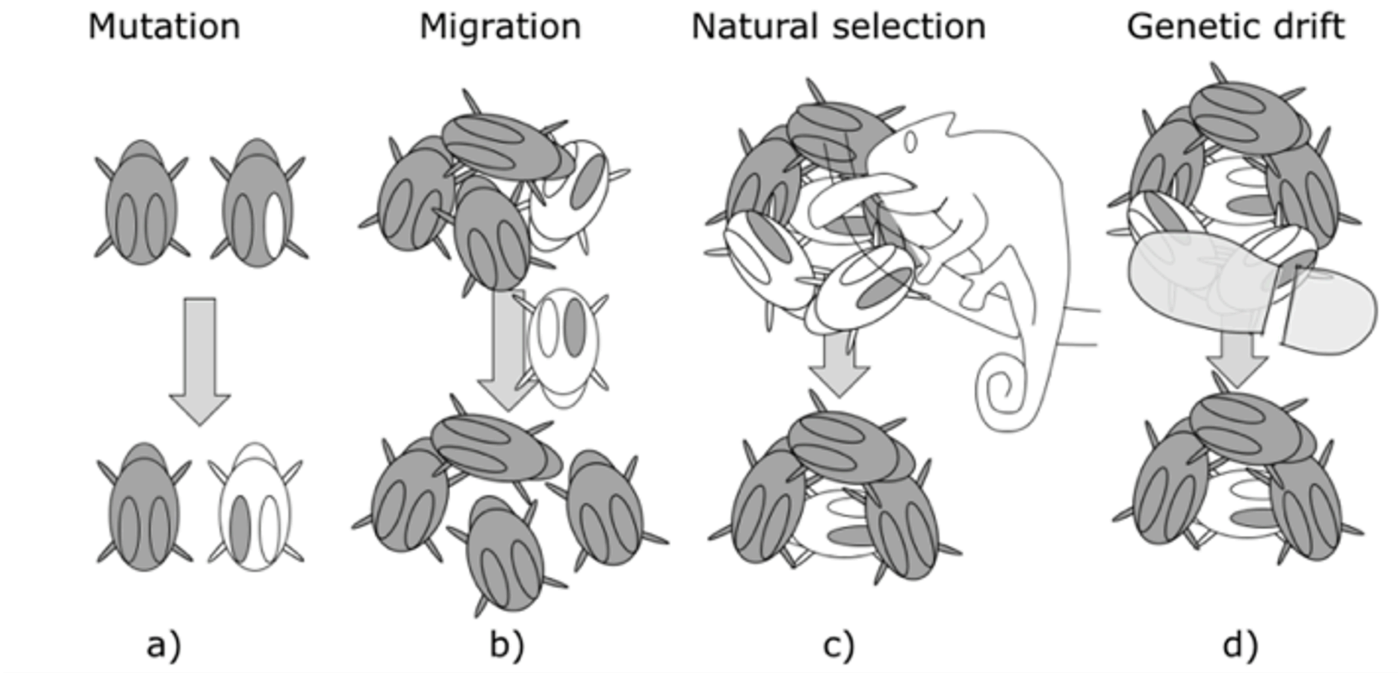
Population Genetics

```
$ echo "Data Sciences Institute"
```

Population genetics

- A field concerned with genetic variation within and between populations over time and space.
- By looking at genetic variation we can learn about population history, migration patterns, and the impact of natural selection on genetic diversity.

Factors that affect patterns of genetic variation

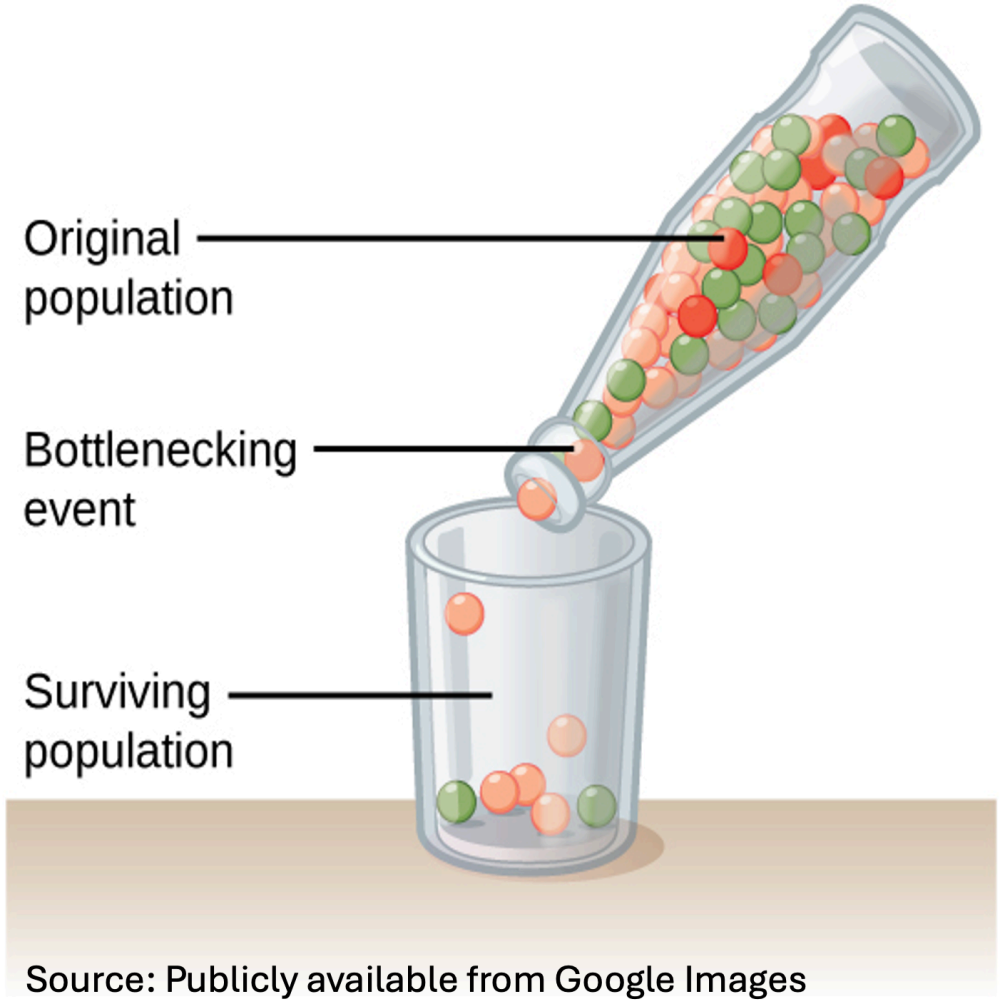


Source: Publicly available from Google Images

- By looking at genetic variation we can learn about **population history**, **migration patterns**, and the **impact of natural selection on genetic diversity**.

Bottleneck effects

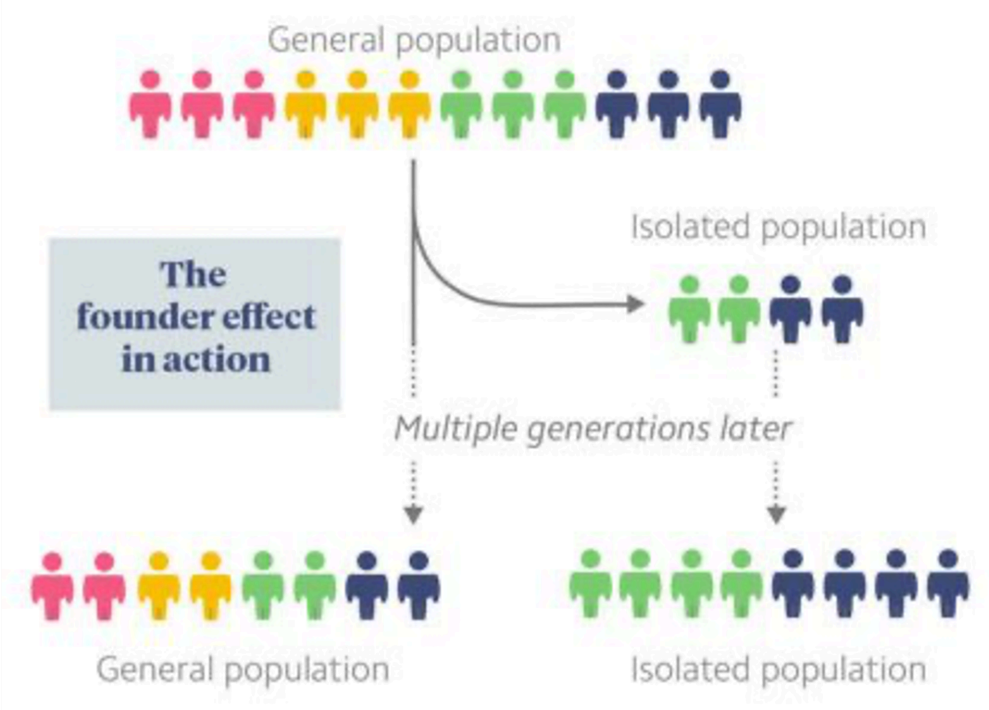
- Natural events like a disaster that kills **at random** a large portion of the population can change the genetic structure of the population.



Source: Publicly available from Google Images

Founder population

- A founder effect occurs when a new colony is started by a few members of the original population.
- For example: the Amish community of Pennsylvania. This population is descended from around 200 German immigrants who started their colony.



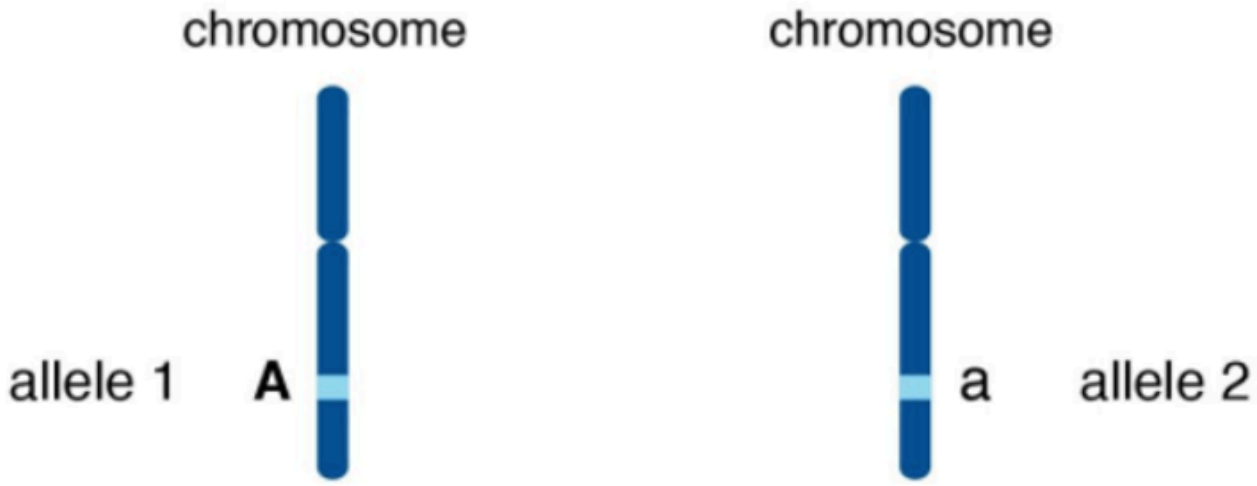
Source: Publicly available from Google Images

Important concepts

- Key principles in population genetics that are important in association analysis:
  - Population Substructure
  - Hardy Weinberg Equilibrium
  - Population Substructure leads to Hardy Weinberg Disequilibrium.

Estimation of allele frequency

- An individual has two copies of each autosomal (not sex) chromosomes.



- Goal: estimate the population proportion of a particular allele A (the other allele is a).

Estimation of allele frequency

- The allele proportion (frequency) in the population is the proportion of chromosomes carrying that allele.
- Suppose we have a sample of n individuals from a population with a proportion p of A alleles.
- We want to estimate p.
- q = 1−p is the frequency of the other allele, a.

Estimation of allele frequency

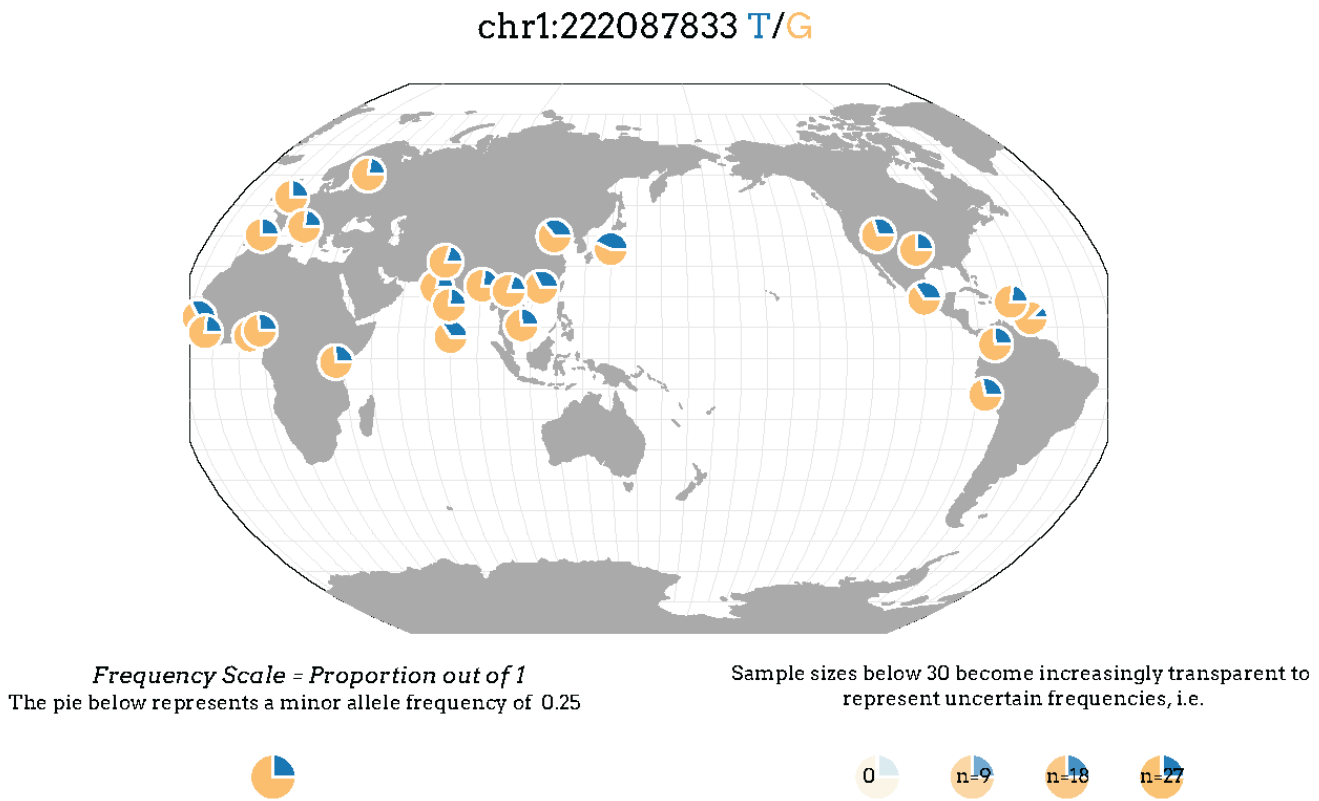
- n_{AA}, n_{Aa}, n_{aa} = number of individuals with genotype AA, Aa, aa.
- $n = n_{AA} + n_{Aa} + n_{aa}$
- $\hat{p} = \frac{2n_{AA}+n_{Aa}}{2n}$
- $\hat{q} = 1 - \hat{p}$
- `plink --bfile your_data --freq --out allele_freq`

Estimation of allele frequency

- $\hat{p}$ is an **unbiased** estimate of p if **random sample with equal probability sampling** (each individual has the same probability of being included in the sample).
- In practice, this means the probability of selection in the sample does not depend on genotype directly or indirectly through a phenotype related to genotype.
    - e.g., some genotypes might be overrepresented if you only sample people with a disease linked to the allele.
- Standard error for proportion $\sqrt{\hat{p}(1-\hat{p})/2n}$ assumes independence -- may not hold.
    - e.g., family-based or structured populations

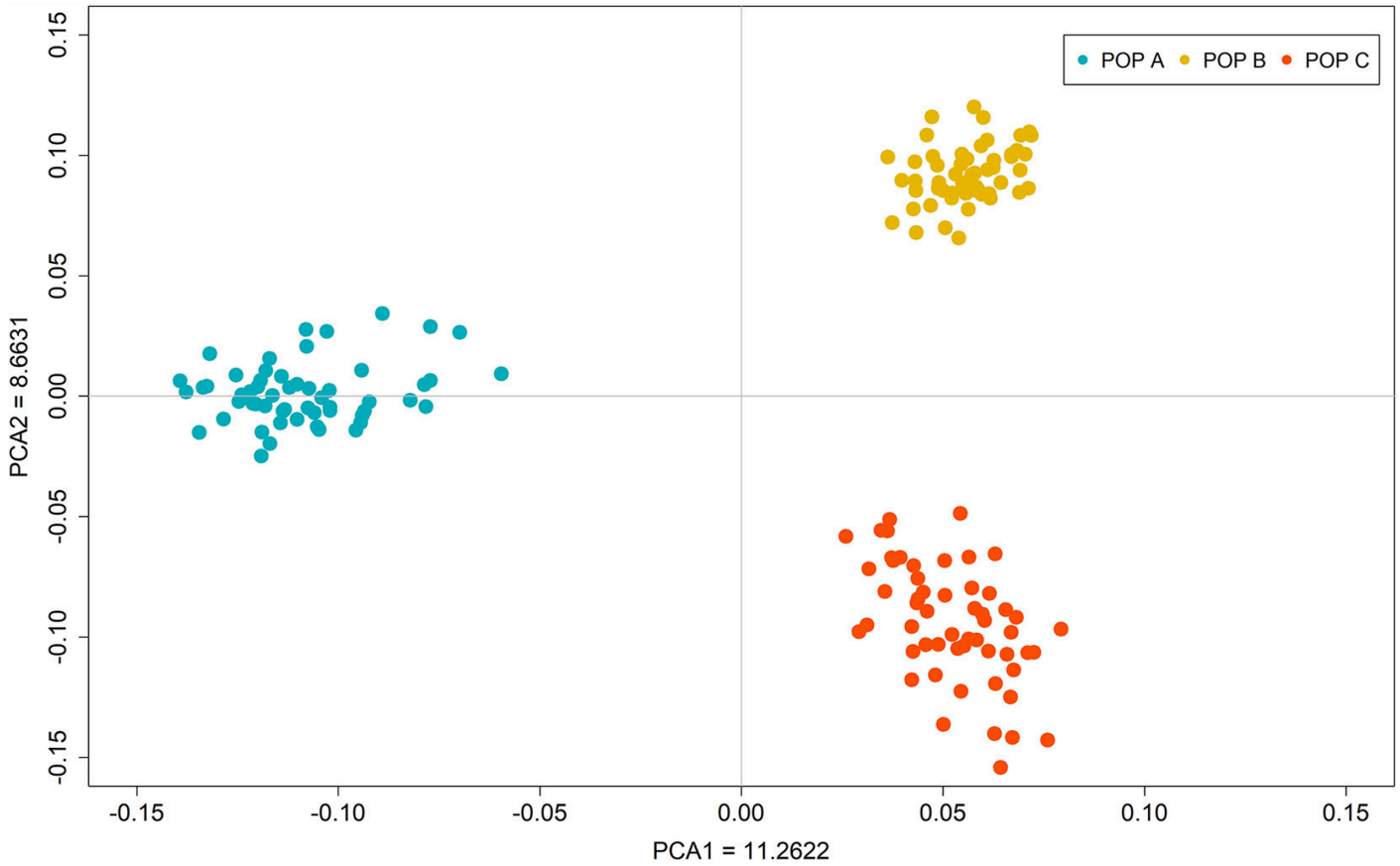
Visualizing the geography of genetic variants

- http://popgen.uchicago.edu/ggv/



Population Substructure

- Different subgroups present within your population



Source: Publicly available from Google Images

Population substructure

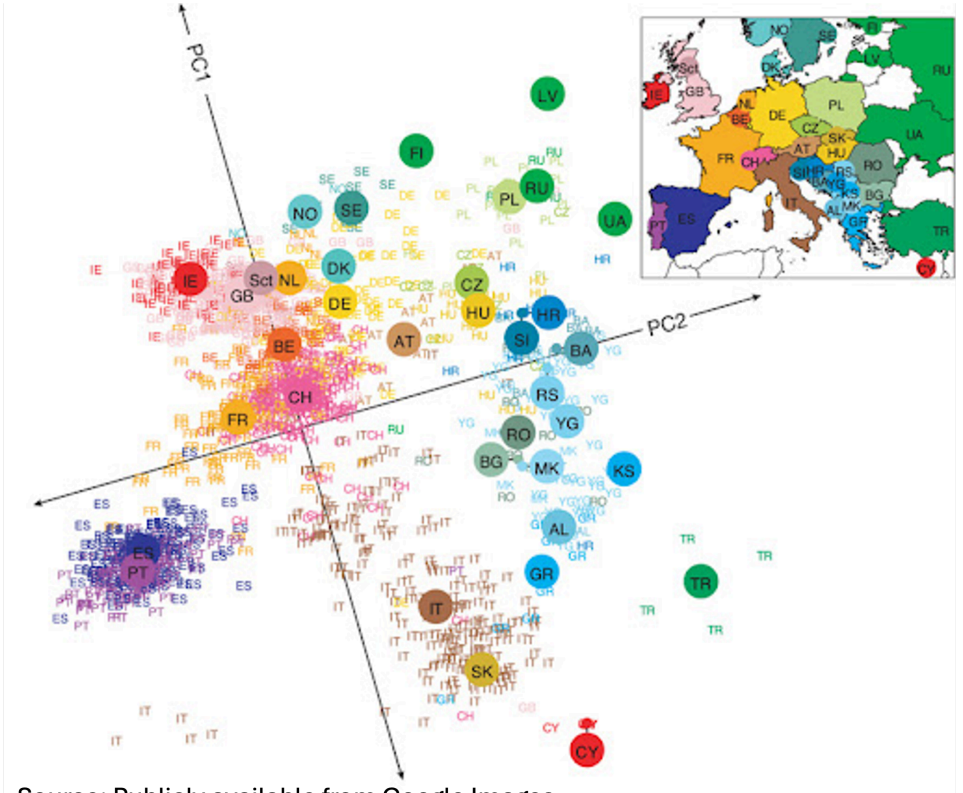
Three common types of population substructure

- Population Stratification
- Population admixture
- Population inbreeding

Population Stratification

- Simplest form of population substructure.
- Individuals in a population can be divided into disjoint strata.
- Strata: ethnic, racial, geographic group.
- Allele frequency can vary among strata.

Population Structure in Europe



Source: Publicly available from Google Images

Population Admixture

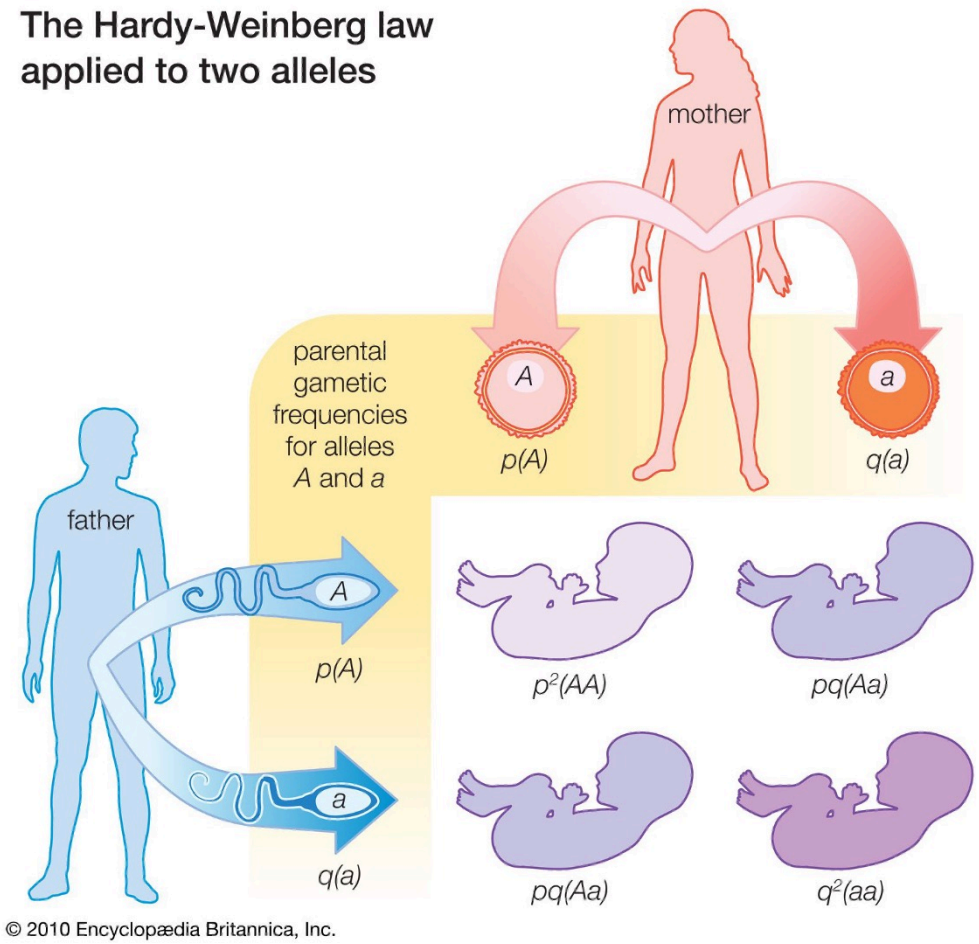
- Individuals in a population have a mixture of different genetic ancestries due to mixing of two or more populations in the past.
- E.g. Hispanic populations:

Hardy Weinberg Equilibrium (HWE)

- In 1908, Hardy and Weinberg independently derived a formula relating the genotype frequencies in offspring to allele frequencies in parents.
- The genotype distribution at a locus is defined by the allele frequencies.
- Assumptions: random mating, no inbreeding, no selection, no mutation, no migration, infinite population size.
- HWE simplifies statistical theory and is often assumed.

Hardy Weinberg Equilibrium

- Let p be **the frequency of A allele**.
- After one generation of random mating: $P(AA) = p^2, P(Aa) = 2pq, P(aa) = q^2$.
- Thus, with random mating, the number of A alleles in the offspring generation $\sim Bin(2, p)$.



Testing for HWE

The Pearson Goodness of Fit Test for HWE

- H_0: HWE holds. Vs H_1: HWE does not hold.
- Given a sample size n from the population:
-

	AA	Aa	aa	
Observed	n_{AA}	n_{Aa}	n_{aa}	n
Expected	$n\bar{p}^2$	$2n\bar{p}\bar{q}$	$n\bar{q}^2$	n

- $\bar{p} = (2n_{AA} + n_{Aa})/(2n)$
- $T = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i}; T \sim \chi^2_{(1)}$ under H_0.

Testing for HWE

- Common first step in any genetic analysis
- Compare the observed genotype frequencies with those expected under HWE (under H_0).

Testing HWE - Example

Exercise

- Assume you observe that the proportion of a population affected with sickle cell anemia is 0.01. Assuming an autosomal recessive diseasmodel and HWE, estimate the frequency of the sickle cell mutation at the hemoglobin locus in this population.

Genetic association studies

- Part 1: Given a trait, should we perform genetic studies and under what conditions?
- Part 2: Association tests

How do we know a trait is genetic?

- Most researchers would not undertake a genetic analysis without enough evidence.

Association testing

- **Objective**: establish association between a trait of interest and a genetic marker.
- Study designs: case-control, case-cohort, population-based design.
- Unrelated subjects or **population-based designs**: easy to collect so possible to achieve large sample sizes as in GWAS.
- **Family-based designs**: robust to population stratification, more difficult to collect. Also hard to collect for late-onset diseases.

Types of tests

- SNP: categorical variable with three genotypes
- Possible tests:
  - 2-DF tests that compare all three genotypes.
  - 1-DF tests : some assumption (e.g. monotonicity) about disease and genotype.
- We assume a case-control design: r cases, s controls, $n=r+s$ total sample size.

Association Testing (2-DF test)

- Y: binary phenotype ($Y = 1$: case).
  - $H_0 : P(Y = 1 \mid AA) = P(Y = 1 \mid Aa) = P(Y = 1 \mid aa)$
  - H_A : At least one inequality holds
-

	aa	Aa	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

- **Two df Pearson test of independence**: $\chi^2 = \sum (O - E)^2 / E.$
  - Sum is over all six entries.
    - e.g. $E[\text{ Case \& aa}] = r * (n_0/n)$; $E[\text{ Controls \& AA}] = s * (n_2/n)$.

Association Testing (2-DF test)

- **Most general test**: assume nothing about the relationship between disease and genotype.
  - $H_0 : P(Y = 1 \mid AA) = P(Y = 1 \mid Aa) = P(Y = 1 \mid aa)$
  - H_A : At least one inequality holds
- Let G be the genotype of the Disease Susceptibility Locus.
- $f_0 = P(Y = 1 \mid G = aa), f_1 = P(Y = 1 \mid G = Aa), f_2 = P(Y = 1 \mid G = AA)$.

Penetrance and mode of Inheritance

- Let G be the genotype of the Disease Susceptibility Locus
- **Penetrance function**: $P(Y = 1 \mid G)$
- **Incompelte penetrance**: $0 < P(Y = 1 \mid G) < 1.$
- **Recessive Model**:
  - $P(Y = 1 \mid G = aa) = 0, P(Y = 1 \mid G = Aa) = 0,$ $P(Y = 1 \mid G = AA) = 1.$
- **Dominance model**:
  - $P(Y = 1 \mid G = aa) = 0, P(Y = 1 \mid G = Aa) = 1,$ $P(Y = 1 \mid G = AA) = 1.$
- **Incompelte penetrance**: $0 < P(Y = 1 \mid G) < 1.$

General Mode of Inheritance

- **Recessive model**:
  - $f_0 = P(Y = 1 \mid G = aa) = P(Y = 1 \mid G = Aa) = f_1$
  - Simple Mendelian recessive disease further assumes $f_1 = f_0 = 0$ and $f_2 = 1$.
- **Dominant model**:
  - $f_1 = P(Y = 1 \mid G = Aa) = P(Y = 1 \mid G = AA) = f_2$
- **Co-dominant model**:
  - f_1 is somewhere between f_0 and f_2.

General Mode of Inheritance

- **Additive model** is a special case of co-dominant model: f_1 is average of f_0 and f_2.
  - Linear scale: $f_1 = \frac{f_0+f_2}{2}$.
  - Log (or multiplicative scale) $f_1 = \sqrt{f_0 \times f_2}$
- Heterozygote advantage model (or disadvantage model):
  - $f_1 <$ both f_0 and f_2 (or $>$ both).

Dominant Tests

- Dominant model:

  $ H_0: P(Y=1 \mid A A)=P(Y=1 \mid A a)=P(Y=1 \mid a a) \
  H_A: P(Y=1 \mid A A \text { or } A a) \neq P(Y=1 \mid a a)$

1 df chi-square test

Optimal when the true disease model is dominant but not for recessive:

$$H_A : P(Y = 1 \mid AA) \neq P(Y = 1 \mid \text{ Aa or } aa)$$

What's next

- Fundamental principles of population genetics
- Basic genetic models and genotype coding frameworks
- Principles of inheritance

What questions do you have about anything from today?