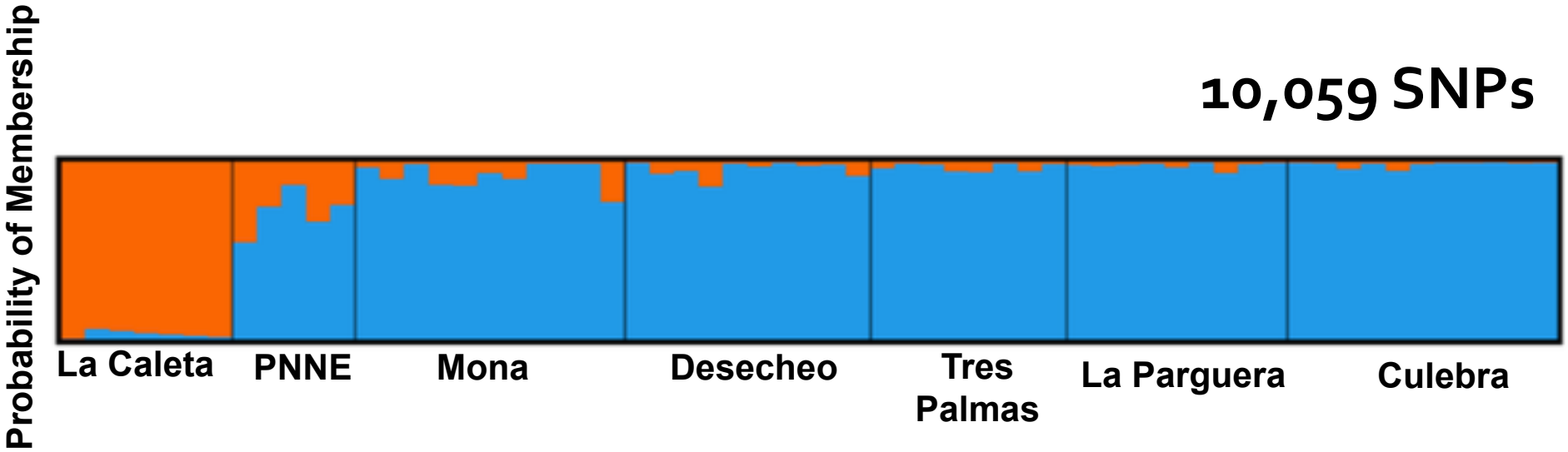
Adapted from John Novembre slide

Ancestry Admixture



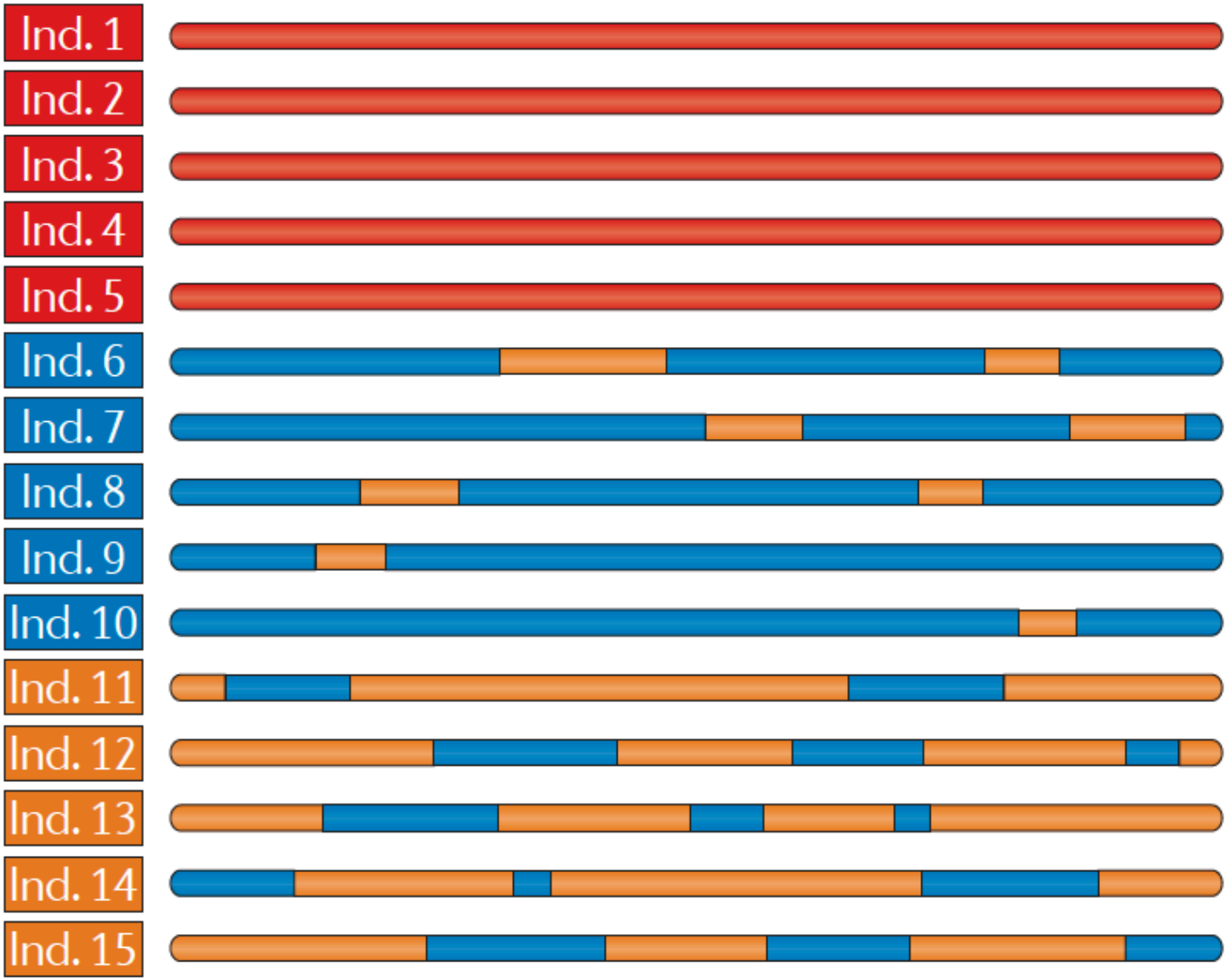
- The chromosomes of an admixed individual represent a mosaic of chromosomal blocks from the ancestral populations.


A
Mean L(K) ± SD
-8700-8400-8100-7800-7500
2
3
4
5
6
7
K
ΔK
0
100
200
300
400
500
2
3
4
5
6
7
K


10,059 SNPs
Probability of Membership
La Caleta
PNNE
Mona
Desecheo
Tres Palmas
La Parguera
Culebra

Chromosome painting

FineSTRUCTURE, HapMix, Globettroter

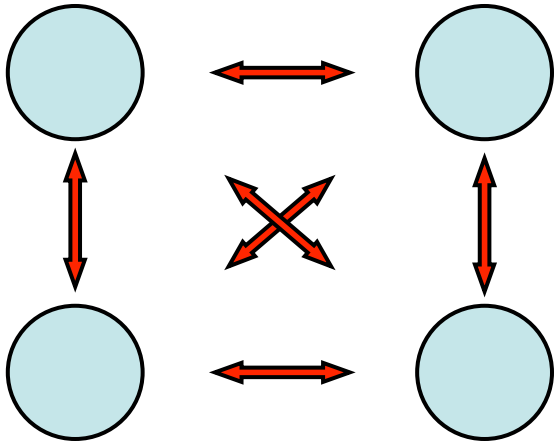


Schraiber and Akey (2015) *Nat Rev*

Methods to estimate migration

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

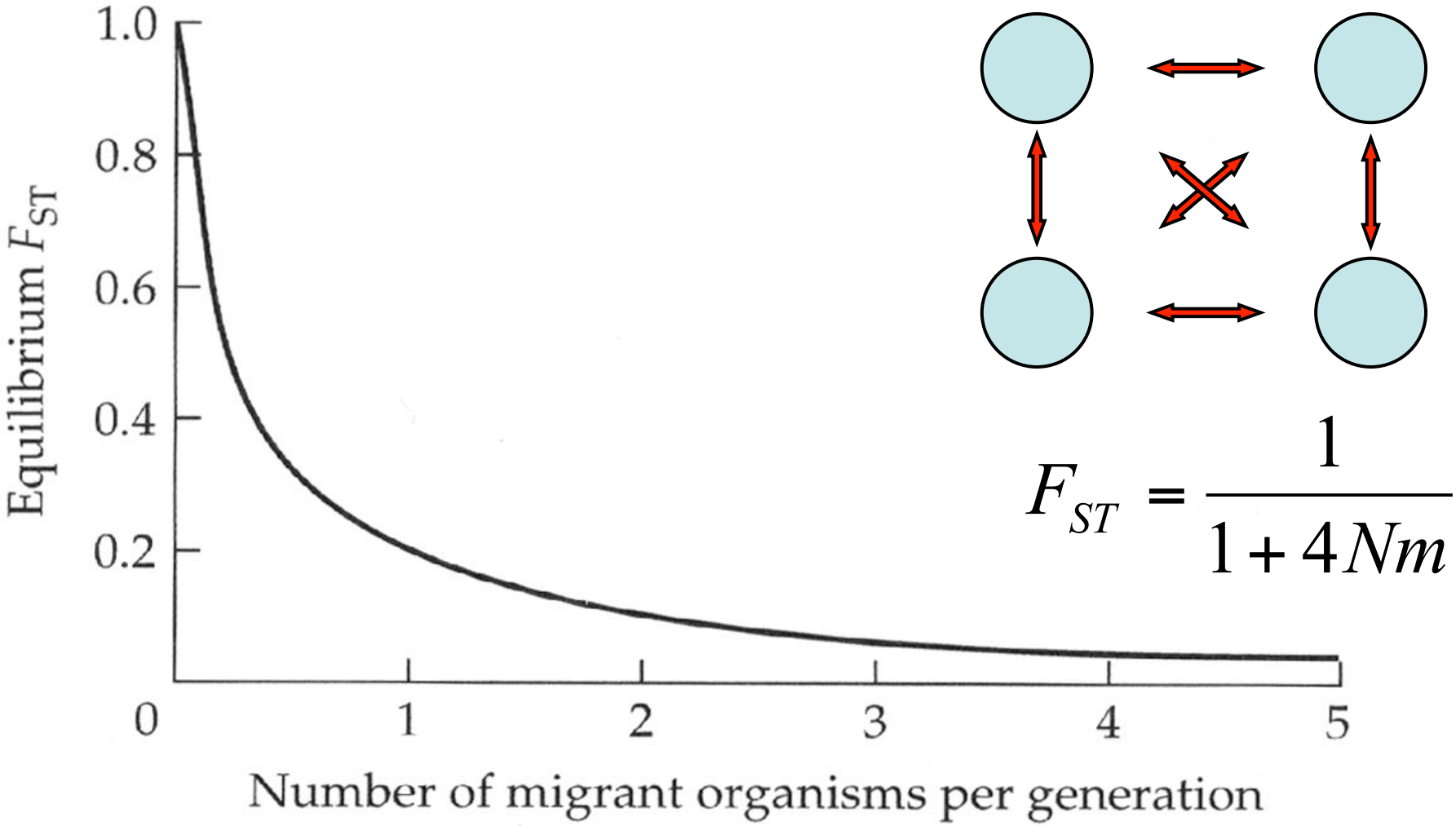
Nm is the absolute number of migrant organisms that enter each subpopulation per generation



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

RELATIONSHIP BETWEEN F_{ST} AND N*m* IN THE ISLAND MODEL

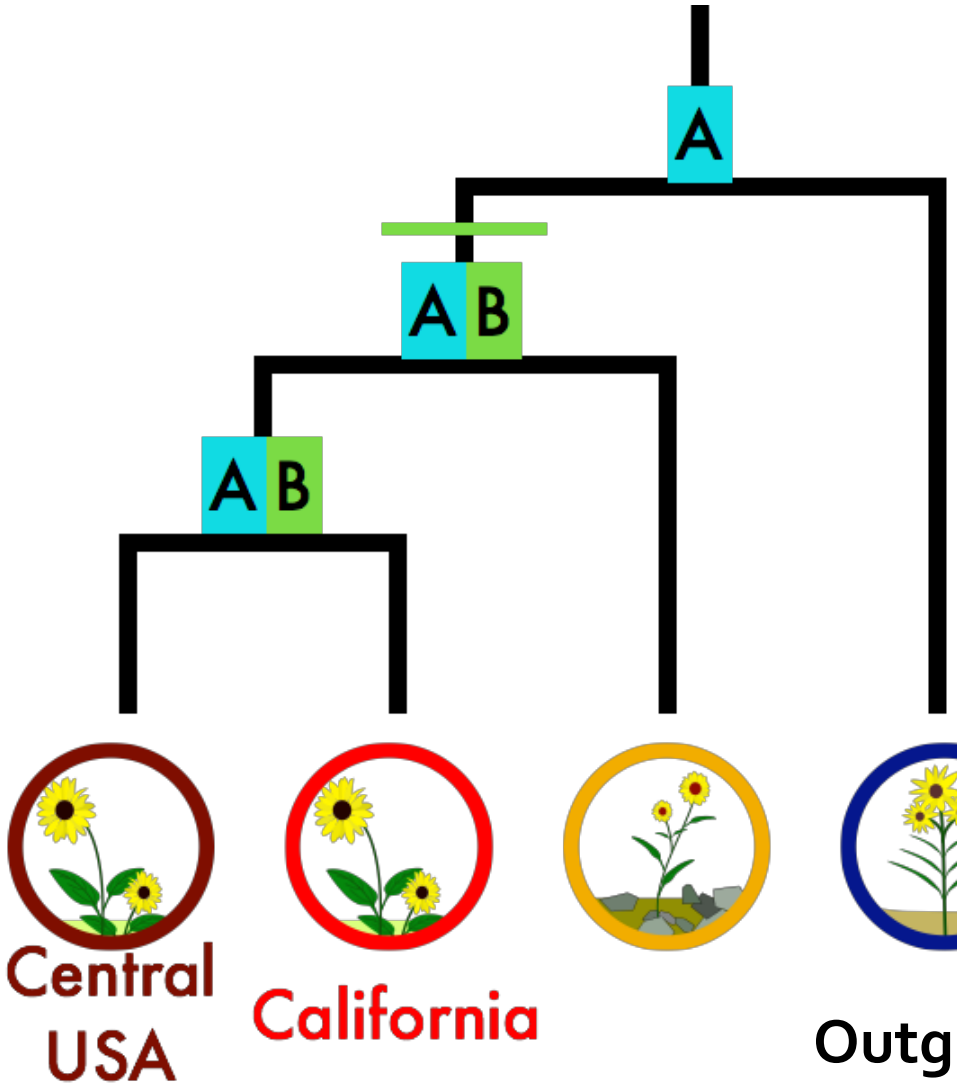
N*m* is the absolute number of migrant organisms that enter each subpopulation per generation



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

Hartl and Clark. 1997. *Principles of Population Genetics.* Sinauer Assoc.

ABBA-BABA test



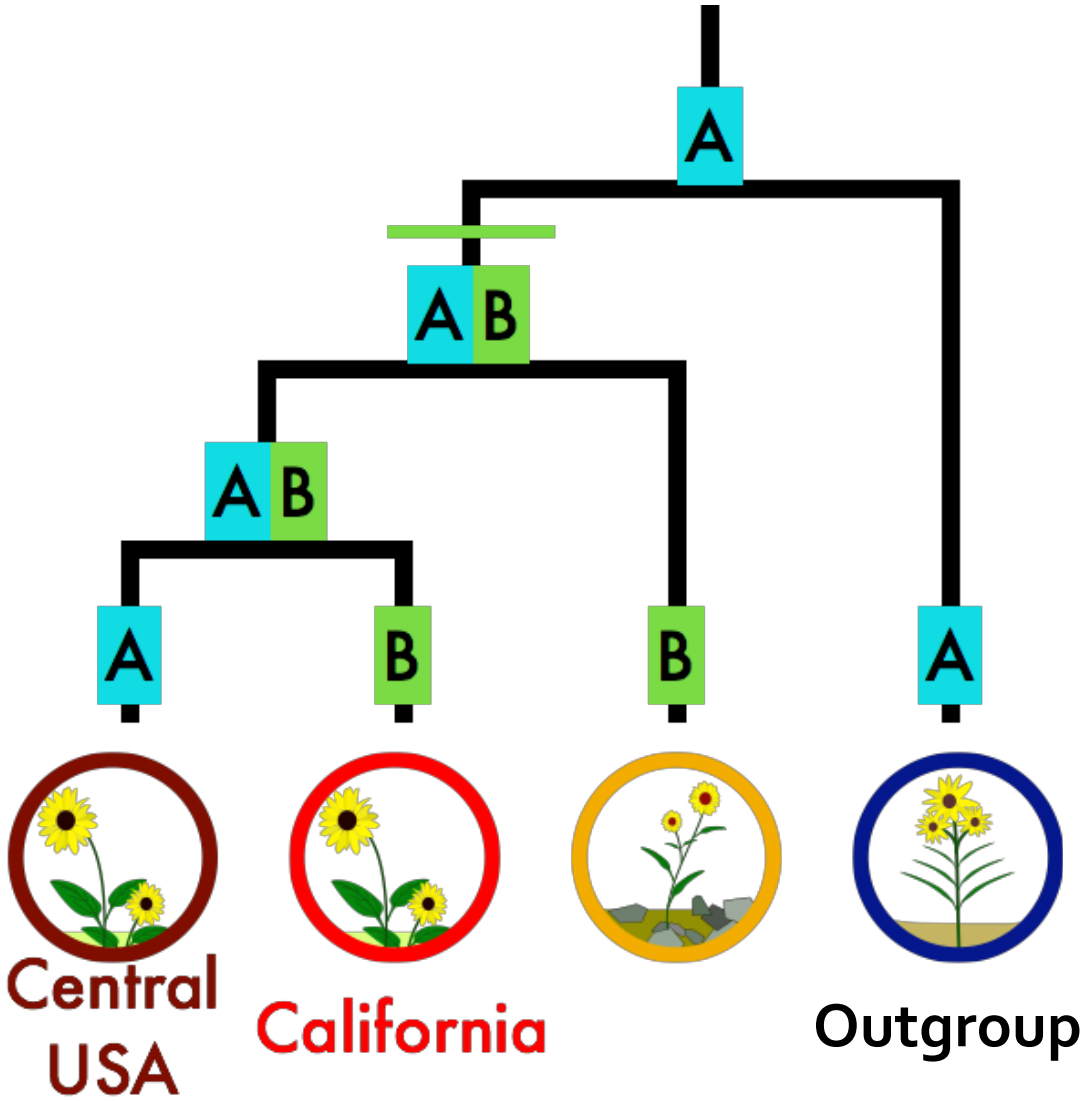
A = Ancestral

B = Derived

H. bolanderi

H. annuus

ABBA-BABA test



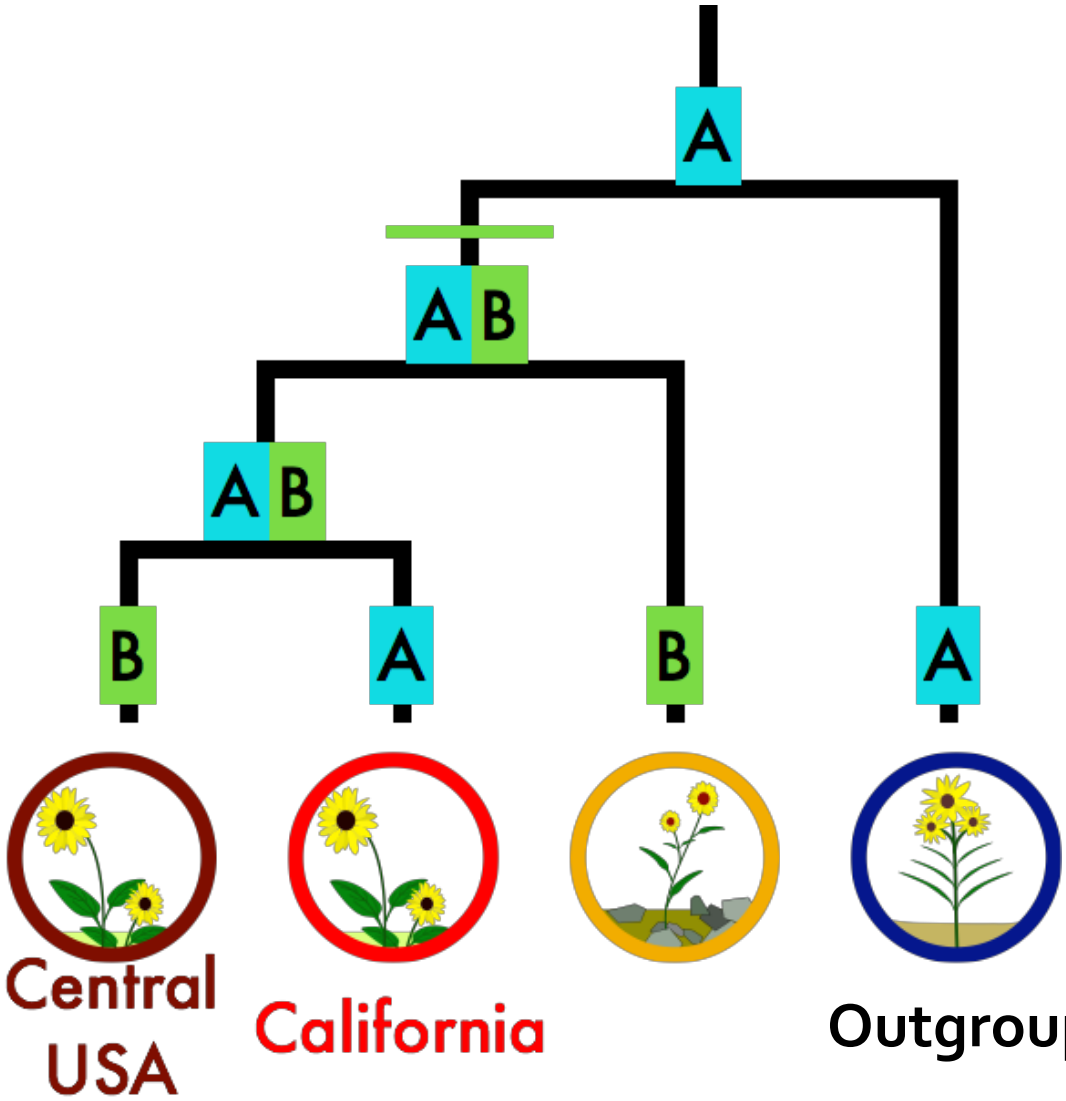
A = Ancestral

B = Derived

H. bolanderi

H. annuus

ABBA-BABA test



A = Ancestral

B = Derived

H. bolanderi

H. annuus

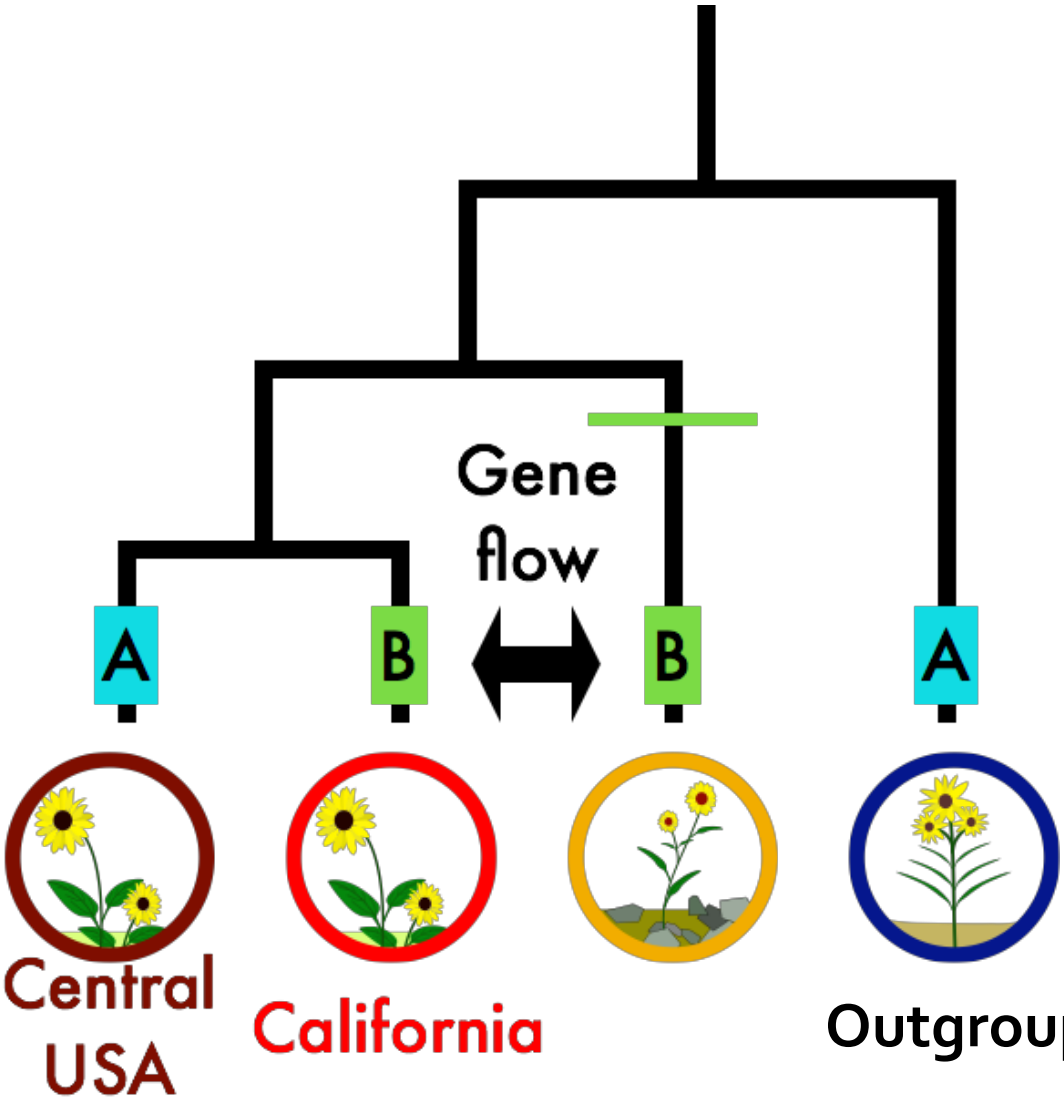
ABBA-BABA test

D-statistic = #ABBA - #BABA

Neutrality: D = 0

Gene flow: D = +

ABBA-BABA test



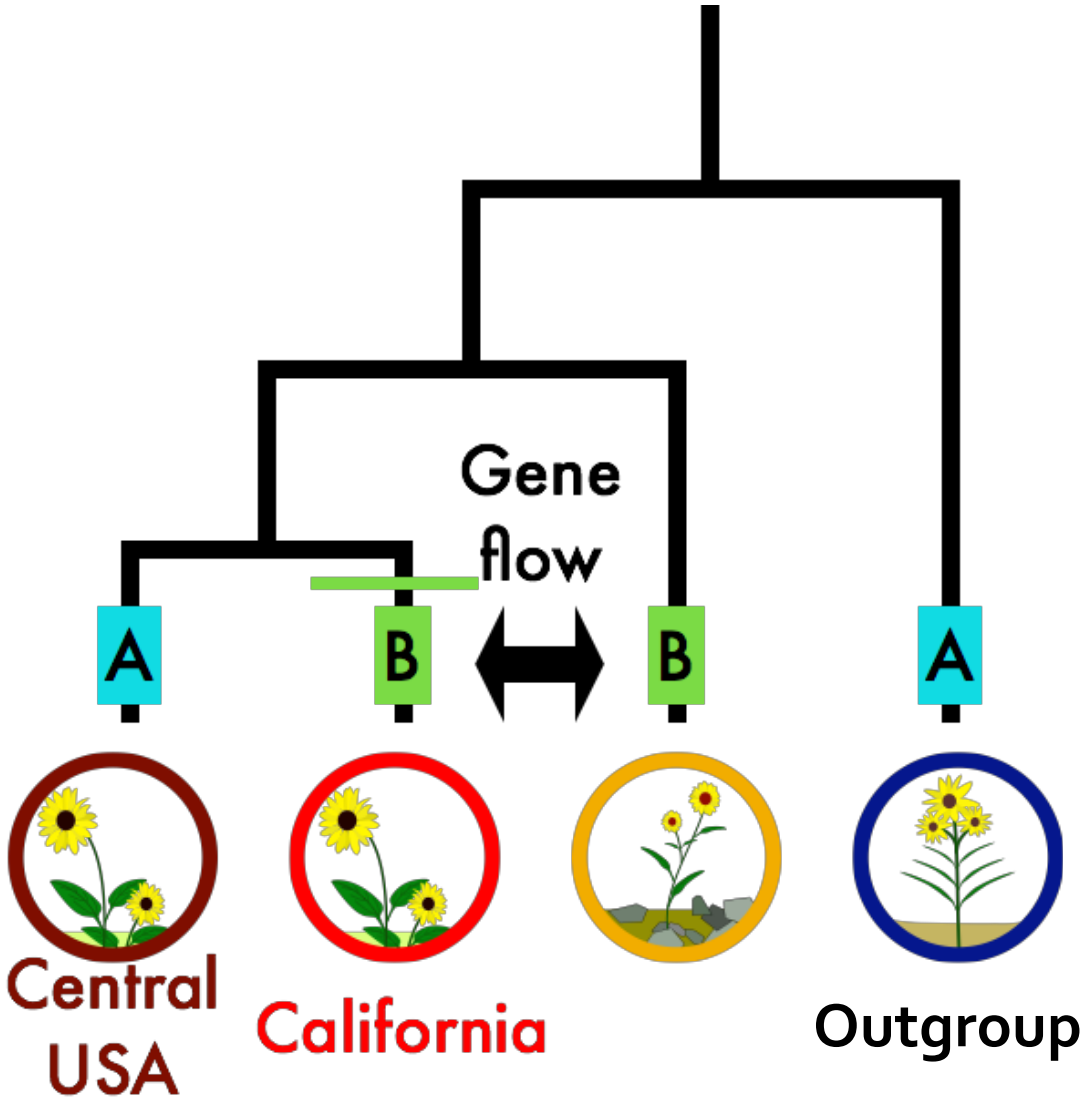
A = Ancestral

B = Derived

H. bolanderi

H. annuus

ABBA-BABA test

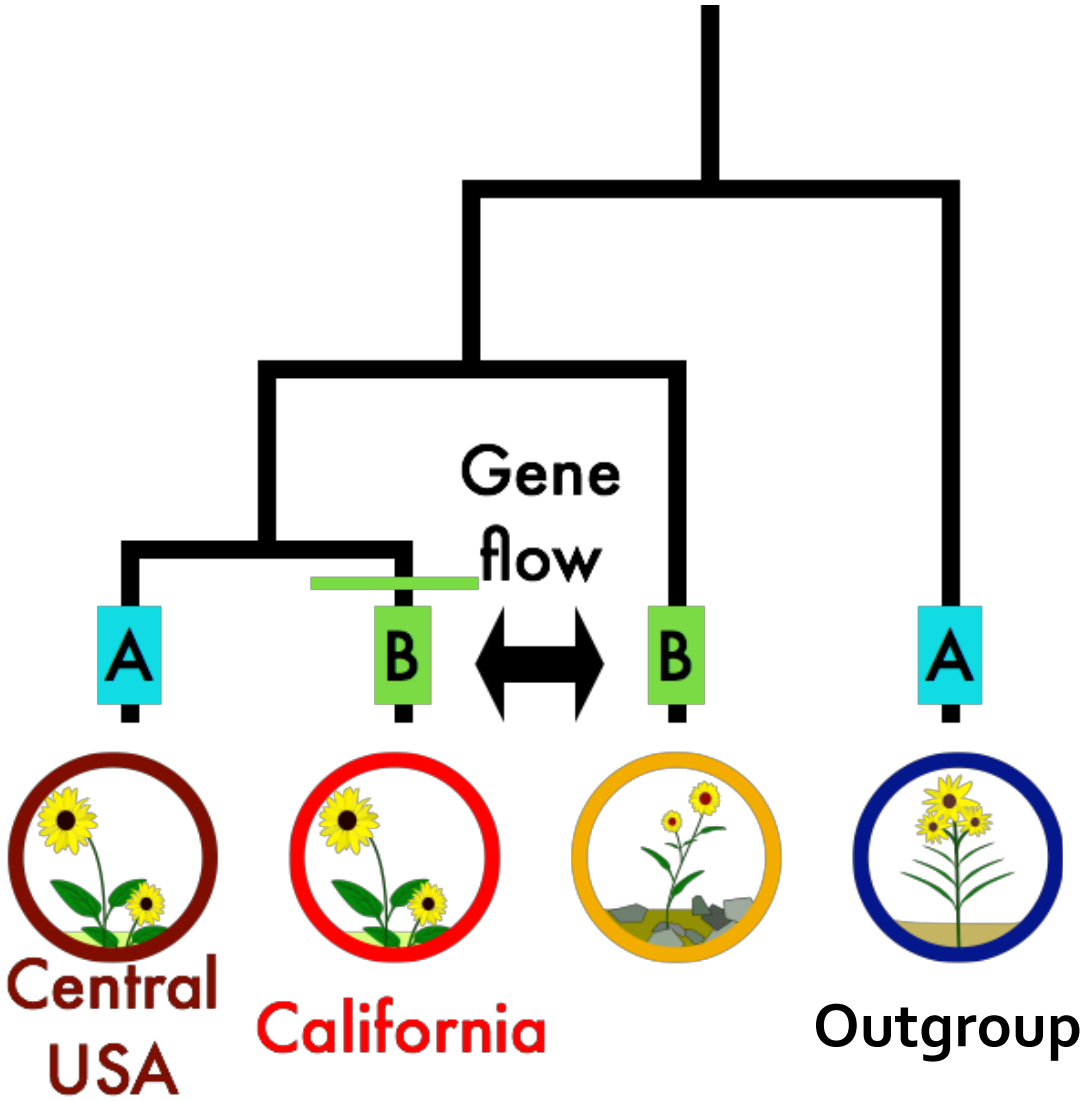


A = Ancestral
B = Derived

H. bolanderi

H. annuus

ABBA-BABA test

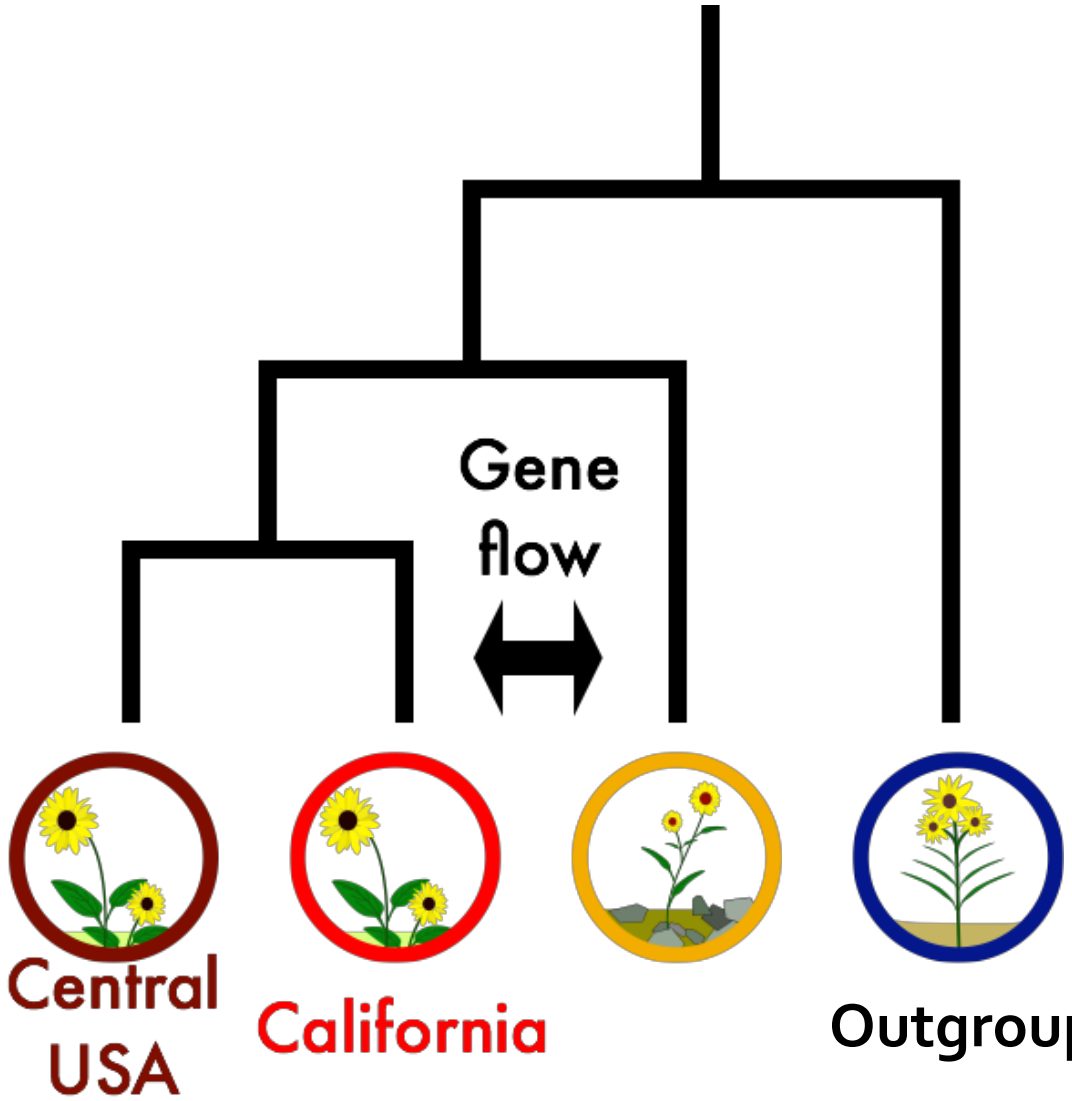


A = Ancestral
B = Derived

H. bolanderi

H. annuus

D statistic = 0.123 ± 0.033



ABBA > BABA

6-8% of the genome introgressed

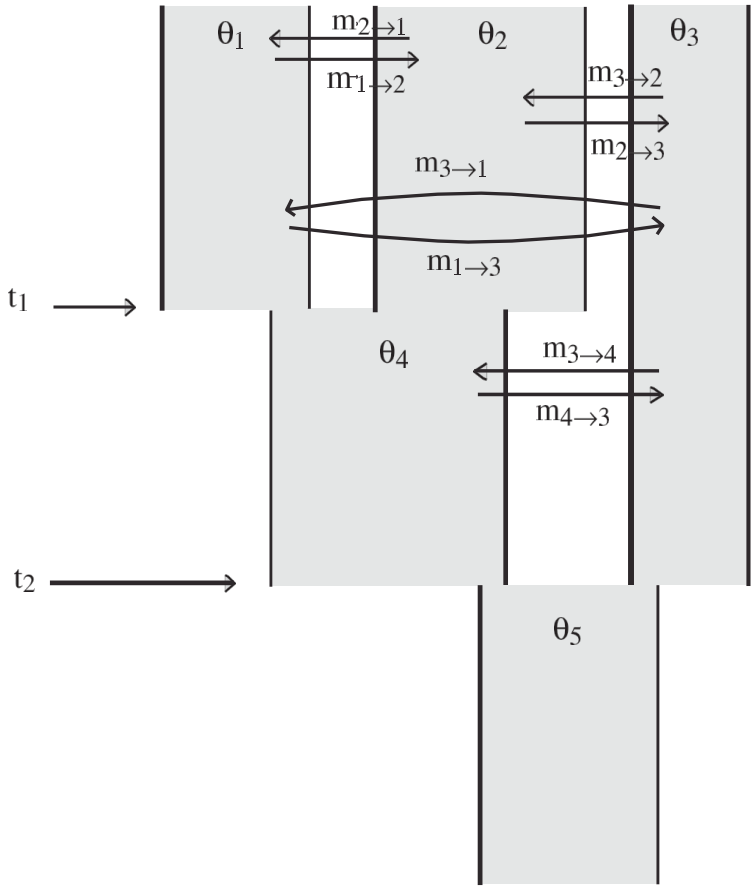
H. bolanderi

H. annuus

Methods to estimate both structure and migration

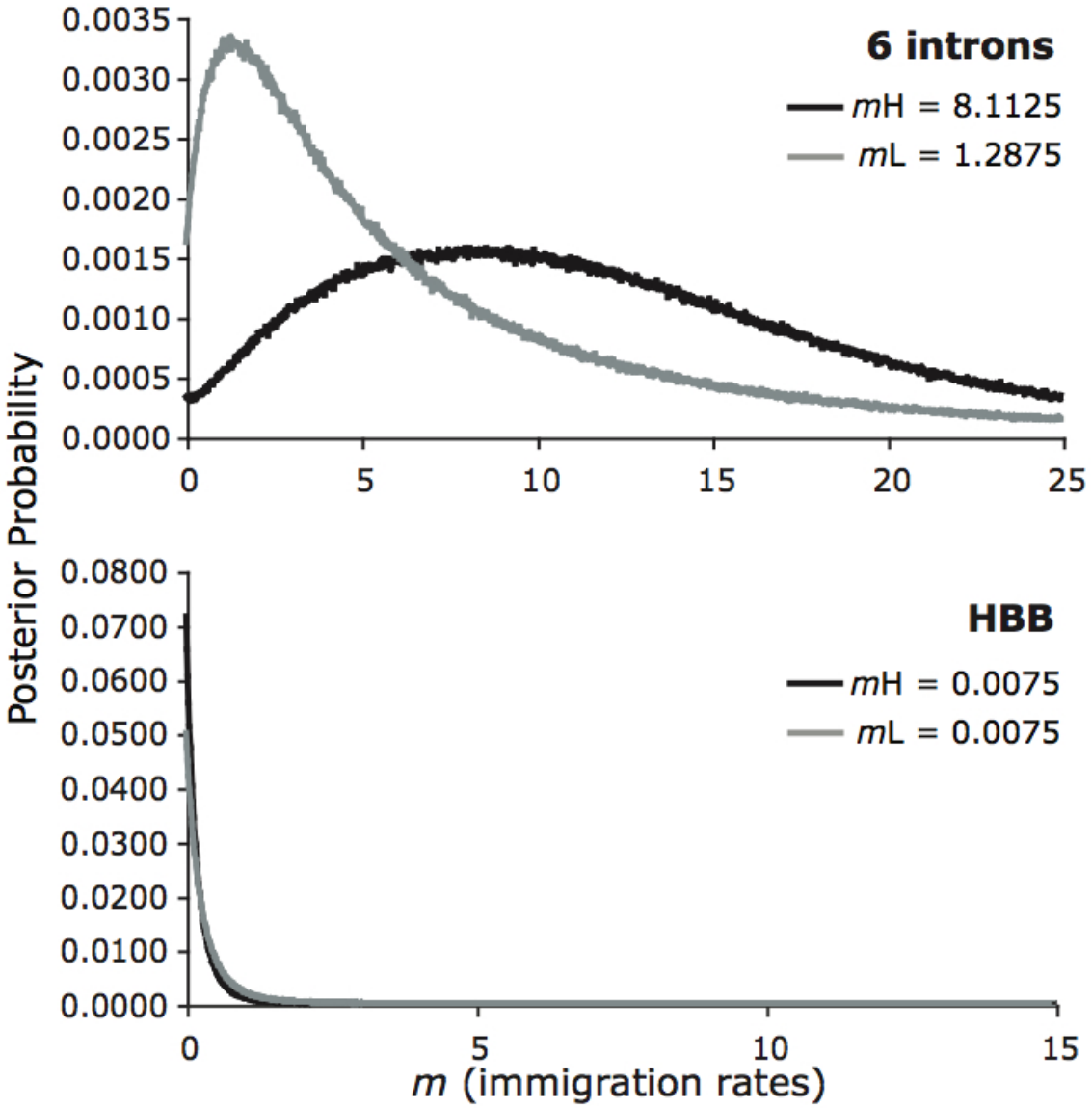
IM/IMa/IMa2

- 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.



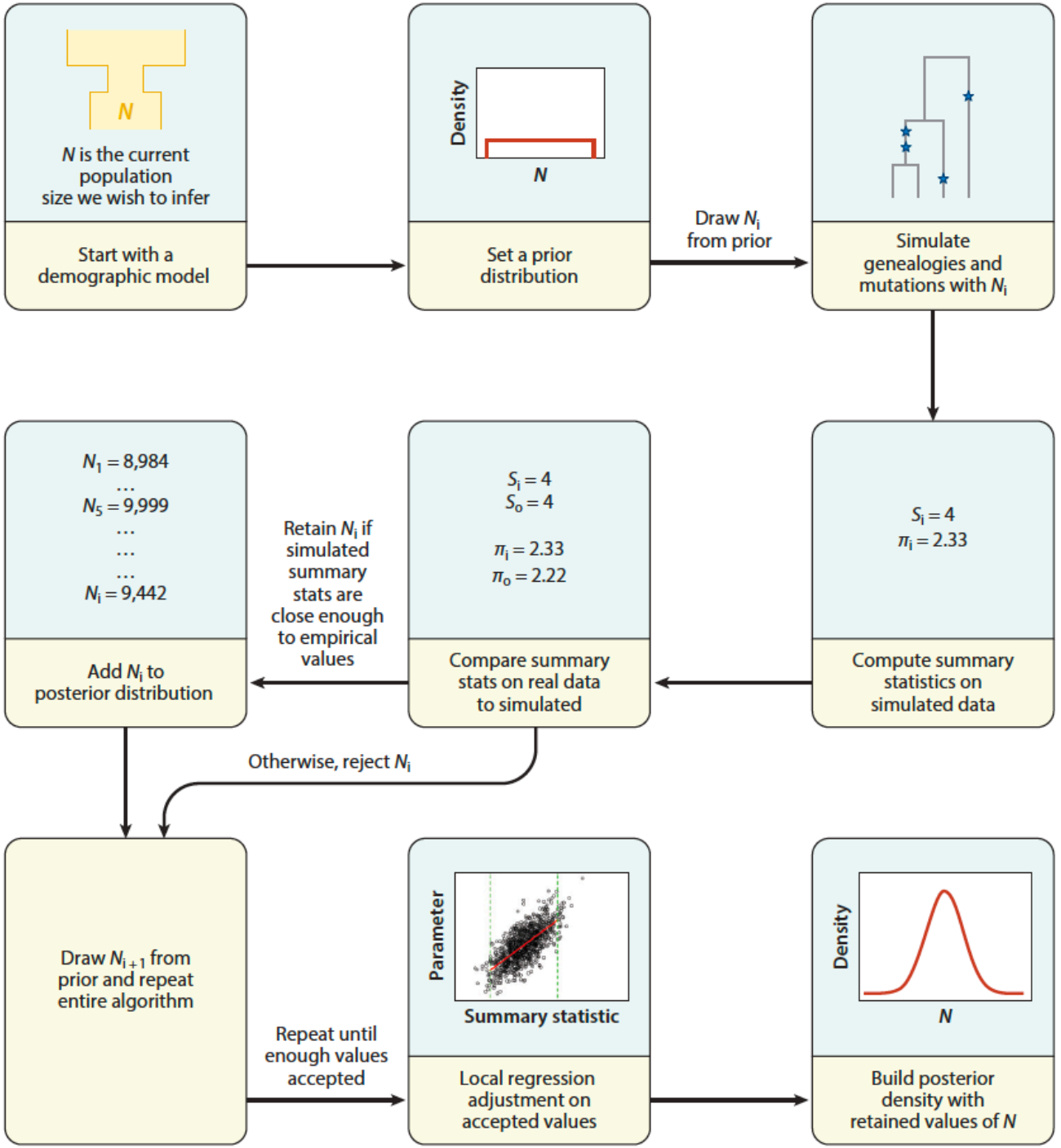
Hey and Nielsen (2004) *Genetics*
Hey (2010) *Mol Biol Evol*

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

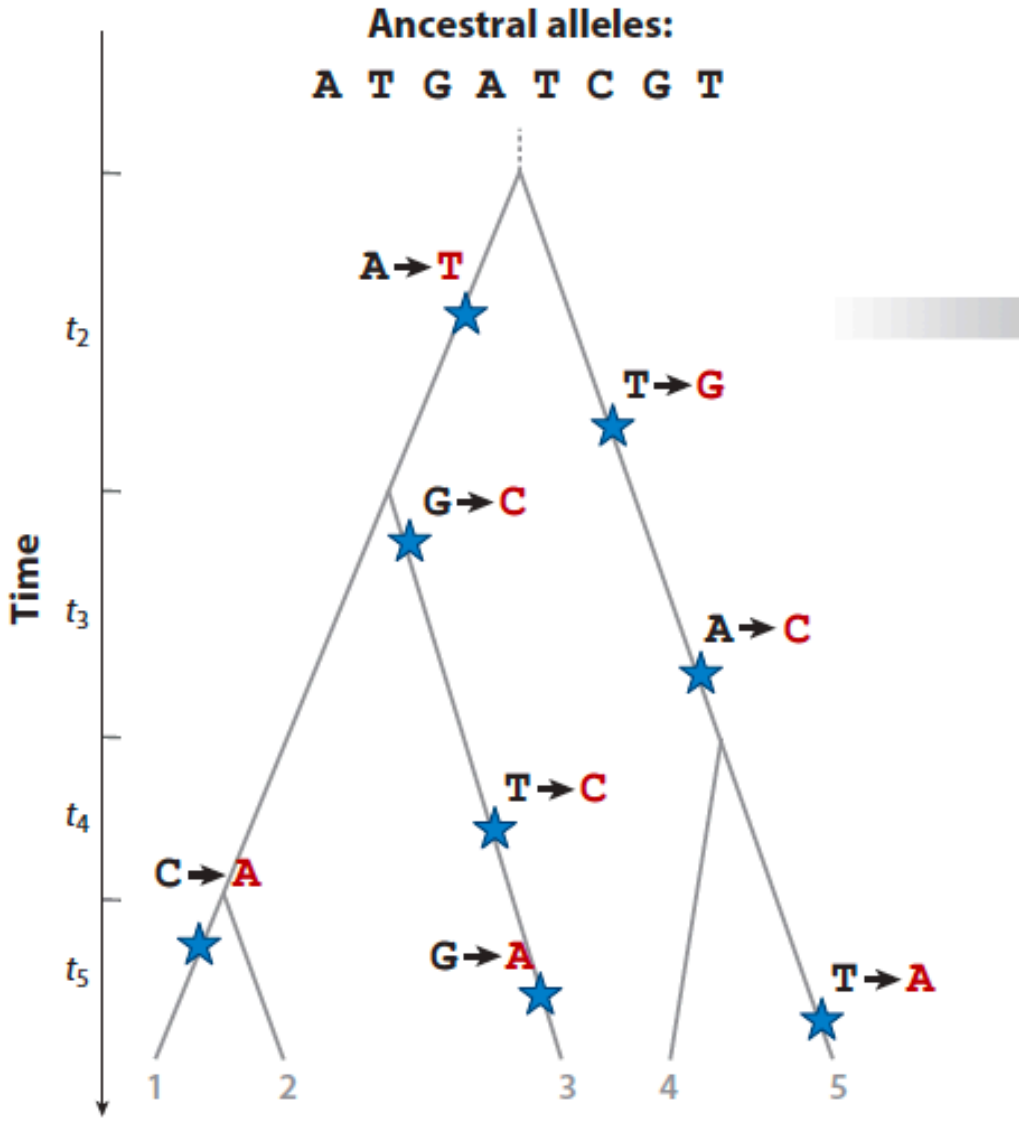


Bulgarella et al. (2012) showing inter-locus contrasts of migration rates in Crested Duck.

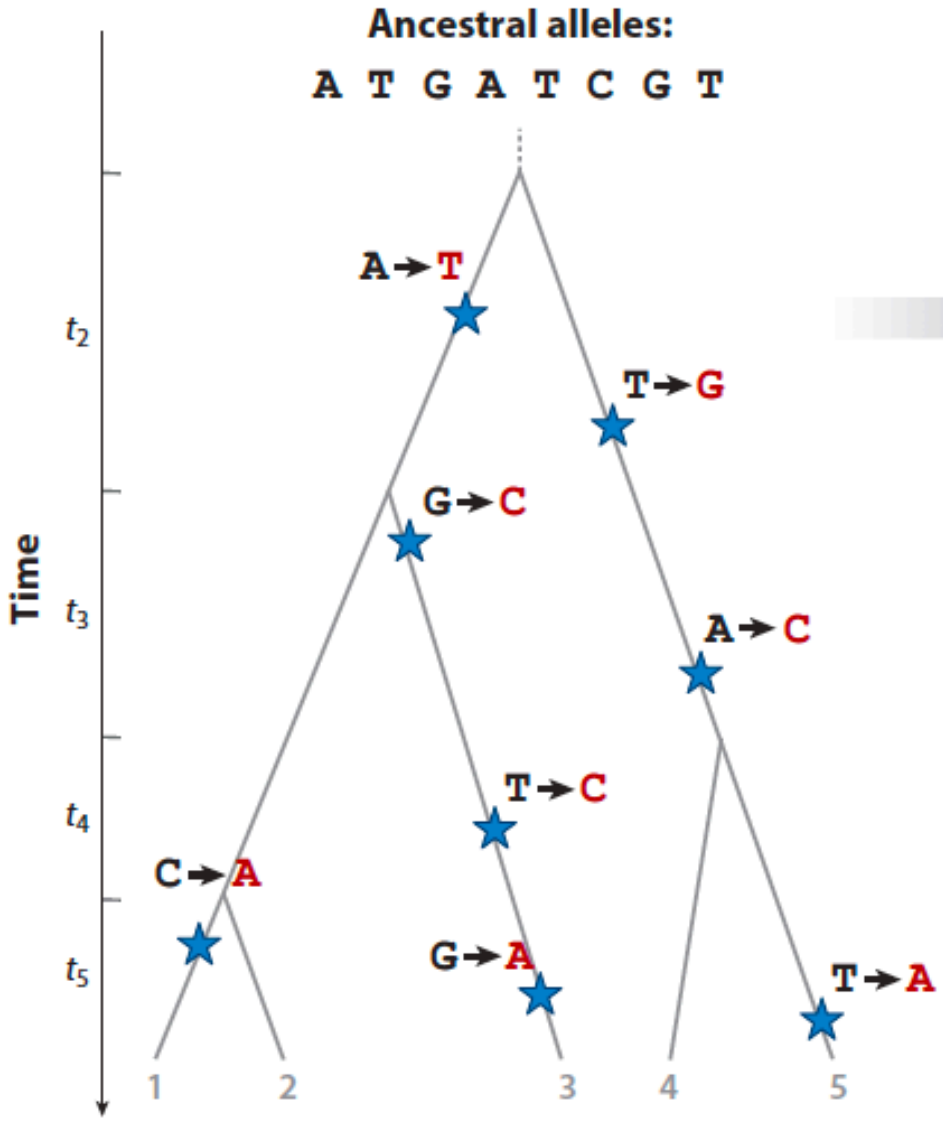
Approximate Bayesian computation



Allele Frequency Spectrum



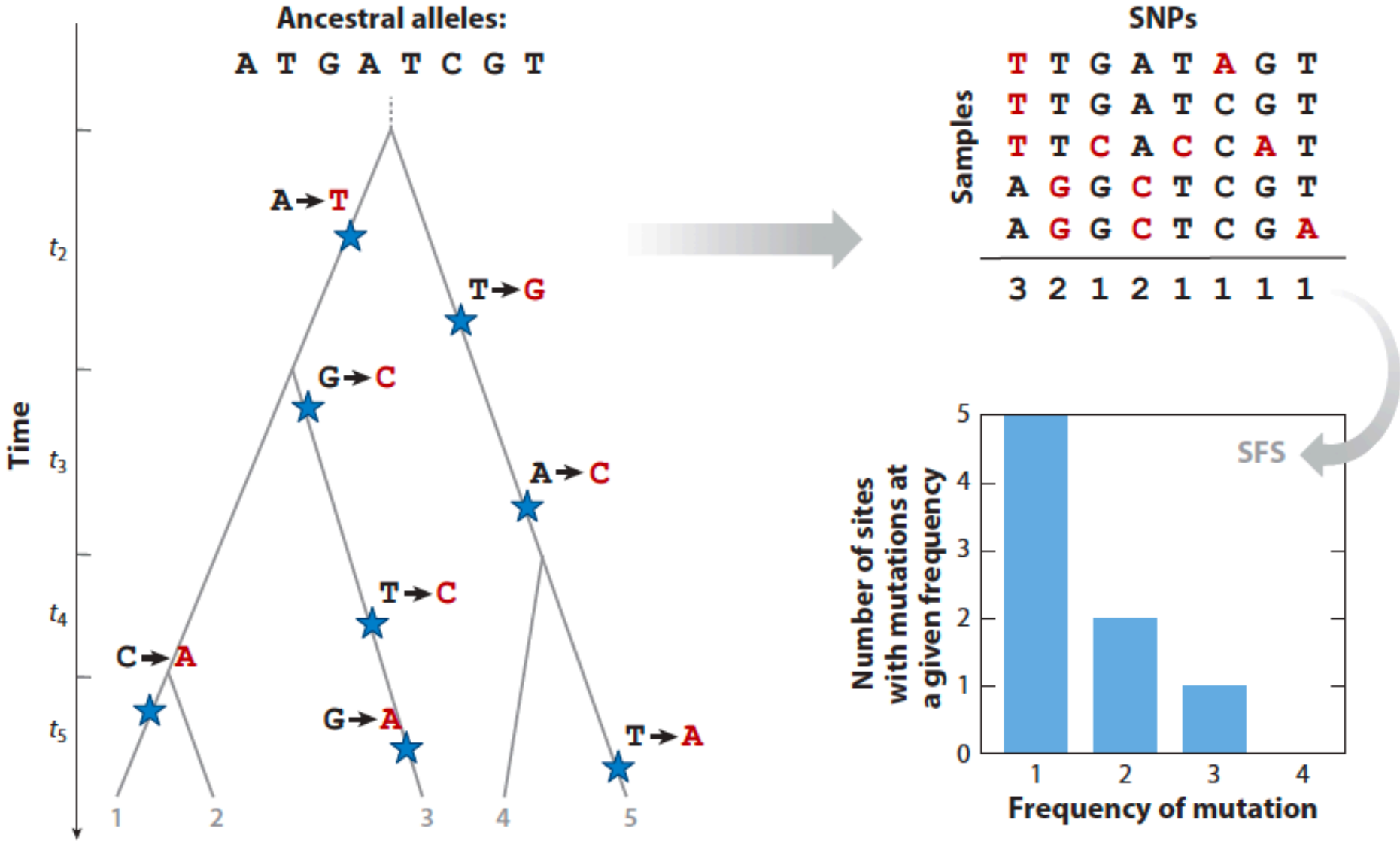
Allele Frequency Spectrum



SNPs

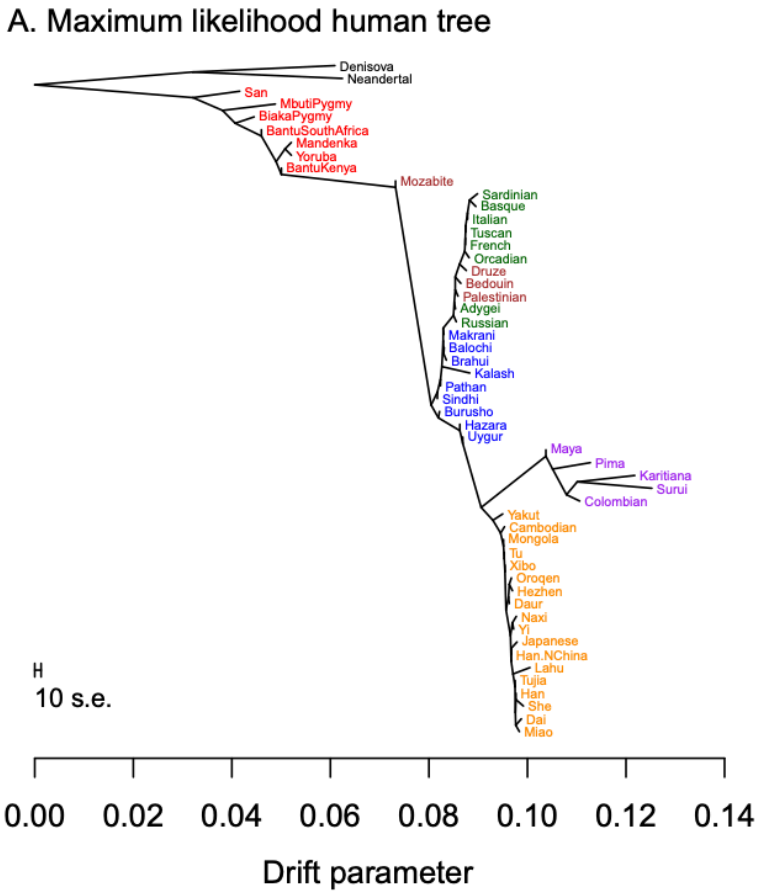
Samples								
	T	T	G	A	T	A	G	T
	T	T	G	A	T	C	G	T
	T	T	C	A	C	C	A	T
	A	G	G	C	T	C	G	T
	A	G	G	C	T	C	G	A
	3	2	1	2	1	1	1	1

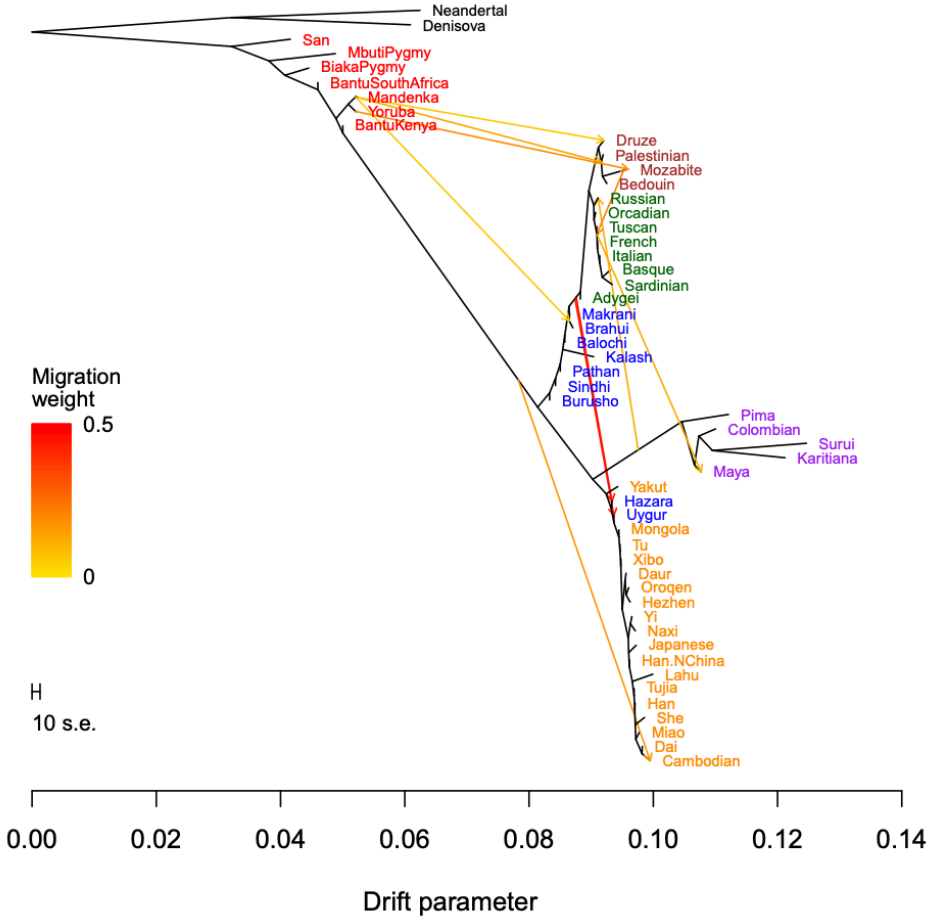
Allele Frequency Spectrum



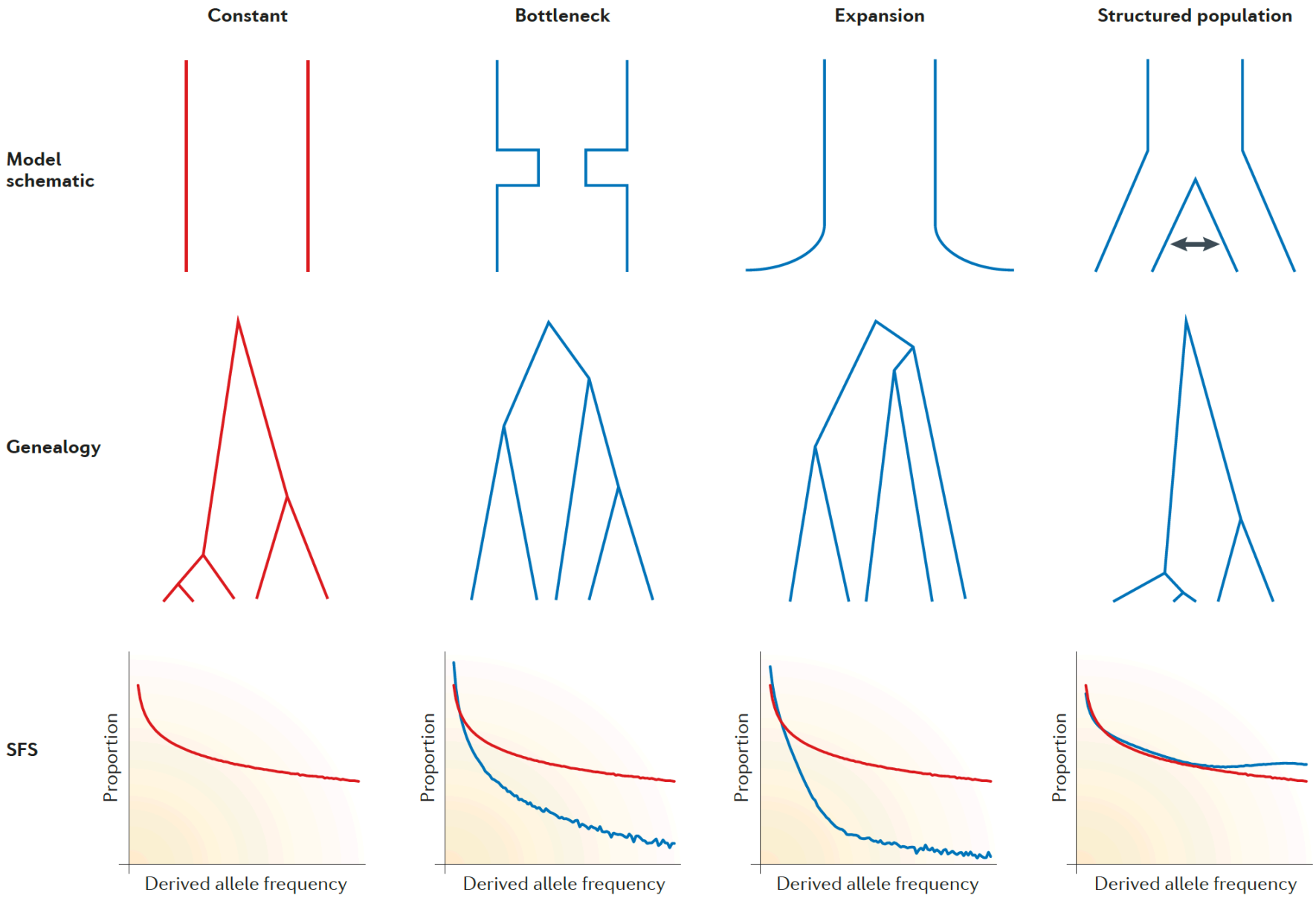
Beichman et al. 2018 Annu. Rev. Ecol. Evol. Syst. 2

TreeMix: Structure with migration bands





Allele Frequency Spectrum



Schraiber and Akey (2015) *Nat Rev Gen*

Model based methods

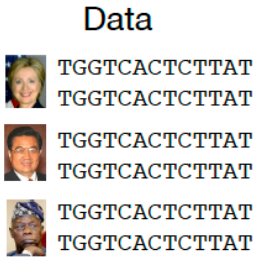
1. Collect data
2. Develop models of evolution
3. Calculate likelihood support for models based on your data

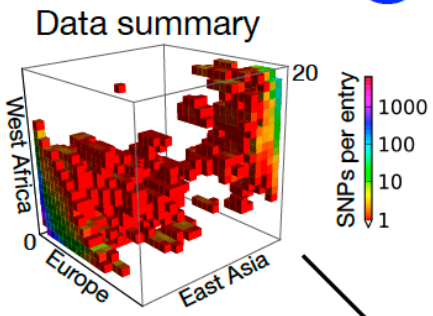
Model based methods

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

Modeling workflow

∂a∂i
fastsimcoal2

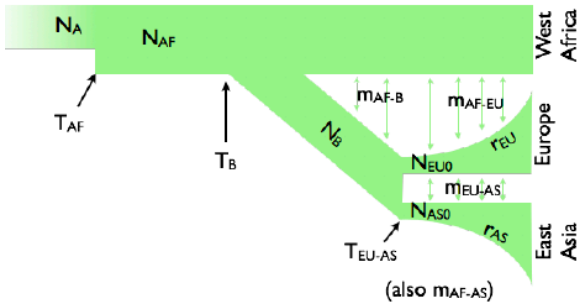




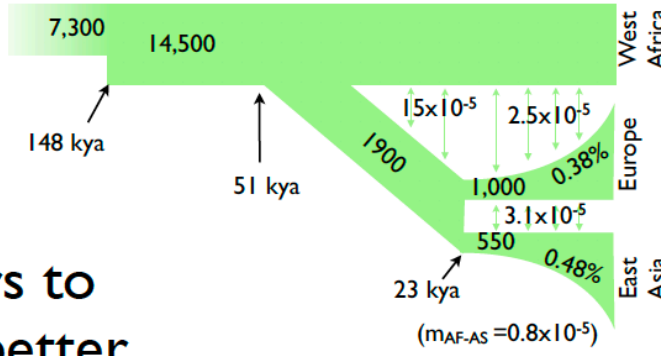
Compare simulations with data.

Choose candidate parameter values, simulate data.

Input parameterized model



Converged output model



Update parameters to (hopefully) fit data better.

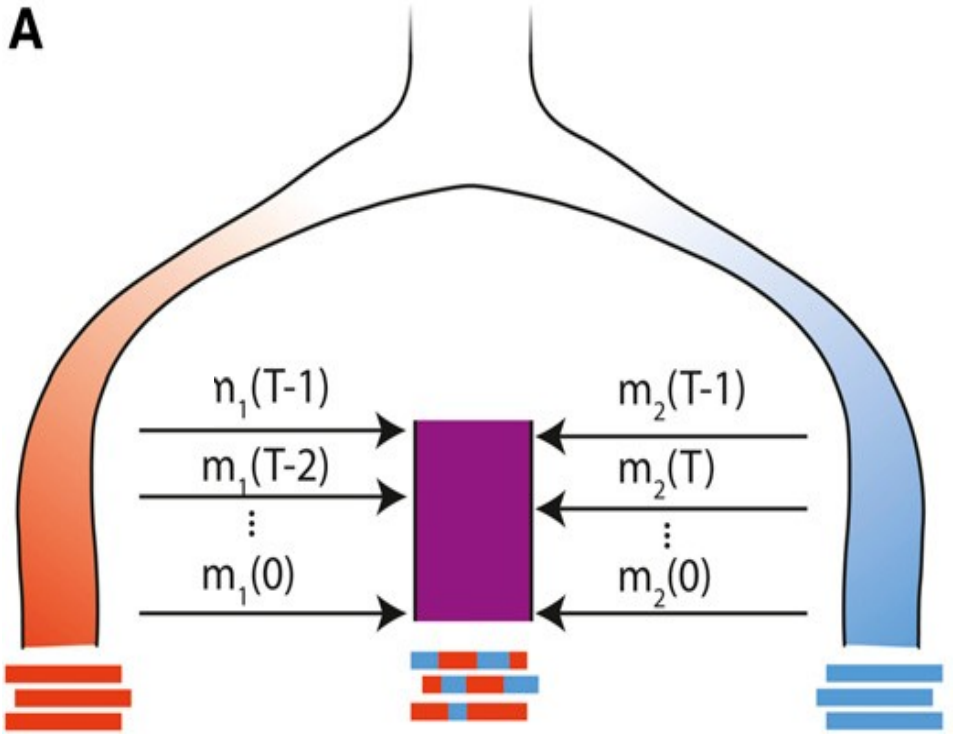
Gutenkunst et al.
PLoS Genet (2009)

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



Gravel et al.
Genetics (2012)

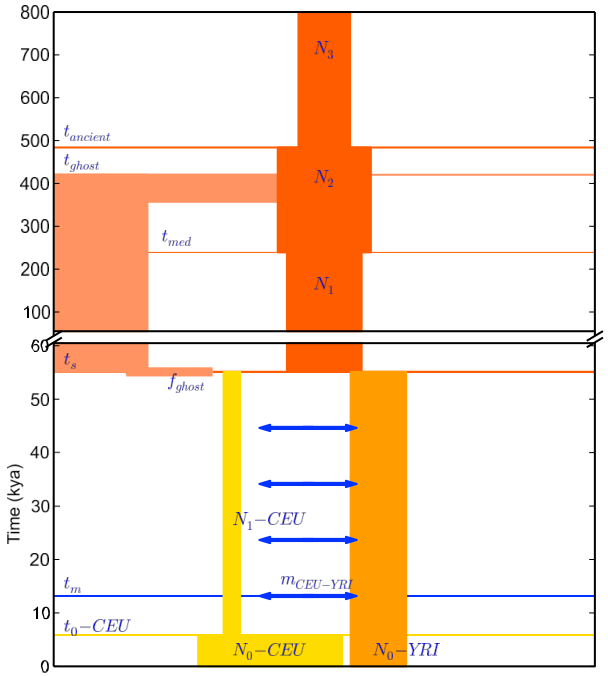
- Sequences that are Identical By State (IBS) with and between populations are informative about demographic history

- Calculate expected spectrum of IBS tract lengths using coalescent theory

- Can fit very complex models

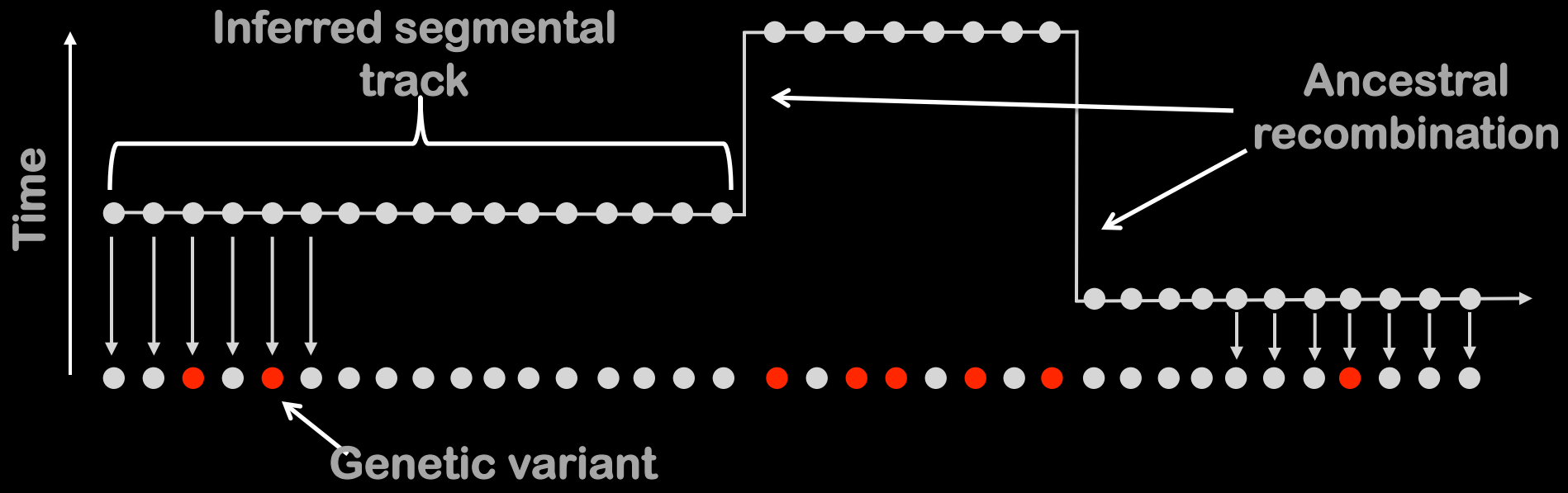
AGGTCGAGCTTG
ACGTCGAGCTGG

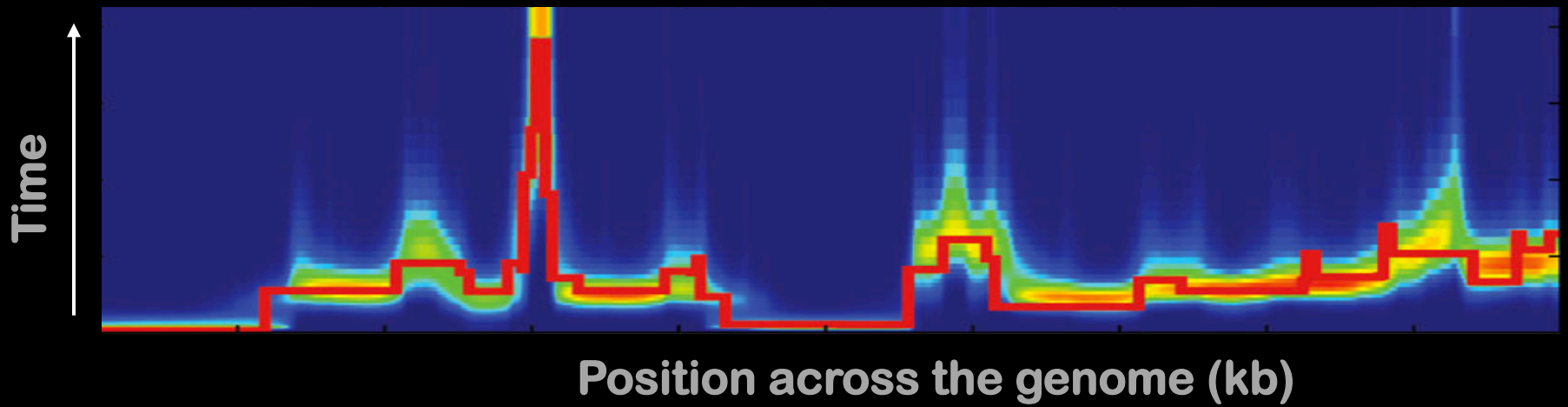
L bases of IBS



Harris and Nielsen (2013) *PLoS Genet*

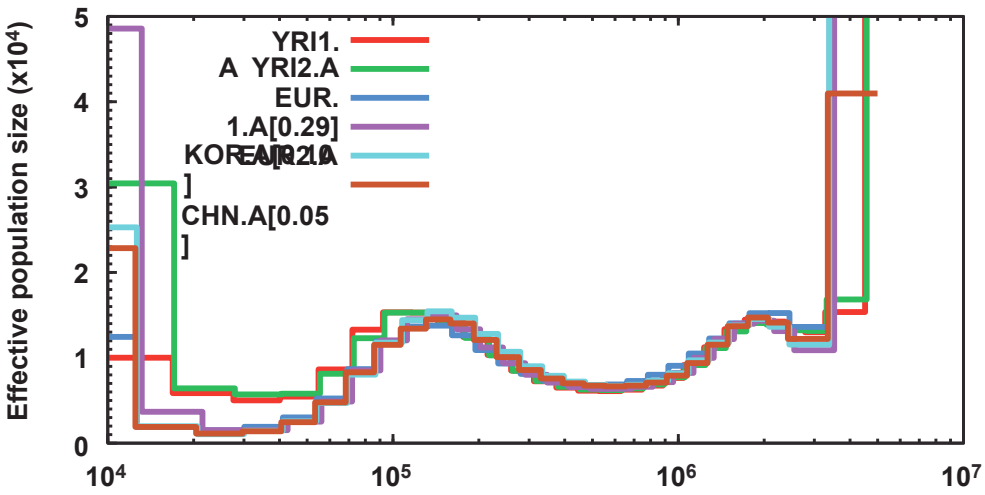
We can infer population sizes using genetic variation across the genome





Li and Durbin 2011 Nature 475

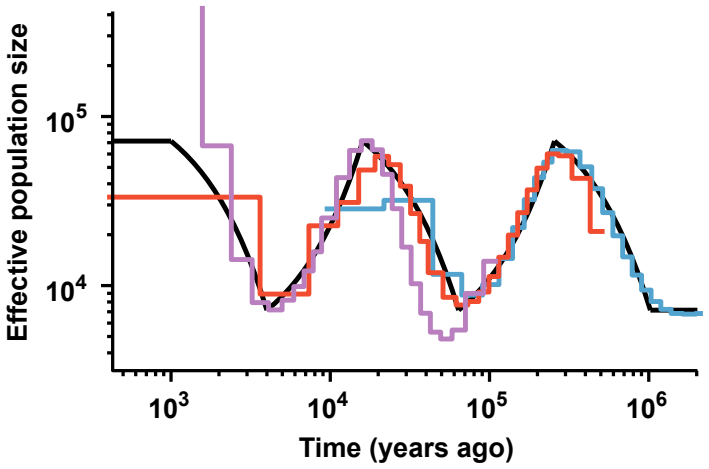
PSMC

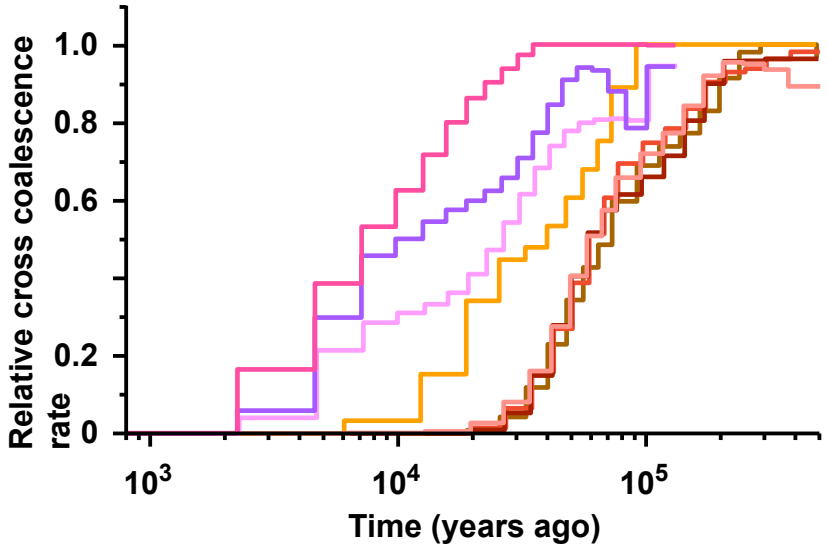


- **Estimate effective population size over time from a single unphased genome.**
- **No parametric model (e.g. exponential growth) assumed.**

Li and Durbin (2011) *Nature*

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.

Schiffels and Durbinn *Nature Genet* (2014)

For Tuesday we will be discussing:

Moreno-Mayar *et al.* 2018

and

Rougemont *et al.* 2018