Molecular Ecology II

(AND MIGRATION)

George Bonsall
Old Dominion University
Biology Graduate Program
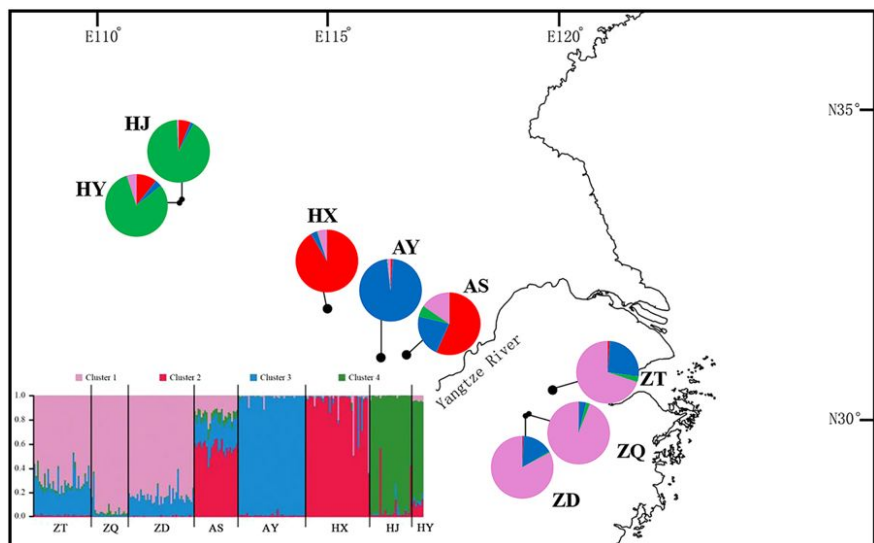
OLD DOMINION
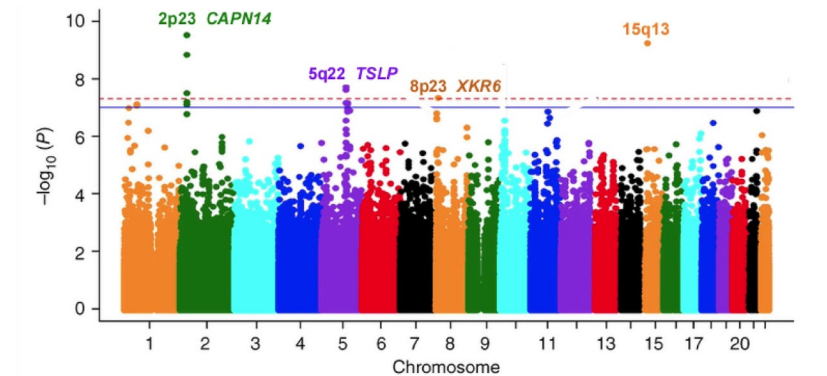UNIVERSITY

Where we have been:

- Molecular Ecology I
- DNA preservation and extraction methods
- Data filtering and Quality Control
- Sequencing
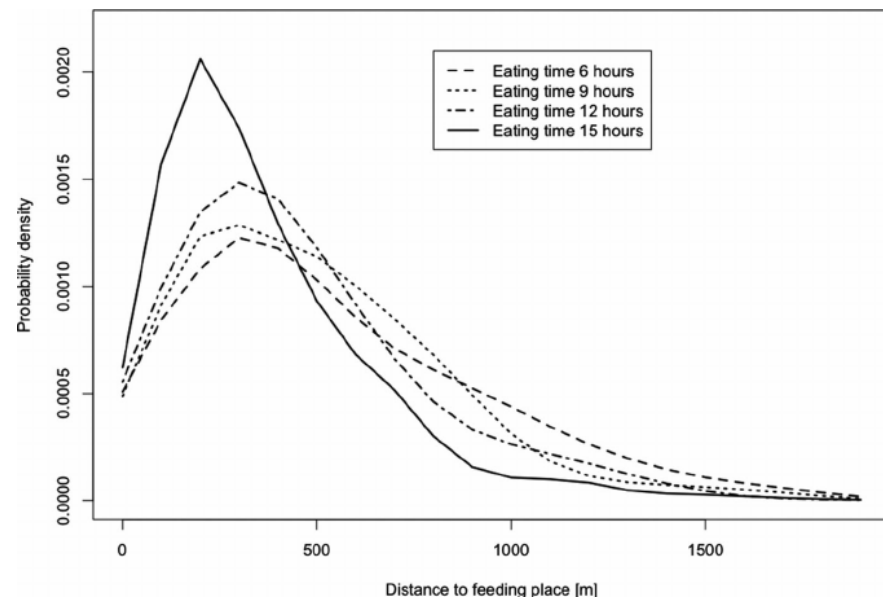- Genomic Assembly

Where we are going

Geng et al. 2017 Nature Scientific Reports 7:17343

Kottyan et al. 2014 Nature Genetics 46:895-900

Stevenson et al. 2014 International Journal of Primatology 35:1021-1036

Outline

1. Allele frequencies & genotype frequencies
2. Hardy Weinberg Equilibrium (HWE)
3. Genetic drift
4. Heterozygosity and other diversity statistics
5. Inbreeding coefficient
6. Linkage disequilibrium

Allele frequencies

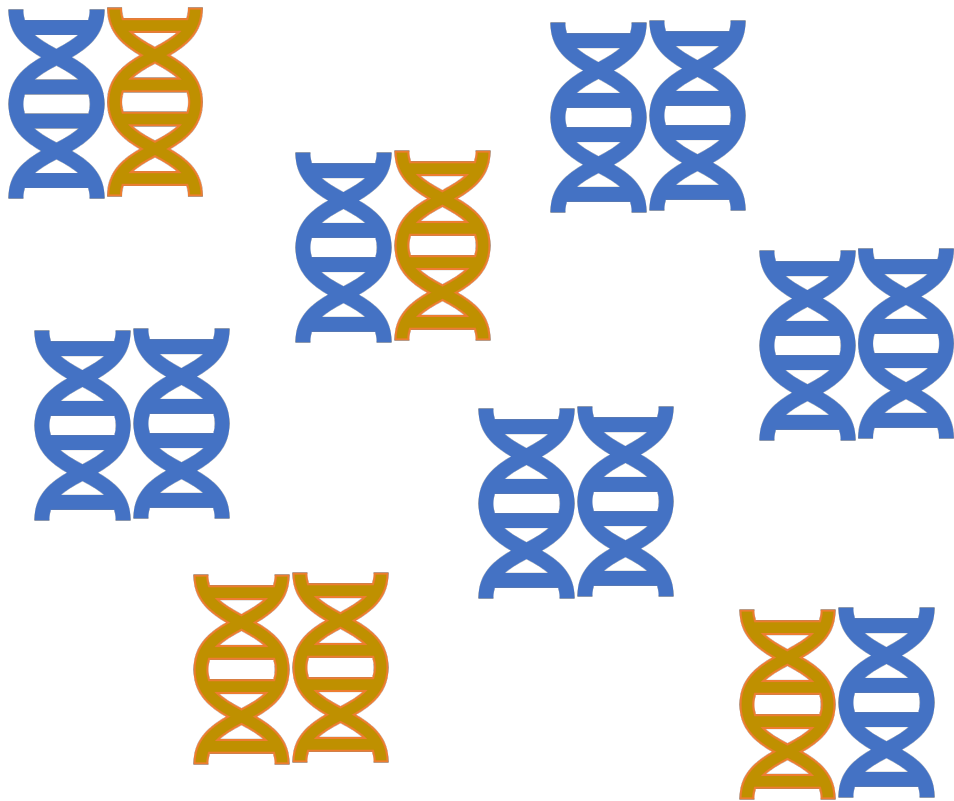Fraction of all **chromosomes** in a population that carry a given allele

Allele - one of two or more versions of DNA sequence at a given genomic location.

An individual inherits two alleles, one from each parent, for any given genomic location where such variation exists.

If the two alleles are the same, the individual is homozygous for that allele, if not, heterozygous.
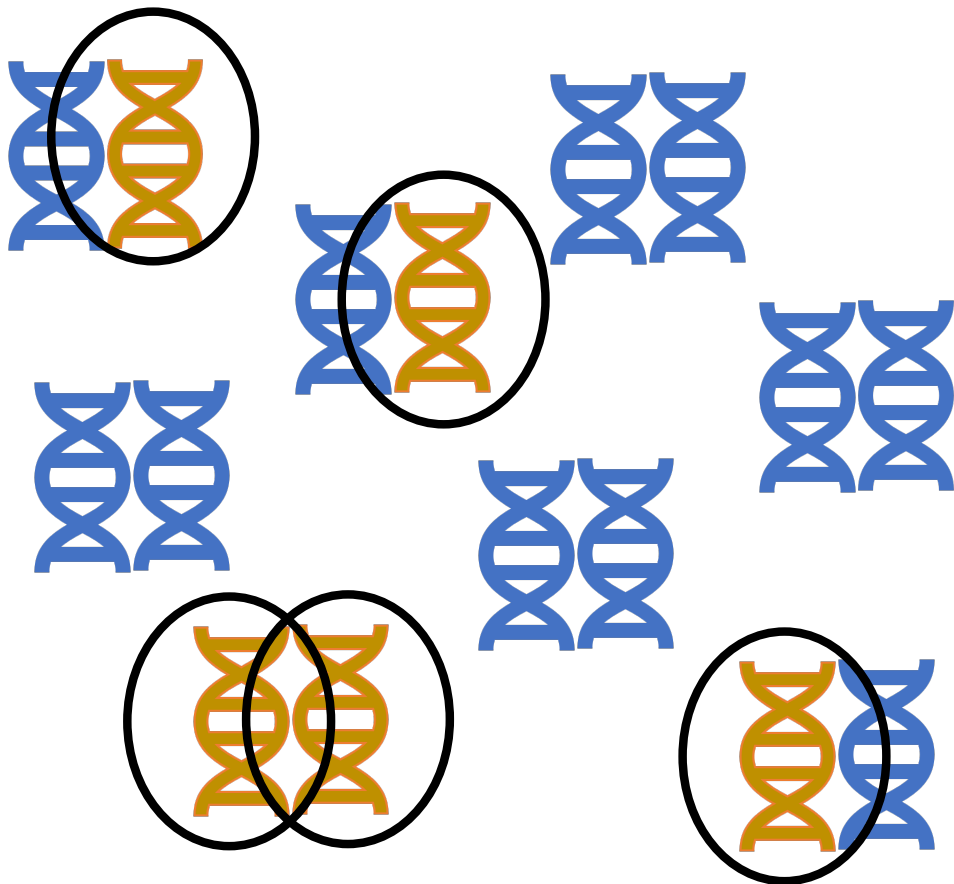
Allele frequencies

Fraction of all **chromosomes** in a population that carry a given allele

$$\frac{\text{\# copies of given allele}}{\text{\# of chromosomes}}$$

Allele frequencies

Fraction of all **chromosomes** in a population that carry a given allele
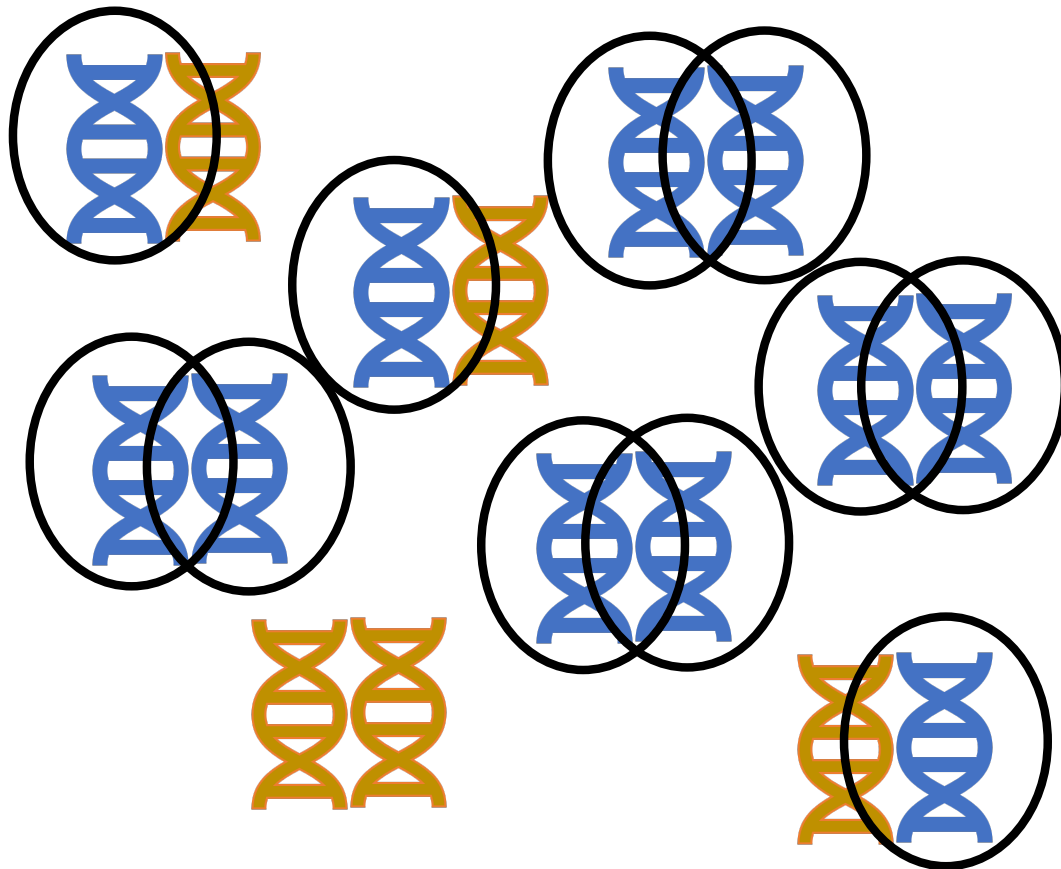

$$\frac{\text{\# copies of given allele}}{\text{\# of chromosomes}}$$

$$p = \frac{5}{16}$$

Allele frequencies

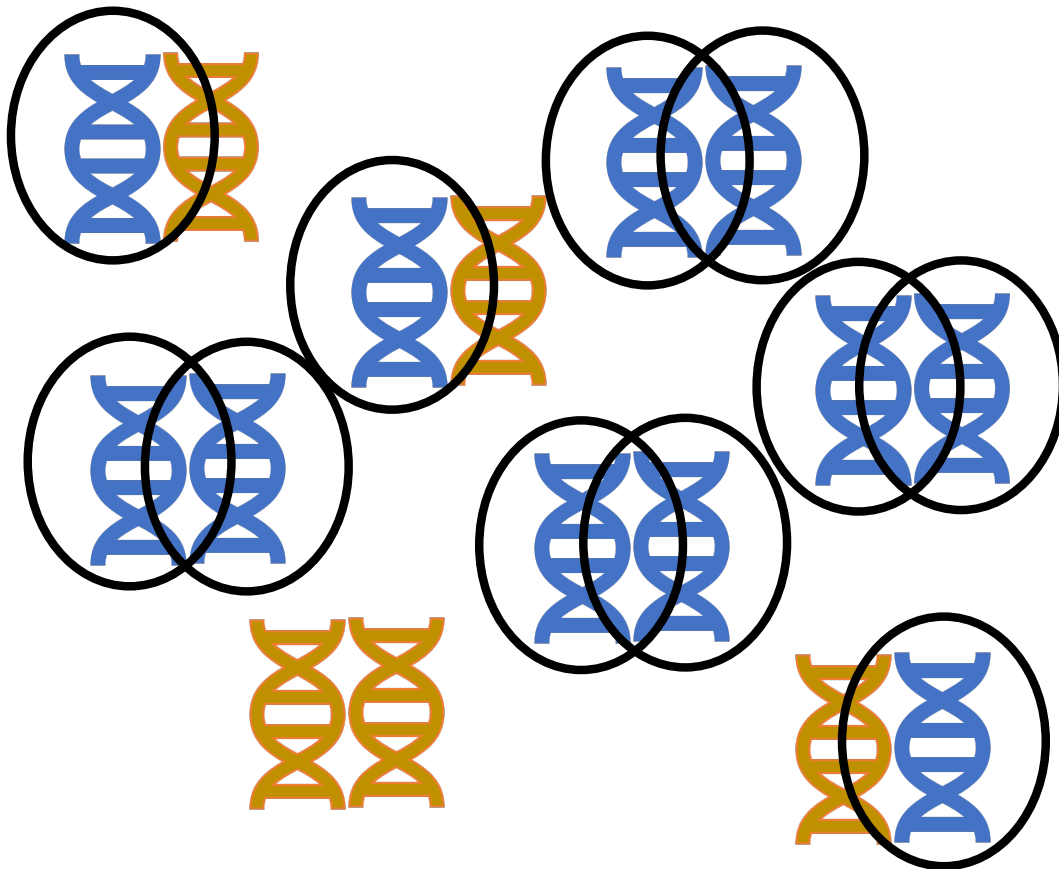Fraction of all **chromosomes** in a population that carry a given allele

$$\frac{\text{\# copies of given allele}}{\text{\# of chromosomes}}$$

$$p = \frac{5}{16}$$

$$q = \frac{11}{16}$$

Allele frequencies

Fraction of all **chromosomes** in a population that carry a given allele

$$\frac{\text{\# copies of given allele}}{\text{\# of chromosomes}}$$
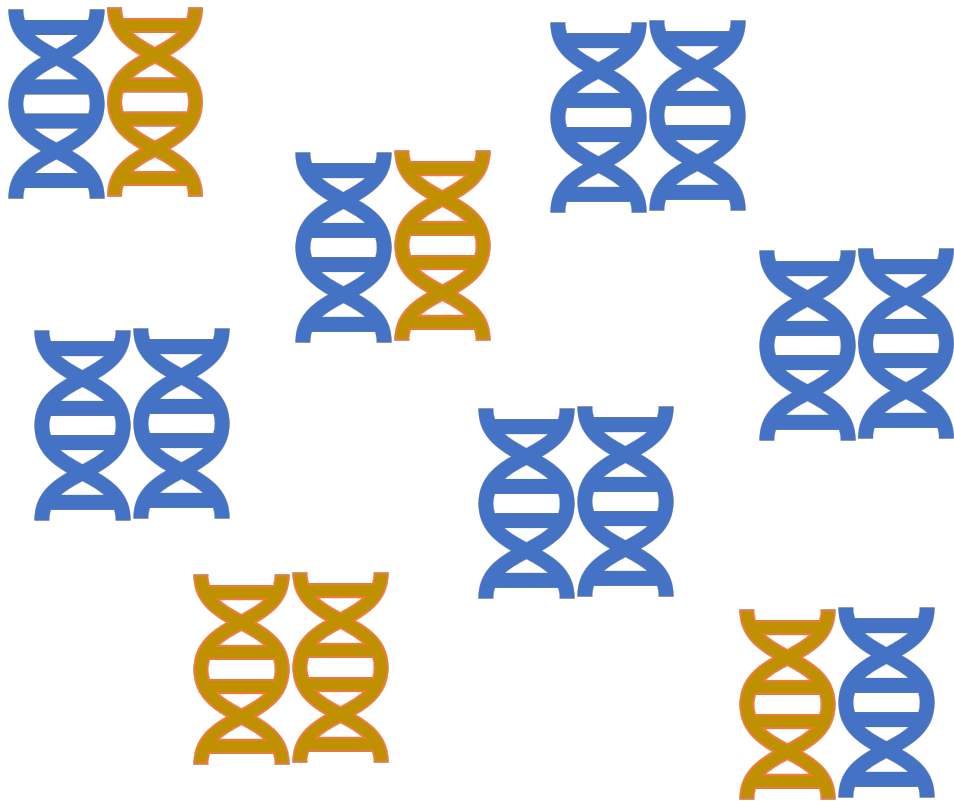
$$p = \frac{5}{16}$$

$$q = \frac{11}{16}$$

p + q = 1

Genotype frequencies

Fraction of all **individuals** in a population that have a given genotype

$$\frac{\text{\# homozygous individuals}}{N \ (\text{\# of individuals})}$$

Genotype frequencies

Fraction of all **individuals** in a population that have a given genotype

$$\frac{\text{\# homozygous individuals}}{N \text{ (\# of individuals)}}$$

$$O(Hom)_1 = \frac{1}{8}$$
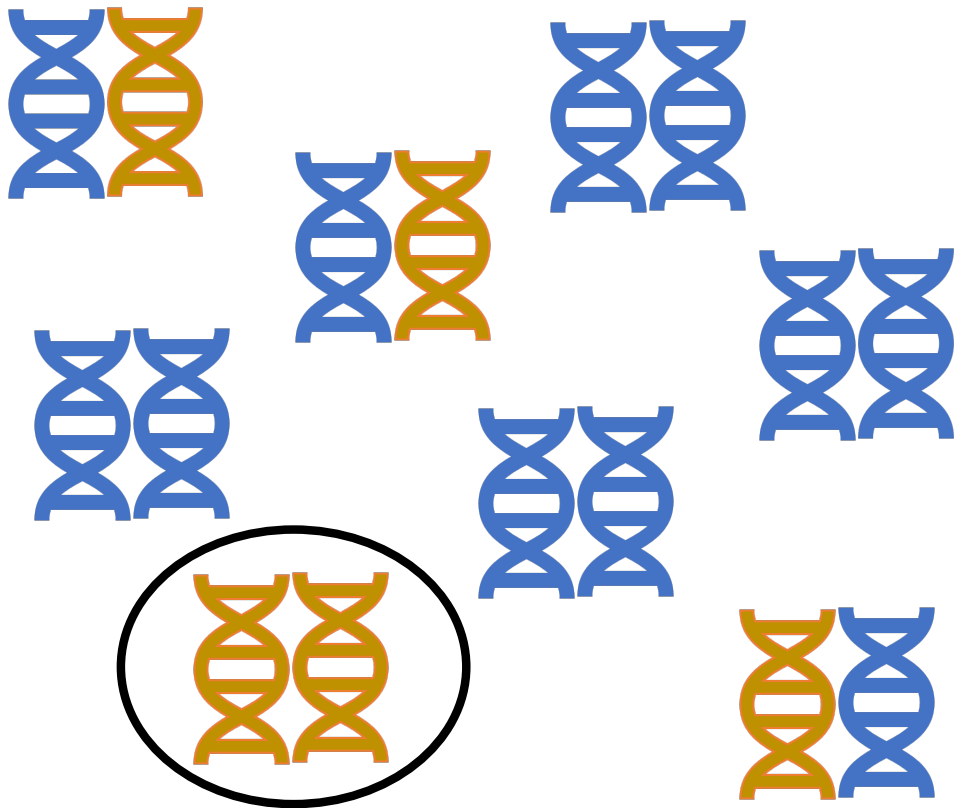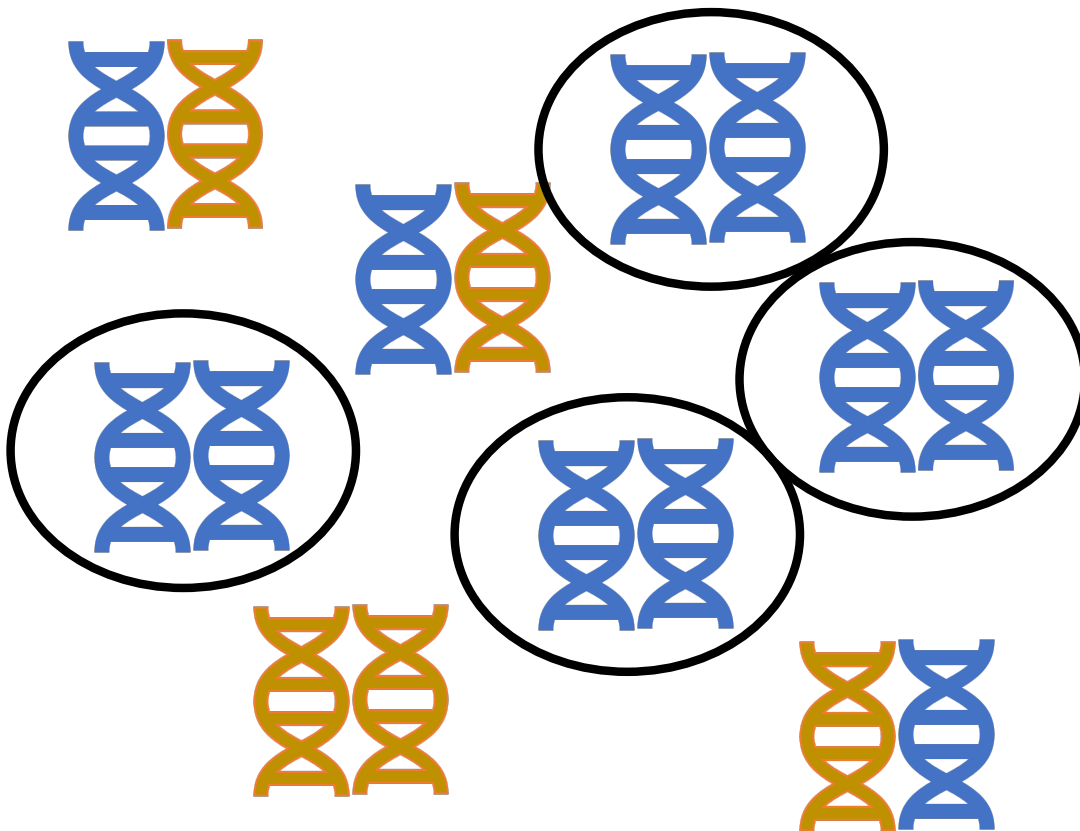
Genotype frequencies

Fraction of all **individuals** in a population that have a given genotype

$$\frac{\text{\# homozygous individuals}}{N \; (\text{\# of individuals})}$$

$$O(Hom)_1 = \frac{1}{8}$$

$$O(Hom)_2 = \frac{4}{8}$$

Genotype frequencies

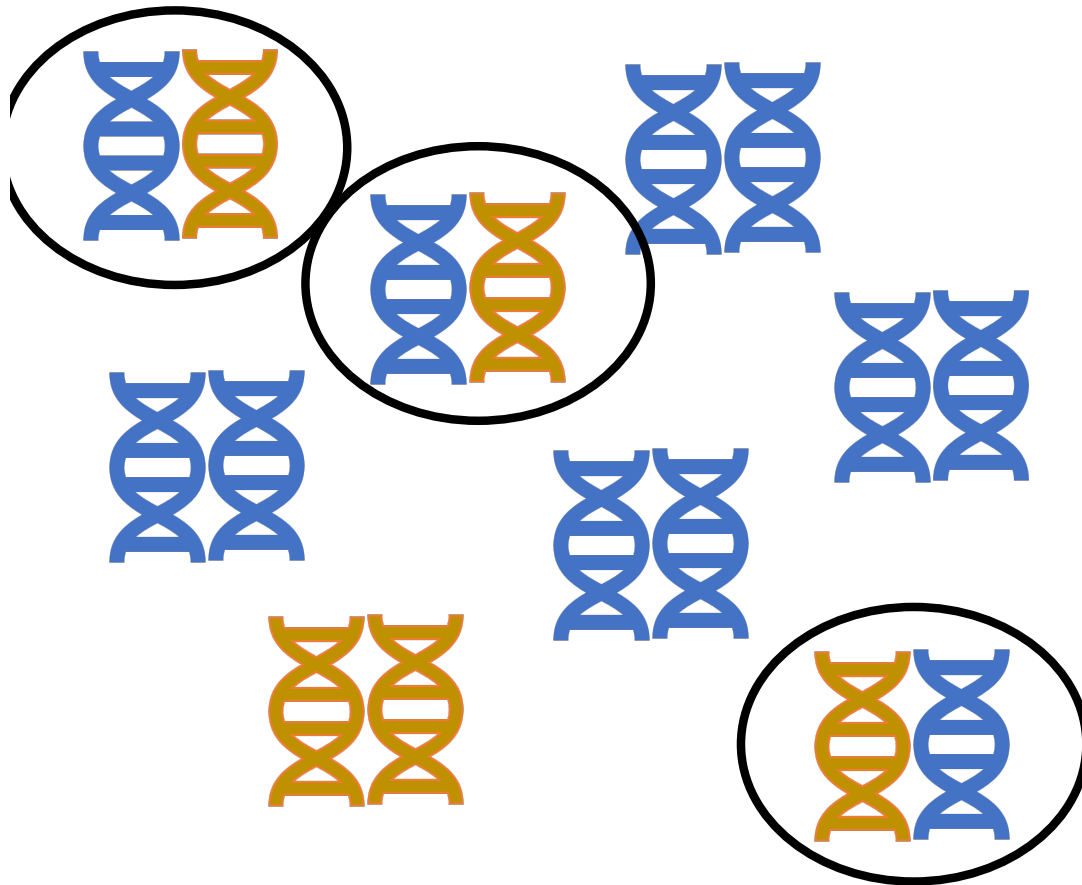Fraction of all **individuals** in a population that have a given genotype

$$\frac{\text{\# heterozygous individuals}}{N\ (\text{\# of individuals})}$$

$$O(\text{Hom})_1 = \frac{1}{8}$$

$$O(\text{Hom})_2 = \frac{4}{8}$$

$$O(\text{Het}) = \frac{3}{8}$$

Multiplicative rule of probability

Probability that two events A and B will both occur is

P (A and B) = P(A) x P(B)

Additive rule of probability

Probability that event A or event B will happen is

P (A or B) = P(A) + P(B)

Expected genotype frequencies
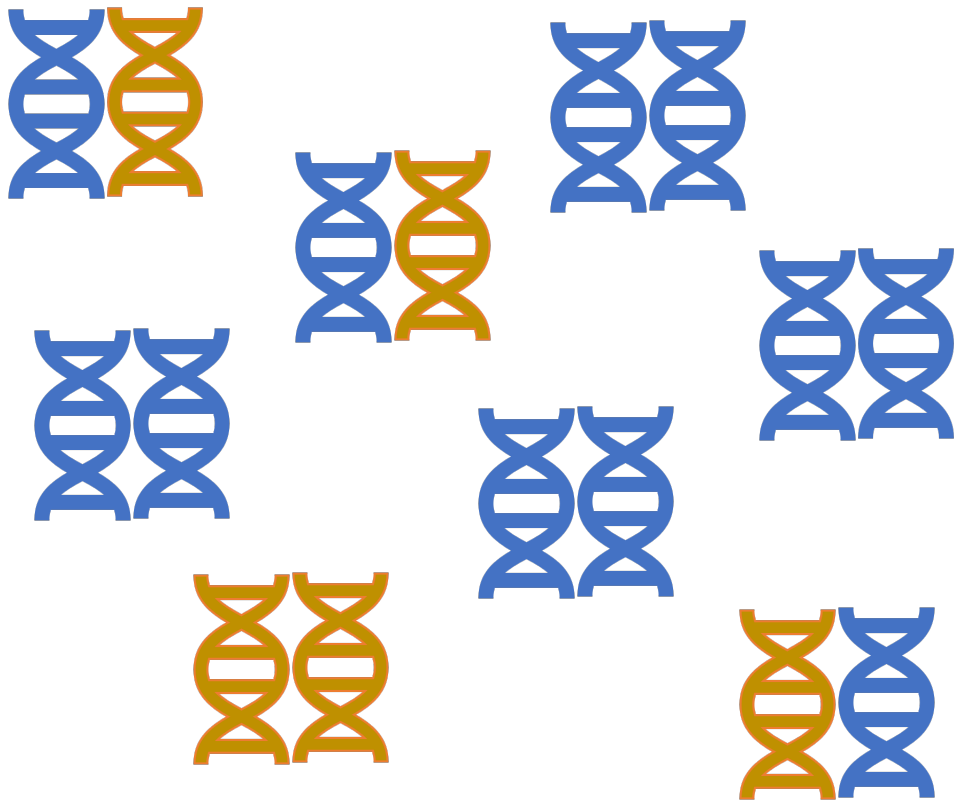
$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

Expected genotype frequencies

$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Hom)_1 = \frac{5}{16} \times \frac{5}{16}$$

Expected genotype frequencies

$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Hom)_1 = \frac{5}{16} \times \frac{5}{16}$$

$$E(Hom)_1 = \frac{25}{256}$$

Expected genotype frequencies

$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Hom)_1 = \frac{5}{16} \times \frac{5}{16}$$

$$E(Hom)_1 = \frac{25}{256}$$

$$E(Hom)_2 = \frac{11}{16} \times \frac{11}{16}$$

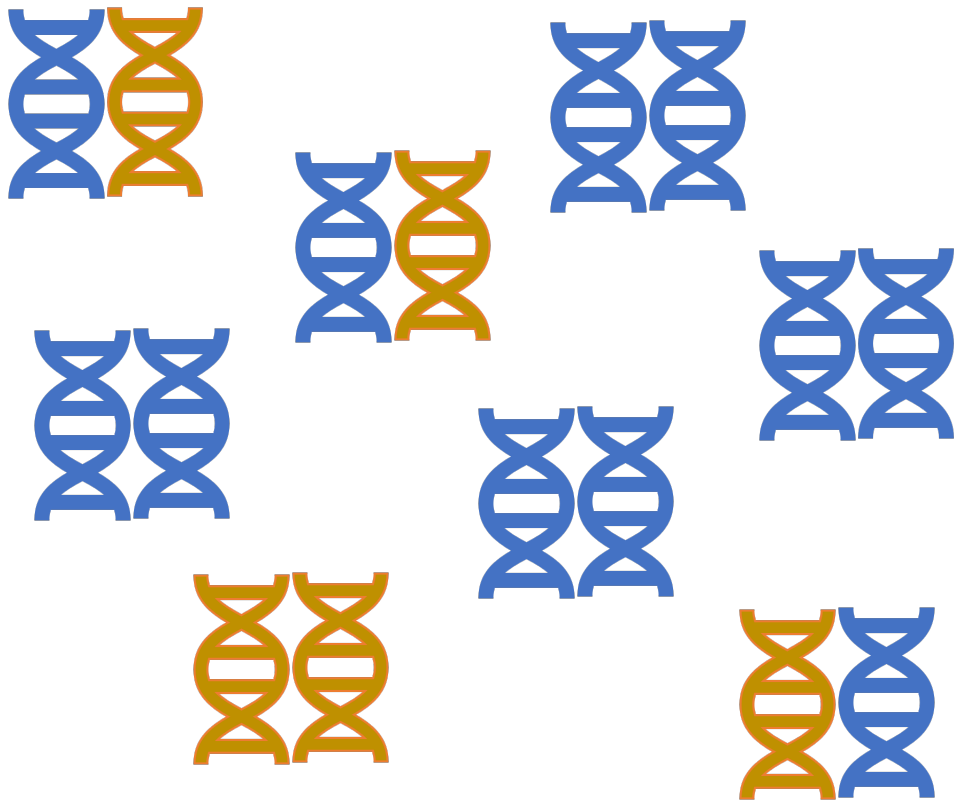$$E(Hom)_2 = \frac{121}{256}$$

Expected genotype frequencies

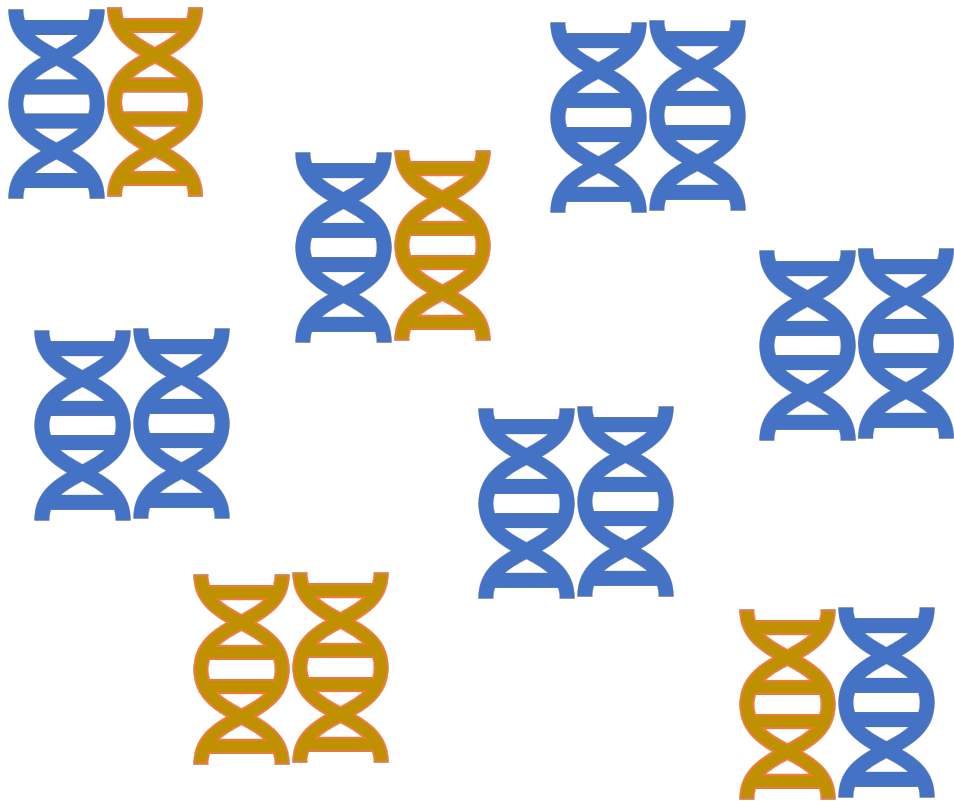$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

Expected genotype frequencies

$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16}$$

Expected genotype frequencies
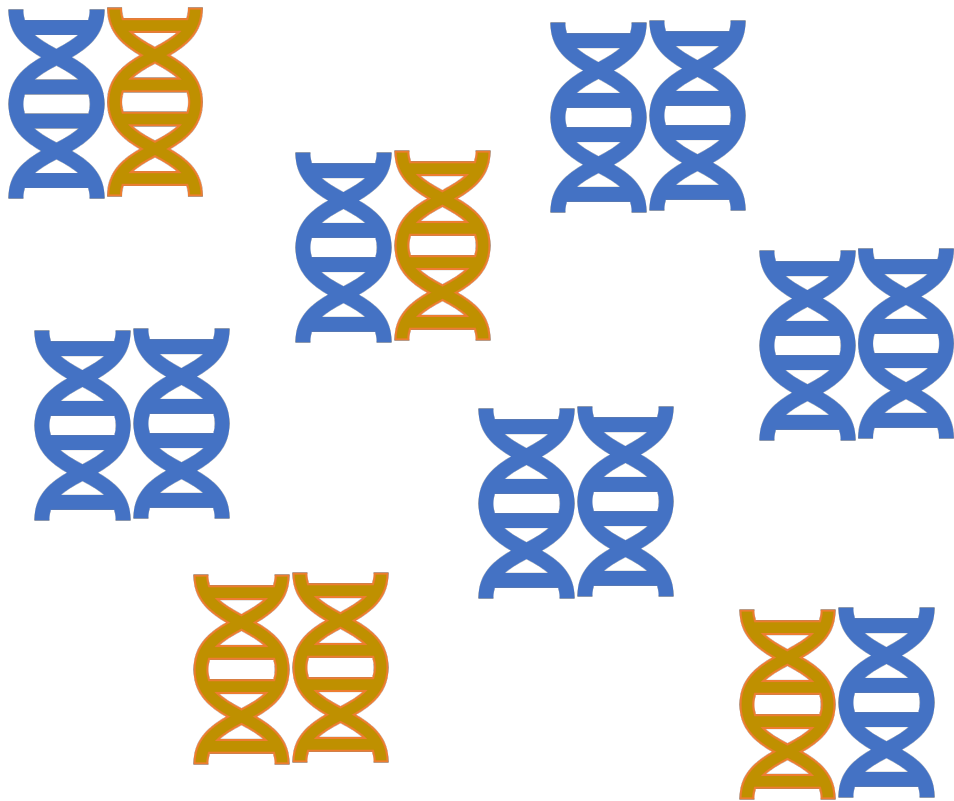
$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} \quad OR \quad \frac{11}{16} \times \frac{5}{16}$$

Expected genotype frequencies

$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} \quad OR \quad \frac{11}{16} \times \frac{5}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} \quad + \quad \frac{11}{16} \times \frac{5}{16}$$

Expected genotype frequencies

$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} \quad OR \quad \frac{11}{16} \times \frac{5}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} + \frac{11}{16} \times \frac{5}{16} = 2 \times \frac{11}{16} \times \frac{5}{16}$$

Expected genotype frequencies

$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} \quad OR \quad \frac{11}{16} \times \frac{5}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} + \frac{11}{16} \times \frac{5}{16} = 2 \times \frac{11}{16} \times \frac{5}{16}$$

$$E(Het) = \frac{55}{256} + \frac{55}{256}$$

Expected genotype frequencies

$$p = \frac{5}{16} \qquad q = \frac{11}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} \quad OR \quad \frac{11}{16} \times \frac{5}{16}$$

$$E(Het) = \frac{5}{16} \times \frac{11}{16} + \frac{11}{16} \times \frac{5}{16} = 2 \times \frac{11}{16} \times \frac{5}{16}$$

$$E(Het) = \frac{55}{256} + \frac{55}{256}$$

$$E(Het) = \frac{110}{256}$$

Expected genotype frequencies

	Females	
	p	**q**
p	p^2	pq
q	qp	q^2

Males

Expected genotype frequencies

	Females	
	p	**q**
p	p^2	pq
q	qp	q^2

Males

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

Hardy-Weinberg equilibrium (HWE)

Allele and genotype frequencies will remain constant from generation to generation assuming the following:

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

Hardy-Weinberg equilibrium (HWE)

Allele and genotype frequencies will remain constant from generation to generation assuming the following:

1. Random mating

Hardy-Weinberg equilibrium (HWE)

Allele and genotype frequencies will remain constant from generation to generation assuming the following:

1. Random mating

2. No mutation

Hardy-Weinberg equilibrium (HWE)

Allele and genotype frequencies will remain constant from generation to generation assuming the following:

1. Random mating

2. No mutation

3. Infinitely large population size

Hardy-Weinberg equilibrium (HWE)

Allele and genotype frequencies will remain constant from generation to generation assuming the following:

1. Random mating

2. No mutation

3. Infinitely large population size

4. No natural selection

Hardy-Weinberg equilibrium (HWE)

Allele and genotype frequencies will remain constant from generation to generation assuming the following:

1. Random mating

2. No mutation

3. Infinitely large population size

4. No natural selection

5. No migration

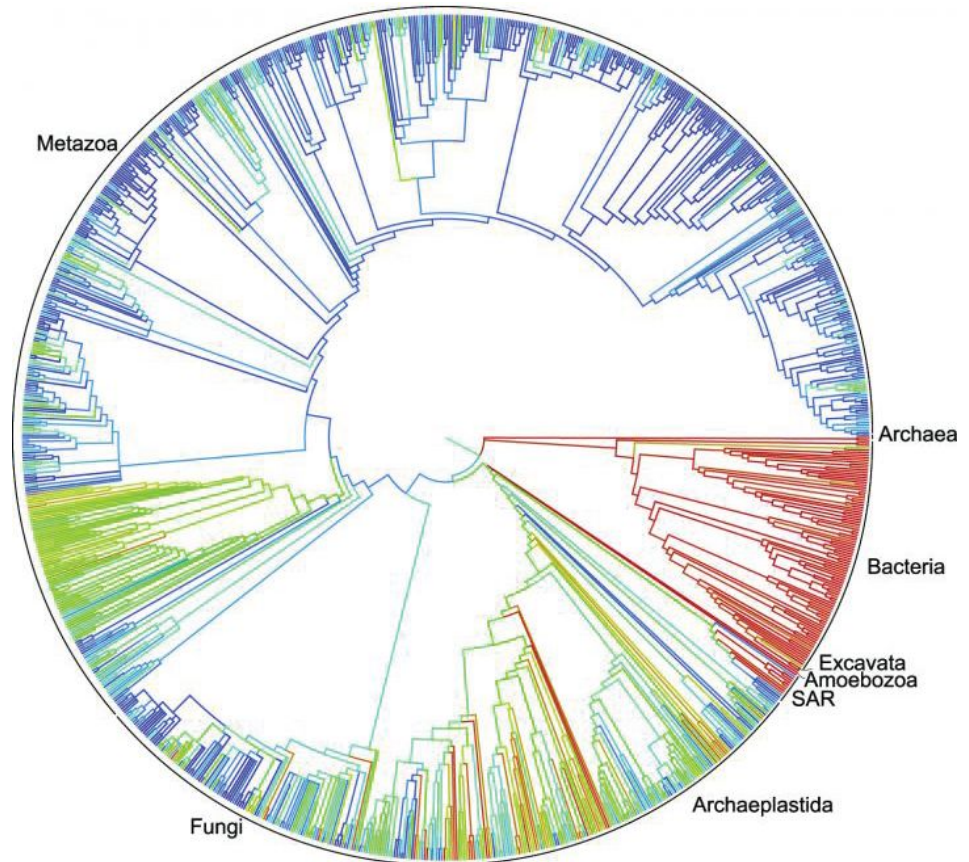# Hardy-Weinberg proportions (HWP)

Genotype frequencies of population at HWE

- As long as population is **randomly mating** we should have HWP
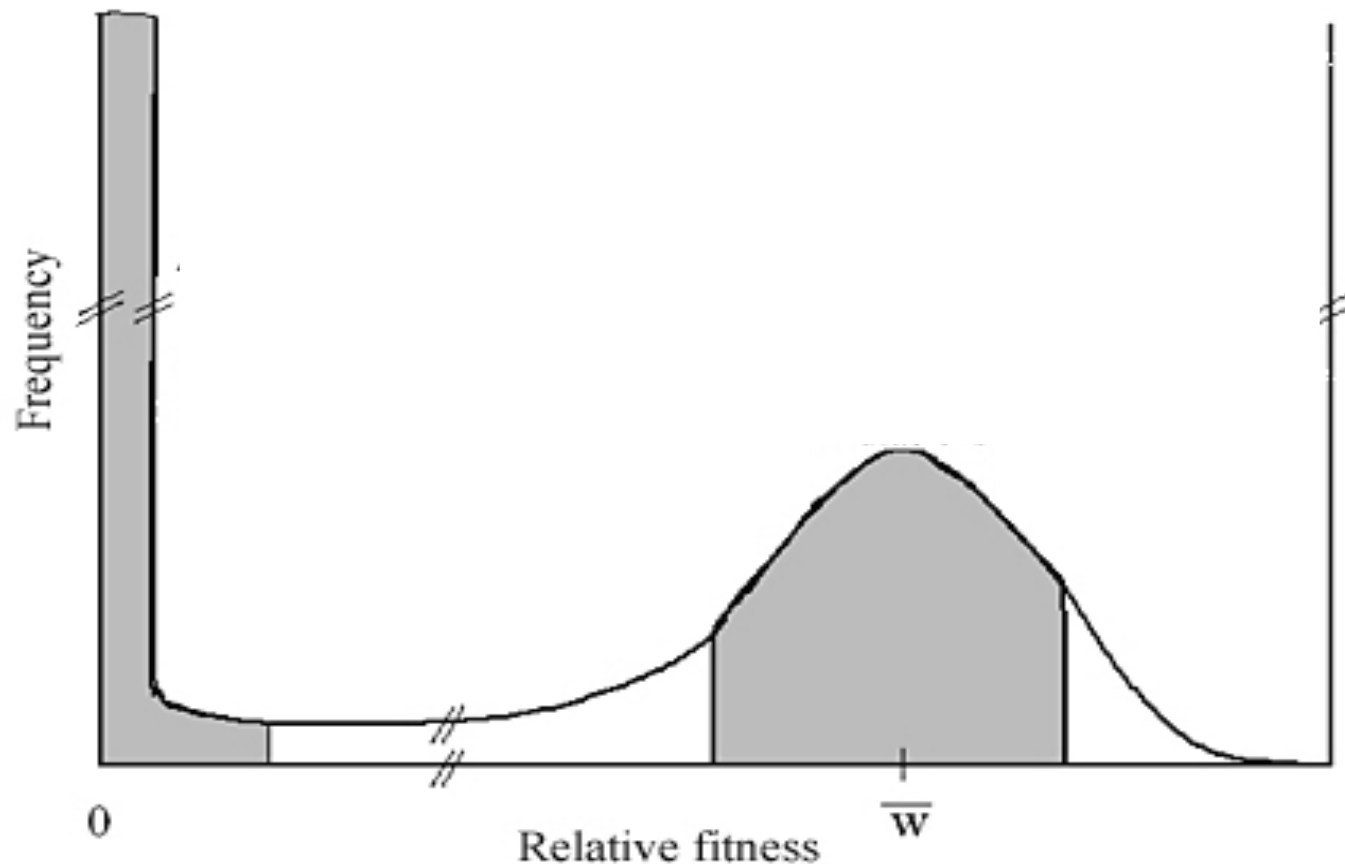
$$p^2 + q^2 + 2pq = 1$$

Evolutionary processes

1. Selection
2. Migration
3. Mutation
4. Genetic drift

Mutation

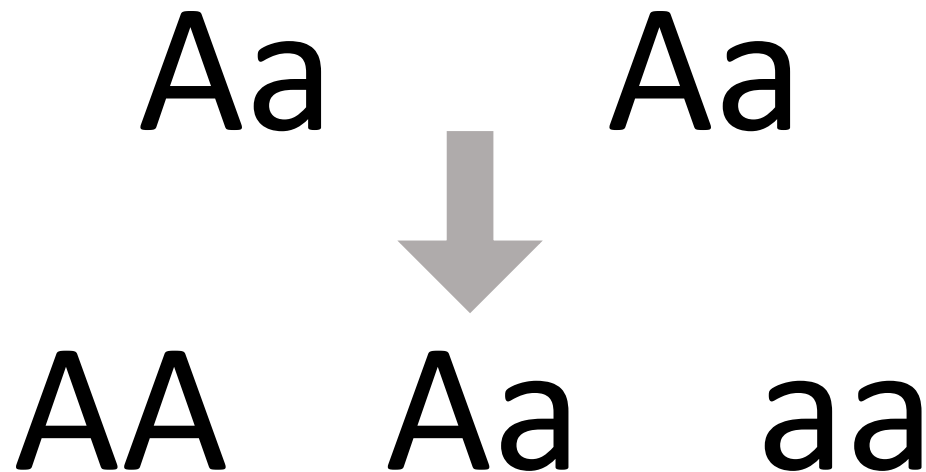Most mutations reduce fitness (or are selectively neutral)

Genetic drift

Change in allele frequencies from generation to generation due to random sampling (sampling error)
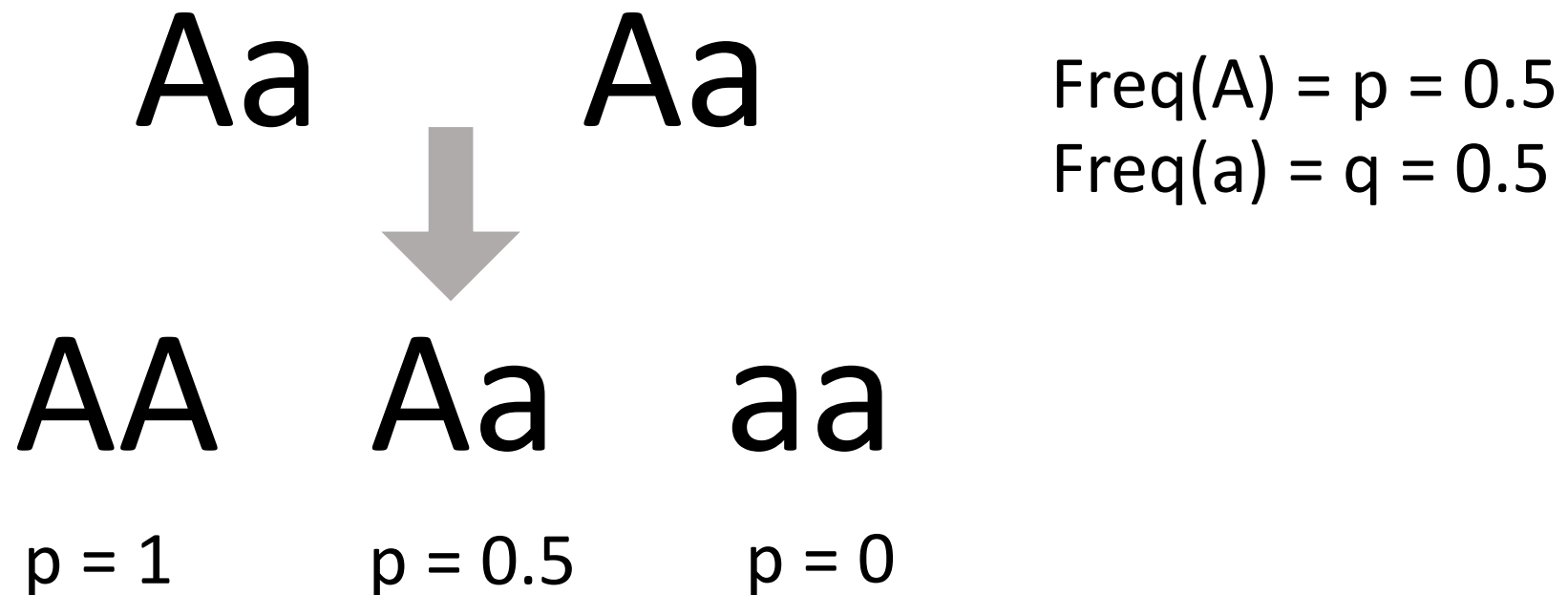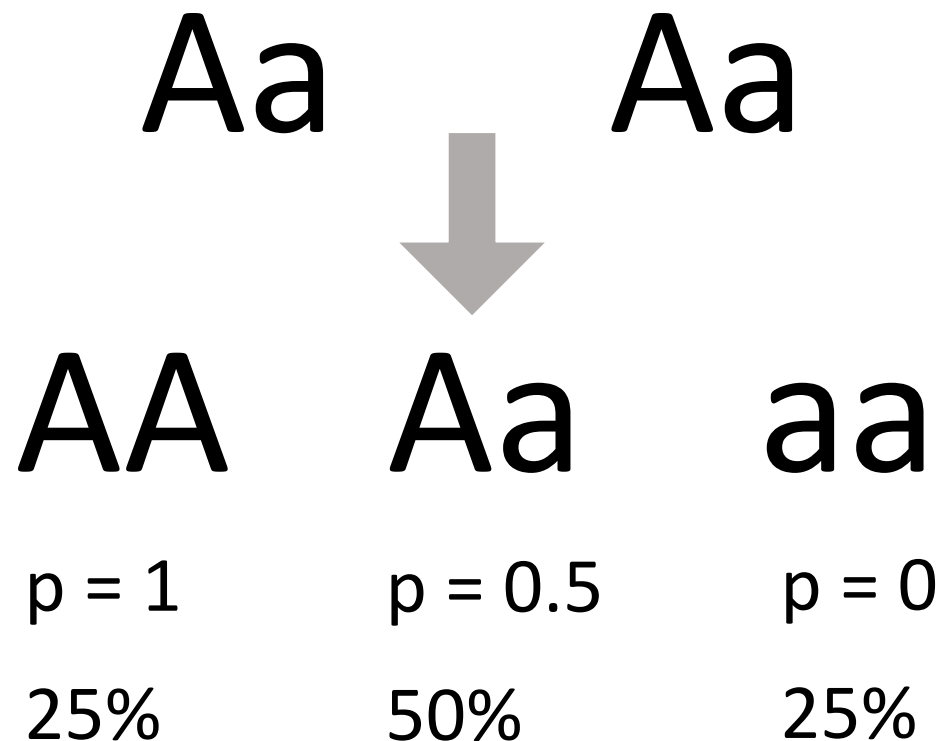
Genetic drift
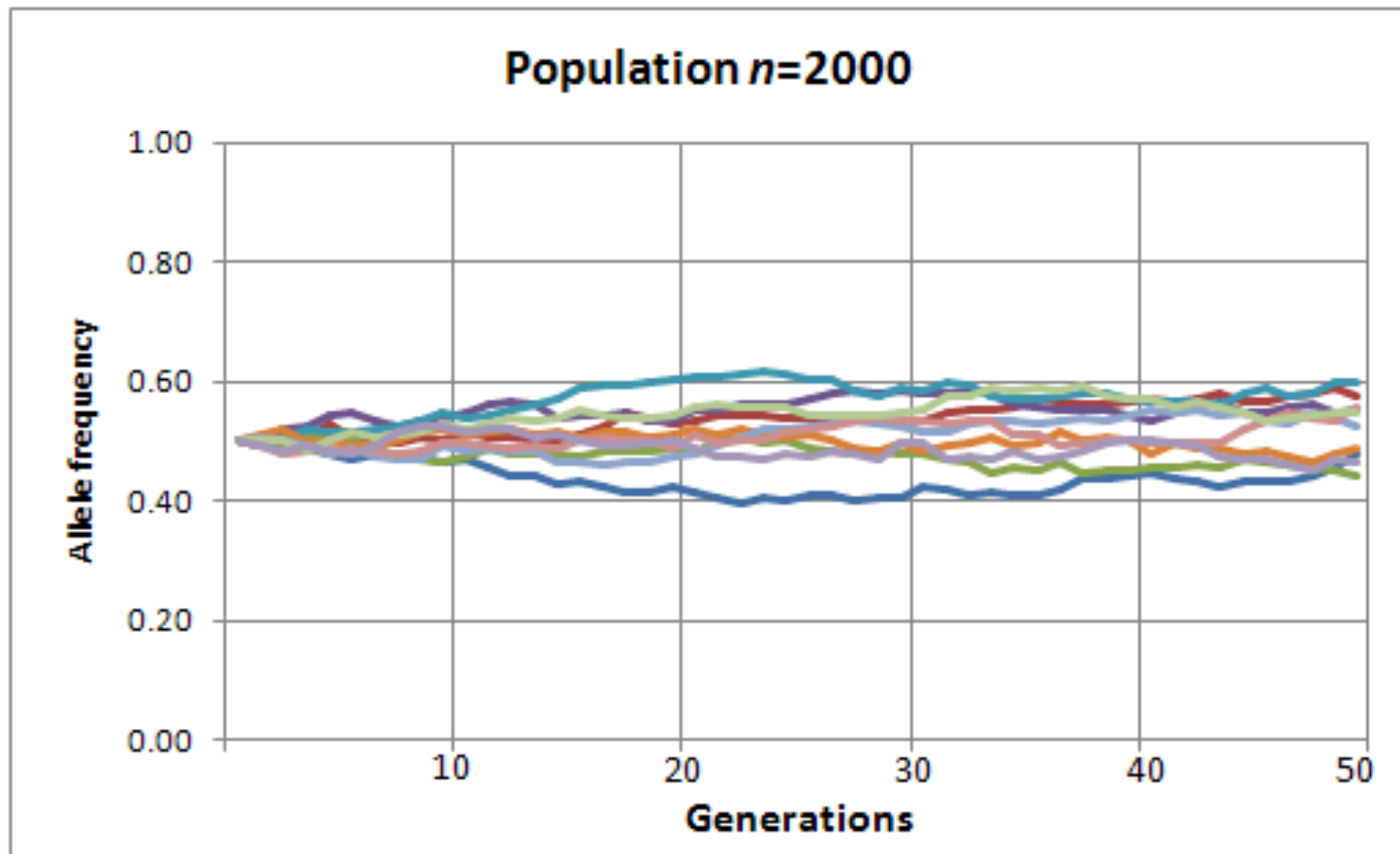
Change in allele frequencies from generation to generation due to random sampling (sampling error)

Aa   Aa

AA   Aa   aa

Freq(A) = p = 0.5
Freq(a) = q = 0.5

Genetic drift

Change in allele frequencies from generation to generation due to random sampling (sampling error)

Aa   Aa

Freq(A) = p = 0.5
Freq(a) = q = 0.5

AA   Aa   aa

p = 1   p = 0.5   p = 0

Genetic drift
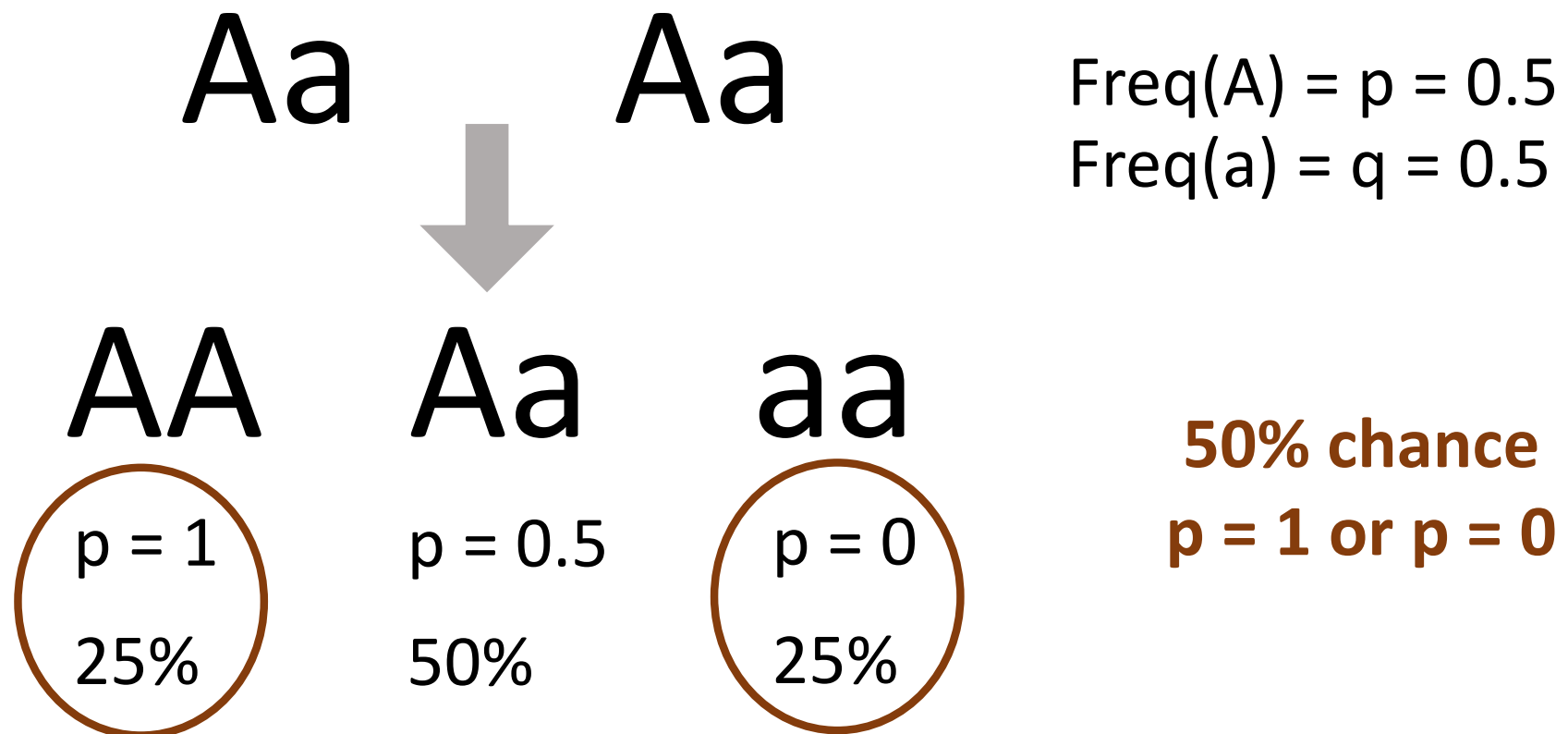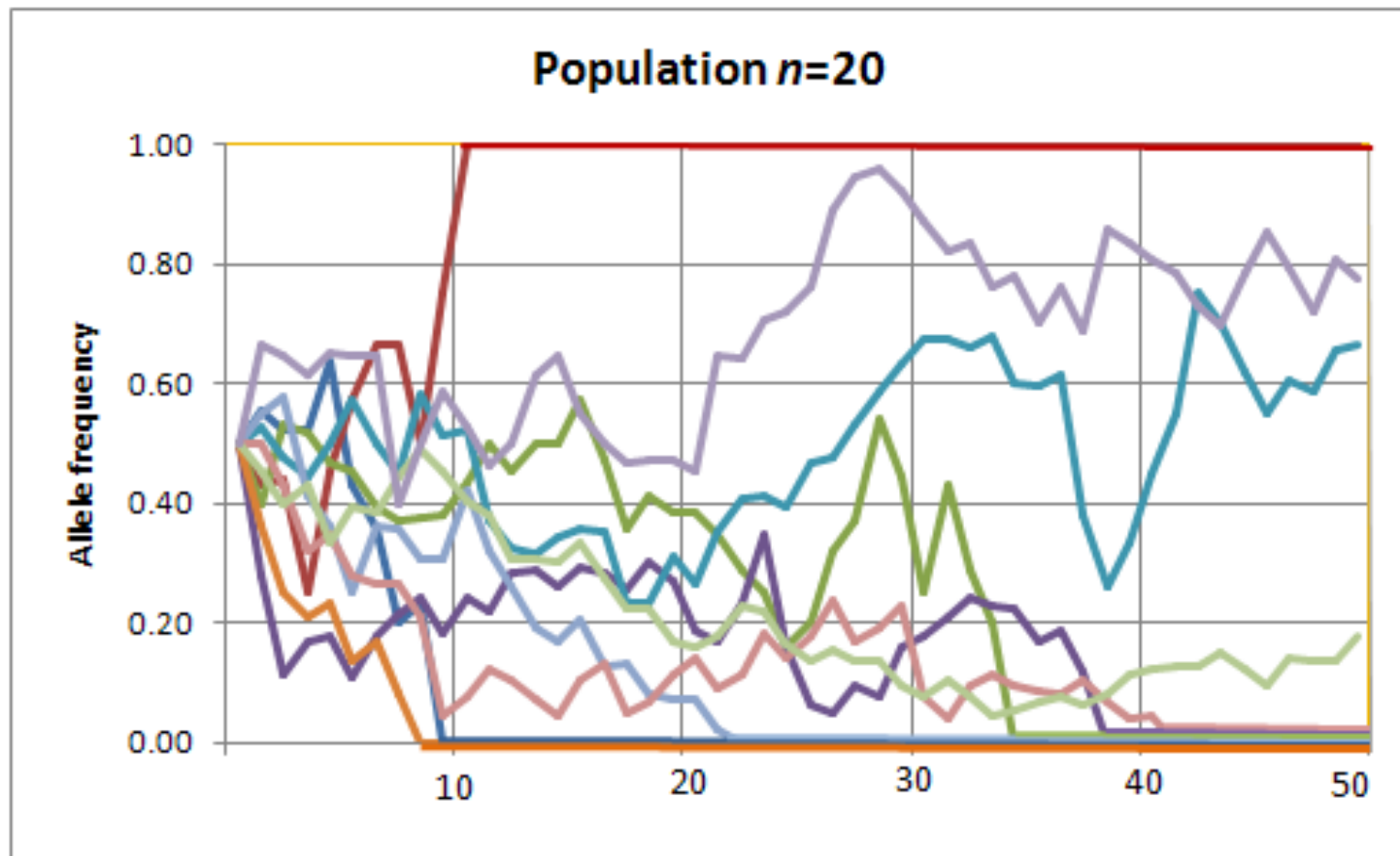
Change in allele frequencies from generation to generation due to random sampling (sampling error)

Aa   Aa   Freq(A) = p = 0.5
                    Freq(a) = q = 0.5

AA   Aa   aa

p = 1   p = 0.5   p = 0

25%   50%   25%

Genetic drift

Change in allele frequencies from generation to generation due to random sampling (sampling error)

Aa   Aa

Freq(A) = p = 0.5
Freq(a) = q = 0.5

AA   Aa   aa

p = 1   p = 0.5   p = 0

25%   50%   25%

50% chance
p = 1 or p = 0

Genetic drift

Population $n=20$
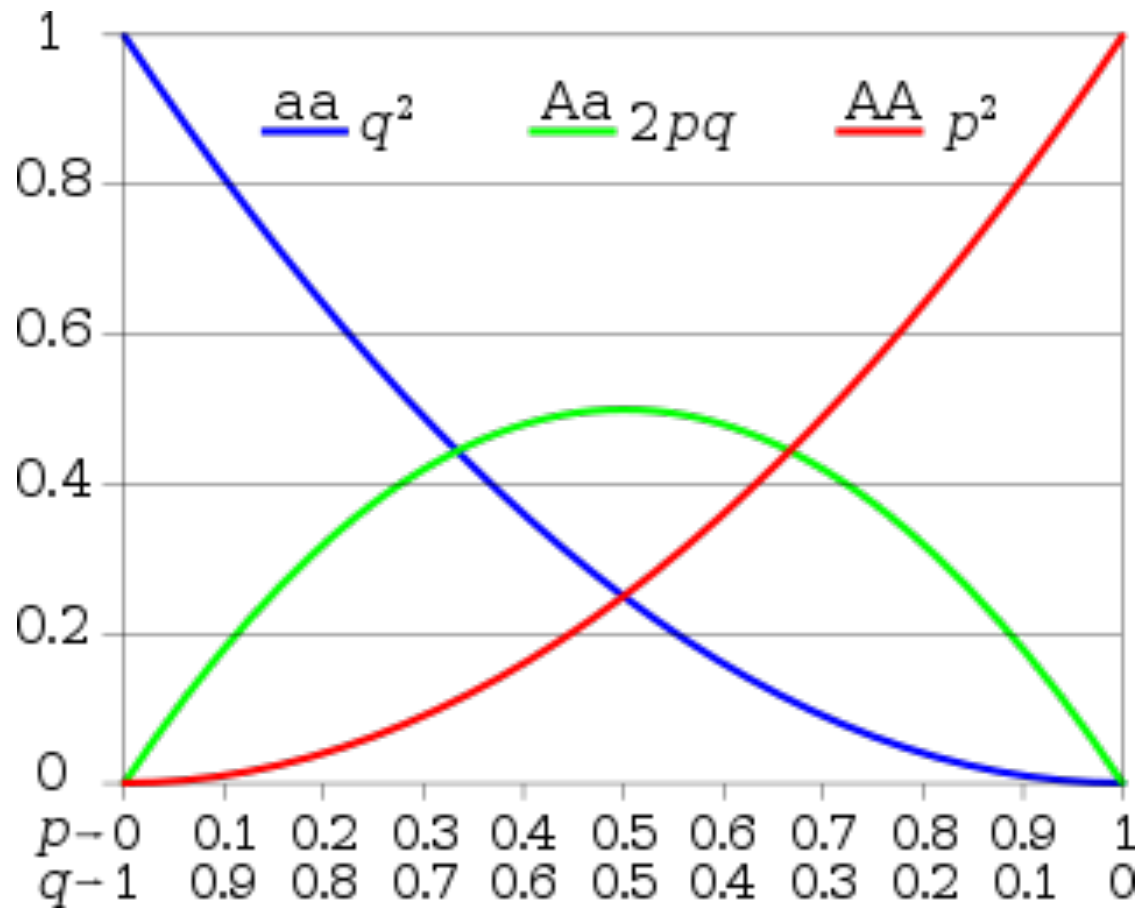
Genetic drift

Population *n*=2000

Outline

Heterozygosity

Expected probability that an individual has two different alleles at a particular locus

- General measure of genetic variance

$$He = 2pq = 1 - \sum_{i=1}^{n} p_i^2$$

Heterozygosity

Other measures of genetic diversity

Allelic Richness

Total number of different alleles at a given locus

Other measures of genetic diversity

Allelic Richness

Total number of different alleles at a given locus

π

Average number of nucleotide differences/site between two DNA sequences in all possible pairs

Outline

Inbreeding coefficient (F_{IS})

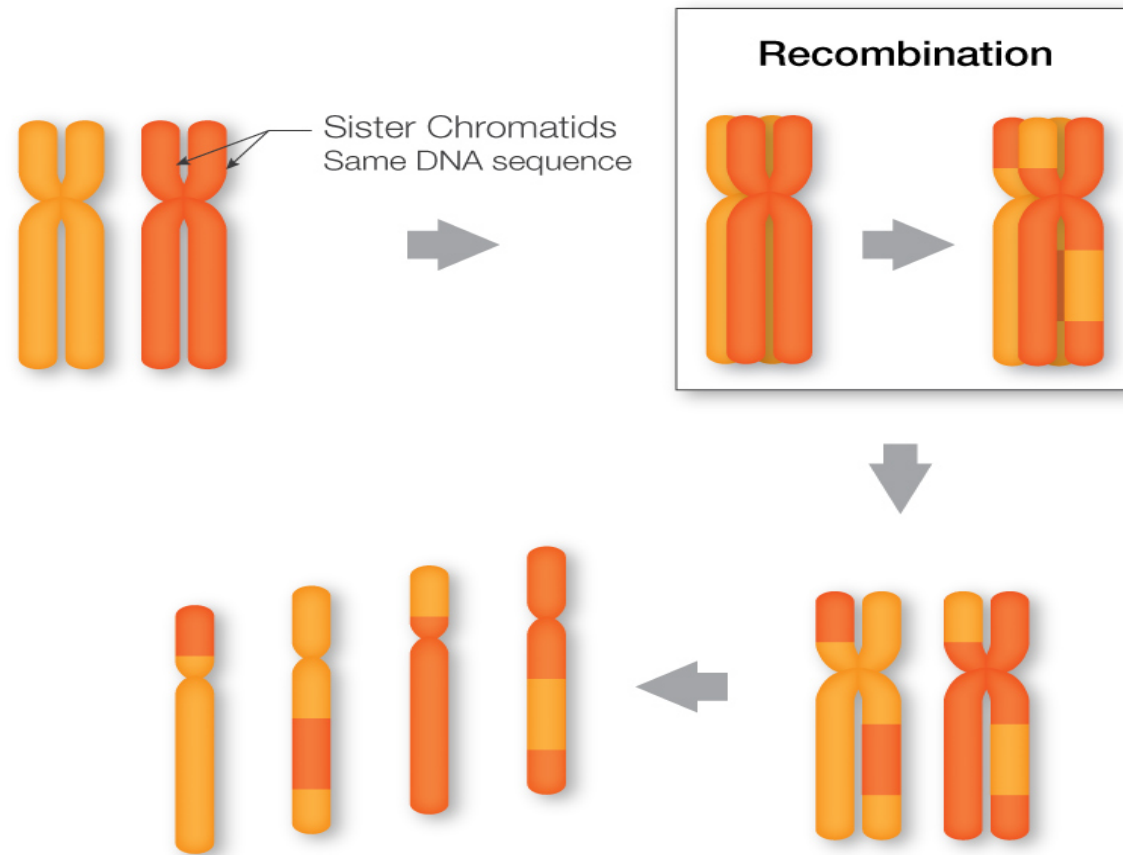Probability that two alleles in an individual are identical, relative to the probability in a random population
- Consequence of small Ne and genetic drift

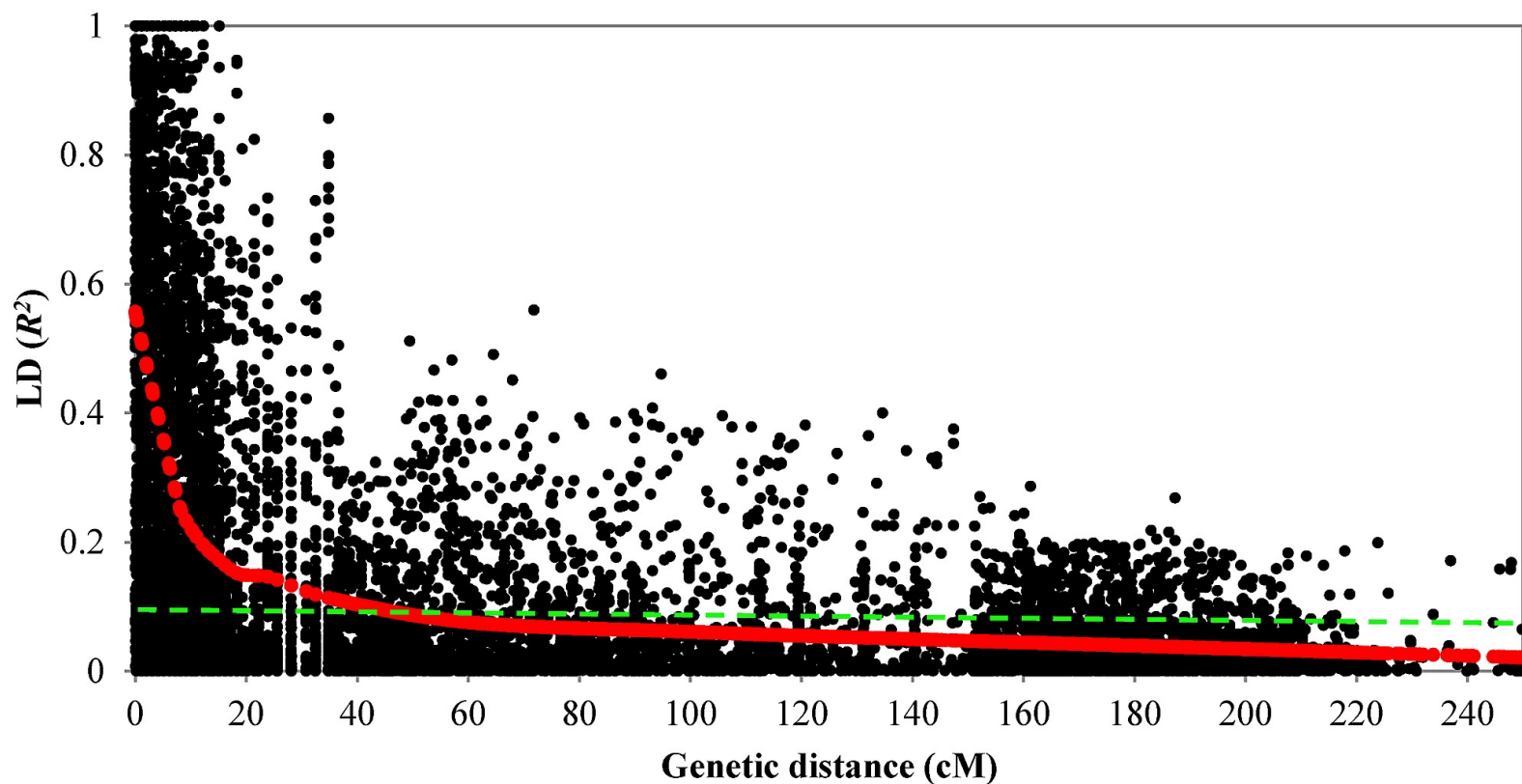$$F_{IS} = 1 - \frac{H_o}{H_e}$$

Outline

Recombination & independent assortment

Sister Chromatids
Same DNA sequence

Recombination

Linkage disequilibrium

Nonrandom association between alleles at two loci

Zegeye et al. 2014 PLOS ONE doi:0105593

Migration, aka Gene Flow

- Migration is the movement of genetic diversity, usually within a species.

- Occurs both with the advancing front of a population, when it is colonizing new areas . . .

- . . . and when genes of two or more populations mix

- Rate of migration is related to the frequency of reproduction and the distances of dispersal

Dispersal

- Life history (survival, growth, reproduction)
- Population dynamics
- Local adaptation
- Range limits
- Reserve networks (size and spacing)

Migration

- Measures of divergence
- Drift-migration balance
- Island model
- Isolation-by-distance
- Methods

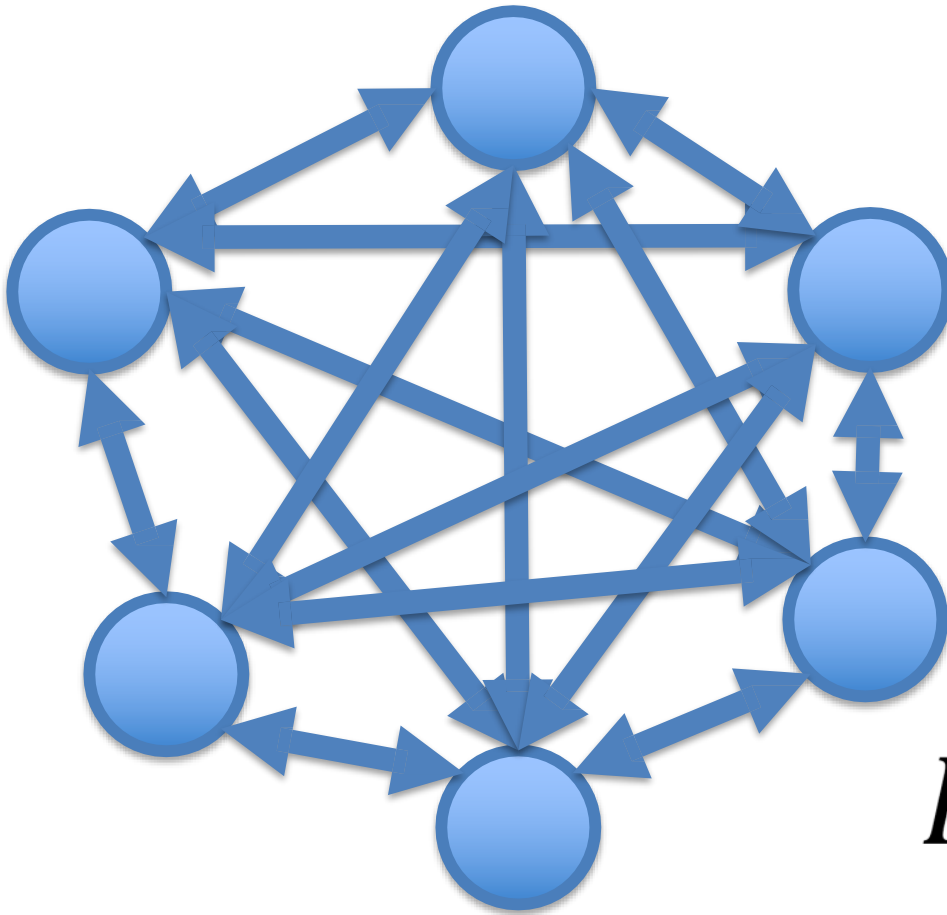# A measure of individual divergence

- Fraction of alleles that two individuals share, identical by descent (IBD) or state (IBS)

- Examples (IBD)
  - parent-offspring: 0.5
  - siblings: 0-1
  - cousins: 0-0.5
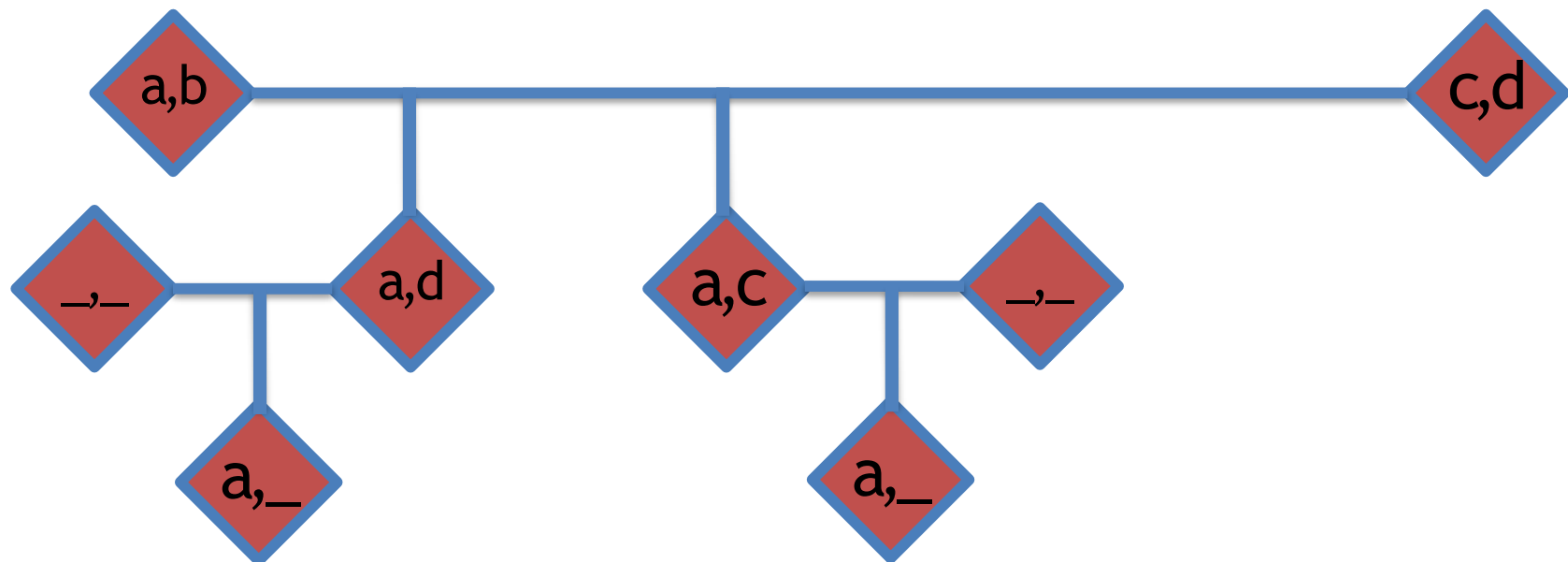
Drift-migration balance

- Drift causes allele frequencies in populations to become less similar

- Migration causes allele frequencies to become more similar
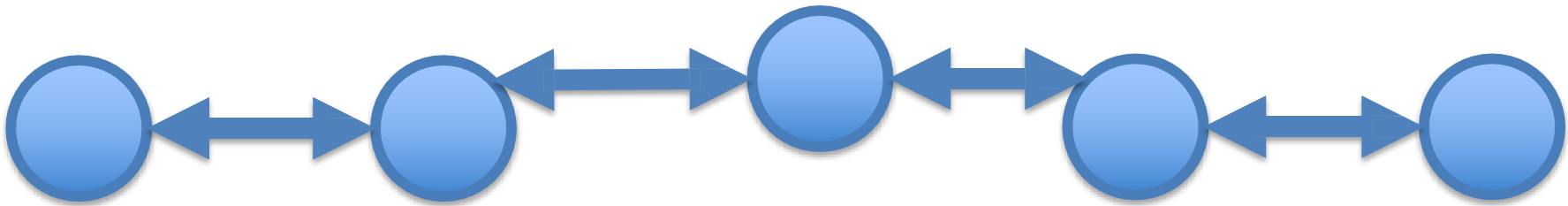
Island model

$$F_{st} \approx \frac{1}{4Nm + 1}$$

Population size   migration rate

Isolation by distance

- Migration among neighbors

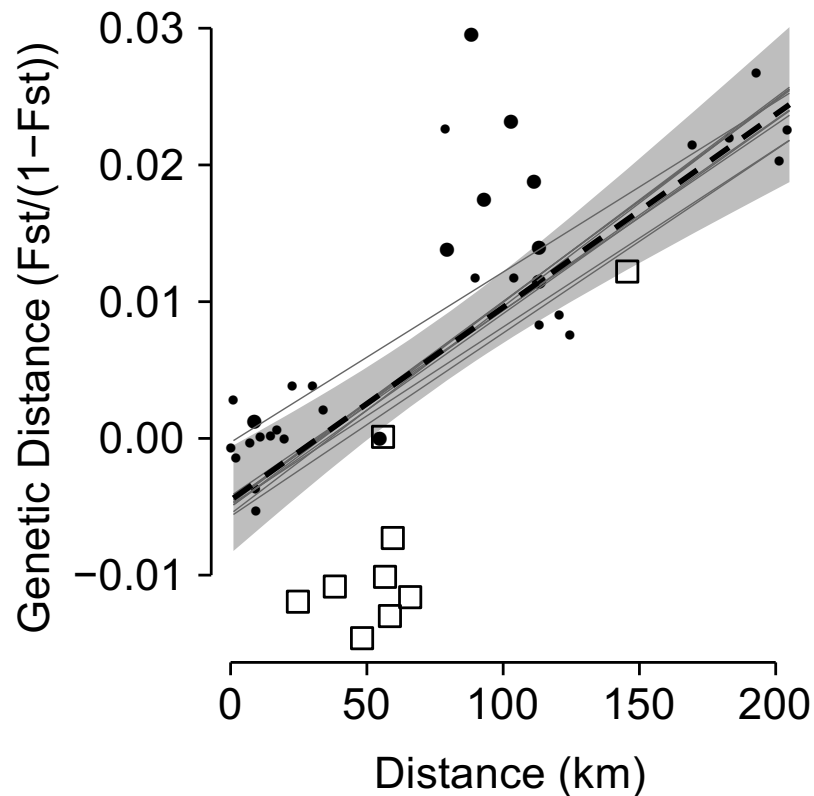$$m = \frac{1}{4D\sigma^2}$$

Population density   dispersal spread

Isolation by distance

- Linear relationship between FST/(1-FST) and geographic distance
- Distance measured in km (1D) or ln(km) (2D)

Migration

- Measures of divergence
- Drift-migration balance
- Island model
- Isolation-by-distance
- **Methods**
  - Fixation Index (Fst)
  - Relatedness
  - PCA
  - STRUCTURE
  - MIGRATE

Using Fixation Index (Fst)

- Measure pairwise Fst between **populations**
- Lower Fst implies:
  - More gene flow
  - Larger population sizes (less drift)
- Higher Fst implies
  - Less gene flow
  - Smaller population sizes (more drift)
- Identify barriers to dispersal
- Isolation by distance patterns between populations

Using Relatedness

- Measure pairwise relatedness between **individuals**
- Lower relatedness across space implies
  - Less gene flow
  - Smaller population sizes
- and vice versa
- Identify barriers to dispersal
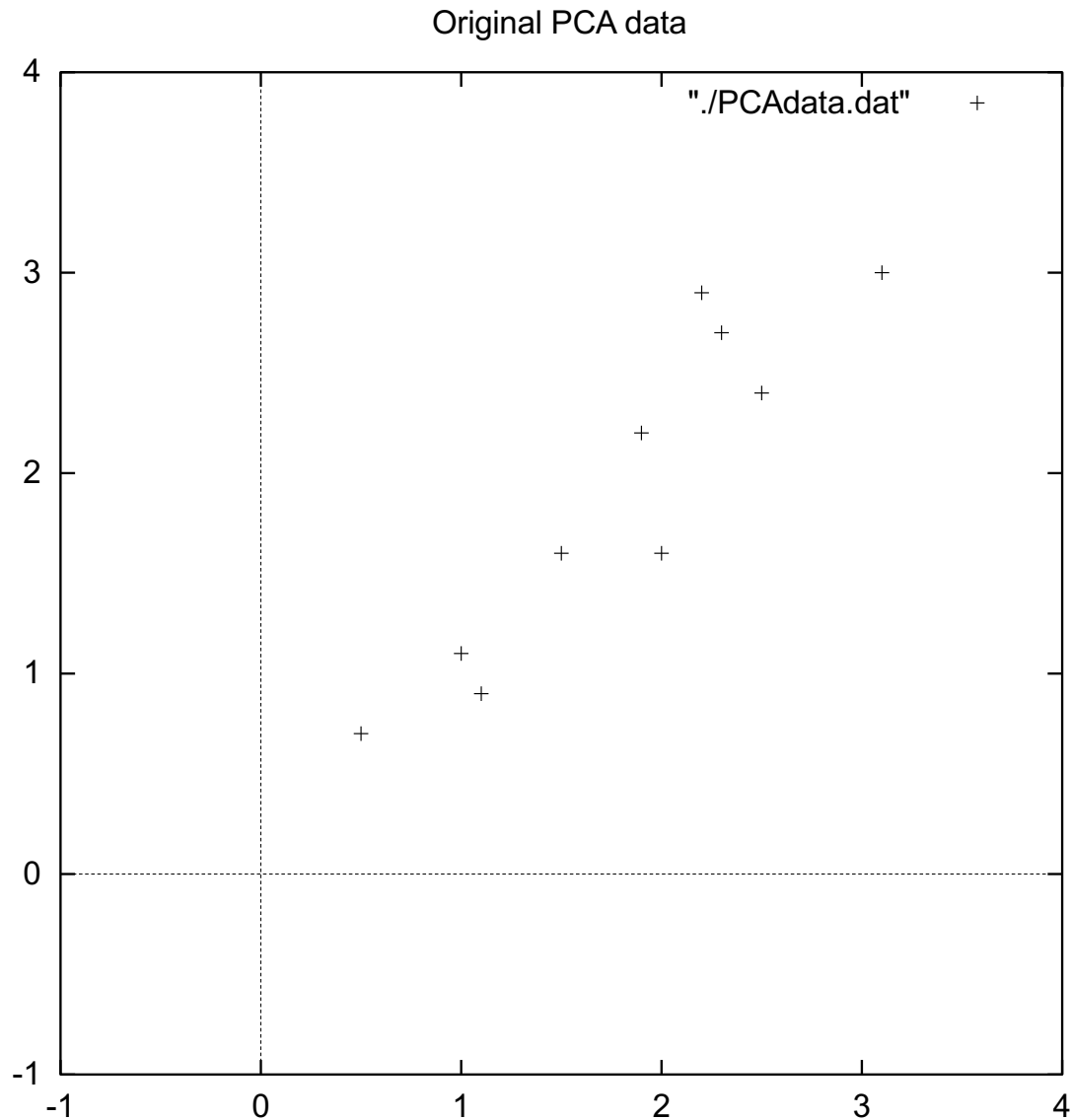- Isolation by distance patterns in continuous space

Principal Components Analysis (PCA)

- Use to simplify a large number of variables (e.g., loci) a few principal axes of variation

Principal Components Analysis (PCA)

$$
\text{Data} = \begin{array}{c|c}
x & y \\
\hline
2.5 & 2.4 \\
0.5 & 0.7 \\
2.2 & 2.9 \\
1.9 & 2.2 \\
3.1 & 3.0 \\
2.3 & 2.7 \\
2 & 1.6 \\
1 & 1.1 \\
1.5 & 1.6 \\
1.1 & 0.9 \\
\end{array}
$$

Original PCA data

A tutorial on Principal Components Analysis

Lindsay I Smith

February 26, 2002
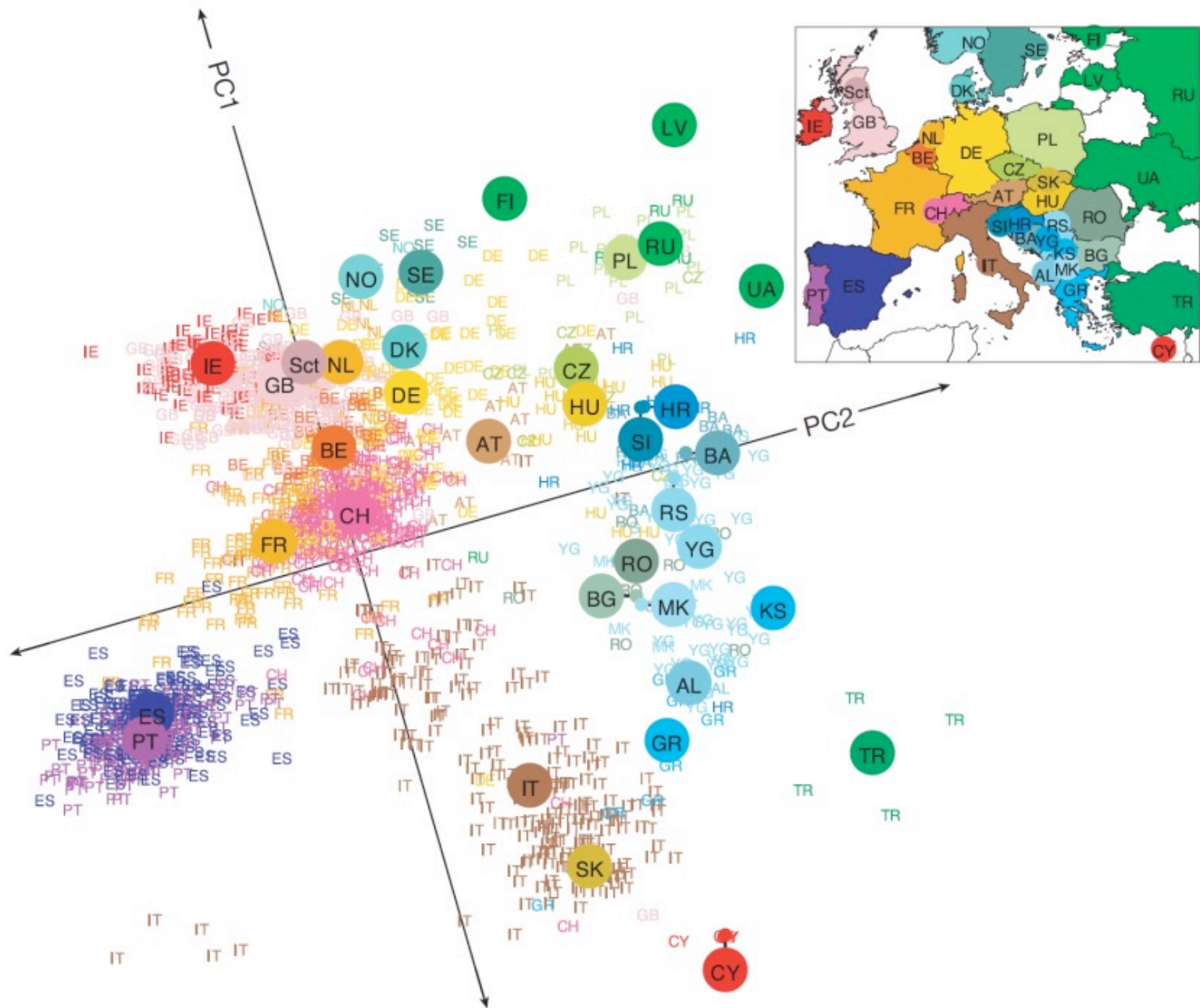
Genes mirror geography within Europe

John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zolta´n Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]

197,146 loci in
1,387 individuals

STRUCTURE

- Clusters individuals without information on sampling location
- Tries to maximize HWP within each population
  - Assumes HWP
- Computationally intensive

Summary

- Drift and migration balance each other

- Island model and isolation by distance models describe simple but useful cases

- Divergence measured between populations with FST or TreeMix

- Divergence between individuals measured with relatedness, PCA, STRUCTURE