Sociogenomics

ACADEMIC YEAR 2021/2022.

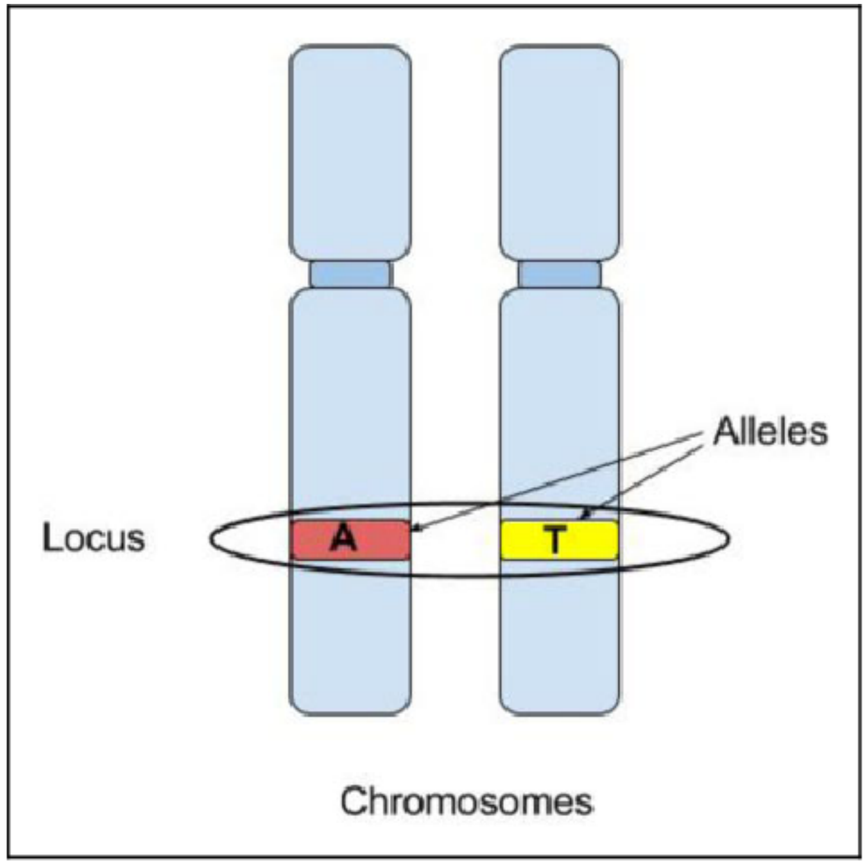
Basic concepts

Nicola Barban

Basic definitions

Loci and alleles

- We will use the term ***locus (pl. loci)*** too refer to a genomic element located in a field position of the genome
- A locus may have different variants called **alleles**
- Although the term gene refers to a functional biological unit, often the terms locus and gene are used interchangeably

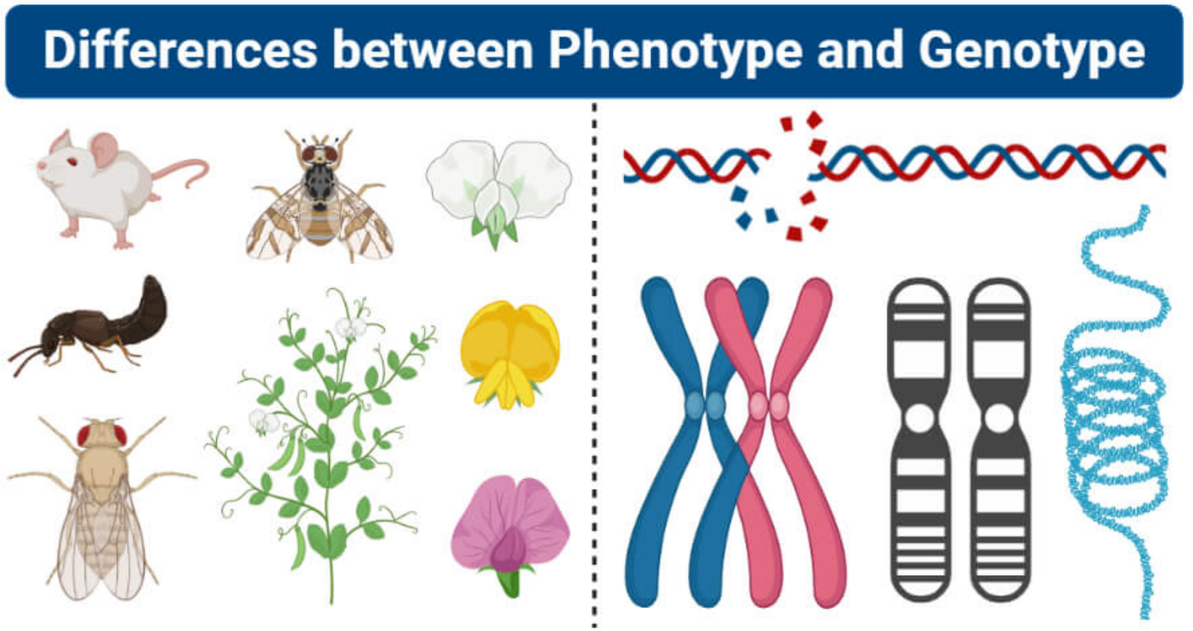


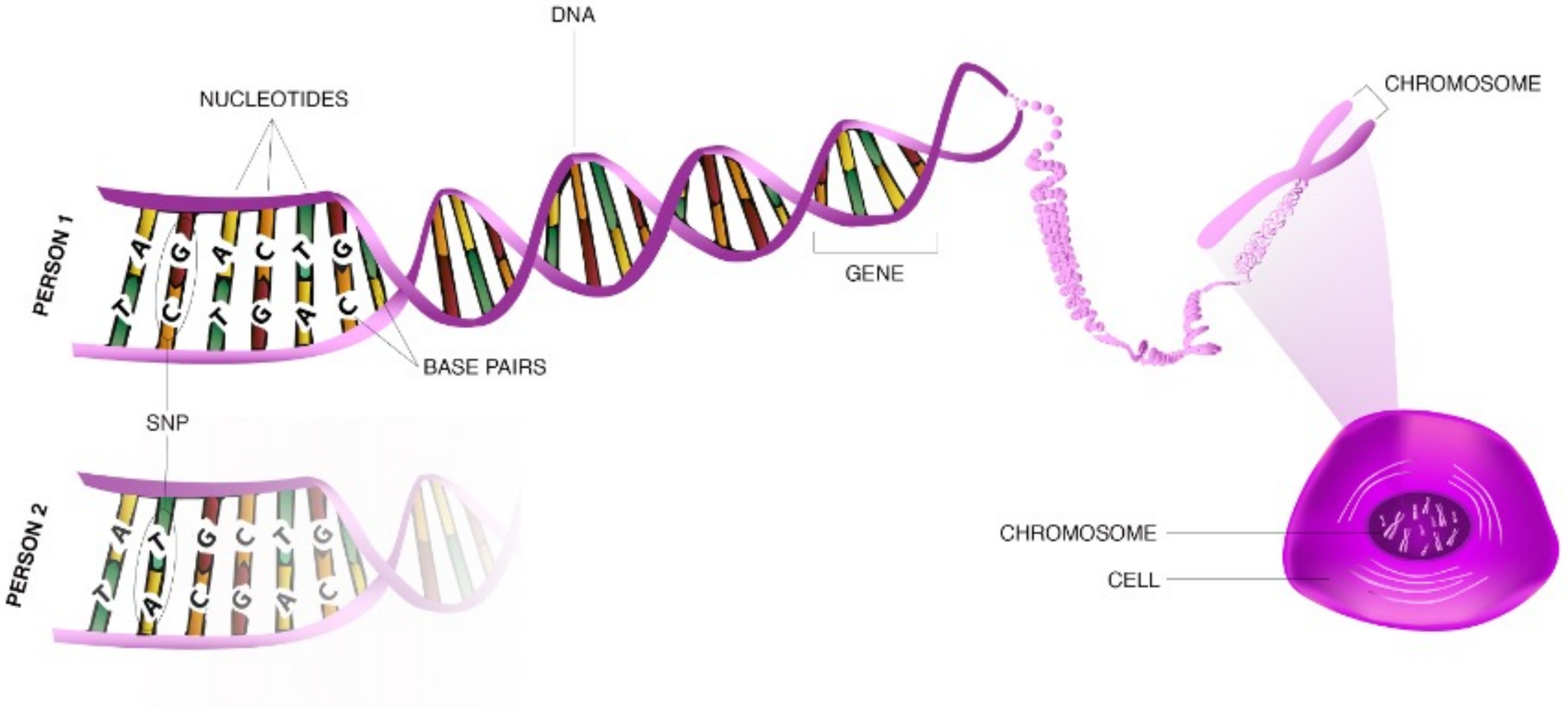
Basic definitions

Phenotype and genotype

Genotype. The complete *heritable* genetic information that can refer to a particular *allele* or set of alleles at a *locus*.

Phenotype. This is the outcome or trait of individuals ranging from physical traits (hair colour, height) to disease status (diabetic) to behaviour (age at first birth, educational attainment).

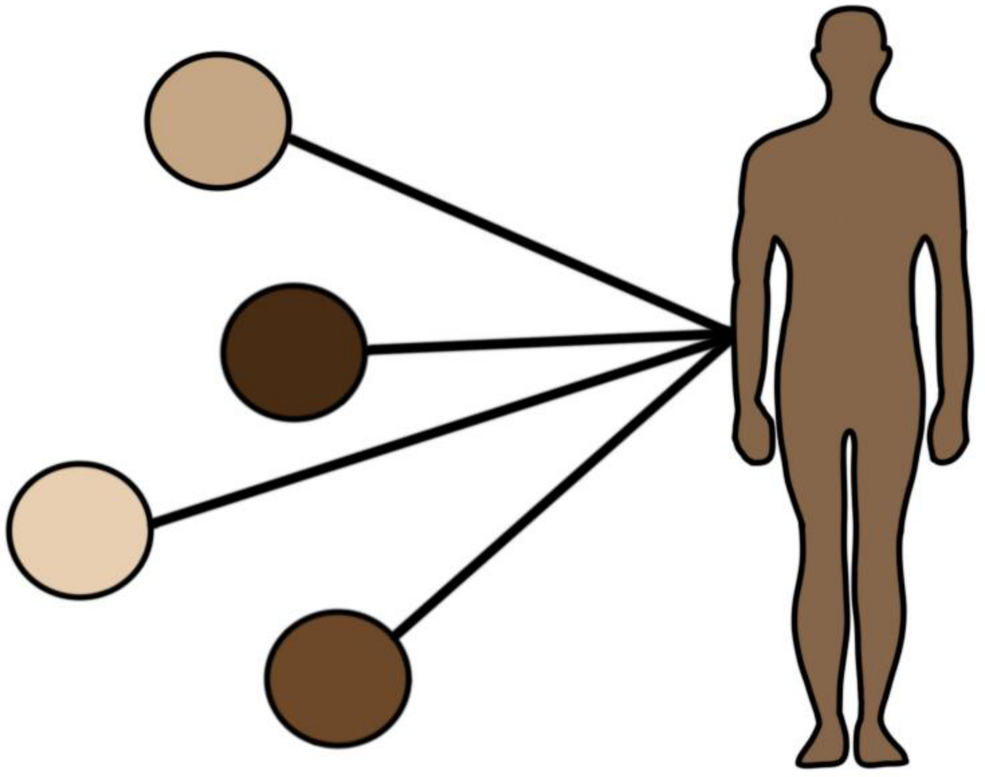



DNA
NUCLEOTIDES
CHROMOSOME
PERSON 1
GENE
BASE PAIRS
SNP
PERSON 2
CHROMOSOME
CELL

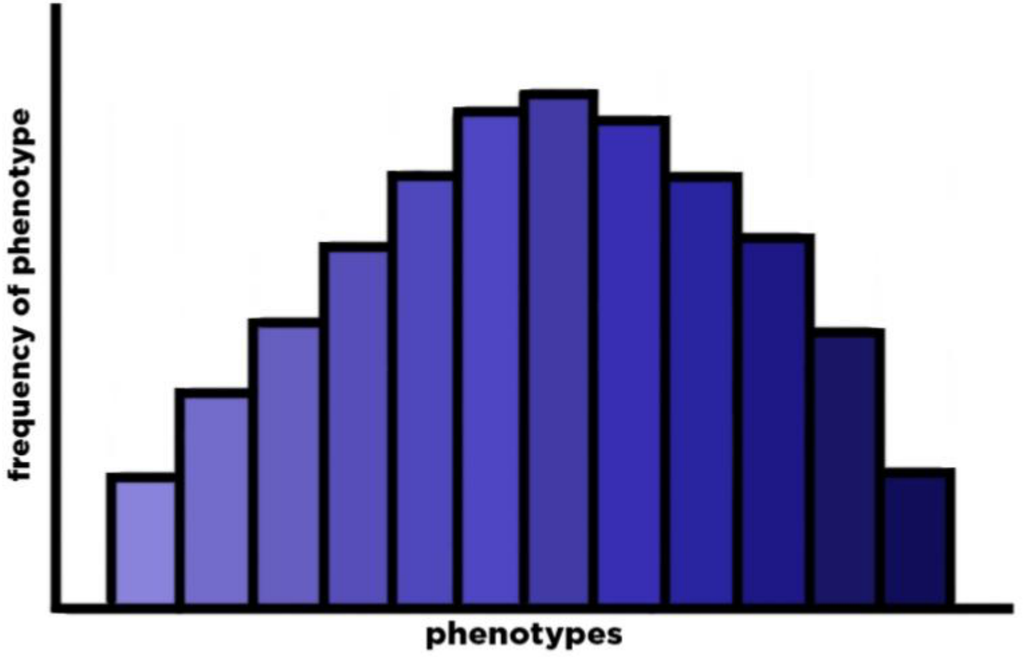
PATTERNS OF INHERITANCE: POLYGENIC TRAITS

Polygenic traits are controlled by more than 1 gene.

Instead, several genes contribute to the final phenotype of a given trait.



Many genes contribute to skin color.



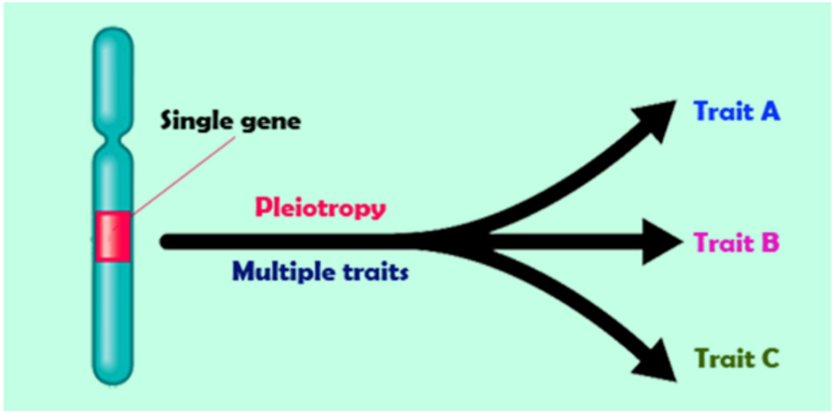
Each gene plays a small role in the expression of the trait.

This allows for a trait to vary greatly from person to person.

Height is a polygenic trait which is why people can be very short, very tall, or anything in between.

Pleiotropy

- The phenomenon of when one *gene* influences two or more seemingly unrelated *phenotypes.*



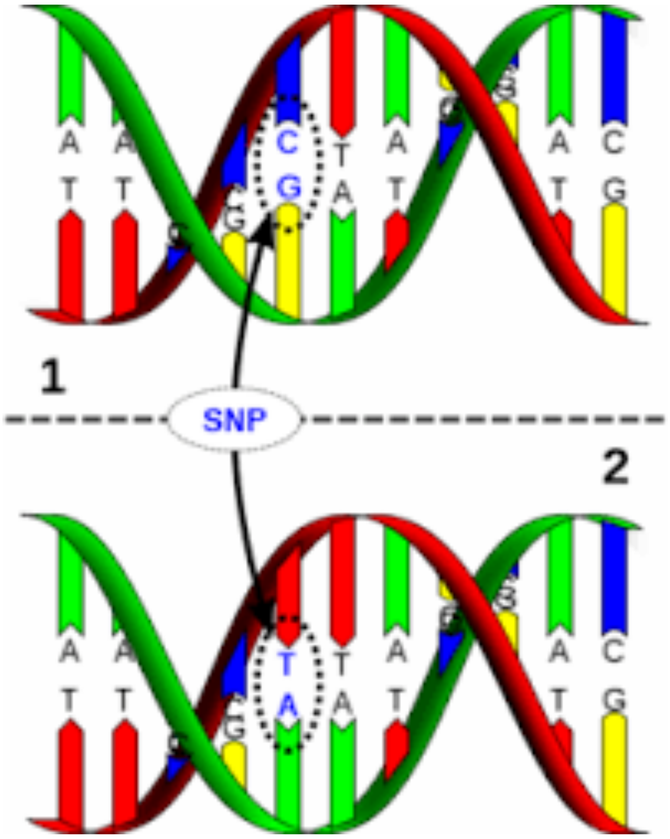
- ***MOST OF HUMAN TRAITS ARE POLYGENIC***
- ***MOST OF GENES HAVE PLEIOTROPIC EFFECTS***

Genetic variability

If all individuals in the population carry the same allele, we say that the locus is **monomorphic**; at this locus there is no genetic variability in the population.

If there are multiple alleles in the population at a locus, we say that this locus is **polymorphic** (this is sometimes referred to as a segregating site).

.



single nucleotide polymorphisms SNP

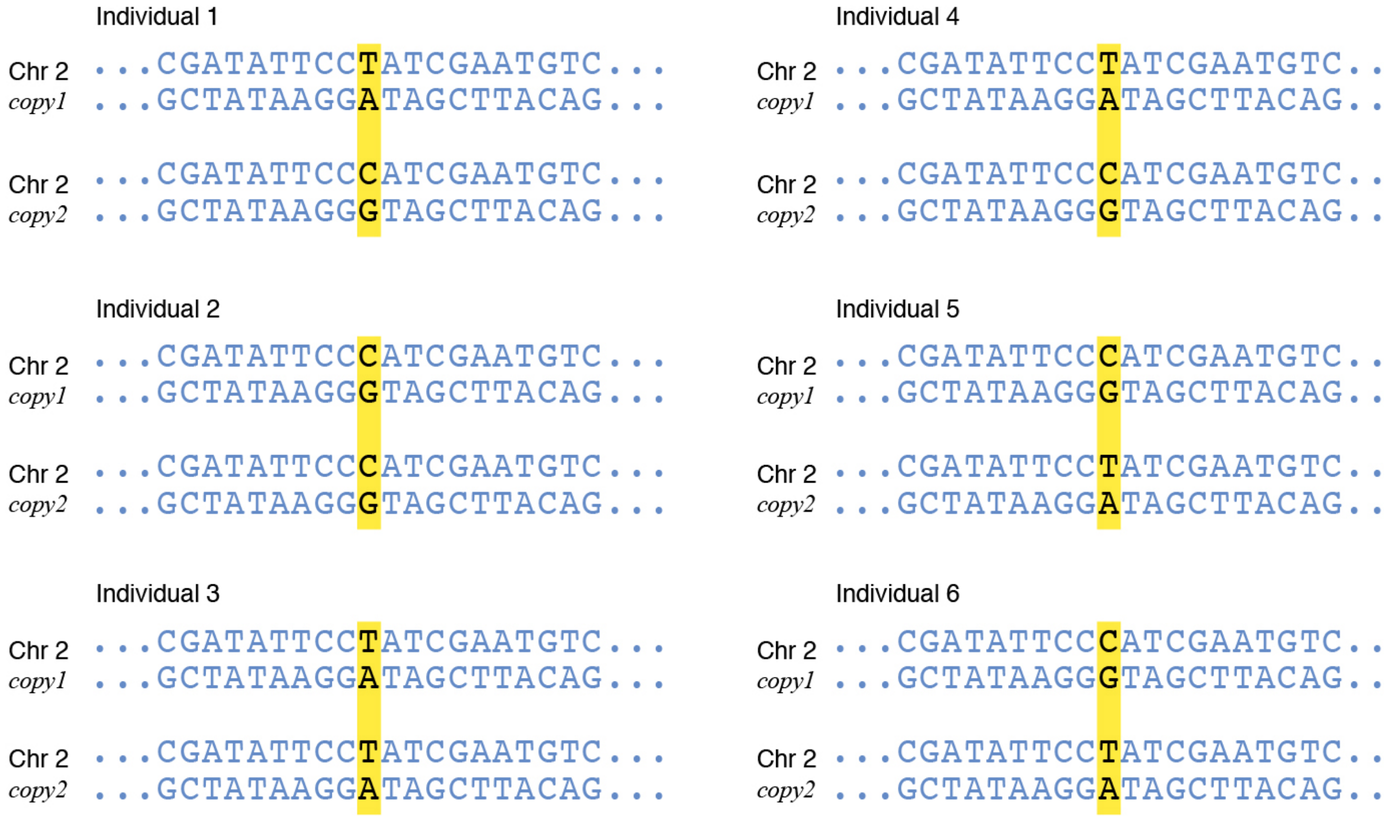
Allele frequencies

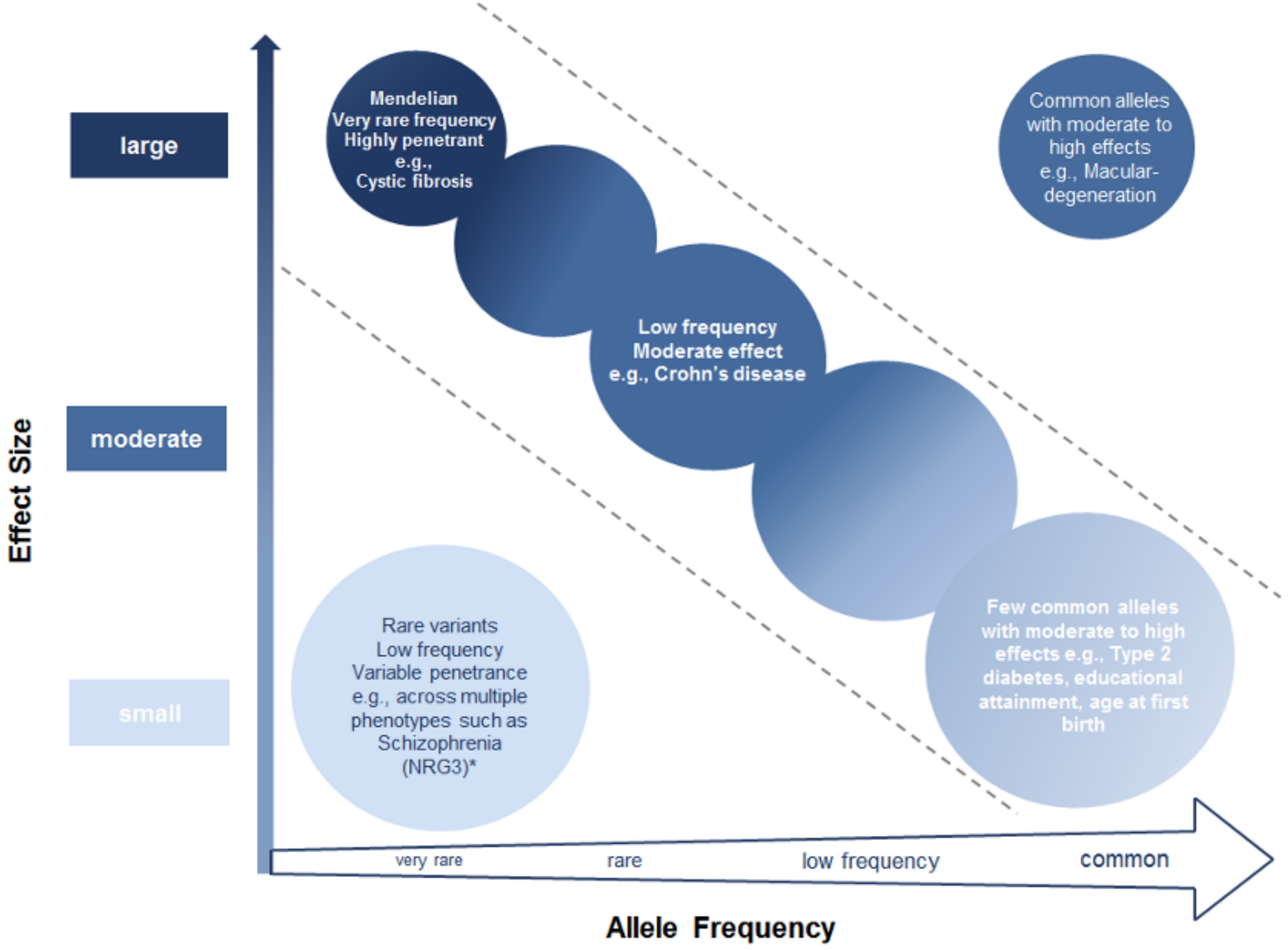
Allele frequencies are a central unit of populati genetics analysis,

The frequency of the least common or minor allele – called **minor allele frequency (MAF)** is the key statistics used to characterize polymorphisms.

In the literature, polymorphisms are distinguished by their MAF, and categorized as **common** (MAF > 0.05), **low-frequency** (0.01 < MAF < 0.05) or **rare** (MAF < 0.01) **variants**.

We only get to observe genotype counts which are used to calculate .




Effect Size
large
moderate
small
Mendelian
Very rare frequency
Highly penetrant
e.g.,
Cystic fibrosis
Low frequency
Moderate effect
e.g., Crohn's disease
Common alleles with moderate to high effects e.g., Macular-degeneration
Rare variants
Low frequency
Variable penetrance
e.g., across multiple phenotypes such as Schizophrenia (NRG3)*
Few common alleles with moderate to high effects e.g., Type 2 diabetes, educational attainment, age at first birth
very rare
rare
low frequency
common
Allele Frequency

When an individual has two of the same allele, regardless of whether it is dominant or recessive, they are called **homozygous.**

Heterozygous refers to having one of each of the different alleles. A person is heterozygous at a gene locus when their cells contain two different alleles. Heterozygosity thus refers to a specific genotype

Consider a locus with two alleles A1 and A2, the possible genotypes are: $\mathbf{A_1A_1}$; $\mathbf{A_1A_2}$ or $\mathbf{A_2A_1}$ and $\mathbf{A_2A_2}$.

Let $\mathbf{N_{11}}$ and $\mathbf{N_{12}}$ be the number of $\mathbf{A_1A_1}$ homozygotes and $\mathbf{A_1A_2}$ heterozygotes, and **N** the number of individuals

The relative frequencies of A_1A_1 is $f_{11} = N_{11}/N$

And $f_{12} = N_{12}/N$

he frequency of allele A_1 in the population is then given by

$$p = \frac{2N_{11} + N_{12}}{2N} = f_{11} + \frac{1}{2} f_{12}$$

The frequency of the alternate allele (A_2) is then just $q = 1 - p$.

Hardy-Weinberg Equilibrium (HWE)

How much genetic variation (allele and genotype frequencies) in a population will remain constant from one generation to the next in the absence of evolutionary influences?

a theoretical model describing the probability and distribution of genotype frequencies in a population

*p = the frequency for the major allele (**A_1**)*

*q = the frequency for the minor allele (**A_2**)*

Let us assume that the allele **A_1** has a frequency of *p* = 0.3 and allele **A_2** has a frequency of *q = 0.7.*

The Hardy-Weinberg equation is thus:

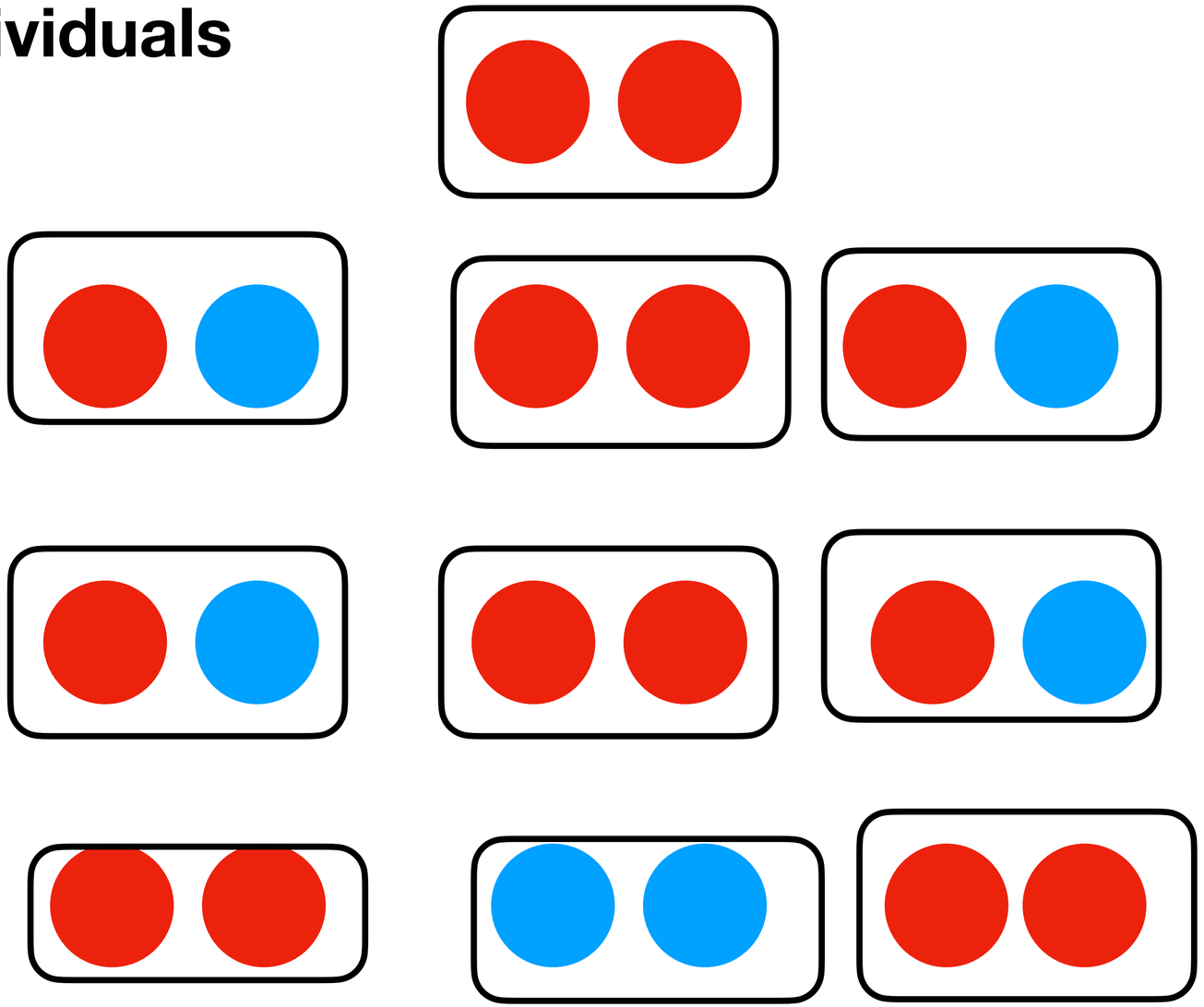
$$p^2 + 2pq + q^2 = 1$$

0.3x0.3 +2x0.3x0.7+0.7x0.7=1

9% **(A_1A_1)** +42% **(A_1A_2)** +49% **(A_2A_2)**

Example

10 individuals



Recessive

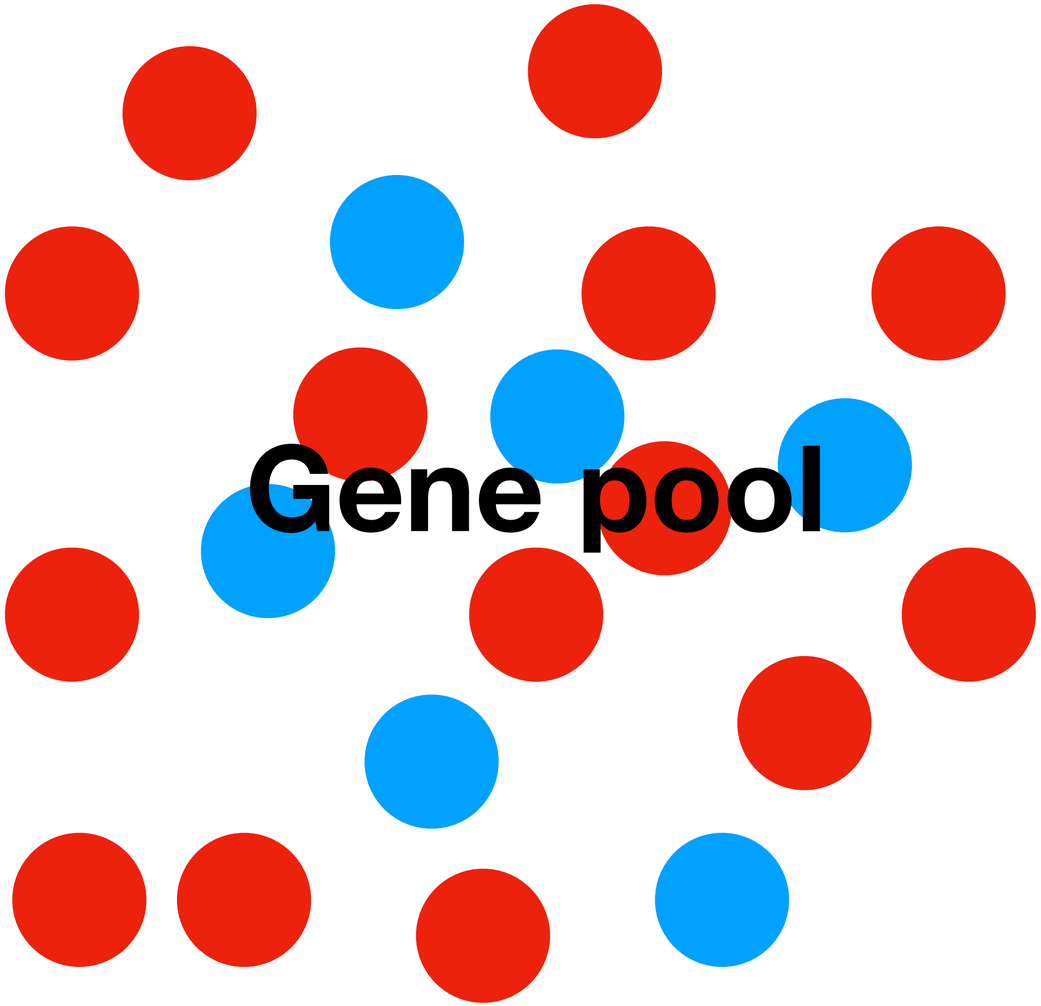
Dominant

If you have two recessive alleles, you have red hair

In this example 50% of the population has red hair

Example

10 individuals

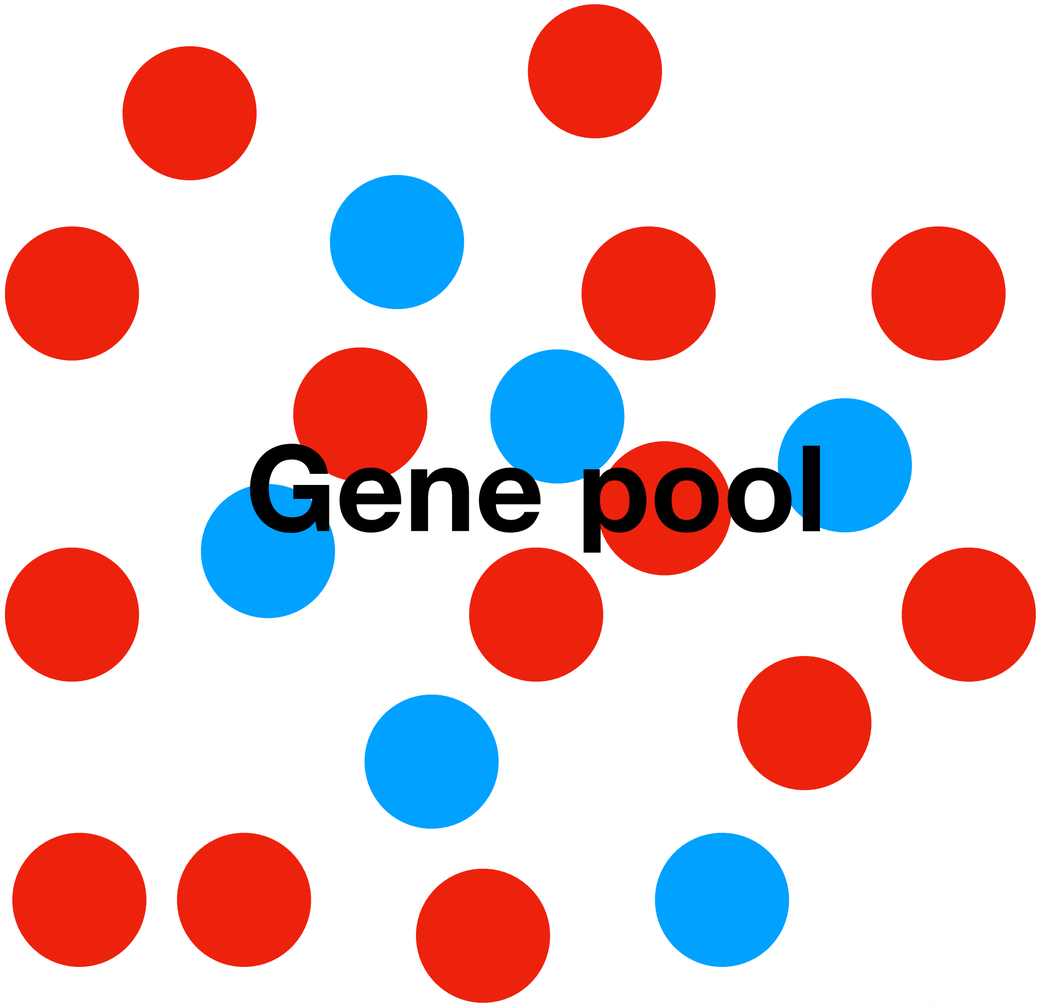


Recessive

Dominant

Example

10 individuals



Recessive

Dominant

$$p = \frac{6}{20} = 0.3$$

$$q = \frac{14}{20} = 0.7$$

$$p + q = 1$$

Example

10 individuals

Recessive

Dominant

$$p = 0.3; q = 0.7$$

What the expected homozygous and heterozygous proportion?

$$p^2 = \frac{7}{10} \times \frac{7}{10} = 0.49$$

$$q^2 = \frac{3}{10} \times \frac{3}{10} = 0.09$$

$$2pq = 2 \times \frac{3}{10} \times \frac{7}{10} = 0.42$$

Exercise 1

- Two percent of the humans in the planet have red hair. What proportion of human are heterozygous for this trait?

Exercise 1

Solution

- Two percent of the humans in the planet have red hair. What proportion of human are heterozygous for this trait?

$$q^2 = 0.02; q = \sqrt{0.02} = 0.14$$

$$p = 1 - 0.14 = 0.86$$

$$2pq = 2 \times 0.86 \times 0.14 = 0.24$$

Exercise 2

- 60 million Italians, 6,000 have cystic fribrosis. How many are carriers?

Exercise 2

- In 2018 5.501 Italians have cystic fibrosis on a population of 60.42 million. How many are expected to be carriers?

- Cystic fibrosis is caused by mutations in the gene that produces the cystic fibrosis transmembrane conductance regulator (CFTR) protein.
- In people with CF, mutations in the CFTR gene can disrupt the normal production or functioning of the CFTR protein found in the cells of the lungs and other parts of the body.
- Cystic fibrosis is an example of a recessive disease. That means a person must have a mutation in both copies of the CFTR gene to have CF.

Exercise 2

Solution

- In 2018 5.501 Italians have cystic fibrosis on a population of 60.42 million. How many are expected to be carriers?

$$q^2 = \frac{5{,}501}{60{,}420{,}000} = 0.0001$$

$$q = \sqrt{0.0001} = 0.01$$

$$p = 1 - 0.01 = 0.99$$

1 in 25 italians

$$2pq = 2 \times 0.99 \times 0.01 = 0.0198$$

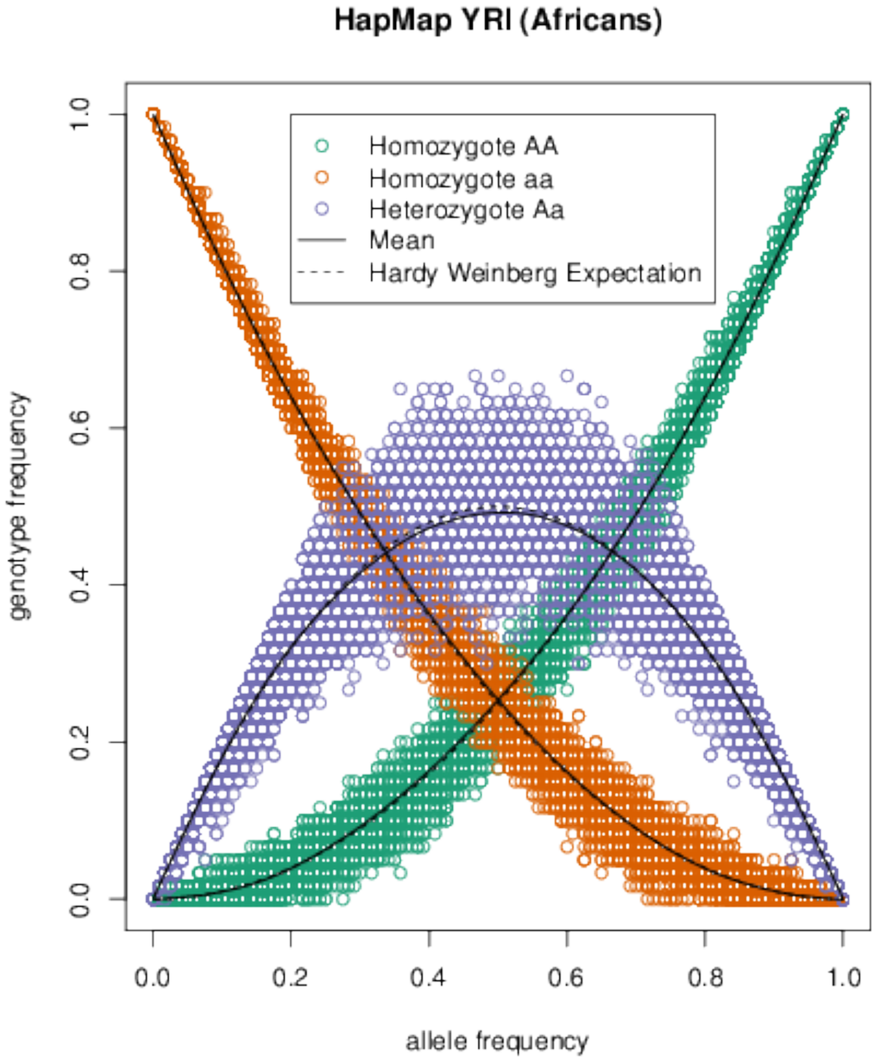
$$60{,}420{,}000, \times 0.02 = 1{,}208{,}400 \text{ carriers}$$

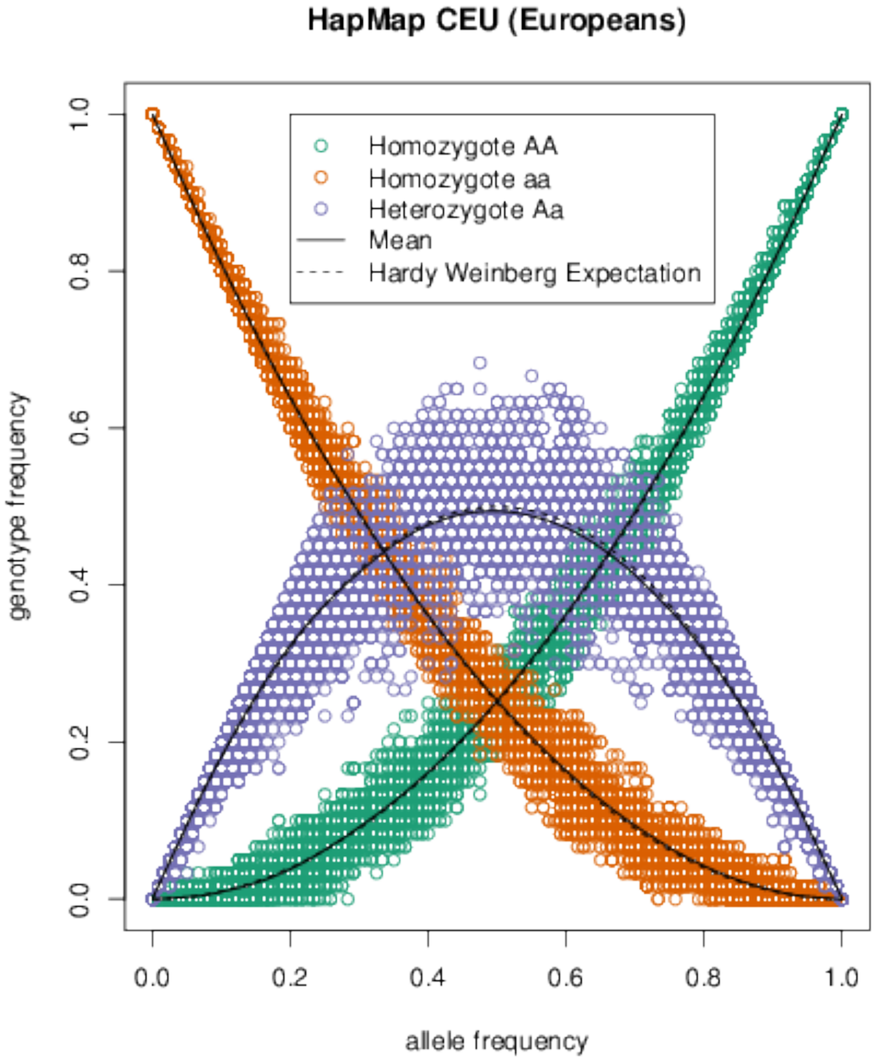
Testing for HWE

The basic idea is to compare the observed genotypes in a sample with those which are expected if HWE holds.

Genotype				
	A_1A_1	A_1A_2	A_2A_2	
Observed	N_{11}	N_{12}	N_{22}	N
Expected	Np^2	Npq	Nq^2	N

$\sum (O-E)^2/E \sim \chi^2$ with one degree of freedom under H_0


HapMap YRI (Africans)
Homozygote AA
Homozygote aa
Heterozygote Aa
Mean
Hardy Weinberg Expectation
genotype frequency
allele frequency
0.0
0.2
0.4
0.6
0.8
1.0


HapMap CEU (Europeans)
Homozygote AA
Homozygote aa
Heterozygote Aa
Mean
Hardy Weinberg Expectation
genotype frequency
allele frequency
0.0
0.2
0.4
0.6
0.8
1.0

Failure of HWE

Rejecting a test of HWE provides some evidence that HWE does not hold in the population. Some cases:

- **Selecting the sample with regard to a** phenotype associated with the **genotype**
- **Population stratification**, i.e. sample comes from heterogenous subpopulations with different allele frequencies
- **Genotyping errors**
- **Inbreeding**, i.e. parents have common ancestors and there is a positive probability that individuals inherit same allele

Genetic drift

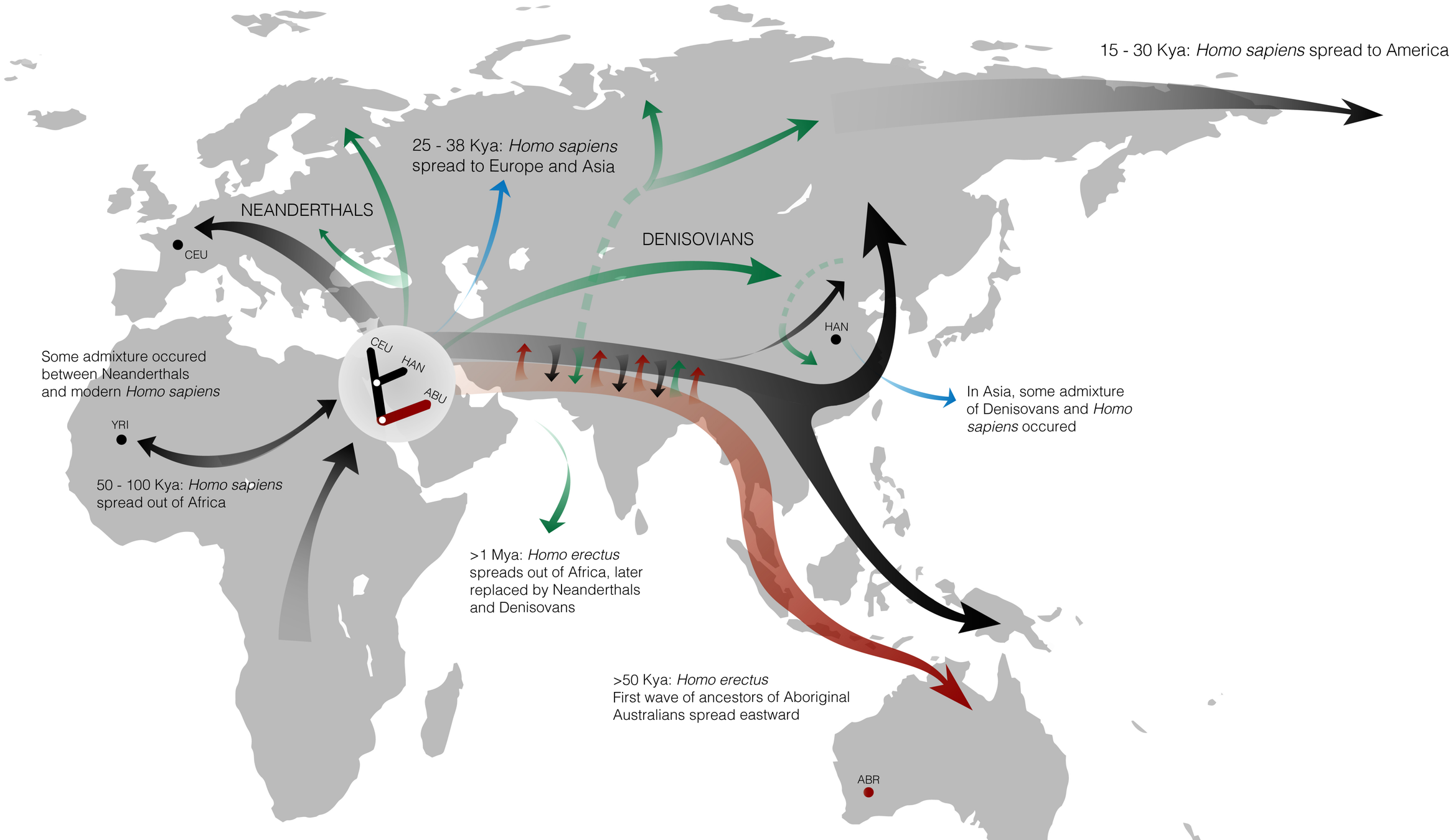
The Hardy–Weinberg principle states that within sufficiently large populations, the allele frequencies remain constant from one generation to the next unless the equilibrium is disturbed by migration, genetic mutations, or selection.

However, in finite populations, no new alleles are gained from the random sampling of alleles passed to the next generation, but the sampling can cause an existing allele to disappear

Genetic drift describes random, non-selective change to the allele frequencies of a population.

https://www.biologysimulations.com/genetic-drift-bottleneck-event

https://keholsinger.shinyapps.io/Genetic-Drift/


15 - 30 Kya: *Homo sapiens* spread to America
25 - 38 Kya: *Homo sapiens* spread to Europe and Asia
NEANDERTHALS
DENISOVIANS
CEU
HAN
Some admixture occured between Neanderthals and modern *Homo sapiens*
CEU
HAN
ABU
In Asia, some admixture of Denisovans and *Homo sapiens* occured
YRI
50 - 100 Kya: *Homo sapiens* spread out of Africa
>1 Mya: *Homo erectus* spreads out of Africa, later replaced by Neanderthals and Denisovans
>50 Kya: *Homo erectus*
First wave of ancestors of Aboriginal Australians spread eastward
ABR

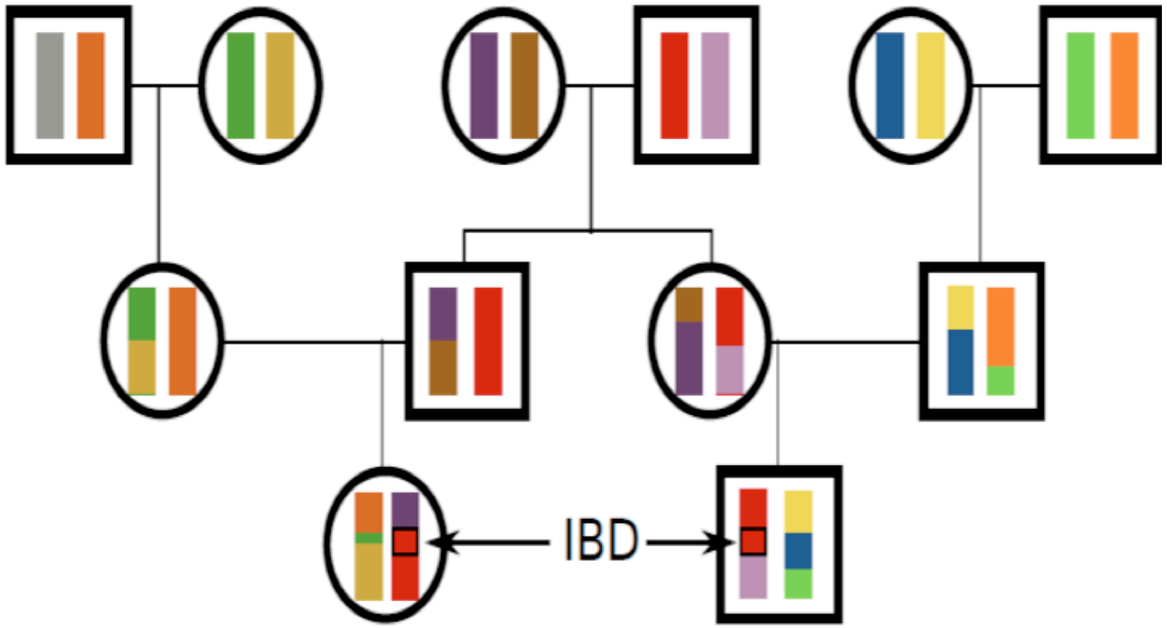
Genetic evolution

Five forces of change in allele frequencies

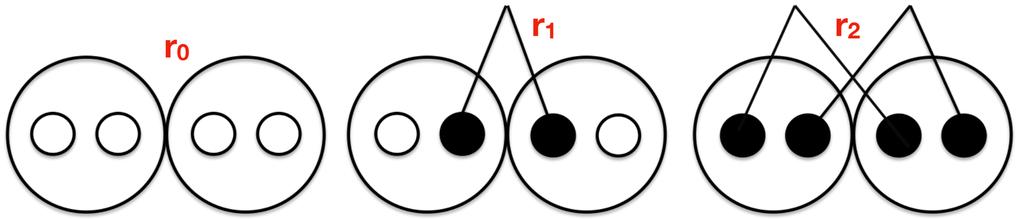
1. Selection
2. Mutation (increases variation)
3. Genetic drift (decrease variation)
4. Migration (gene flow)
5. Non-random Mating (it does not actually cause any change in allele frequencies across generations)

Allele sharing among related individuals and Identity by Descent

- All of the individuals in a population are related to each other by a giant pedigree (family tree)
- Related individuals can share alleles that have both descended from the shared common ancestor.
- We will define two alleles to be **identical by descent (IBD)** if they are identical due to a common ancestor in the past few generations.
- One summary of how related two individuals are is the probability that s pair of individuals share 0, 1, or 2 alleles identical by descent



Identity by descent

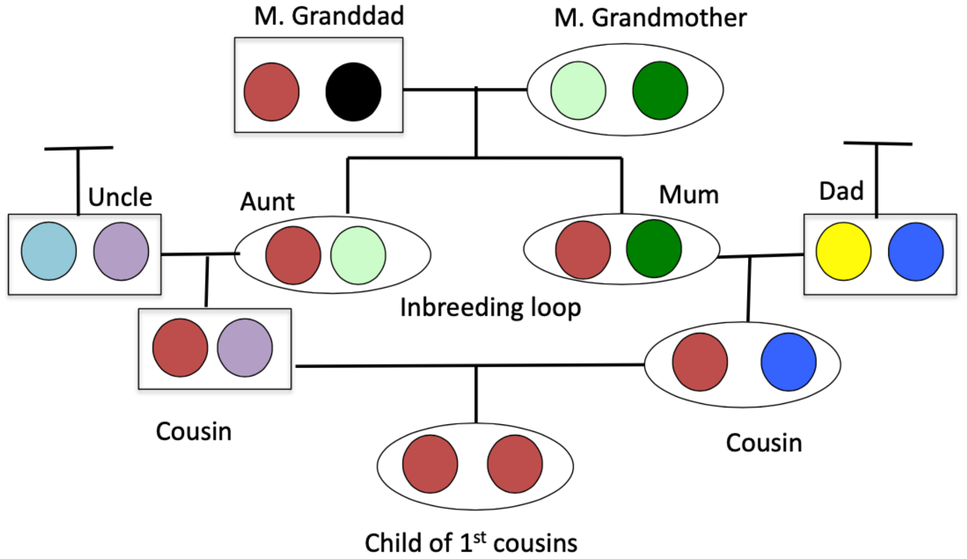


One summary of relatedness is the **kinship coefficient:** i.e. probability that two alleles (I & J) picked at random, one from each of the two different individuals i and j, are identical by descent

he relationship between a parent and a child is the chance that the randomly picked allele in the child is from the parent (probability 1/2) and the probability of the allele that is picked from the parent being the same one passed to the child (probability 1/2)

relationship(i,j)	r_0	r_1	r_2	F_{ij}
Parent-child	0	1	0	1/4
Full siblings	1/4	1/2	1/4	1/4
Identical twins	0	0	1	1/2
1st cousins	3/4	1/4	0	1/16

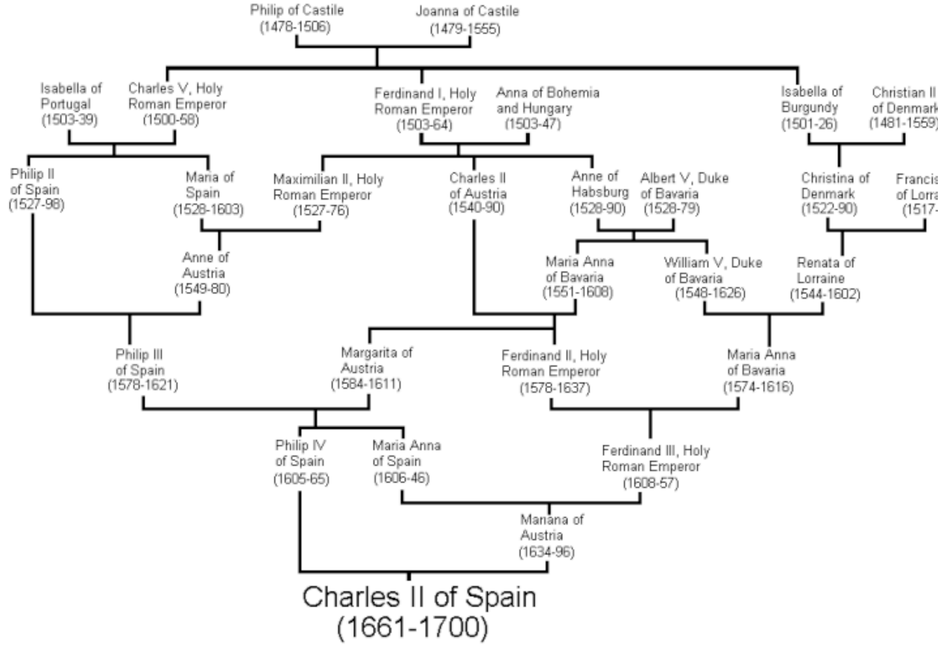
Inbreeding



- We can define an inbred individual as an individual whose parents more closely related to each other than two random individuals dra from some reference population.

- Multiple inbreeding loops increase the probability that a child is homozygous by descent at a locus

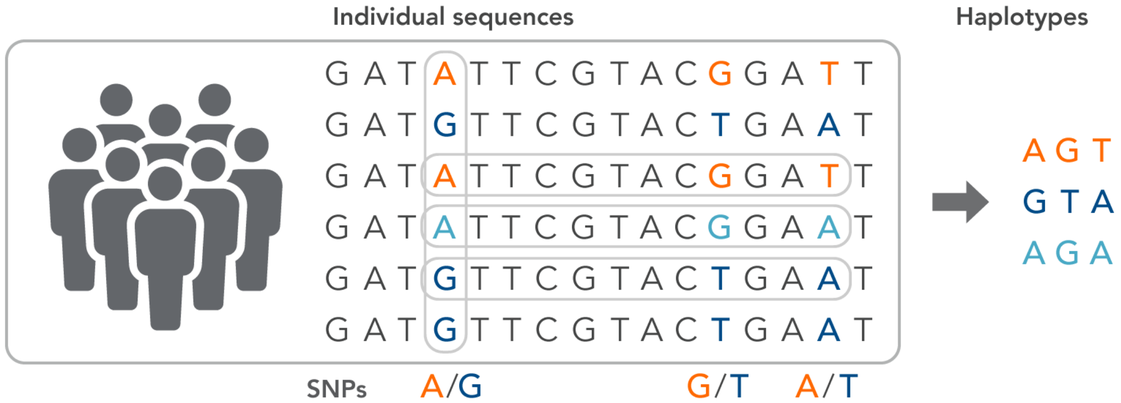
- Alvarez et al. (2009) calculated that Charles II had an inbreed- ing coefficient of 0.254, equivalent to a full-sib mating, thanks to all of the inbreeding loops in his pedigree. Therefore, he is expected to have been homozygous by descent for a full quarter of his genome. As we'll talk about later in these notes, this means that Charles may have been homozygous for a number of recessive disease alleles, and indeed he was a very sickly man who left no descendants due to his infertility. [6] Thus plausibly the end of one of the great European dynasties came about through inbreeding.



Correlations Among Loci

A **haplotype** is a set of DNA variations, that tend to be inherited together. A haplotype can refer to a combination of alleles or to a set of SNPs found on the same chromosome.

Information about haplotypes is being collected by the International HapMap Project and is used to investigate the influence of genes on disease.



Linkage disequilibrium

Linkage disequilibrium (LD) refers to the statistical non-independence (i.e. a correlation)
of alleles in a population at different loci

Consider two loci **A** (alleles A a) and **B** (alleles B b) and allele frequencies $p_A; p_a; p_B; p_b$

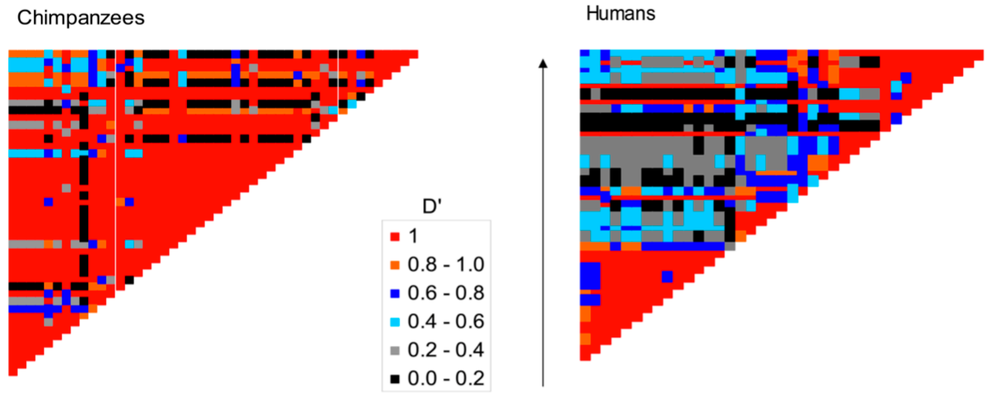
IF THEIR SEGREGATION IS INDEPENDENT: $p_{AB} = p_A\, p_B$*, OTHERWISE THE TWO LOCI ARE IN LINKAGE DISEQUILIBRIUM*

Quantifying linkage disequilibrium

$D = p_{AB} - p_A p_B$

If D = 0 we'll say the two loci are in linkage equilibrium, while if
D > 0 or D < 0 we'll say that the loci are in linkage disequilibrium

physically close SNPs, i.e. those close to the diagonal, have higher absolute values of D as closely linked alleles are separated by recombination less often allowing high levels of LD to accumulate.



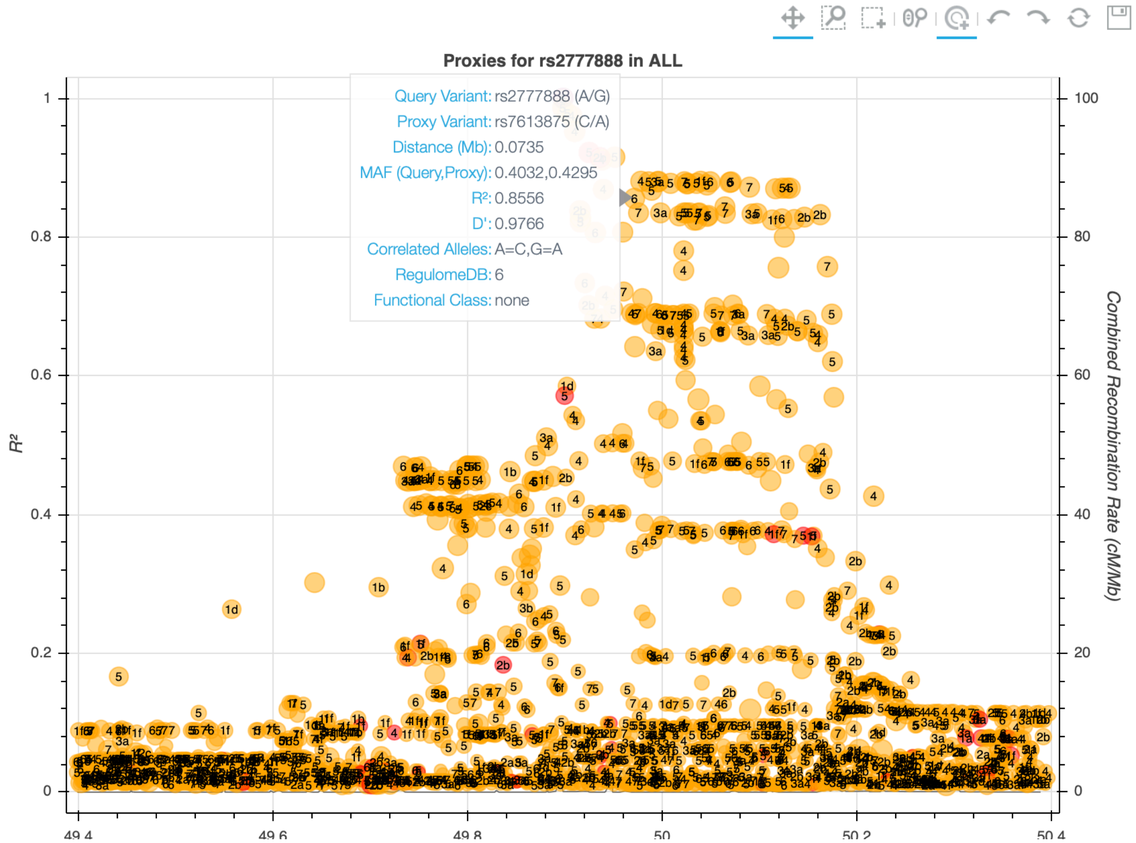
correlation coefficient

As D is a covariance, and $p_A(1 - p_A)$ is the variance of an allele drawn at random from locus A

.

$$r^2 = \frac{D^2}{p_A(1 - P_A)P_B(1 - p_B)}$$

https://ldlink.nci.nih.gov/?tab=home



Why is it relevant?

1. Evolutionary biology

- LD is of importance in evolutionary biology provides information about past events and it constrains the potential response to both natural and artificial selection.
- LD in each genomic region reflects the history of natural selection, gene conversion, mutation and other forces that cause gene-frequency evolution.
- Haplotype blocks vary somewhat among human populations — they tend to be shorter in African populations

Why is it relevant?

2. Association studies

- Most of the studies do not sequence the entire genome, but they genotype a sample of SNPs
- Genotype chips are built to represent a (large) number of SNPs

Figure 1: Omni Family of Microarrays

Omni arrays provide flexibility for timing and budget to help investigators effectively achieve their research goals.

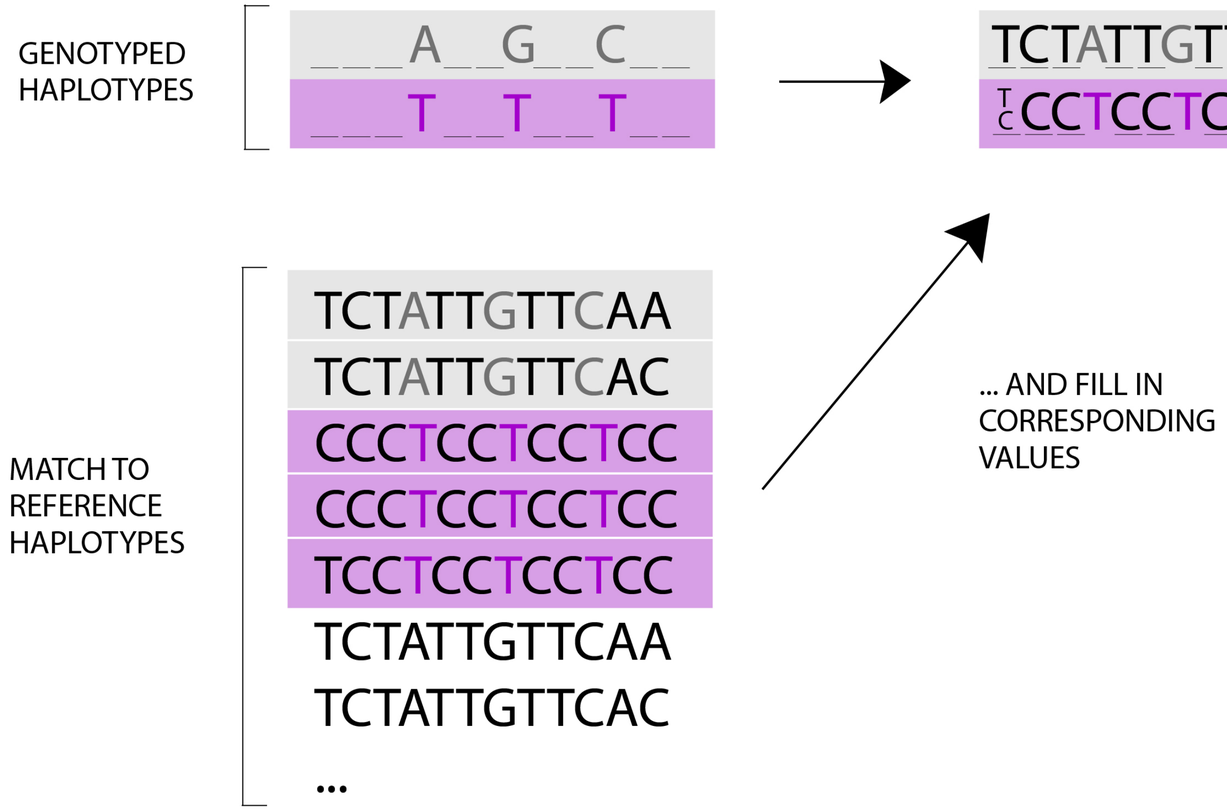
Table 1: Omni BeadChip Performance Parameters

	OmniExpress	Omni2.5	Omni5
Number of Fixed Markers	730,525	2,379,855	4,301,331
Available Custom Markers	up to 200,000	n/a	up to 500,000
Number of Samples	12	8	4
DNA Requirement	200 ng	200 ng	400 ng
Assay	Infinium HD	Infinium LCG	Infinium LCG
Instrument Support	HiScan or iScan	HiScan or iScan	HiScan or iScan
Sample Throughput*	> 1,400 / week	~1,067 samples / week	> 460 samples / week
Scan Time / Sample	5 minutes	6.5 minutes (HiScan) 11.4 minutes (iScan)	15 minutes (HiScan) 25 minutes (iScan)

% Variation Captured (r^2 > 0.8)	**1kGP† MAF > 5%**	**1kGP† MAF > 1%**	**1kGP† MAF > 5%**	**1kGP† MAF > 1%**	**1kGP† MAF > 5%**	**1kGP† MAF > 1%**
CEU	0.73	0.58	0.83	0.73	0.87	0.83
CHB + JPT	0.74	0.62	0.83	0.73	0.85	0.76
YRI	0.40	0.25	0.65	0.51	0.71	0.58

Genetic Imputation

Imputation in genetics refers to the statistical inference of unobserved genotypes.[1] It is achieved by using known haplotypes in a population

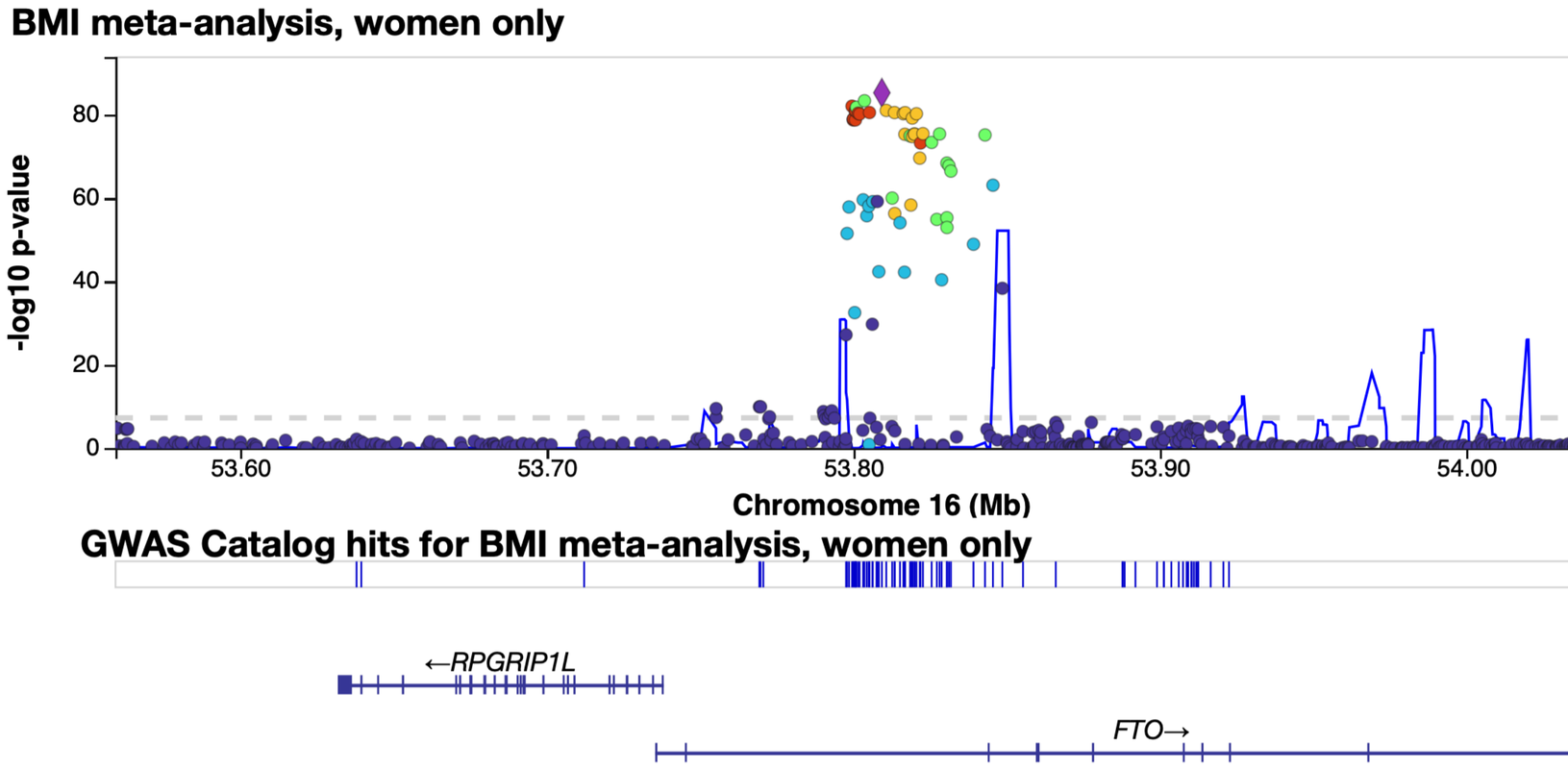


Reference Panels

- HapMap
- 1000Genome
- The Haplotype Reference Consortium(http://www.haplotype-reference-consortium.org/participating-cohorts)

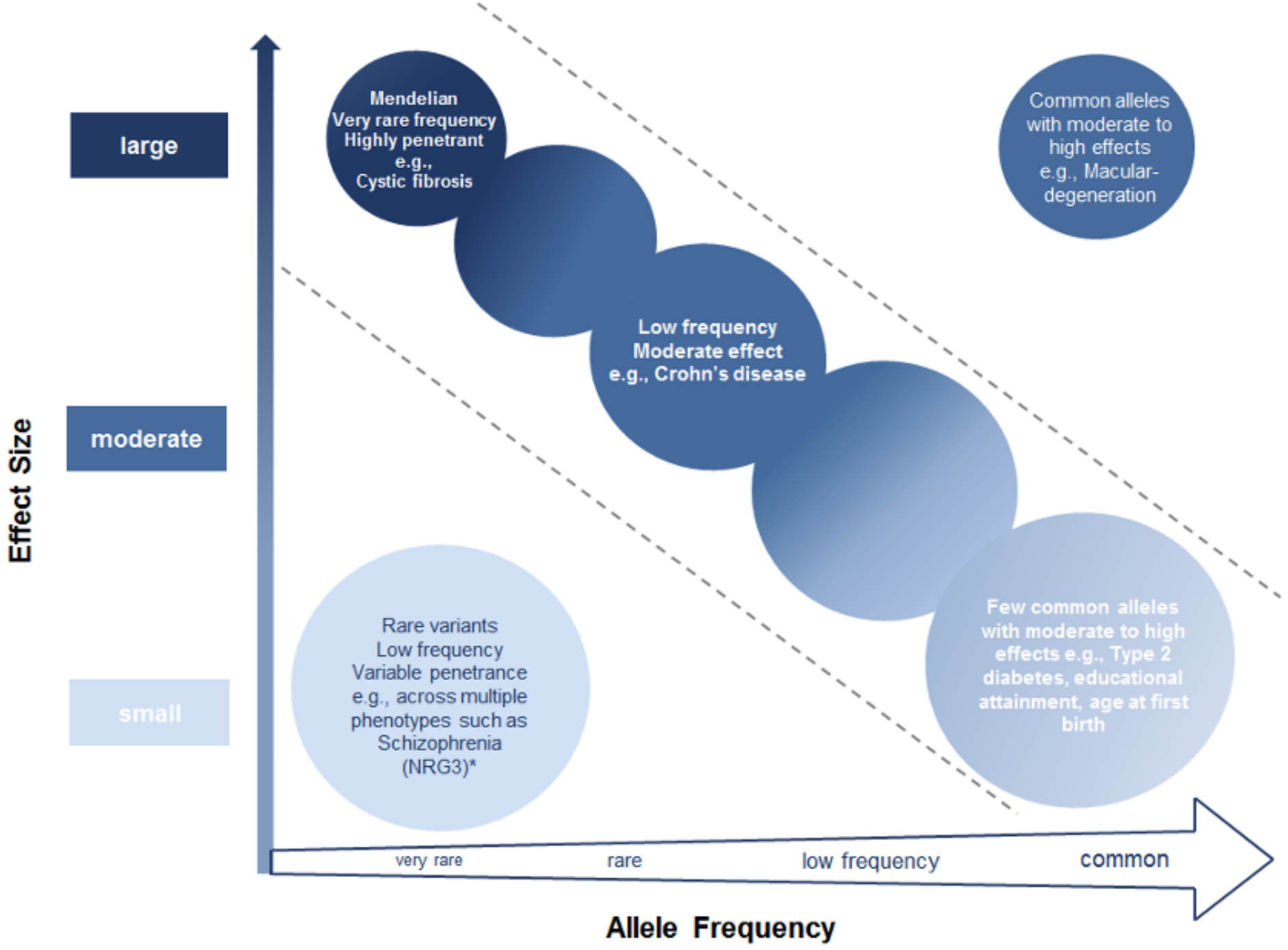
LD and genetic association

- The LD structure allows to identify “genomic regions” in association results.
- As it is not possible to genotype all genetic variants, we identify “markers” that can be in LD with the real causal variants.
-



Consequences of Imputing genotypes

- The imputed genotypes will be affected by imputation probabilities. Each allele will be characterised by a probability *Es. AA 90% AC 5% CC 5%*
- Only common variants can be imputed. Not rare variants (for those mutations necessary sequencing)
- Imputation is highly population-specific! Most of reference panels are based on European Ancestry
- Also, individuals with African Ancestry have higher genetic diversity. More difficult to impute


Effect Size
large
moderate
small
Mendelian
Very rare frequency
Highly penetrant
e.g.,
Cystic fibrosis
Low frequency
Moderate effect
e.g., Crohn's disease
Common alleles with moderate to high effects e.g., Macular-degeneration
Rare variants
Low frequency
Variable penetrance
e.g., across multiple phenotypes such as Schizophrenia (NRG3)*
Few common alleles with moderate to high effects e.g., Type 2 diabetes, educational attainment, age at first birth
very rare
rare
low frequency
common
Allele Frequency