Sociogenomics

Introduction

Nicola Barban

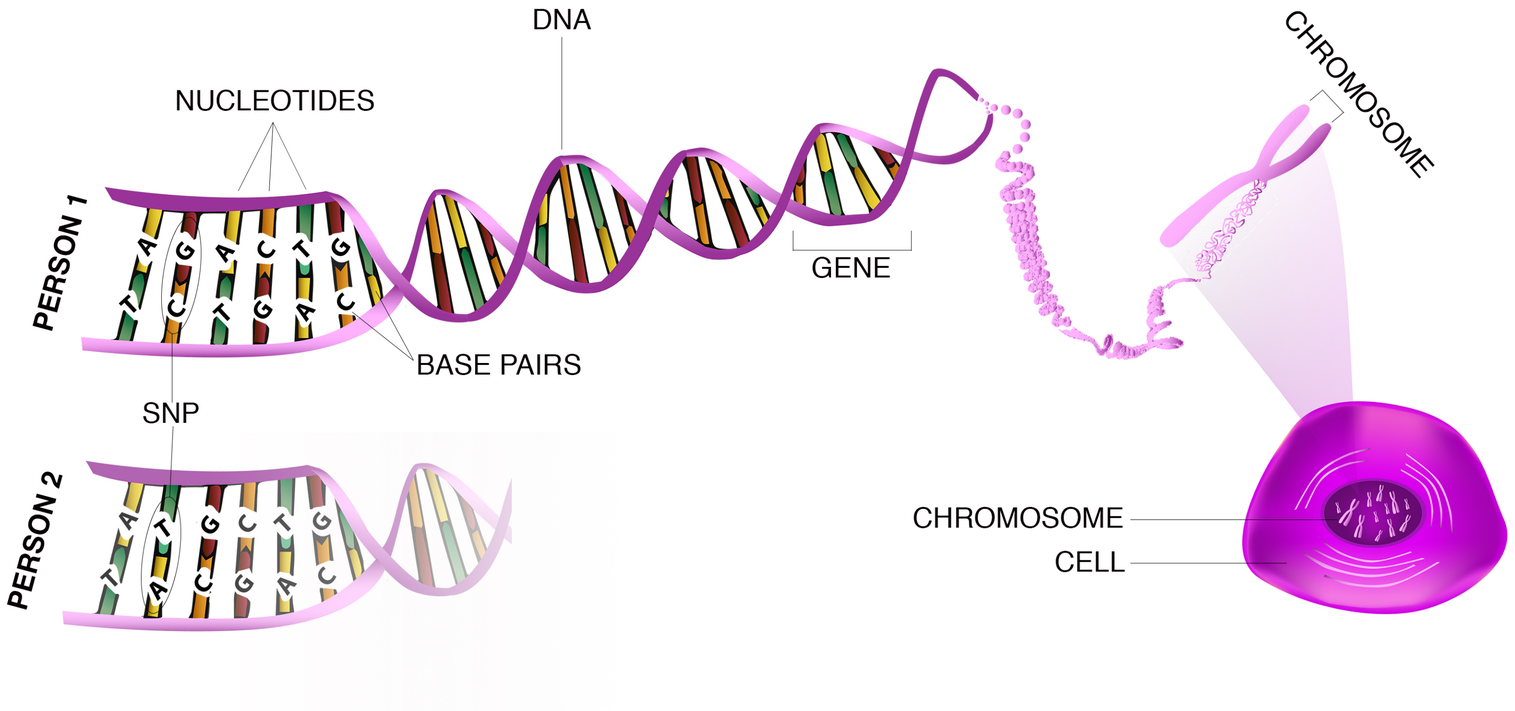
Outline

1. Why a course on “sociogenomics” ?
2. Learning objectives
3. Course Structure
4. Evaluation
5. Resources

1. Why a course on “sociogenomics” ?

The relevance of Human Genetics in the post-genomic era

Human Genetics is now relevant **beyond biology**, epidemiology and the medical sciences, with applications in psychology, psychiatric, statistics, demography, sociology and economics.



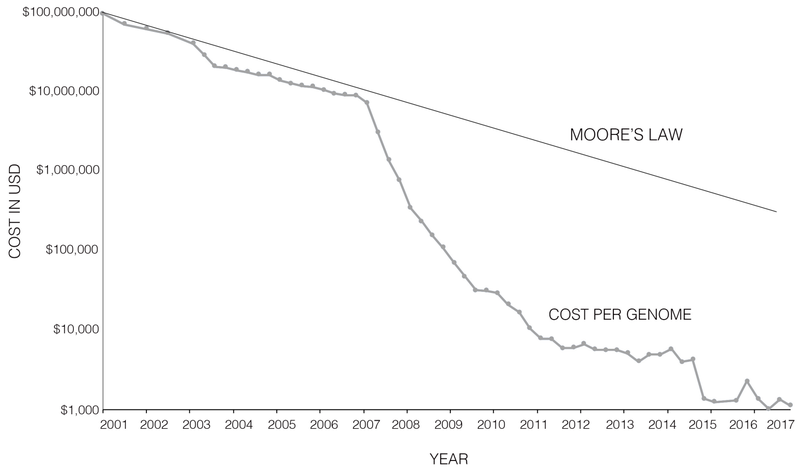
Give the most meaningful gift this season

50% Off

Health + Ancestry Kit

Shop now

Offer ends November 29.
Limit 3 kits.



Ancestry + Traits Service

~~$99~~ $79

SAVE $100

Health + Ancestry Service

~~$199~~ $99

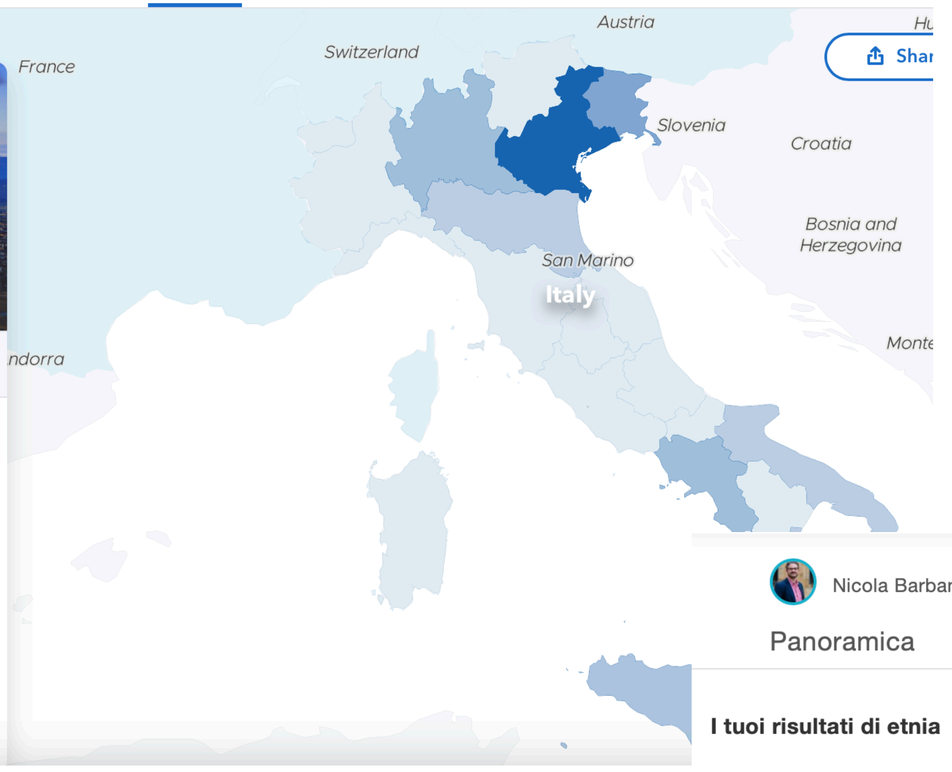
23andMe+ Membership

~~$199~~ $99 kit + ~~$29~~ $9.99 one year prepaid membership

View your ancestry timeline, parental inheritance, and ancestry chromosome painting in **Ancestry Overview**.

Ancestry Composition

Summary | Scientific Details | Frequently Asked Questions



Italy — Highly Likely Match

We did not detect enough evidence of recent ancestry from Malta.

LEARN MORE

Explore your Italian heritage

23andme.com

Nicola Barban, questo sei tu

Myheritage.com

Panoramica | Stima di Etnia | Corrispondenze DNA | Strumenti DNA

Esegui l'ana

Ottieni suggeri

I tuoi risultati di etnia

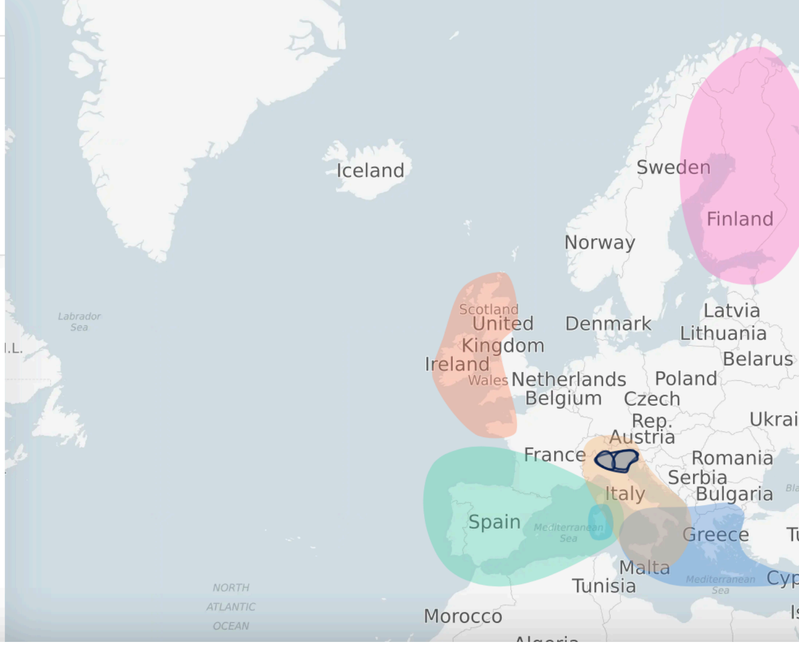
LIVELLO DI CONFIDENZA DEI GRUPPI GENETICI

Elevato — Basso

Visualizzati 3/3

EUROPA

Italiano	**35,6%**
Irlandese, scozzese e gallese	**20,5%**
Greco e italiano meridionale	**14,4%**
Sardo	**13,7%**



NEWS | SCIENCE AND POLICY

We will find you: DNA search used to nab Golden State Killer can home in on about 60% of white Americans

Researchers call for limiting how ancestry databases can be used to protect privacy

11 OCT 2018 • BY JOCELYN KAISER

If you're white, live in the United States, and a distant relative has uploaded their DNA to a public ancestry database, there's a good chance an internet sleuth can identify you from a DNA sample you left somewhere. That's the conclusion of a new study, which finds that by combining an anonymous DNA sample with some basic information such as someone's rough age, researchers could narrow that person's identity to fewer than 20 people by starting with a DNA database of 1.3 million individuals.

https://www.science.org/content/article/we-will-find-you-dna-search-used-nab-golden-state-killer-can-home-about-60-white

Trump’s ‘good genes’ speech echoes racial eugenics

BY GREGORY J. WALLANCE, OPINION CONTRIBUTOR - 09/25/20 8:00 AM ET

Getty Images

Most Popular

1 Breach at Air Force One base under investigation

2 Pentagon ‘did not detect’ previous Chinese spy balloons:...

3 Twitter suspends Sen. Steve Daines’s account

4 Texas governor unveils plan for statewide TikTok ban

https://twitter.com/atrupar/status/1307124621389463553

Belly of the Beast: California's dark history of forced sterilizations

Documentary tells story of state-sanctioned process in prisons as activists fight for a reparations bill

Photograph: Courtesy Belly of the Beast

A new documentary film is shedding light on forced sterilizations in California, reviving a dark chapter in the state’s history that is getting increased scrutiny amid a campaign to secure reparations for survivors.

Sociogenomics

In the social sciences there is a small, but rapidly growing, literature considering how genetic influences vary with institutions, historical contexts, gender, and other environments, sometimes placed under the umbrella terms sociogenomics or social science genetics

2.Learning Objectives (some examples)

1. Understand and work with molecular genetic data

Text PLINK files

***.ped**

FID	ID	F	M	S	P	-GENETIC INFO-							
CH18526	NA18526	0	0	2	1	G	G	C	C	T	T	A	A
CH18524	NA18524	0	0	1	1	G	G	C	C	T	T	A	A
CH18529	NA18529	0	0	2	1	C	G	C	C	T	T	C	A
CH18558	NA18558	0	0	1	1	G	G	C	C	G	T	A	A
CH18532	NA18532	0	0	2	1	G	G	C	C	T	T	A	A

***.map**

Chr	SNP	SNP Position	Base-Pair Coordinate
8	rs17121574	12.7991	12799052
8	rs754238	12.8481	12848056
8	rs11203962	12.8484	12848438
8	rs6999231	12.8623	12862253
8	rs17178729	12.867	12867001

Binary PLINK files

***.bed**

Binary version of the SNP info of the *.ped file which is only readable by your computer

***.fam**

FID	ID	F	M	S	P
CH18526	NA18526	0	0	2	1
CH18524	NA18524	0	0	1	1
CH18529	NA18529	0	0	2	1
CH18558	NA18558	0	0	1	1
CH18532	NA18532	0	0	2	1

***.bim**

Chr	SNP	SNP Position	Base-Pair Coordinate	Allele1	Allele2
8	rs17121574	12.7991	12799052	G	G
8	rs754238	12.8481	12848056	G	G
8	rs11203962	12.8484	12848438	C	G
8	rs6999231	12.8623	12862253	G	G
8	rs17178729	12.867	12867001	G	G

Covariates

FID	ID	Sex	Cohort	PC1	PC2	etc…
CH18526	NA18526	2	1	0.00542	-0.00876	
CH18524	NA18524	1	1	0.04517	-0.00761	
CH18529	NA18529	2	4	0.07776	-0.00231	

2. Fundaments of statistical genetics

how much of the variation in a trait is due to variation in genetic factors?

Heritability (From twin studies to molecular genetics)

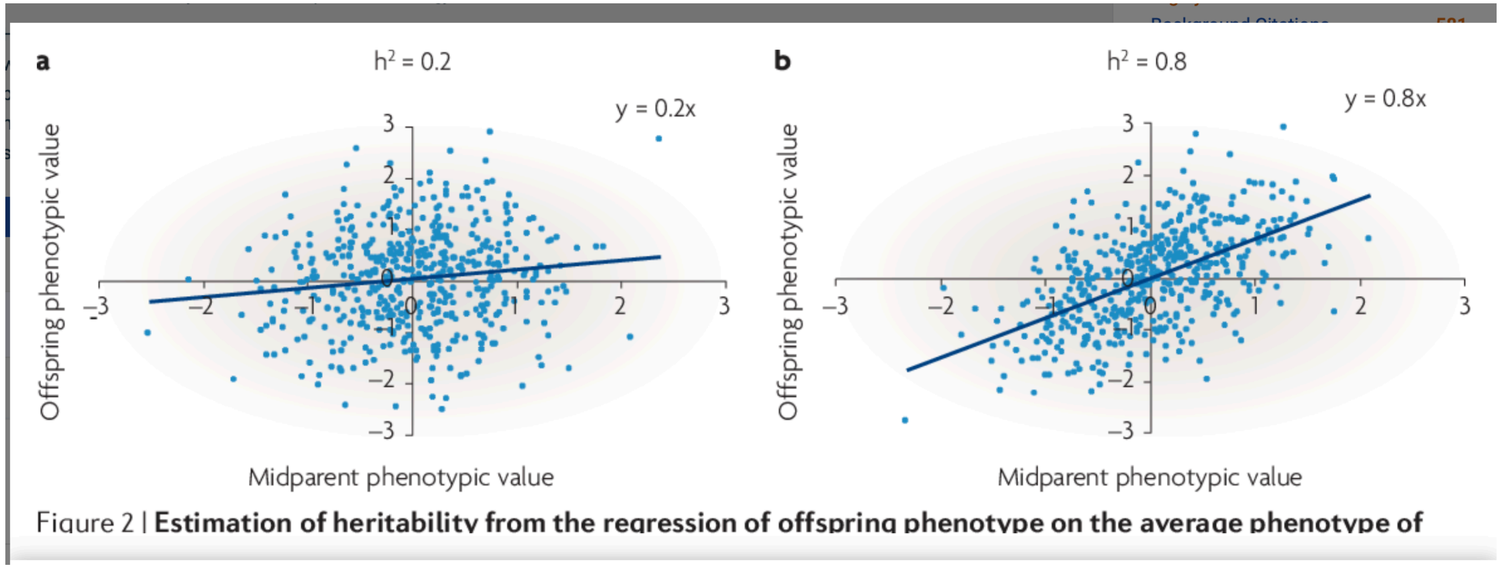
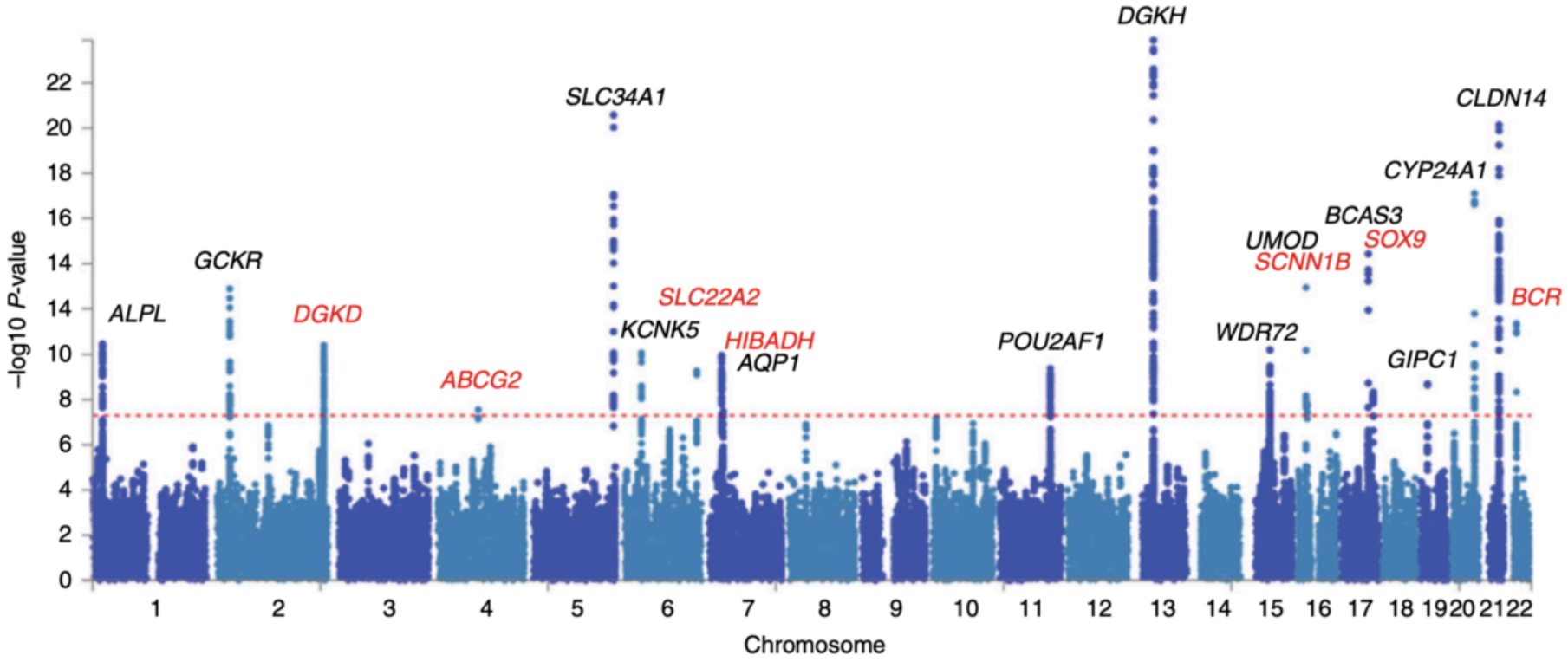


Figure 2 | **Estimation of heritability from the regression of offspring phenotype on the average phenotype of**

3. Genome Wide Association Studies

Associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases, traits, behaviours



4. Polygenic prediction

Can we use genetic information to "predict" a phenotype (trait)?

Genetic Weight

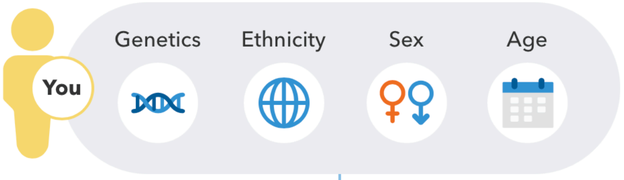
Your genes influence not just your weight, but also the impact of different healthy habits.

Overview | Scientific Details

How we determine your result

1. Collect some details from you.

You tell us your age, sex, height, weight, and ethnicity, so we can customize your result.



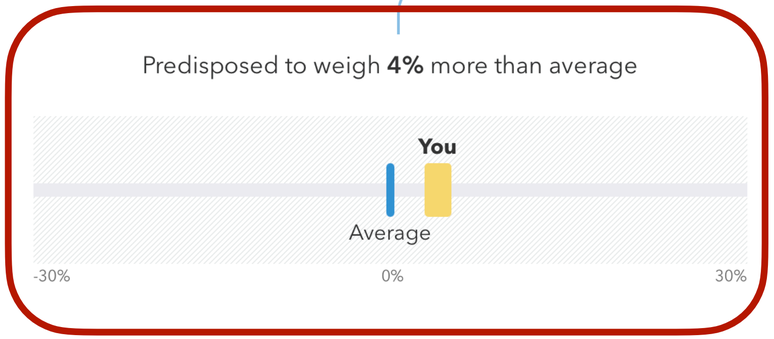
2. Calculate your score.

We use data from 23andMe research participants to create a genetic weight score based on your genotype at over 300 different genetic markers associated with weight. Based on your score, we then make a prediction about your BMI that also factors in your age, sex, and ethnicity.

Check out the details of our model

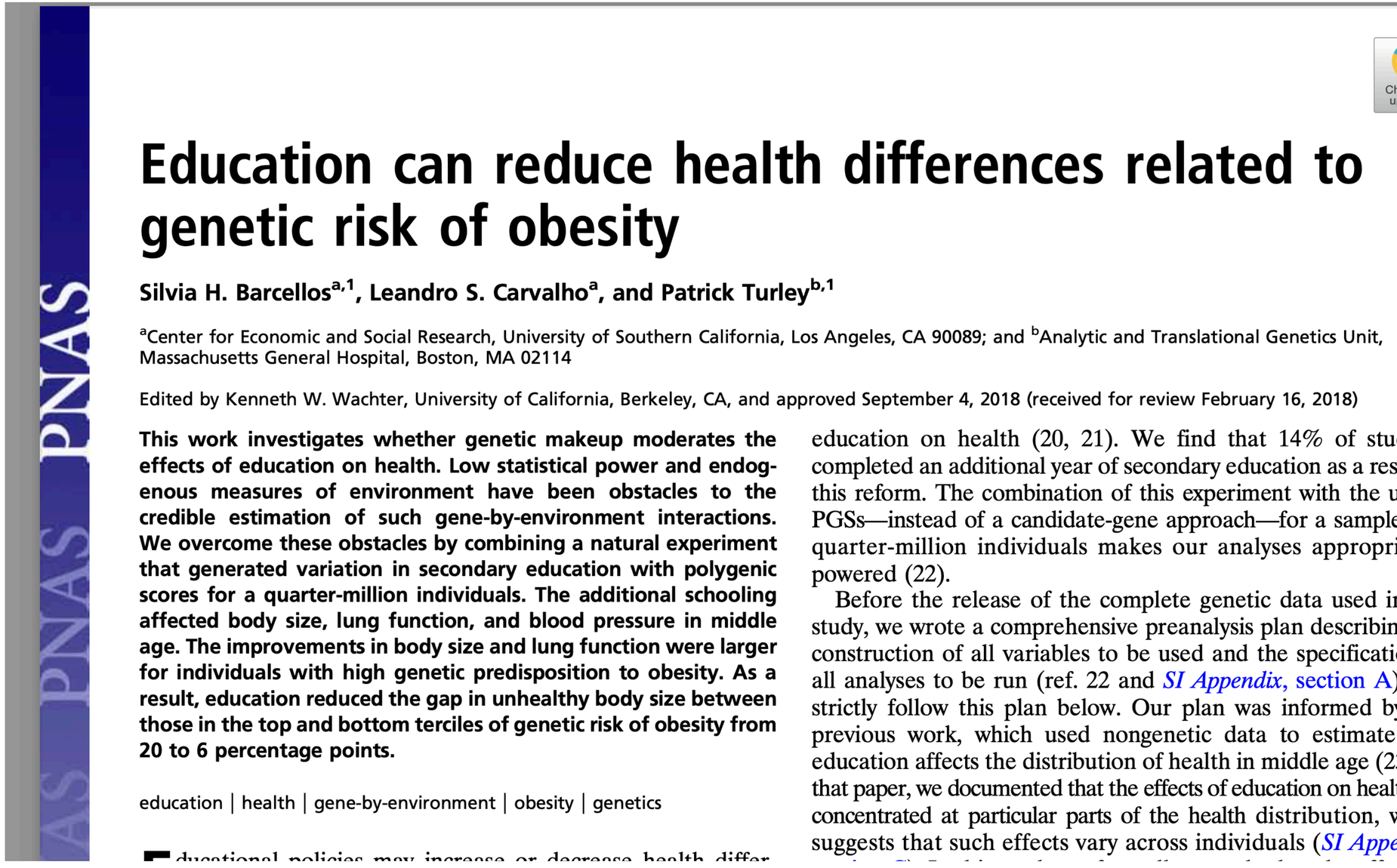
3. Summarize your weight predisposition.

To determine whether you have a genetic tendency to weigh more or less than average, we compare your BMI prediction to other 23andMe participants of your age, sex, and ethnicity. Because average weights change with age, how your predisposition compares to average may also change slightly over time. See our white paper about the science behind this report.



5. GxE interactions

…Educational policies may increase or decrease health differences, depending on whether they reinforce or counteract gene-related differences …



PNAS

Education can reduce health differences related to genetic risk of obesity

Silvia H. Barcellos[a,1], Leandro S. Carvalho[a], and Patrick Turley[b,1]

[a]Center for Economic and Social Research, University of Southern California, Los Angeles, CA 90089; and [b]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved September 4, 2018 (received for review February 16, 2018)

This work investigates whether genetic makeup moderates the effects of education on health. Low statistical power and endogenous measures of environment have been obstacles to the credible estimation of such gene-by-environment interactions. We overcome these obstacles by combining a natural experiment that generated variation in secondary education with polygenic scores for a quarter-million individuals. The additional schooling affected body size, lung function, and blood pressure in middle age. The improvements in body size and lung function were larger for individuals with high genetic predisposition to obesity. As a result, education reduced the gap in unhealthy body size between those in the top and bottom terciles of genetic risk of obesity from 20 to 6 percentage points.

education | health | gene-by-environment | obesity | genetics

Educational policies may increase or decrease health differ

education on health (20, 21). We find that 14% of stu
completed an additional year of secondary education as a res
this reform. The combination of this experiment with the u
PGSs—instead of a candidate-gene approach—for a sample
quarter-million individuals makes our analyses appropri
powered (22).

Before the release of the complete genetic data used i
study, we wrote a comprehensive preanalysis plan describin
construction of all variables to be used and the specificati
all analyses to be run (ref. 22 and *SI Appendix*, section A)
strictly follow this plan below. Our plan was informed by
previous work, which used nongenetic data to estimate
education affects the distribution of health in middle age (2
that paper, we documented that the effects of education on healt
concentrated at particular parts of the health distribution, w
suggests that such effects vary across individuals (*SI Appe*

6. Ethical Implications

- How to conduct responsible research?
- Who is going to benefit?
- Data privacy?
- Can genetic research lead to a more (un)equal society?
- Is there enough diversity in genetic research?
- What are the implication of genetic prediction?
- What are the risk of genetic editing?

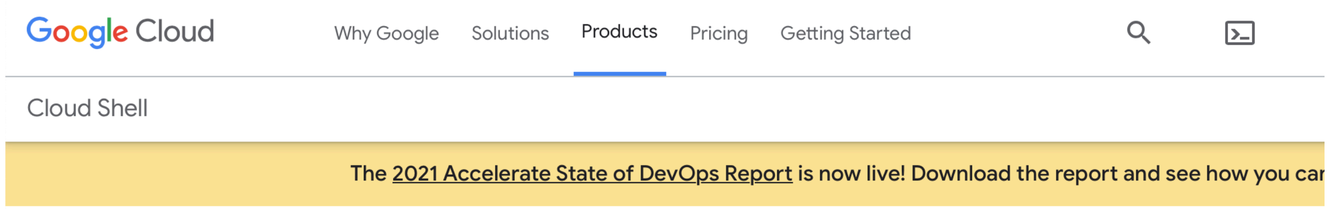
3. Course Structure

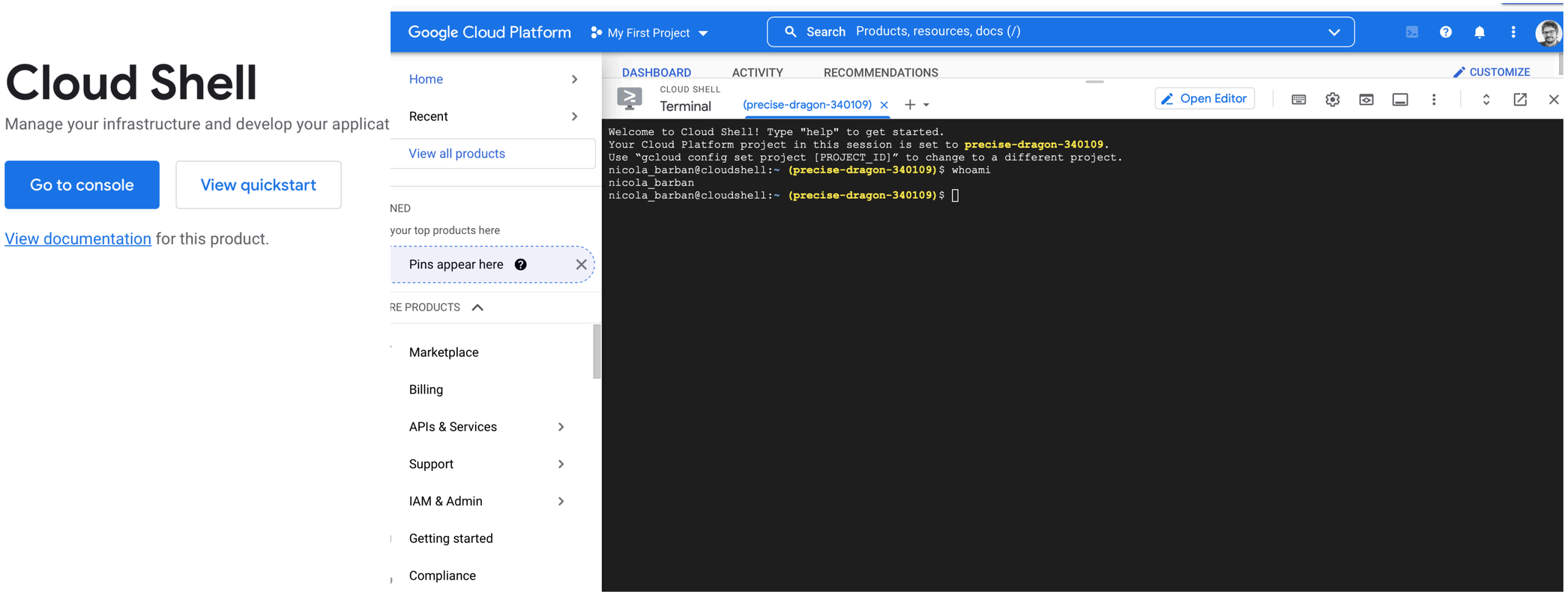
Structure of the course

1. Frontal Lectures (interactions with students)
2. Lab sessions
3. Students’ presentations
4. Invited lectures

Open the command-line interface

Opening: Online Cloud Shell





Data, syntax etc.

- In virtuale and here:

http://nicolabarban.com/sociogenomics2025/

4. Evaluation

Course evaluation a.k.a exam

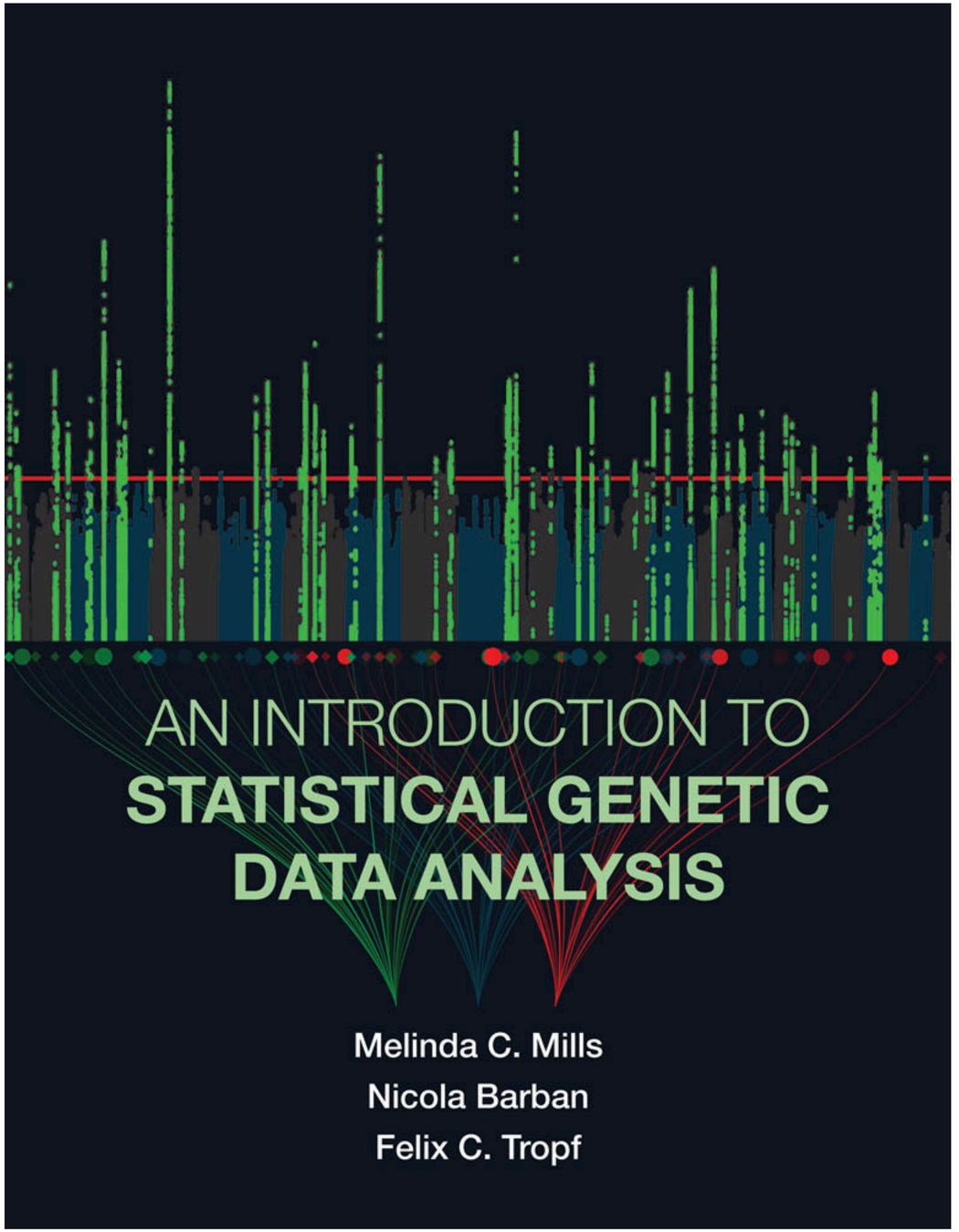
For students attending class regularly, the final evaluation will be composed by two parts:

1. Group or individual presentations (max 3 people) **(30% of the final grade)**
2. Group or individual project/assignment (max 3 people). Instructions on the project will be distributed in class. **(30% of the final grade)**
3. Oral examination **(40% of the final grade)**

Students not attending class are invited to contact the instructor to discuss the examination.

4. Resources

Textbook

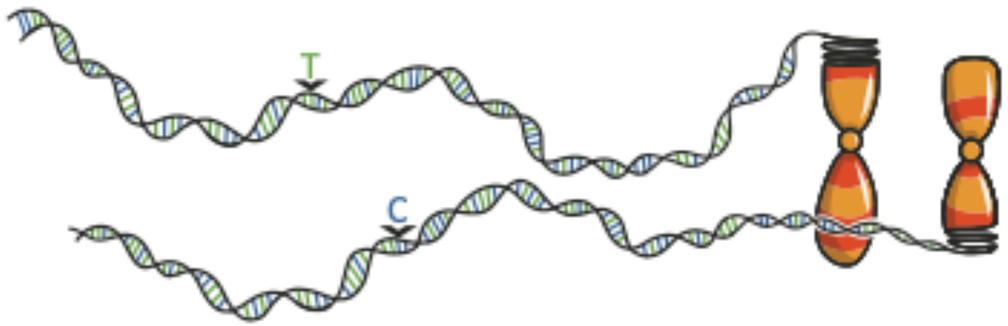

AN INTRODUCTION TO
STATISTICAL GENETIC
DATA ANALYSIS
Melinda C. Mills
Nicola Barban
Felix C. Tropf

An Owner's Guide to the Human Genome

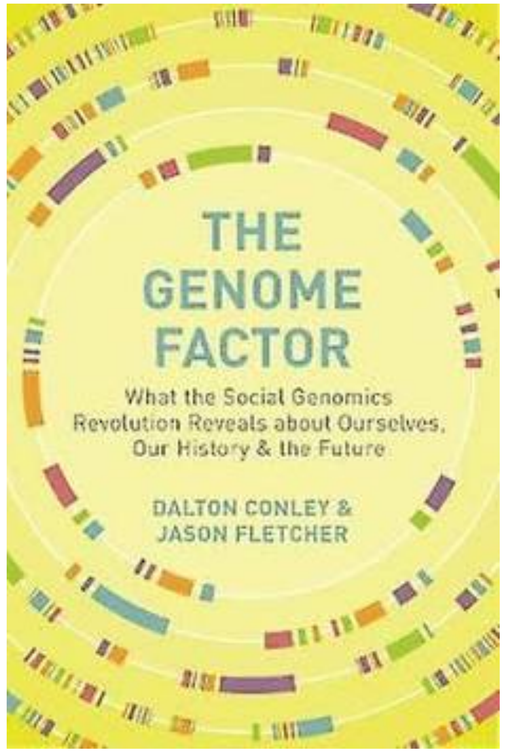
An introduction to human population genetics, variation, and disease

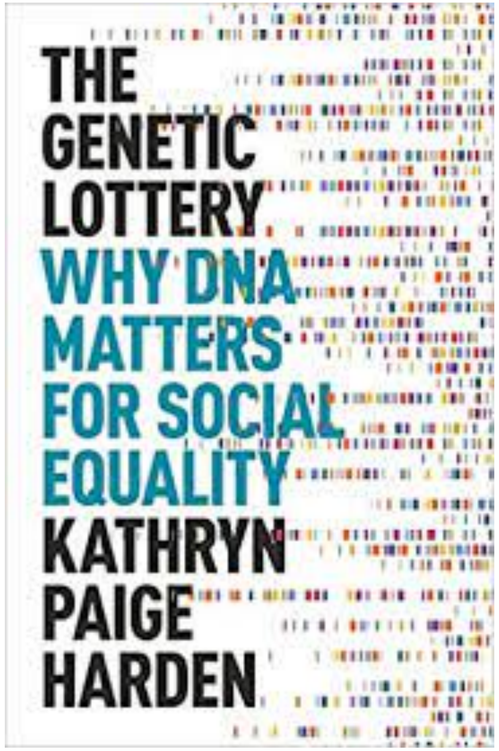
Jonathan K Pritchard
Stanford University
Release 1.0: September 30, 2023

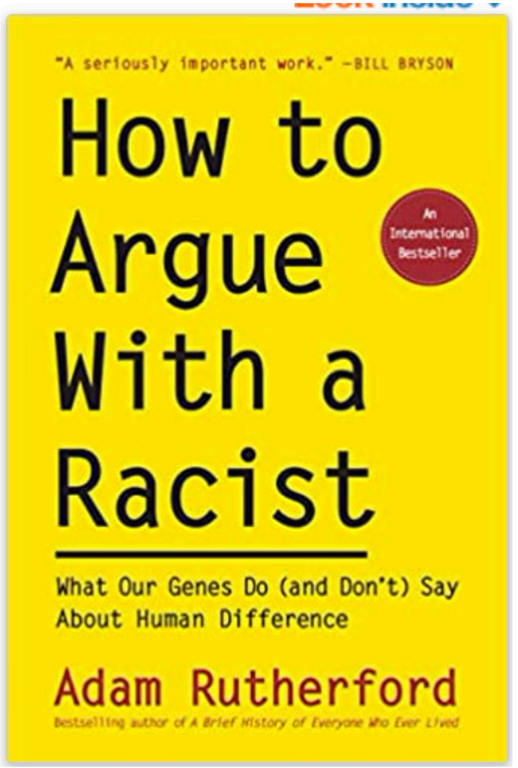
https://web.stanford.edu/group/pritchardlab/HGbook.html

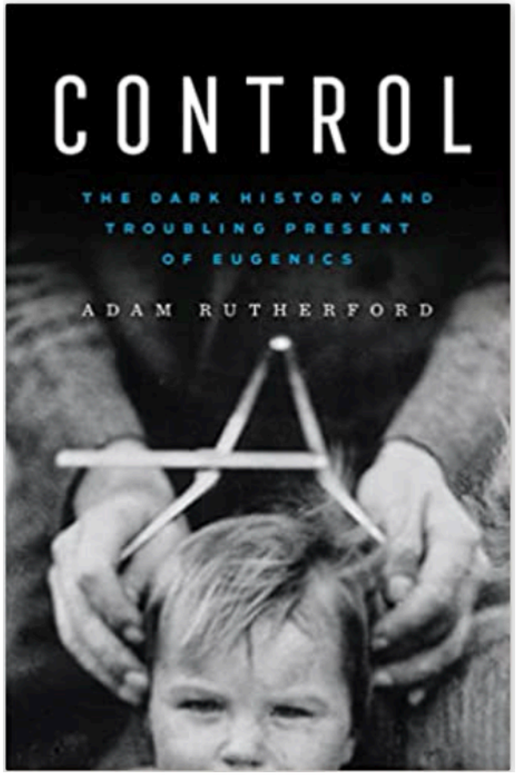


Addional Textbooks (Optional) + Assigned readings during class (not optional)









Additional Material

https://www.bbc.co.uk/sounds/brand/m001fd39

Sociogenomics

History of genetics

Nicola Barban

Where do similarities and differences between living organisms come from?

Genetics in ancient times

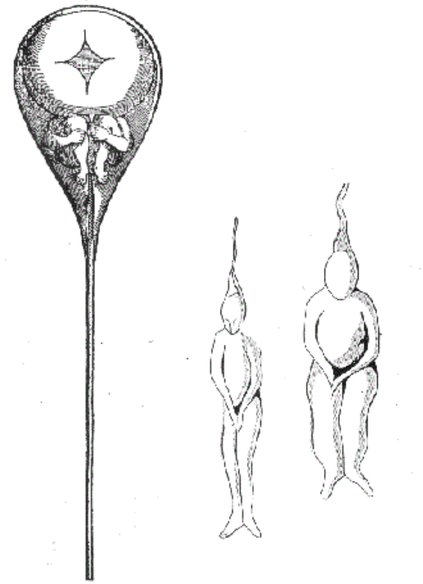
- Since the dawn of civilization, humankind has recognized the influence of heredity and applied its principles to the improvement of cultivated crops and domestic animals.

- Hippocrates (c. 460–c. 375 BCE), believed in the inheritance of acquired characteristics, and postulated that all organs of the body of **a parent gave off invisible “seeds,” which were like miniaturized building components and were transmitted during sexual intercourse**, reassembling themselves in the mother’s womb to form a baby.

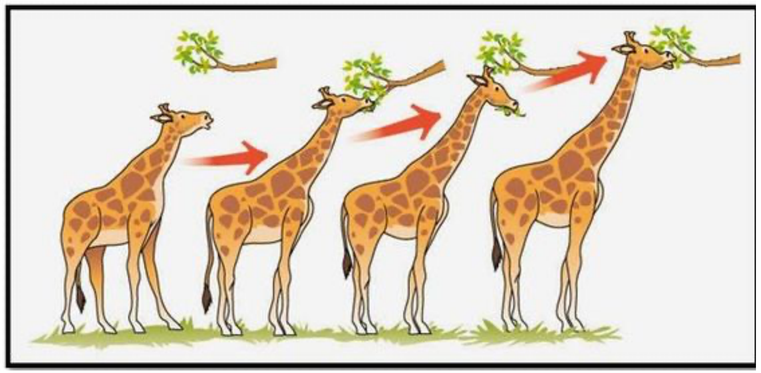
- Aristotle (384–322 BCE) emphasized the importance of blood in heredity. He believed that the **male’s semen was purified blood and that a woman’s menstrual blood was her equivalent of semen. These male and female contributions united in the womb to produce a baby.** The blood contained some type of hereditary essences, but he believed that the baby would develop under the influence of these essences, rather than being built from the essences themselves.

Genetics during 17-18th century

During the 1600s, Dutch microscopist Anton van Leeuwenhoek (1632-1723) discovered "animalcules" in the sperm of humans and other animals. **Some scientists speculated they saw a "little man" (homunculus) inside each sperm.**



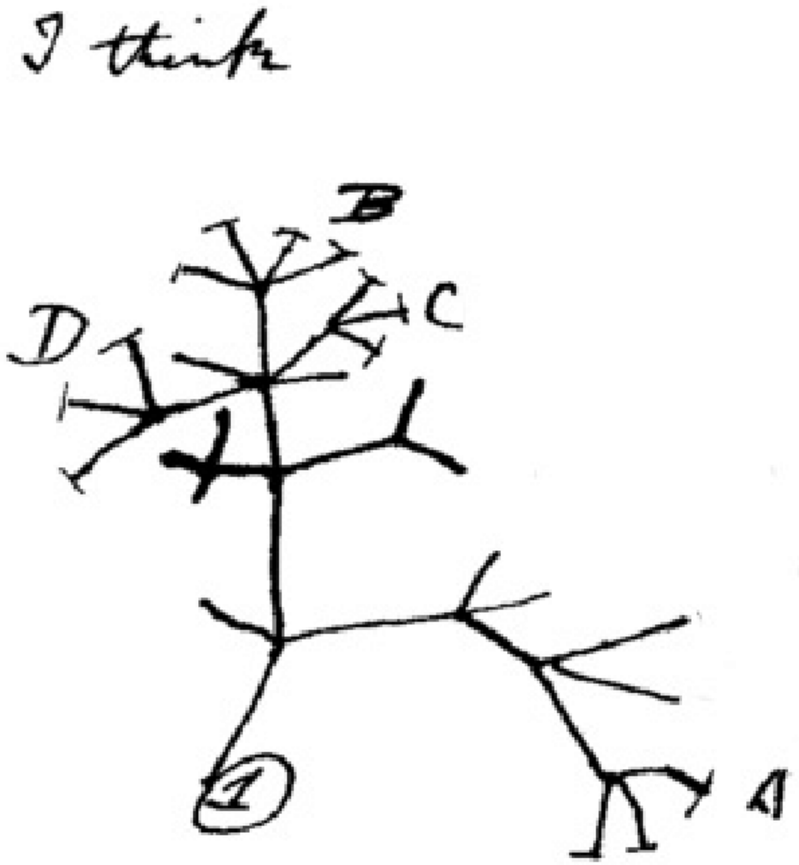
Jean-Baptiste Lamarck (1744-1829) invoked the idea of “the inheritance of acquired characters,” not as an explanation for heredity but as a model for evolution. **He lived at a time when the fixity of species was taken for granted**, yet he maintained that this fixity was only found in a constant environment.



1859 Darwin publishes “On the origin of Species”

Theory of evolution: all species arose through the natural selection of small, inherited variations that increase the individual's ability to compete, survive, and reproduce.

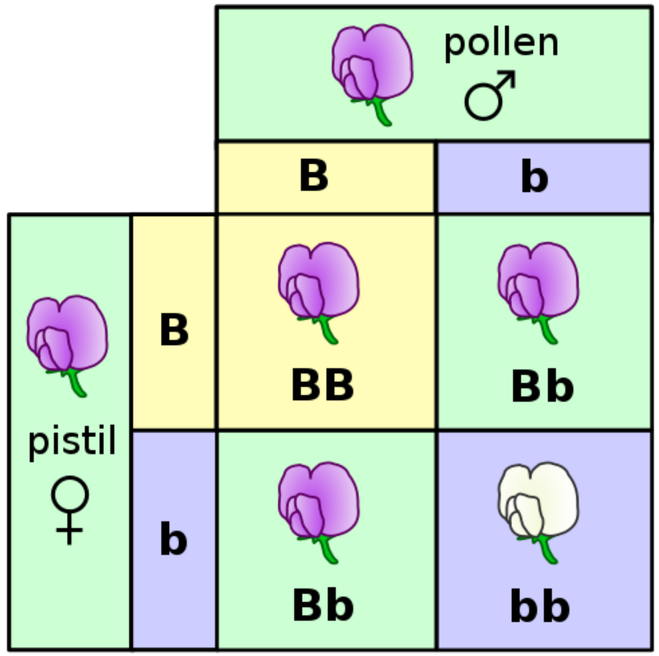
Darwin’s observations during his circumnavigation of the globe aboard the HMS *Beagle* (1831–36) provided evidence for natural selection and his suggestion that humans and animals shared a common ancestry.



Mendel’s laws of inheritance

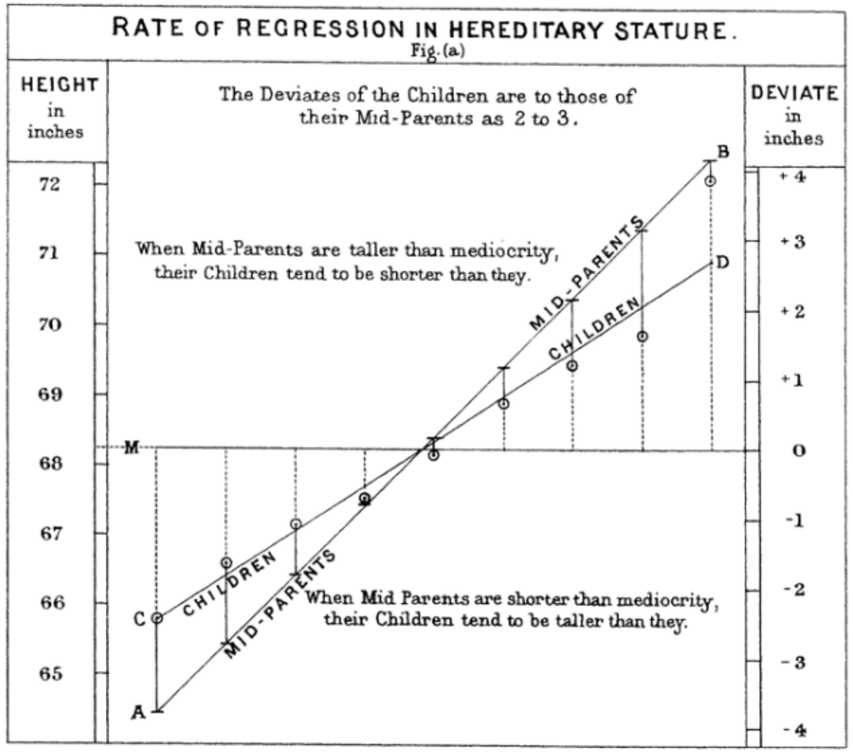
Published in 1866, but largely ignored until early 1900

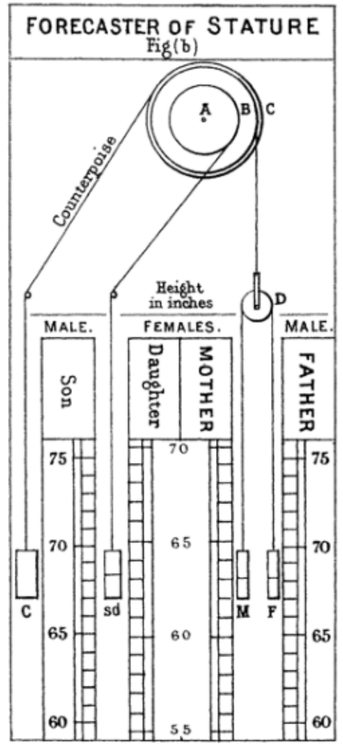
Law	Definition
Law of dominance and uniformity	**Some alleles are dominant while others are recessive**; an organism with at least one dominant allele will display the effect of the dominant allele.
Law of segregation	During gamete formation, **the alleles for each gene segregate from each other** so that each gamete carries only one allele for each gene.
Law of independent assortment	**Genes of different traits can segregate independently** during the formation of gametes.



Galton and quantitative genetics

- Francis Galton (1889) was interested in the transmission of those characteristics that presented a continuous variation



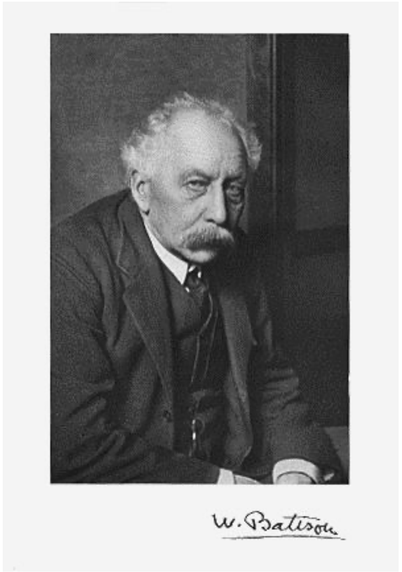


Mendelians and biometricians

In the beginning of the 1900s, controversy arose between:

- Mendelians: **qualitative traits** that show Mendelian patterns of inheritance. Quantitative traits reflect environmental differences.

- Biometricians: **quantitative traits that are normally distributed.** They doubted that mendelians laws could apply to continuous traits



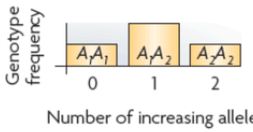
William Bateson 1861-1923

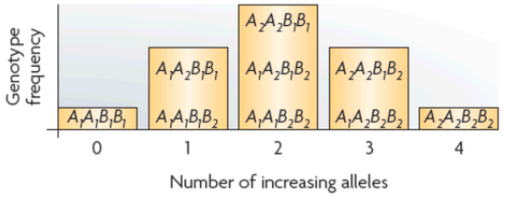
Karl Pearson 1857-1936

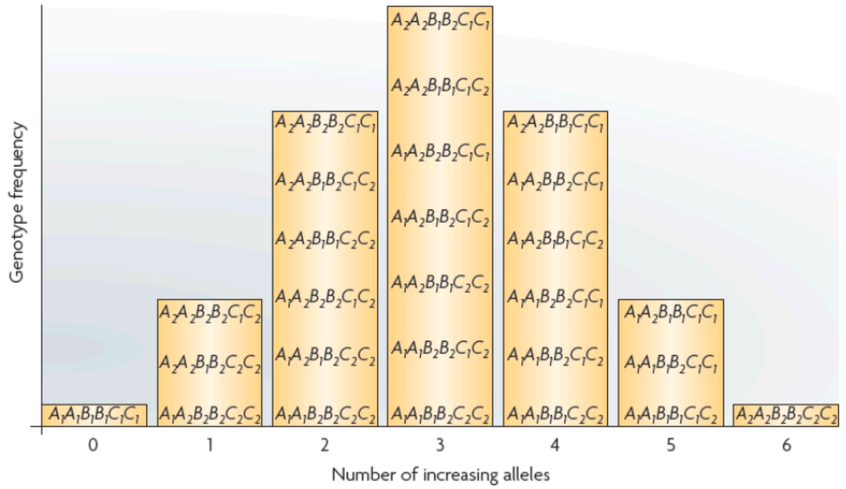
The infinitesimal model

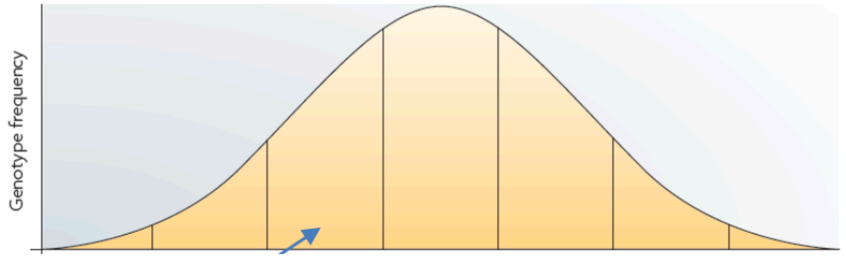
Ronald Fisher (1918)

The **infinitesimal model**, also known as the **polygenic model**, is based on the idea that variation in a **quantitative trait** is influenced by an infinitely large number of **genes**, each of which makes an infinitely small (infinitesimal) contribution to the phenotype, as well as by environmental factors.







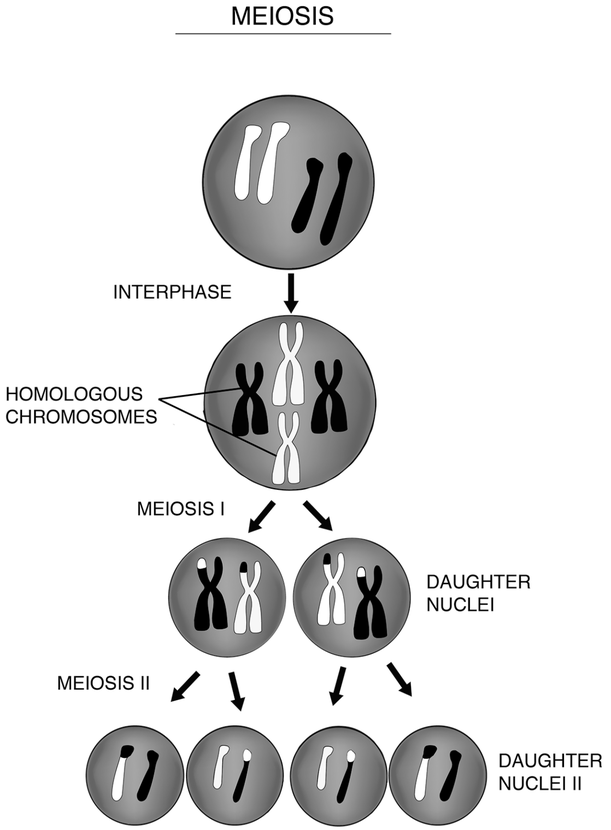


Most Human Traits (phenotypes) are polygenic! Meaning that a very large number of genetic variants contributes in explaining phenotypic variation

Genes and heredity

Mendel's genes were only hypothetical entities, factors that could be inferred to exist in order to explain his results

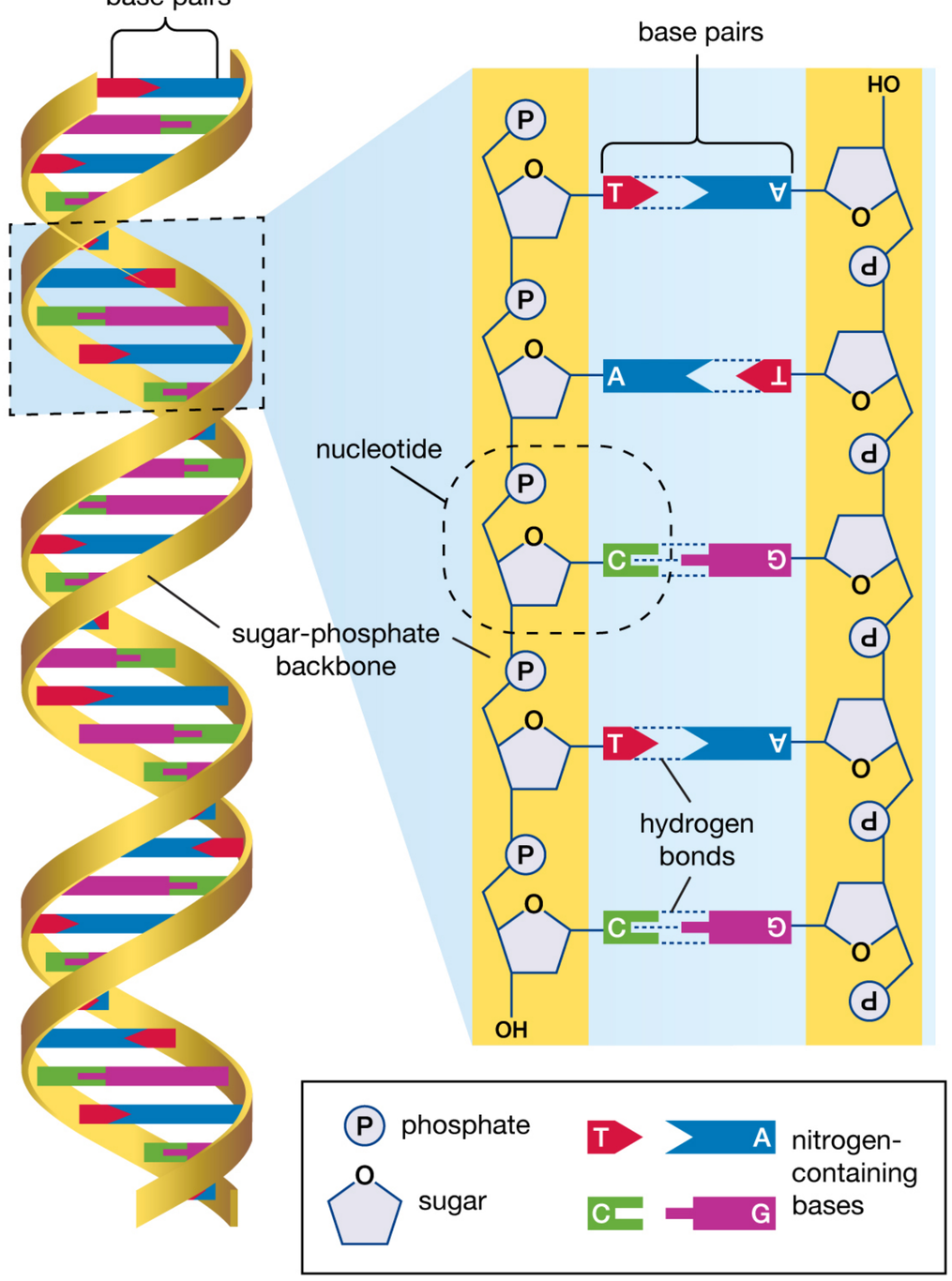
It seemed that genes were parts of chromosomes. In 1910 this idea was strengthened through the demonstration of parallel inheritance of certain Drosophila (a type of fruit fly) genes on sex-determining chromosomes by American zoologist and geneticist Thomas Hunt Morgan.



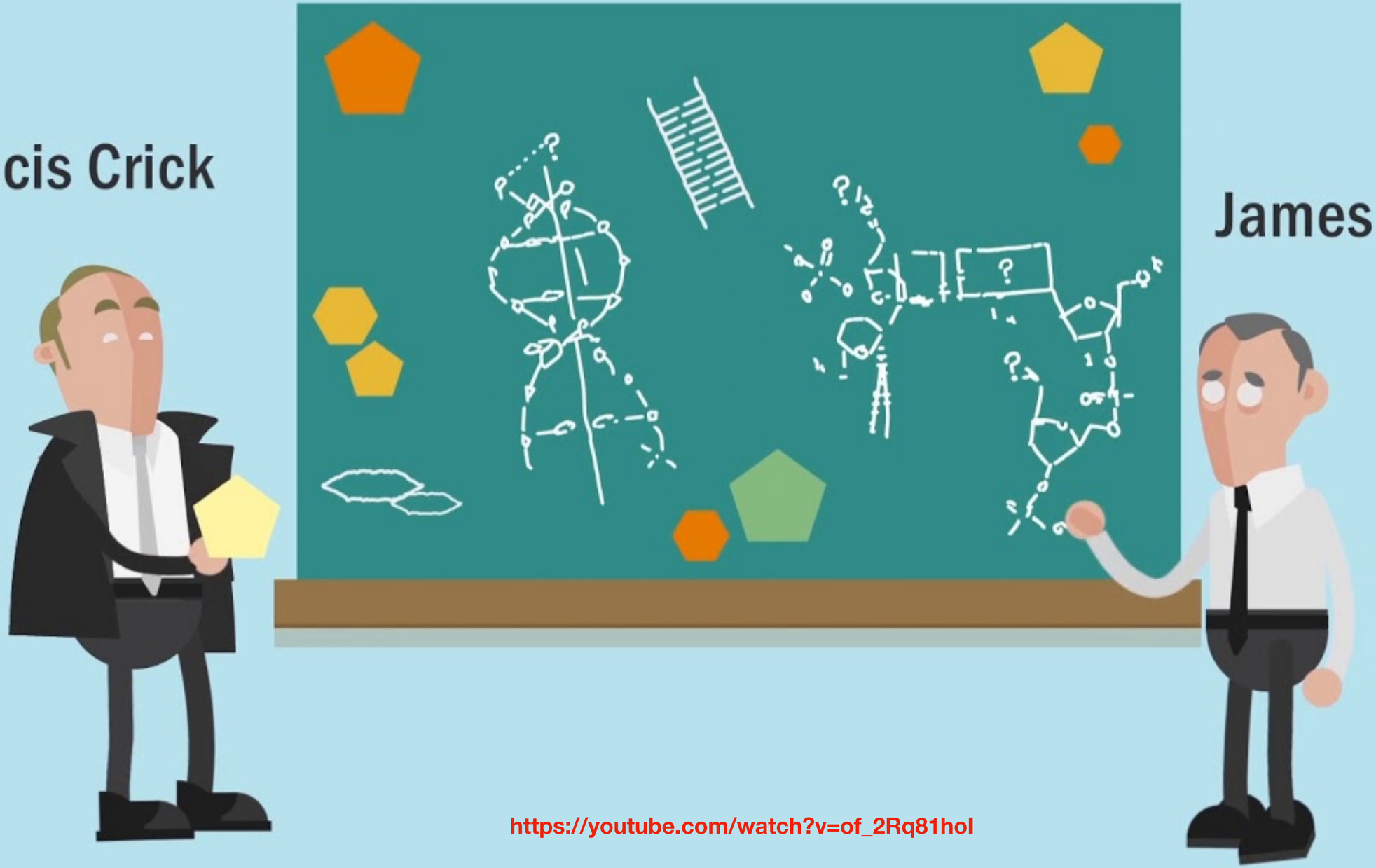
The discovery of DNA structure

A major landmark was attained in **1953** when James D. Watson and Francis Crick and Maurice Wilkins devised a double helix model for DNA structure.

Their breakthrough was made possible by the work of Rosalind Franklin, whose X-ray diffraction studies of the DNA molecule shed light on its helical structure.



Francis Crick



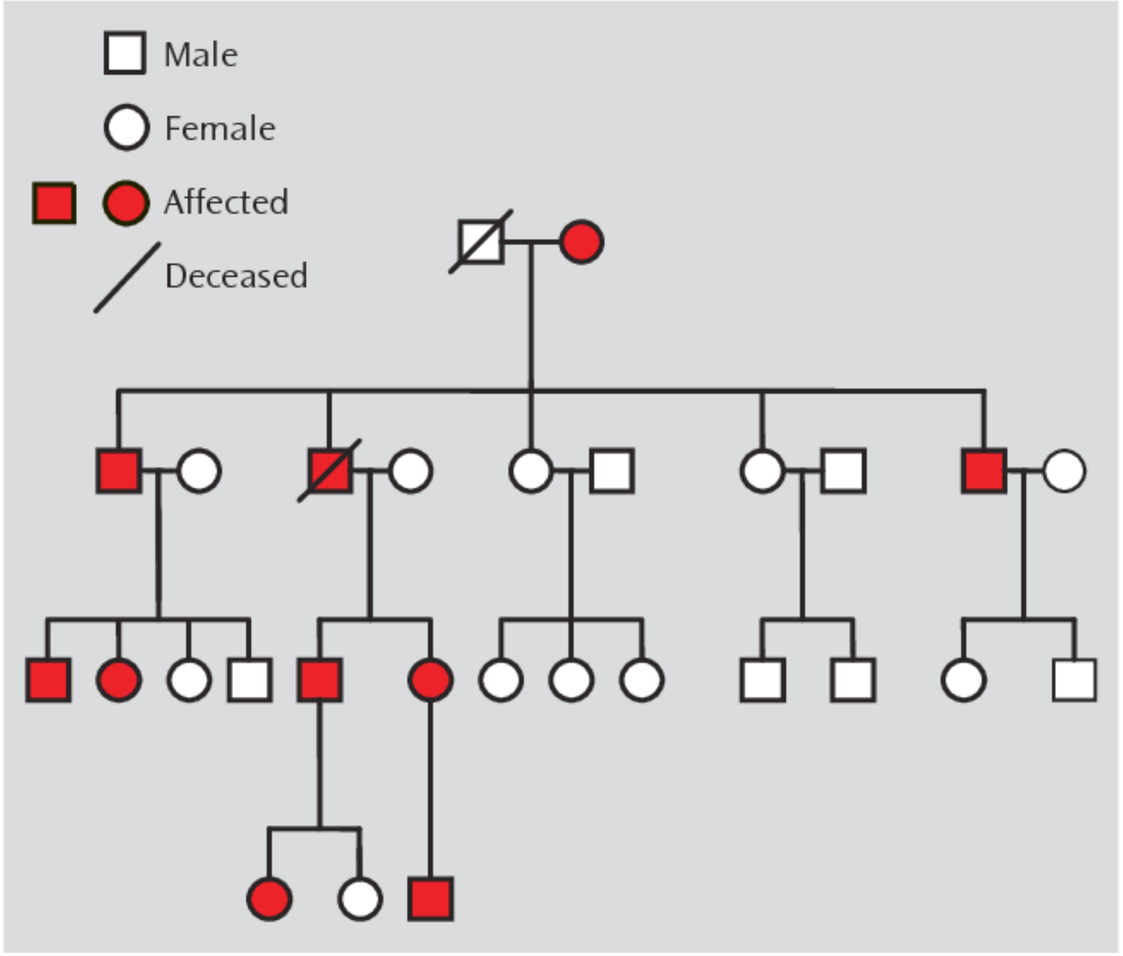
James D. Watson

https://youtube.com/watch?v=of_2Rq81hoI

Linkage Studies

A gene-hunting technique that traces patterns of disease in high-risk families. It attempts to locate a disease-causing gene by identifying genetic markers of known chromosomal location that are co-inherited with the trait of interest.

suitable for large effects: genes were found for many single gene disorders (Mendelian traits)

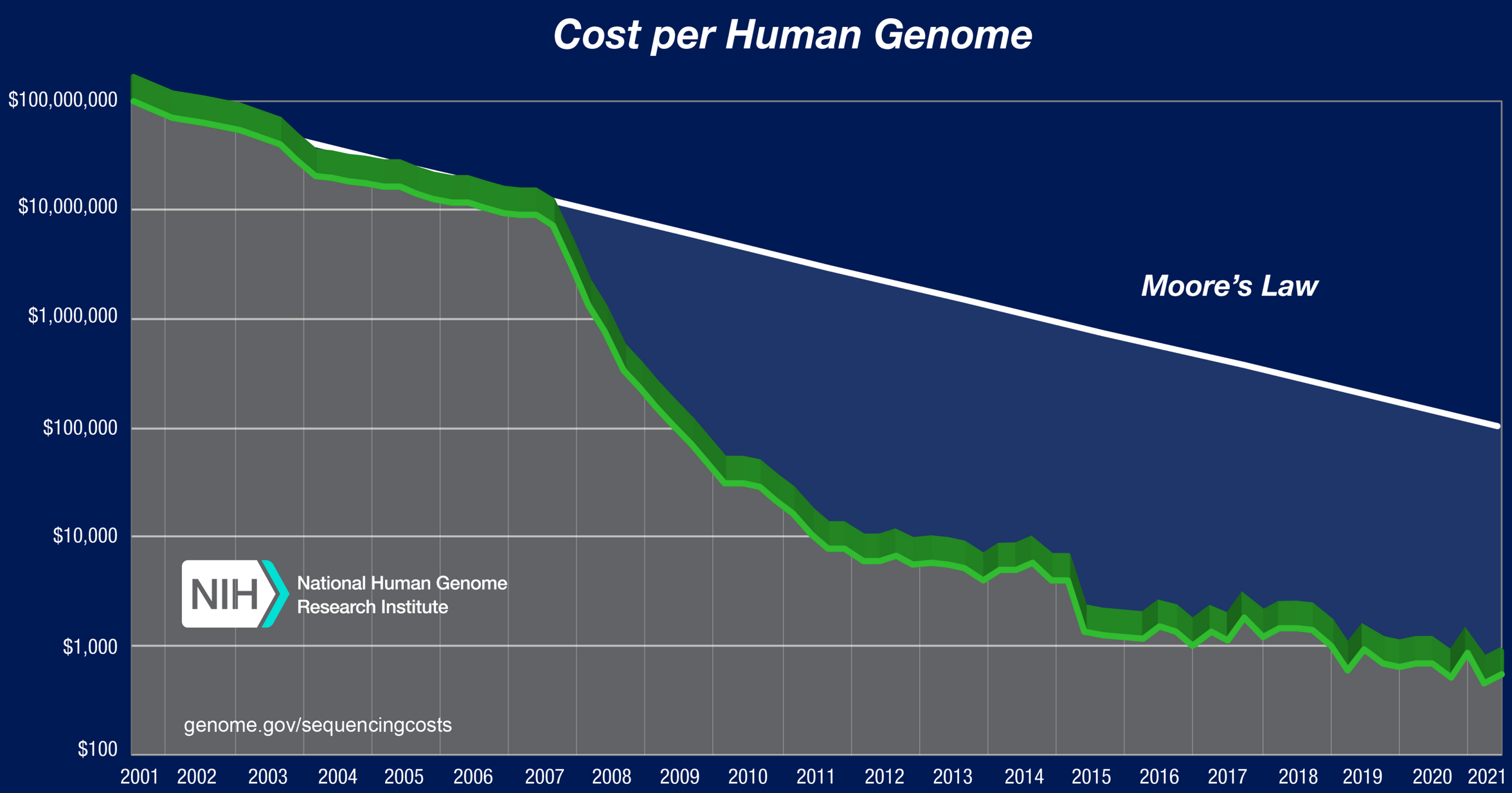


The Human Genome Project

The Human Genome Project (HGP), which operated from 1990 to 2003, provided researchers with basic information about the sequences of the three billion chemical base pairs (i.e., adenine [A], thymine [T], guanine [G], and cytosine [C]) that make up human genomic DNA (deoxyribonucleic acid).

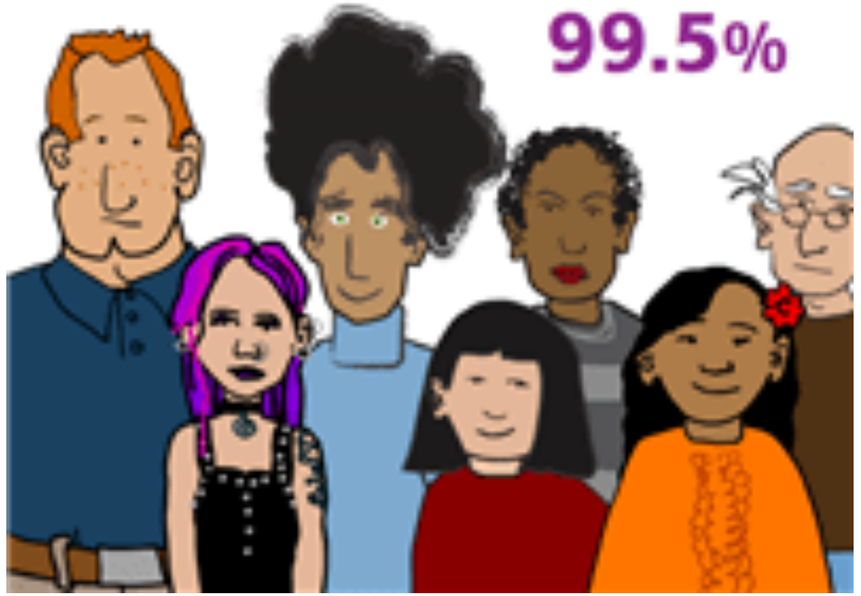
Announced on June 26, 2000 by Bill Clinton in a joint announcement with Tony Blair

Once significant human genome sequencing began for the HGP, a 'draft' human genome sequence (as described above) was produced over a 15-month period (from April 1999 to June 2000). The estimated cost for generating that initial 'draft' human genome sequence is ~**$300 million** worldwide.


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Human Genetic Variation

- Human Genome consists of approx 3 billion base pairs
- 99.5% similarity among individuals
- Only about 1.5% of the genome codes for proteins
- Each person has the same set of genes - about 20,000 in all. The differences between people come from slight variations in these genes



Mapping genomic variations

Cells sometimes make mistakes during the copying process. These typos lead to variations in the DNA sequence at particular locations, called single nucleotide polymorphisms, or SNPs (pronounced “snips”).

Need to build a map of genetic variation

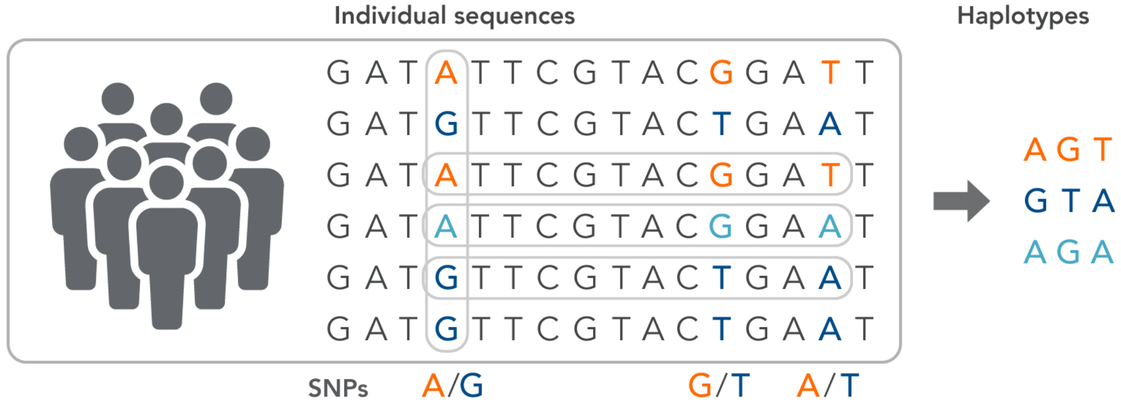
The HapMap Project

The **International HapMap Project** was an organization that aimed to develop a haplotype map (**HapMap**) of the human genome

The HapMap project focuses only on common SNPs, those where each allele occurs in at least 1% of the population.

The complete data obtained in Phase I were published on 27 October 2005

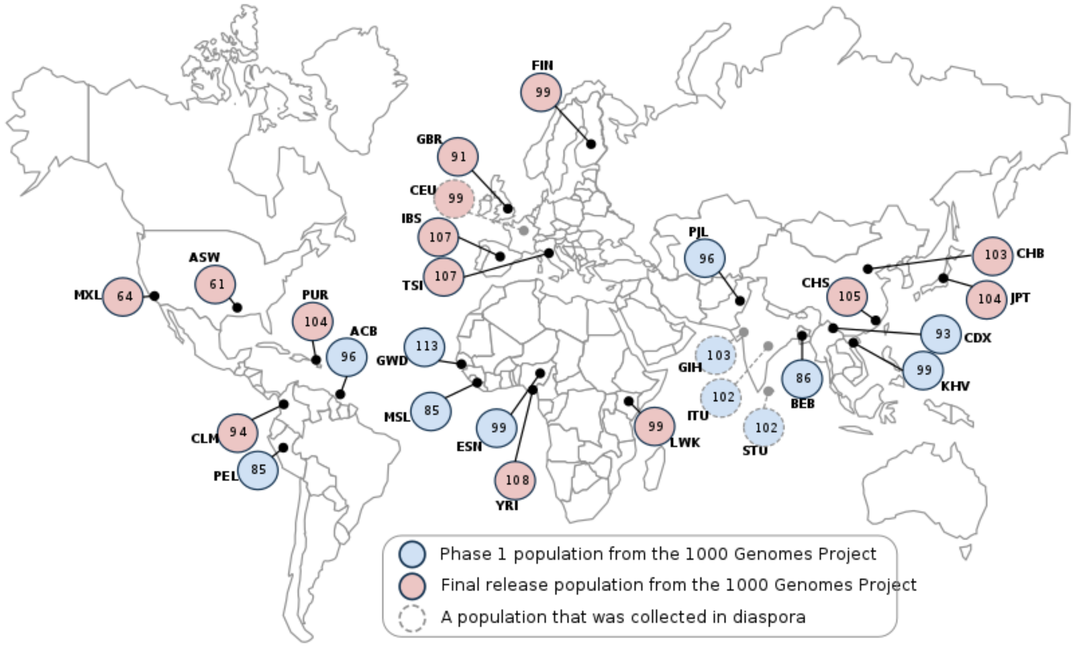
The HapMap project chose a sample of 269 individuals and selected several million well-defined SNPs, genotyped the individuals for these SNPs, and published the results.



The 1000 Genome Project

Launched in 2008, to extend the HapMap

goal to create a complete and **detailed catalogue of human genetic variations**



Genome Wide association Studies

- Data from HapMap and 1000Genomes help scientists to go from linkage to association studies
- A Genome Wide Association Study is the “gold standard” techniques to find association between a genetic variant and a trait/disease

- https://www.ebi.ac.uk/gwas/diagram

What’s next?

- Understanding interaction between **environment and genetics**
- **Genetic editing (CRISPR)**
- **Epigenetic**
- **Precision medicine**
- **Genomic privacy**

Sociogenomics

Basic concepts

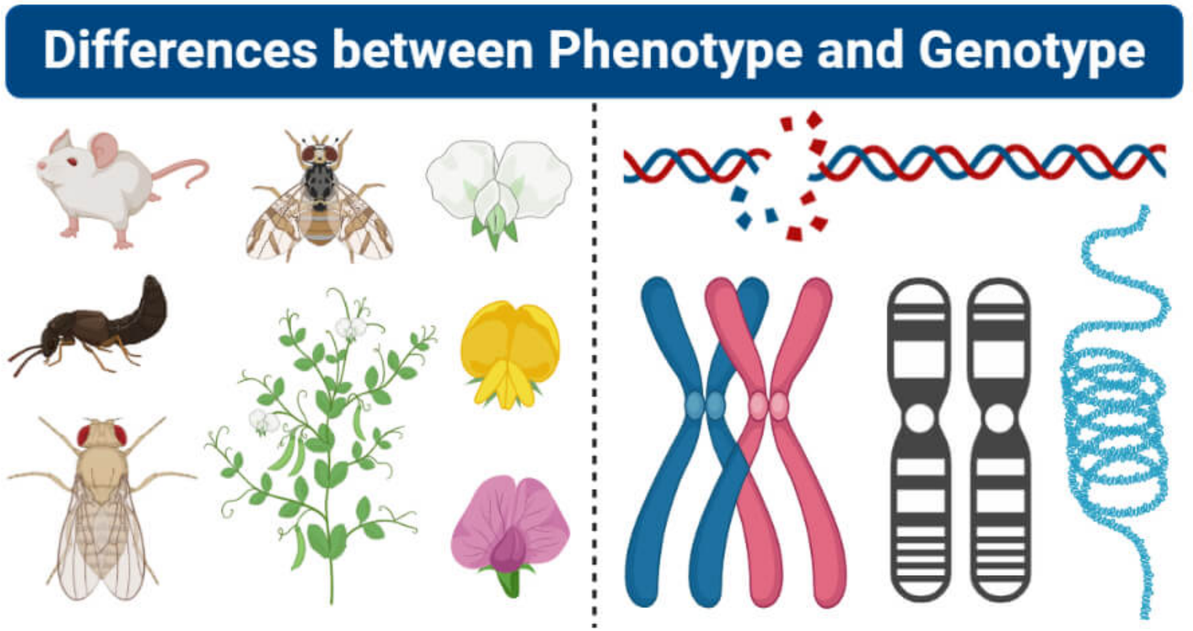
Nicola Barban

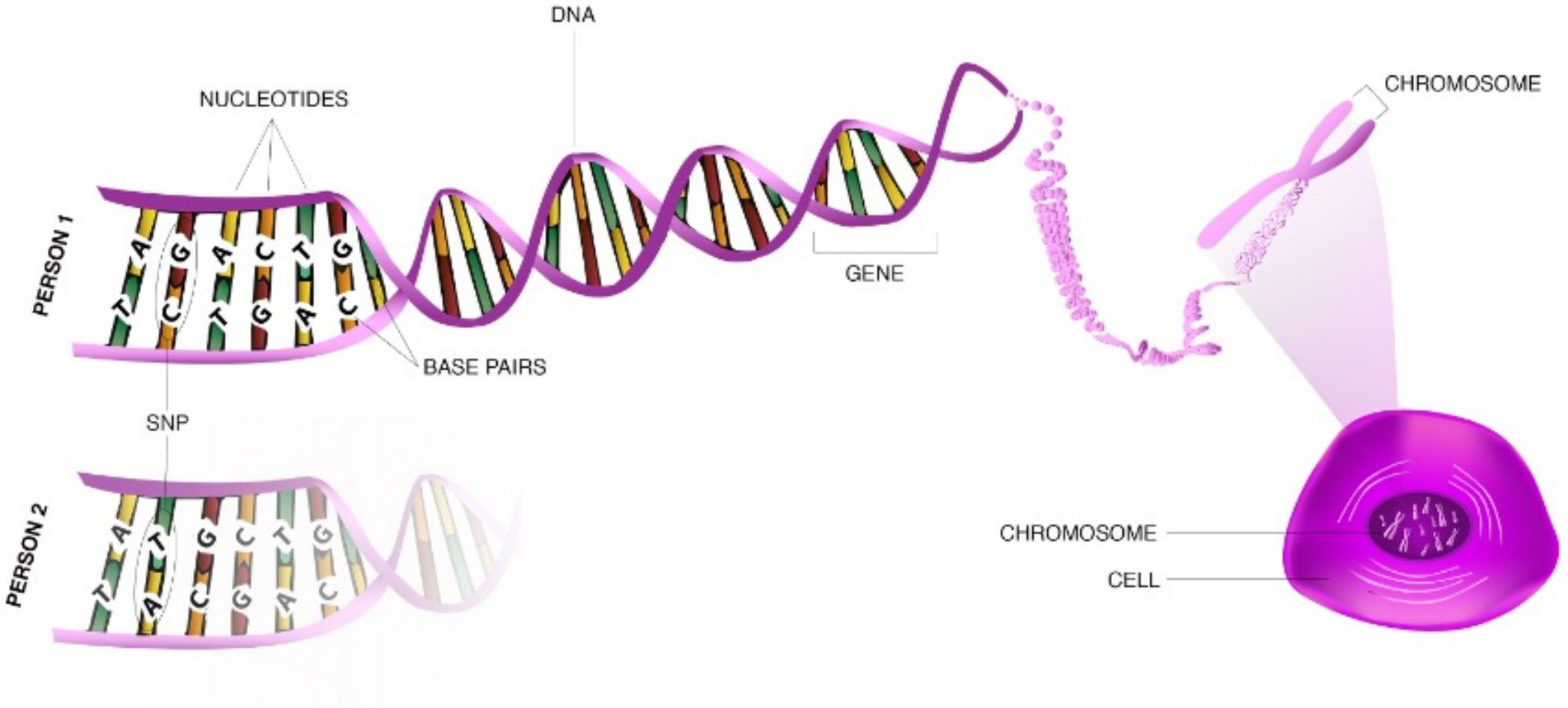
Basic definitions

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).

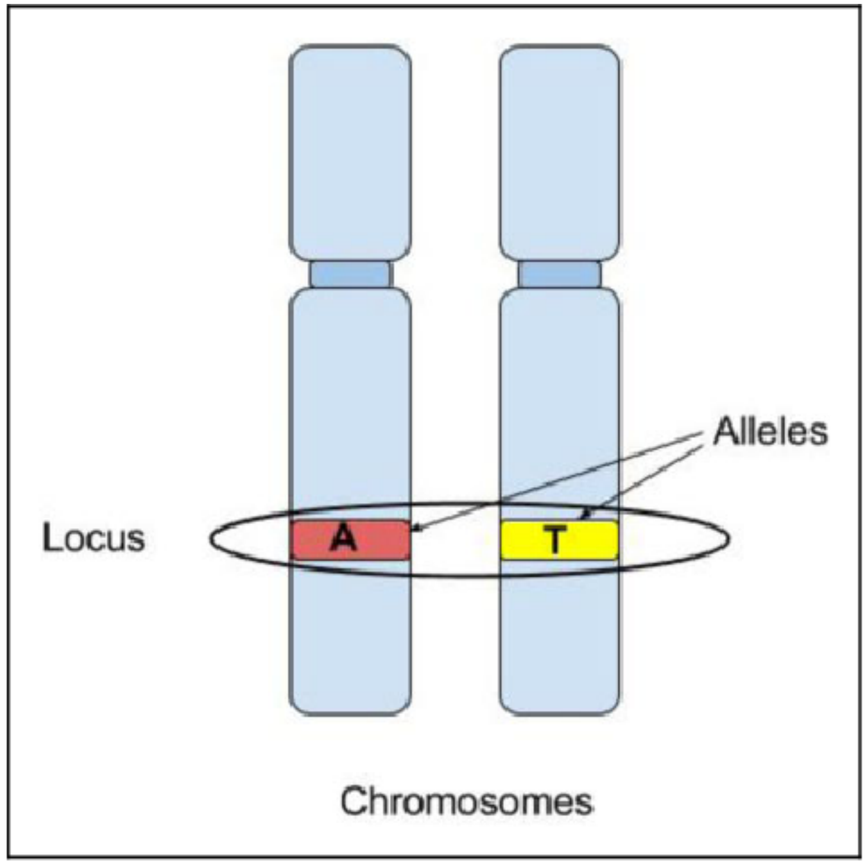



DNA
NUCLEOTIDES
CHROMOSOME
PERSON 1
GENE
BASE PAIRS
SNP
PERSON 2
CHROMOSOME
CELL

Basic definitions

Loci and alleles

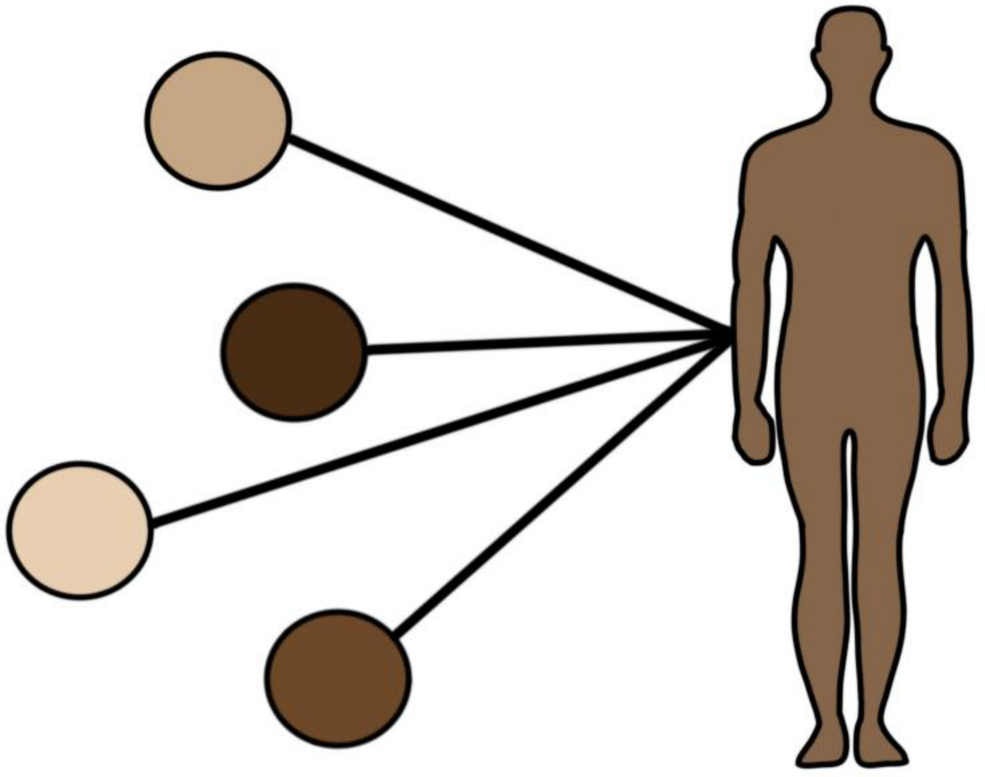
- We will use the term ***locus (pl. loci)*** too refer to a genomic element located in a field position of the genome
- A locus may have different variants called **alleles**
- Although the term gene refers to a functional biological unit, often the terms locus and gene are used interchangeably



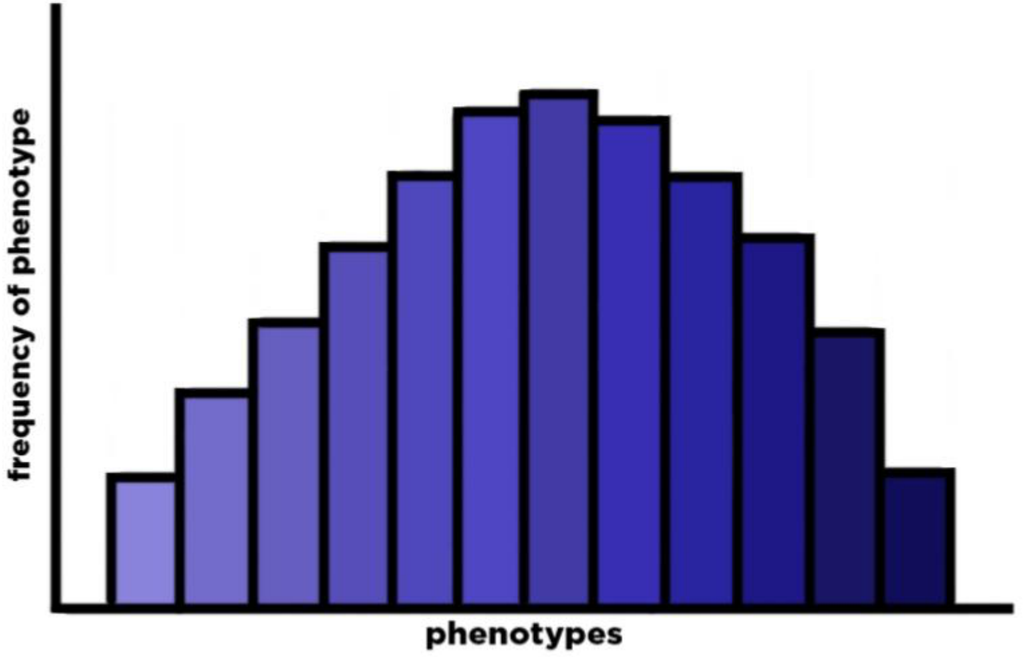
PATTERNS OF INHERITANCE: POLYGENIC TRAITS

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.



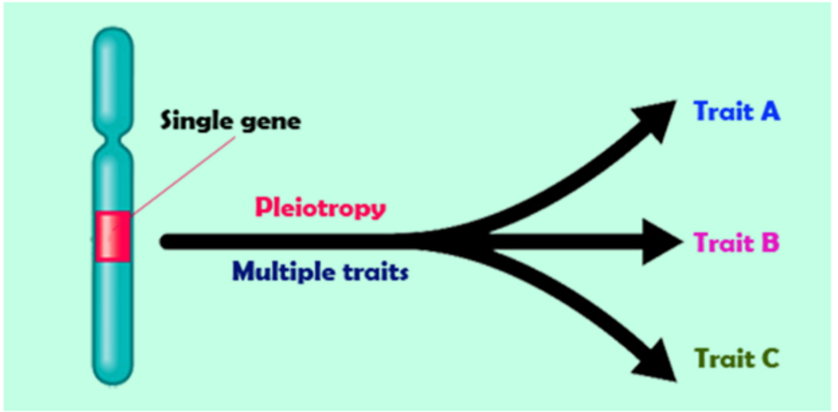
Each gene plays a small role in the expression of the trait.

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.*



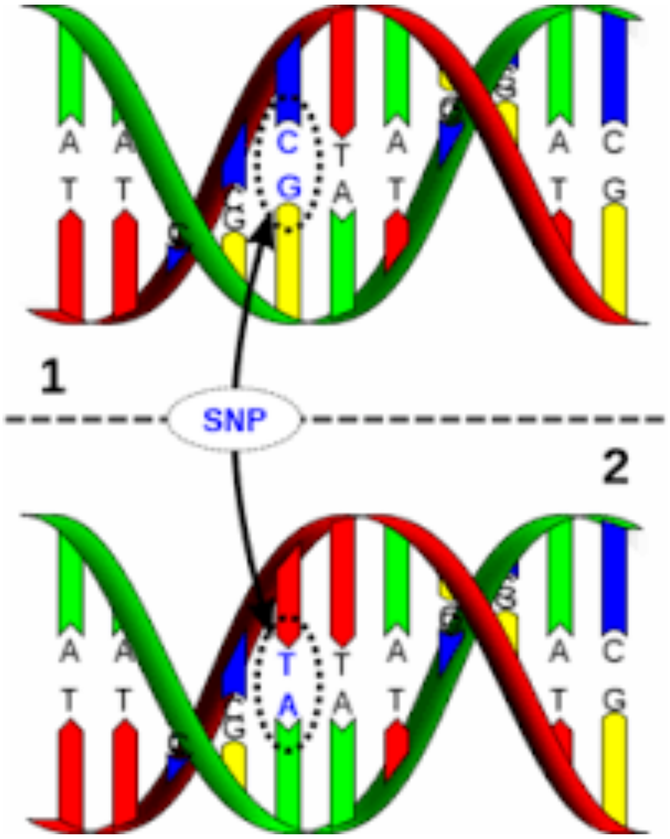
- ***MOST OF HUMAN TRAITS ARE POLYGENIC***
- ***MOST OF GENES HAVE PLEIOTROPIC EFFECTS***

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

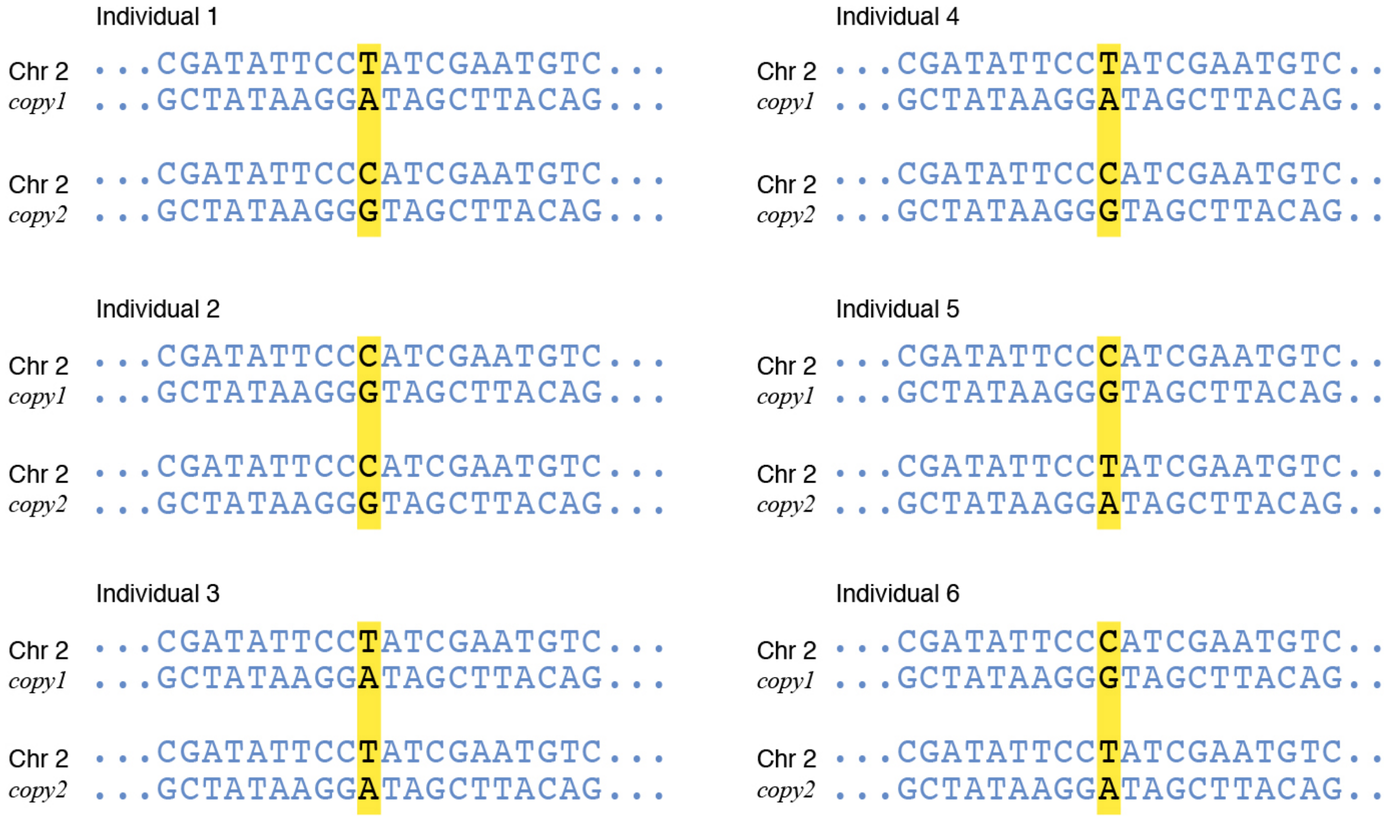
Allele frequencies

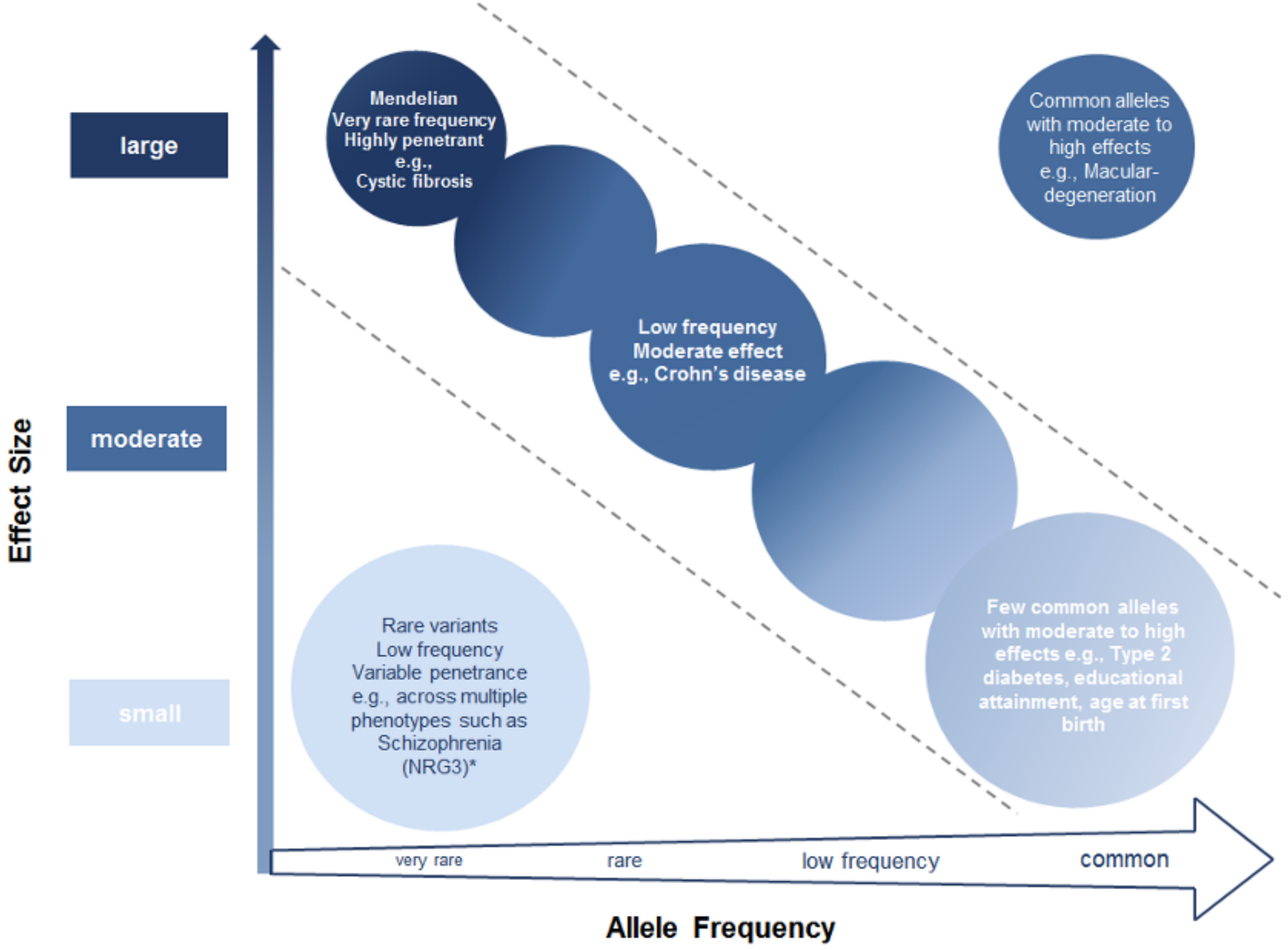
Allele frequencies are a central unit of populati genetics analysis,

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:

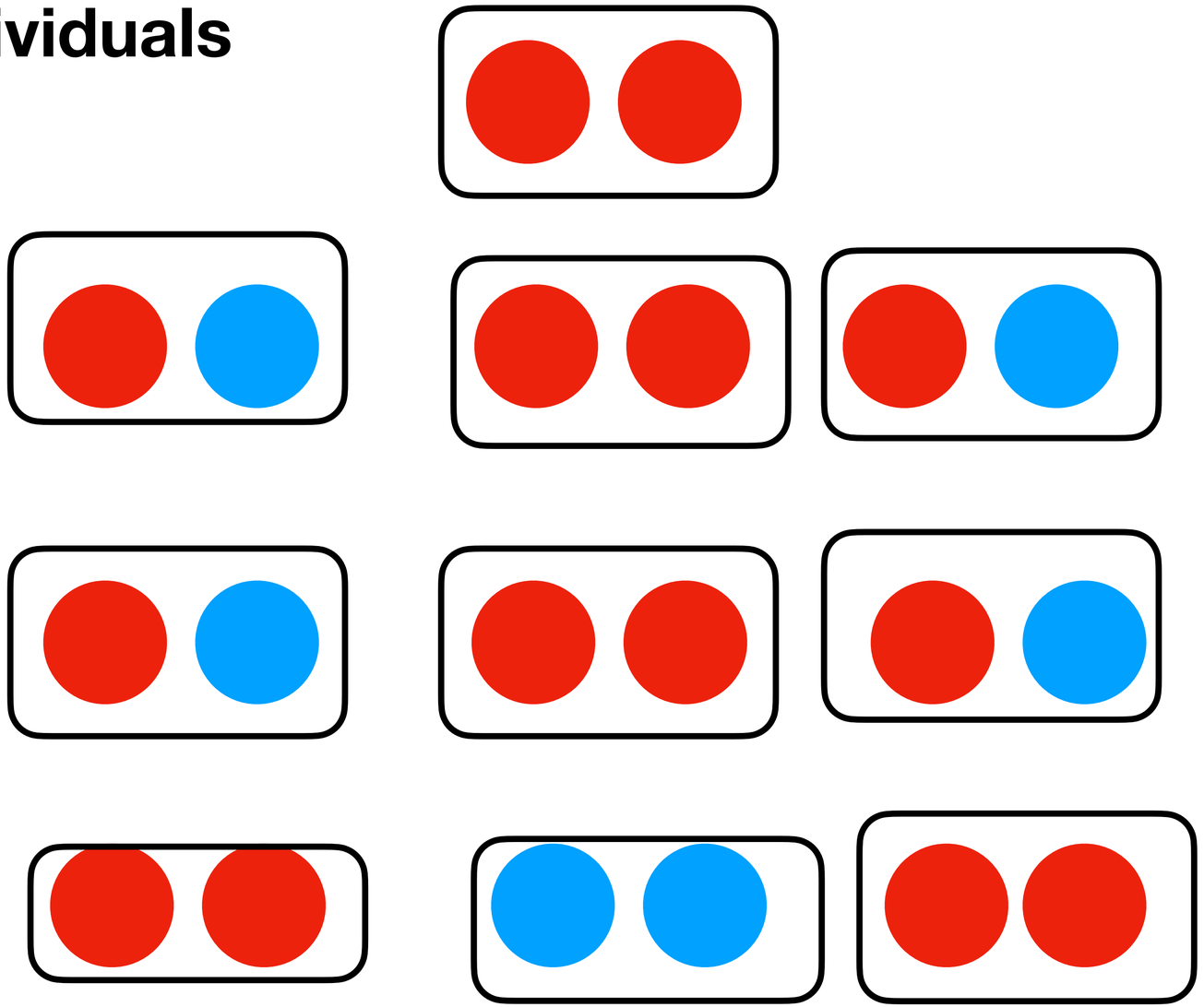
$$p^2 + 2pq + q^2 = 1$$

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

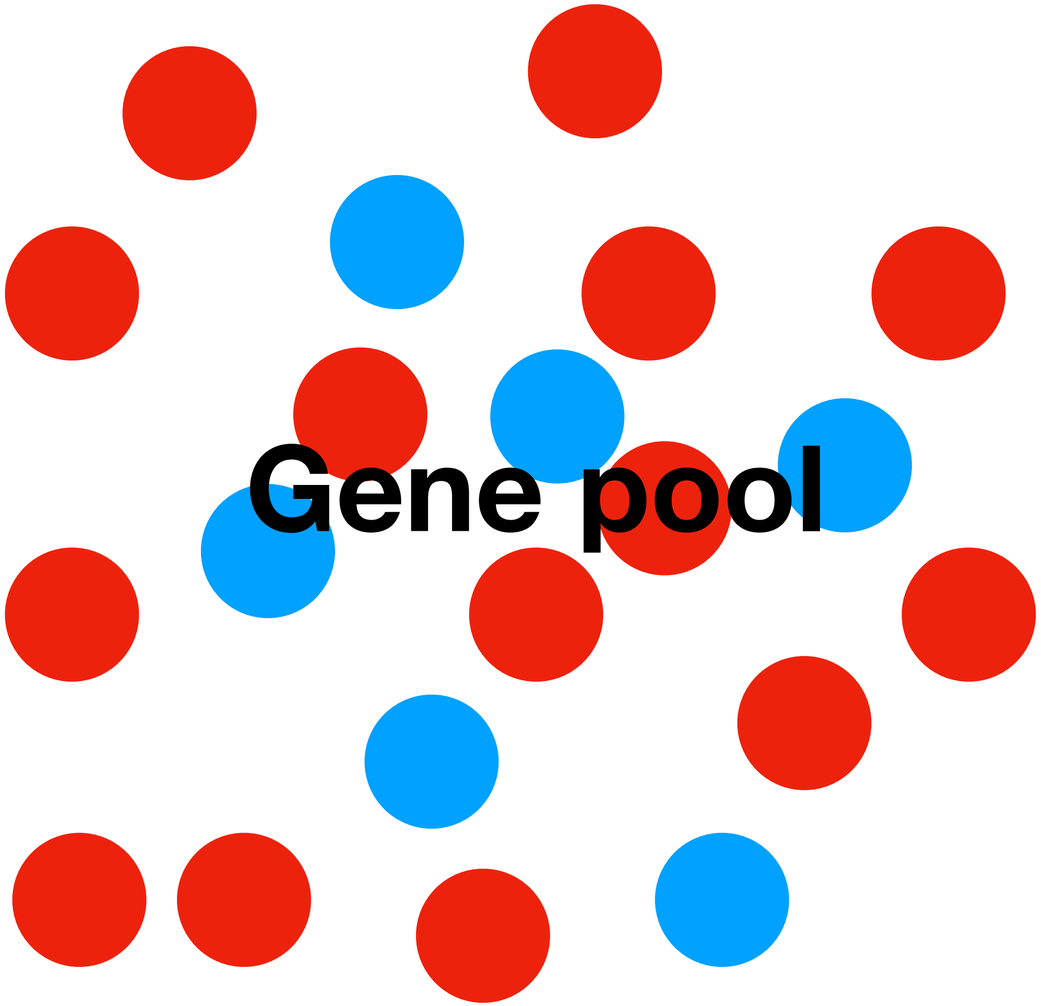
Dominant

If you have two recessive alleles, you have red hair

In this example 50% of the population has red hair

Example

10 individuals

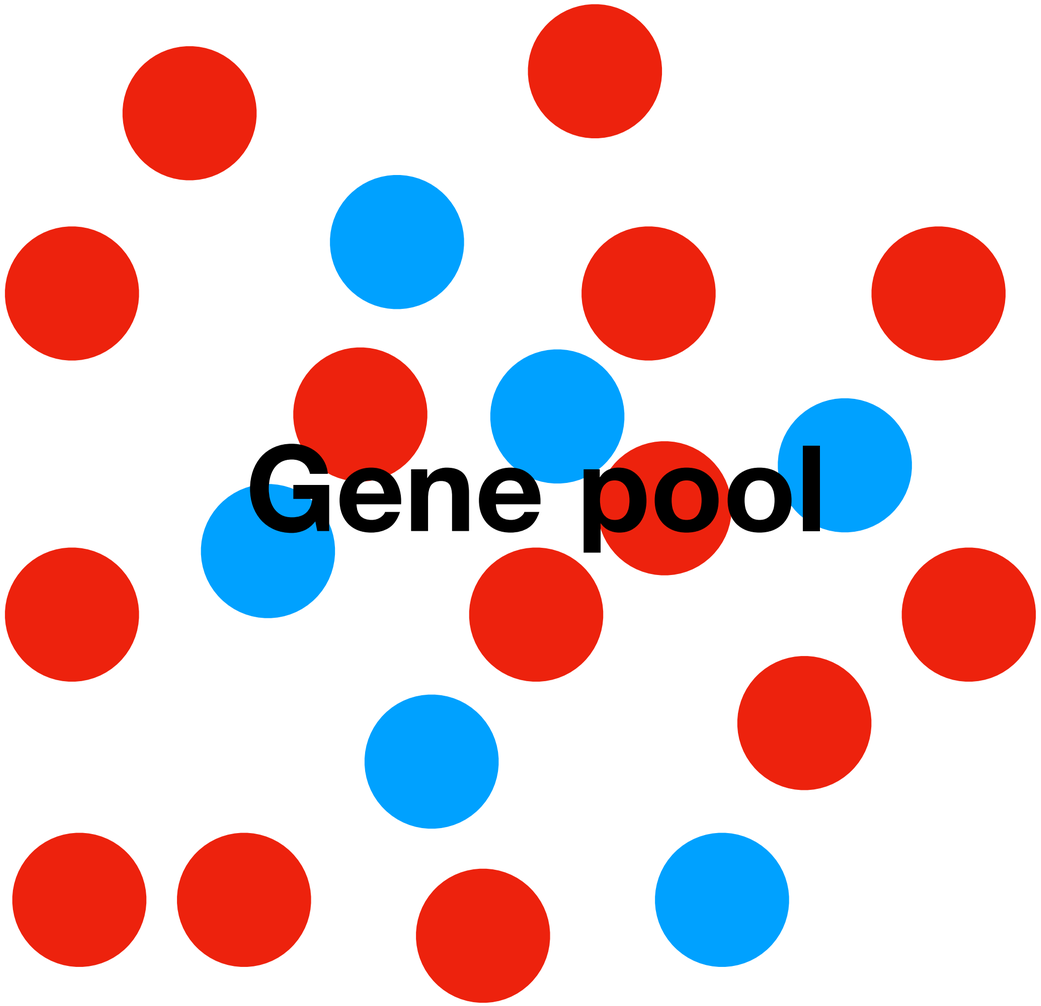


Recessive

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?

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

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

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

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

Variance of Allele Frequency under HWE:

- In a **diploid** population, each individual carries two alleles.
- The total number of **alleles** in the population is 2N.
- The variance follows a **binomial distribution**, where each allele is independently sampled.
- Since each diploid individual contributes two alleles, the variance is reduced by a factor of **2N**.

$$Var(p) = \frac{p(1-p)}{2N}$$

Variance of Genotypic Frequencies Under HWE

Homozygous dominant (p^2)

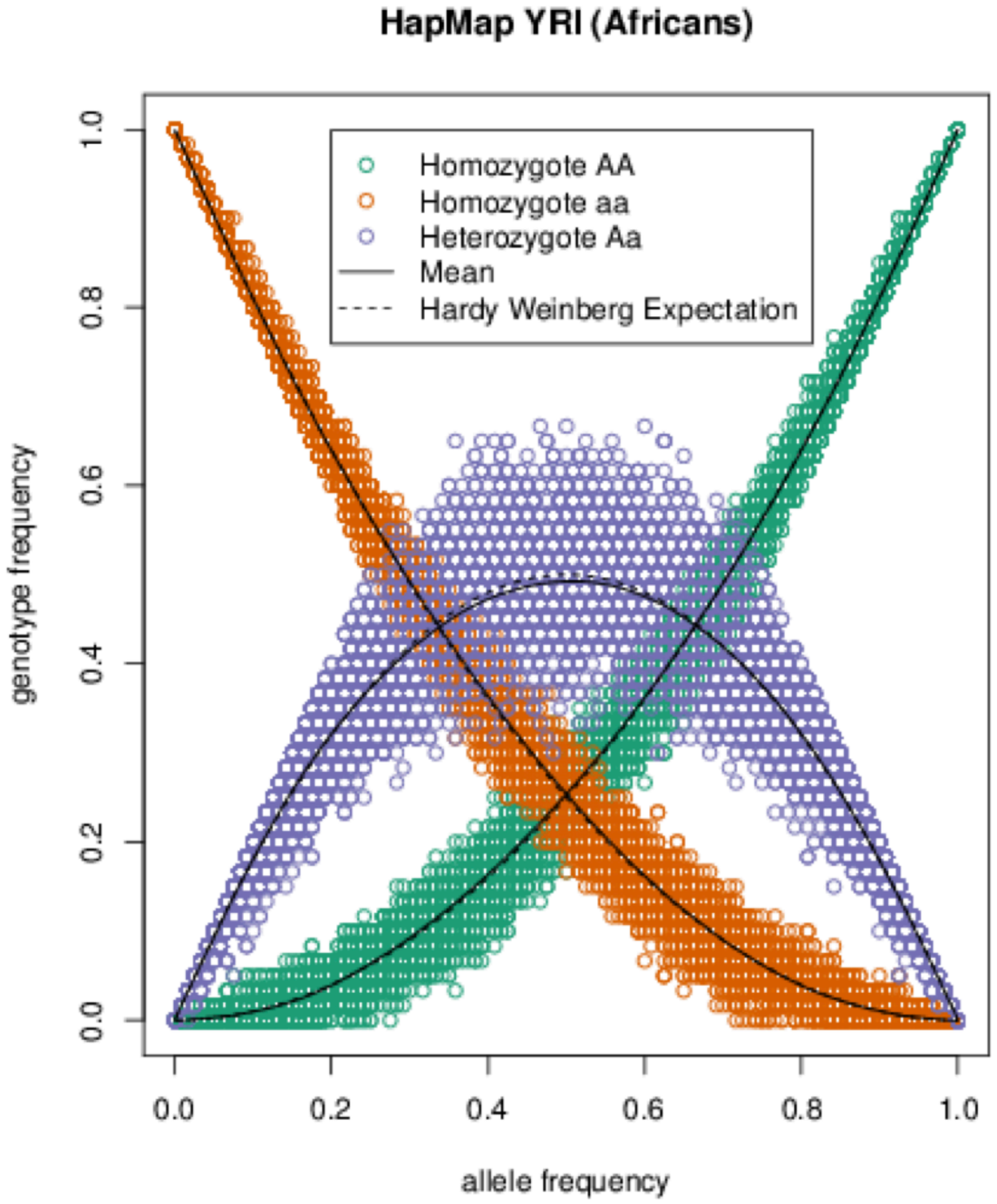
$$Var(p^2) = \frac{p^2(1-p^2)}{N}$$

Homozygous recessive (q^2)

$$Var(q^2) = \frac{q^2(1-q^2)}{N}$$

Heterozygous ($2pq$)

$$Var(2p) = \frac{4p^2q^2}{N}$$



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 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)

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

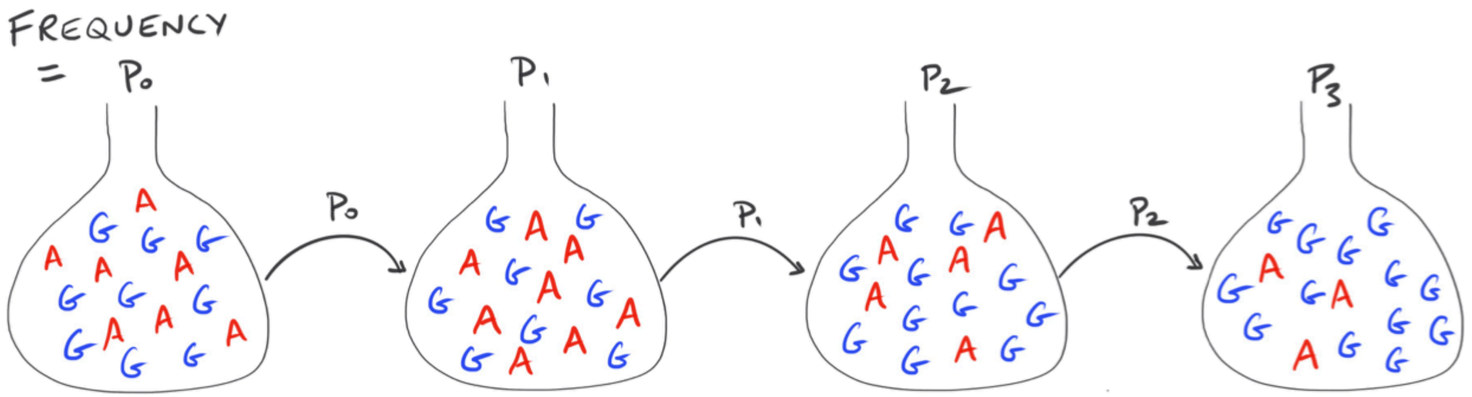
Genetic drift describes random, non-selective change to the allele frequencies of a population.

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

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

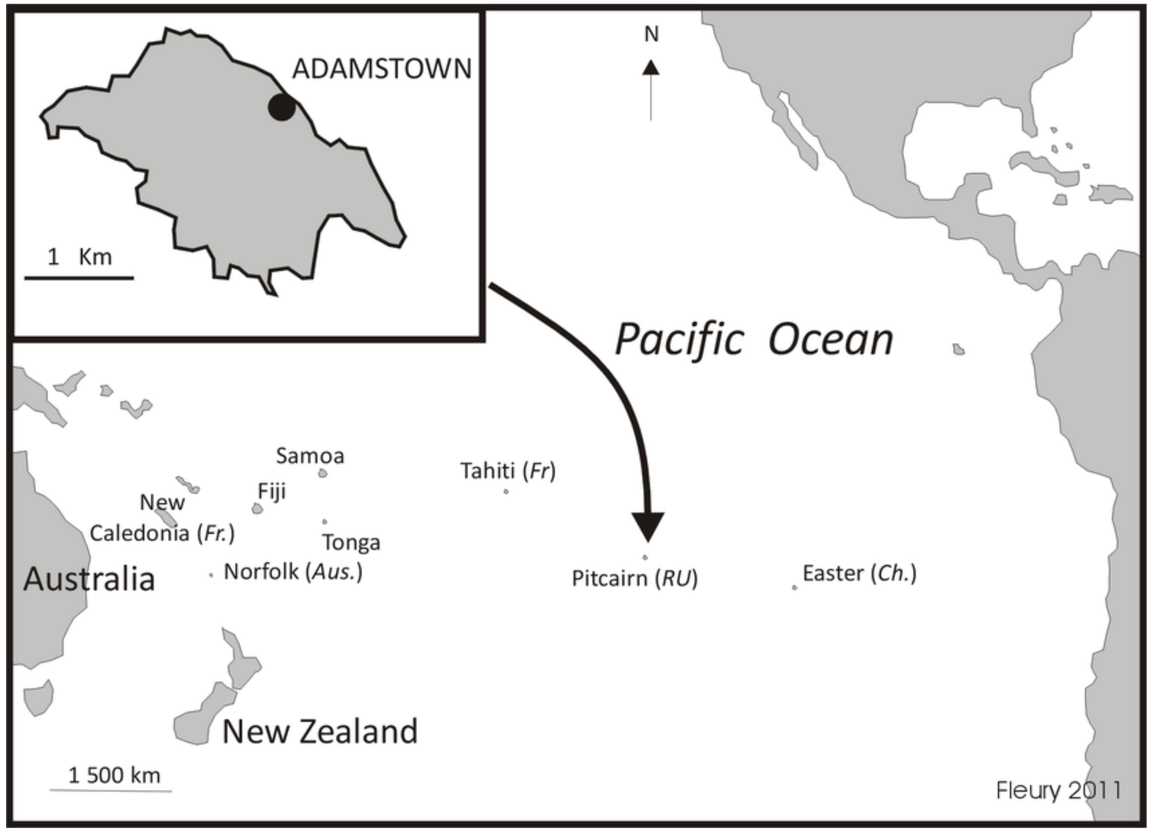
Genetic drift

- Binomial sampling over successive generations produces genetic drift.
- This will produce a series of allele frequency changes over time called a random walk
- Eventually, the G allele will either reach 100% frequency, in which case we say that it has **fixed**, or 0% in which case we say it has been **lost**.



Genetics of Pitcairn Island: A Case Study in Genetic Drift and Founder Effects

- In **1790**, the **HMS Bounty mutineers**, led by **Fletcher Christian**, settled on Pitcairn Island along with **nine Tahitian men and women**.
- This small **founder population** consisted of only **27 individuals**.
- The island remained almost completely **isolated** from outside genetic influences for over a century.



Genetic consequences

Founder Effect

- The entire population descended from a tiny genetic pool, meaning only a fraction of the genetic diversity of the original European and Polynesian populations was retained.
- Rare alleles in the founding group became overrepresented in the descendants.

Genetic Drift

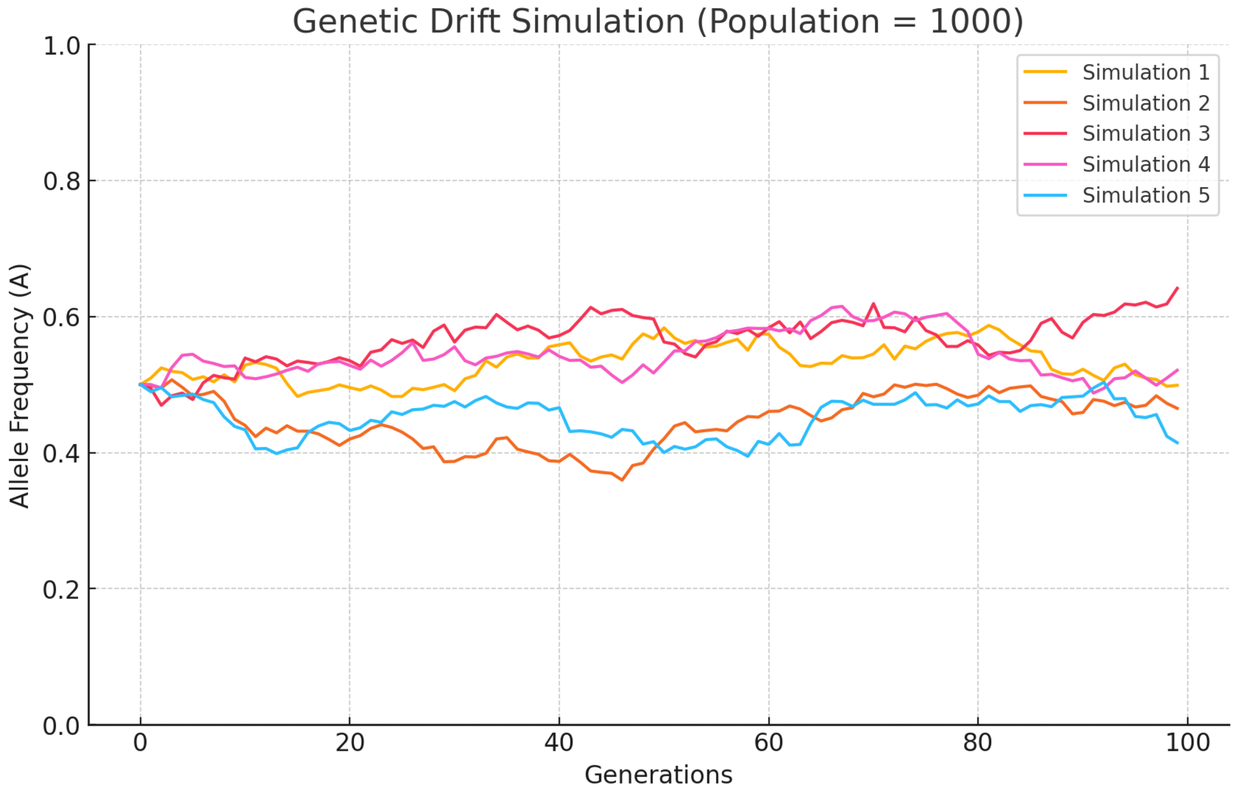
- Since the population remained small (often under 100 individuals), random fluctuations in allele frequencies occurred.
- Some alleles were lost, and others became fixed by chance.

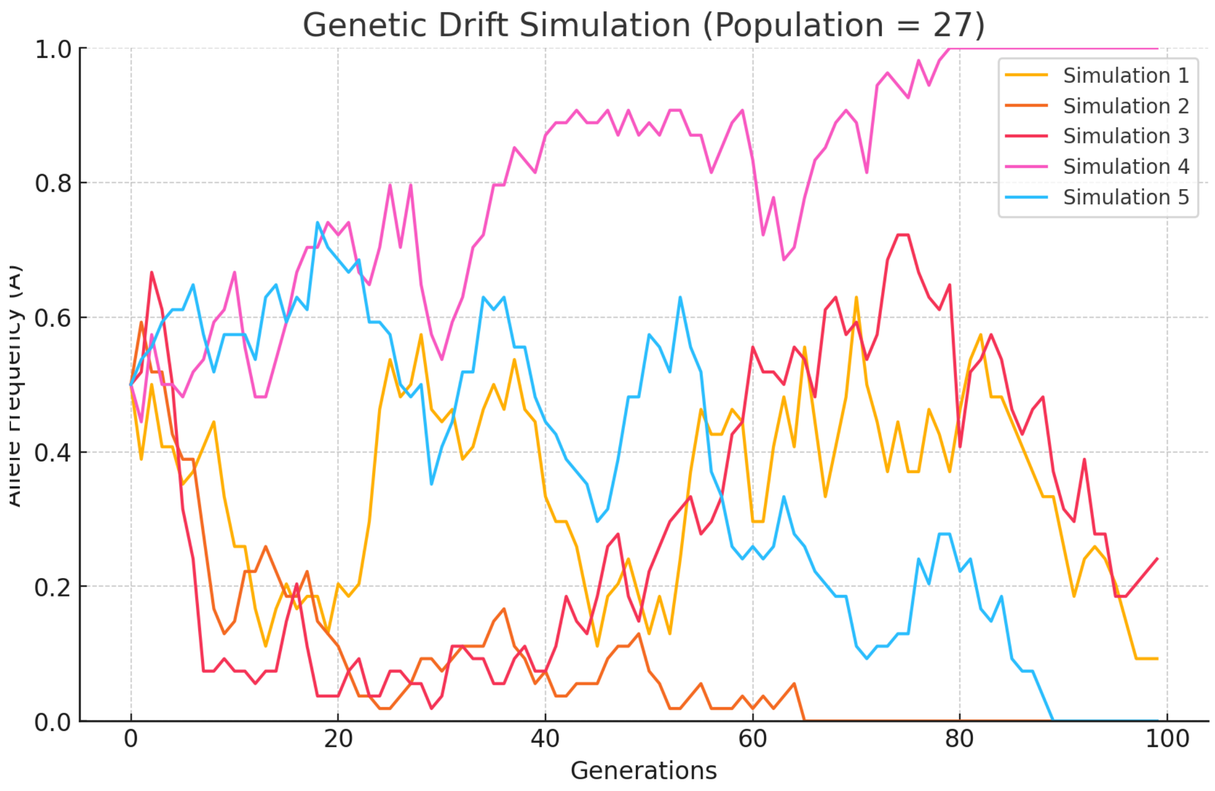
Inbreeding

- Due to the limited number of families, many individuals were distantly (or closely) related.
- Inbreeding coefficients increased over generations, which led to a higher chance of homozygosity (having two copies of the same allele).

Loss of Genetic Diversity

- Compared to the populations from which they originated (British and Tahitian), the Pitcairn Islanders have lower genetic variation.
- This makes them more susceptible to genetic diseases and low adaptability to environmental changes.


Genetic Drift Simulation (Population = 1000)
Allele Frequency (A)
Generations
Simulation 1
Simulation 2
Simulation 3
Simulation 4
Simulation 5
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100


Genetic Drift Simulation (Population = 27)
Allele Frequency (A)
Generations
Simulation 1
Simulation 2
Simulation 3
Simulation 4
Simulation 5
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

The Role of Genetics in Human Prehistory

- Over the past 30 years, **population genetics** has revolutionized our understanding of human evolution.
- Genetic variation in modern humans reflects our species' history.
- Combination of fossil evidence, genetic data, and ancient DNA provides insights into deep human history.

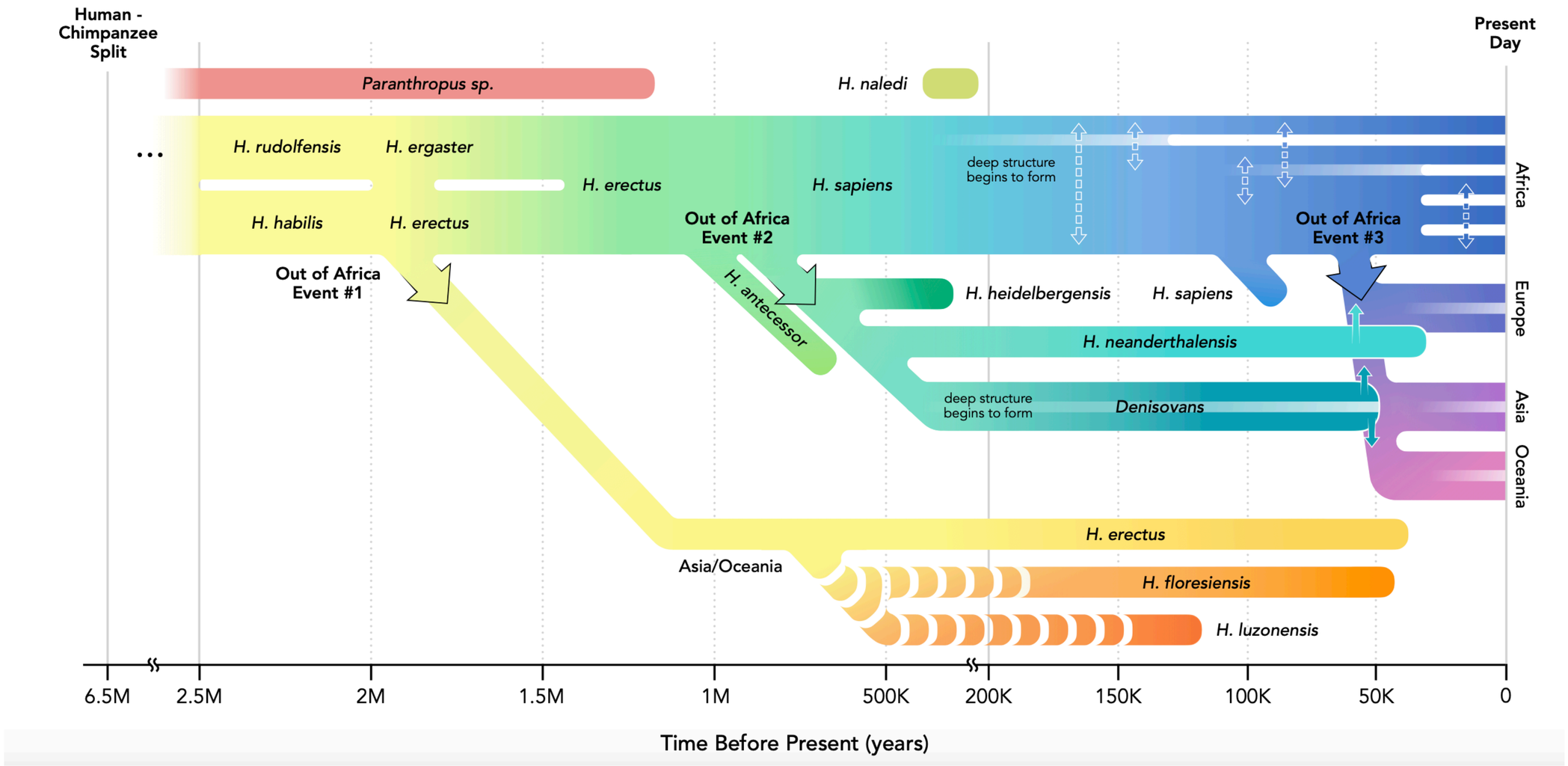
- **1856 Neanderthal Discovery**: First archaic hominin found in Neander Valley, Germany.
- **Diversity of Hominins: Multiple species coexisted in Africa, Eurasia, and Oceania over 2M years.**
- **Key Migration Events: Three major out-of-Africa dispersals of hominins.**

Evolutionary Tree and Hominin Diversity

Homo sapiens emerged ~200-300KYA in Africa.

Other contemporaneous species:

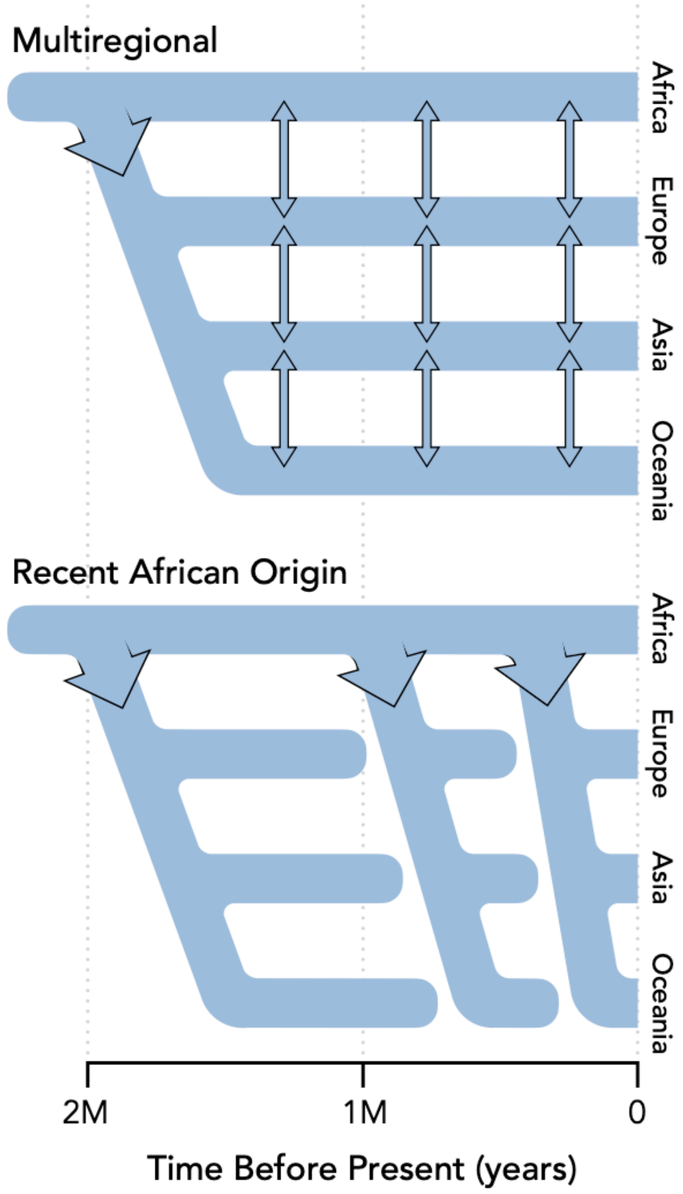
- **H. erectus** (first to leave Africa, ~2MYA).
- **H. neanderthalensis** (Europe, ~400-40KYA).
- **Denisovans** (Asia, ~400-40KYA).
- **H. floresiensis** (Indonesia, ~190-50KYA).


Human - Chimpanzee Split
Present Day
Paranthropus sp.
H. naledi
H. rudolfensis
H. ergaster
H. habilis
H. erectus
H. erectus
H. sapiens
deep structure begins to form
Out of Africa Event #1
Out of Africa Event #2
H. antecessor
Out of Africa Event #3
H. heidelbergensis
H. sapiens
H. neanderthalensis
deep structure begins to form
Denisovans
Africa
Europe
Asia
Oceania
H. erectus
Asia/Oceania
H. floresiensis
H. luzonensis
6.5M
2.5M
2M
1.5M
1M
500K
200K
150K
100K
50K
0
Time Before Present (years)

When, and where did modern Homo sapiens evolve?

Multiregional Hypothesis: Continuous evolution across Africa, Asia, and Europe with regional continuity.

Recent African Origin (Out-of-Africa) Model: Modern humans evolved in Africa and replaced archaic populations globally.



The Genetic Evidence for the Out-of-Africa Model

- Genetic diversity decreases with distance from Africa.
- All non-African populations share ~1.5-2% Neanderthal DNA, proving interbreeding events.

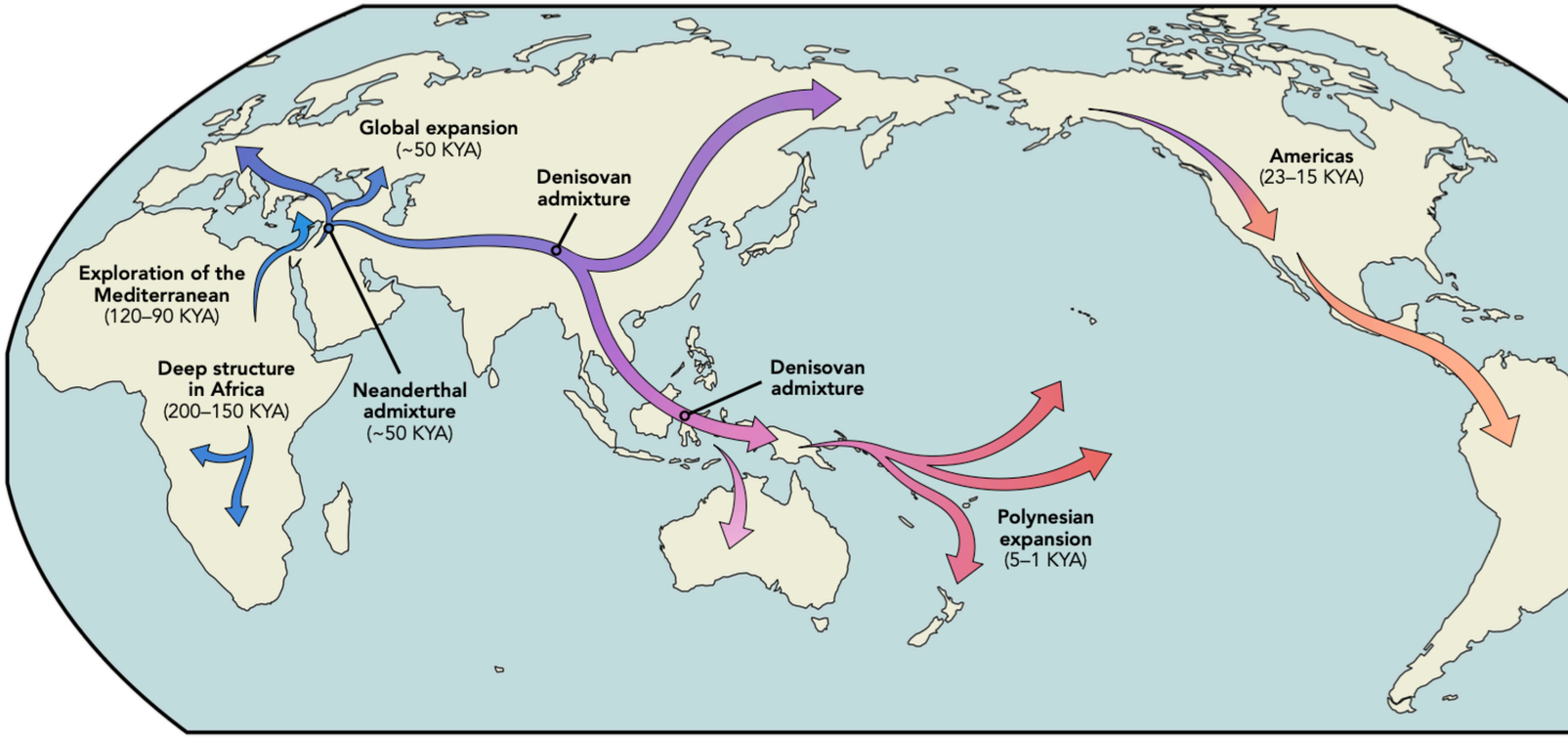
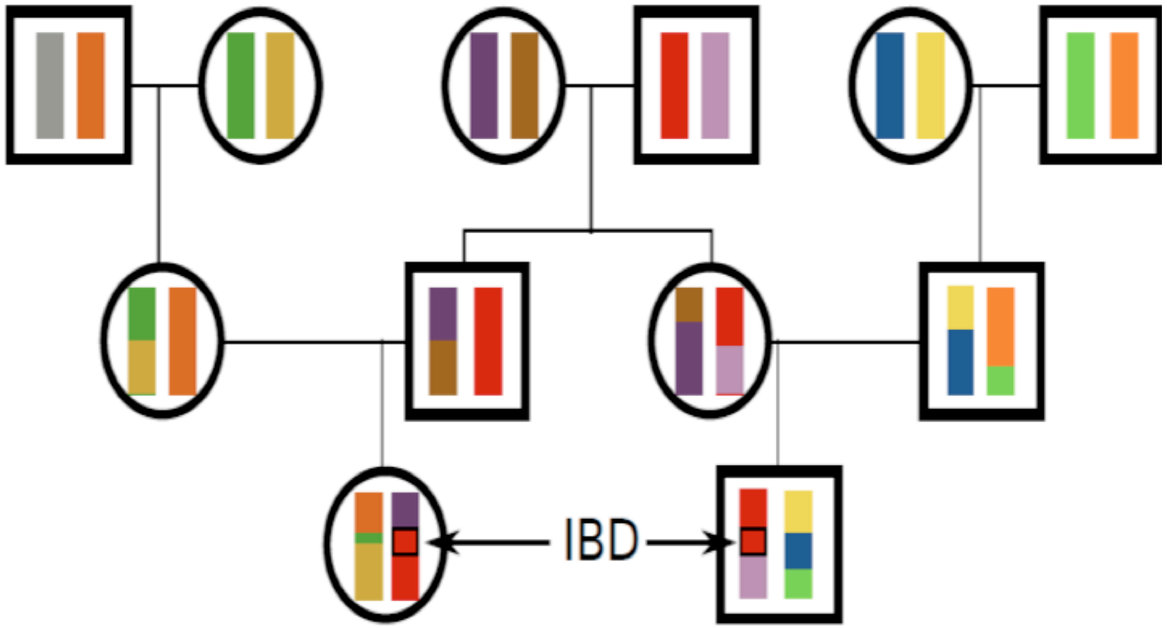


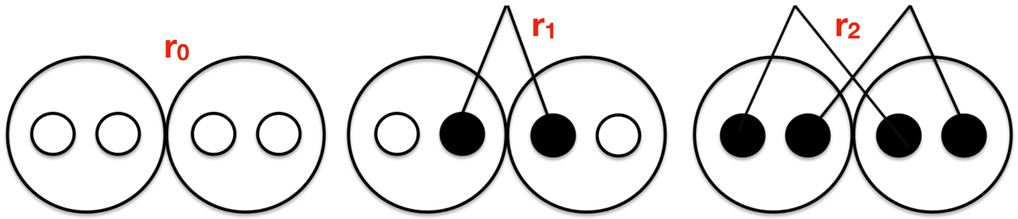
Figure 3.59: Global spread of modern humans. *The earliest fossils with modern features appear in Africa* 200–3
dispersals of modern humans into Eurasia (120–50 *KYA) were likely replaced by a single major dispersal after* 50 *KYA*
locations of arrows and admixture events are generally not known in detail. ***All times are approximate.***

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 a 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)

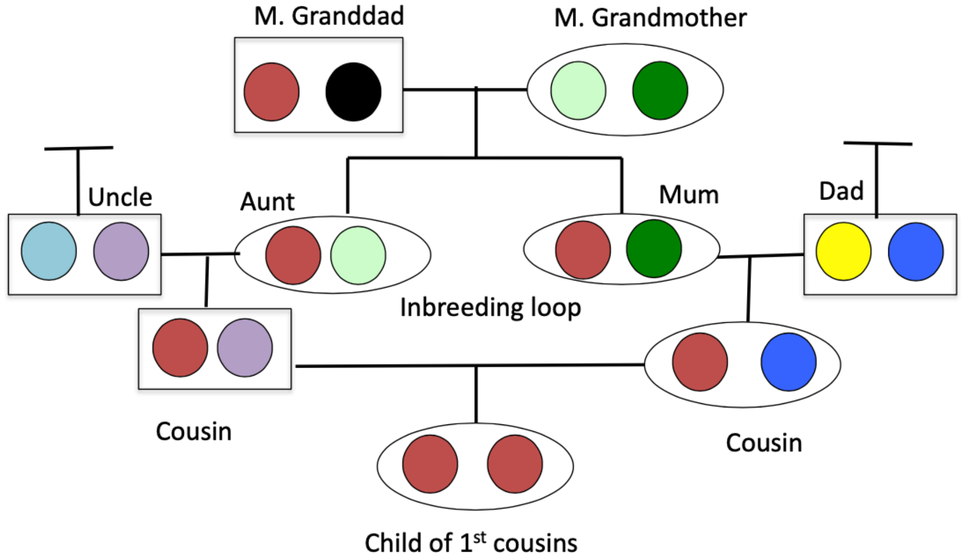
relationship(i,j)	r_0	r_1	r_2	Φ_{ij}
Parent-child	0	1	0	1/4
Full siblings	1/4	1/2	1/4	1/4
Identical twins	0	0	1	1/2
1st cousins	3/4	1/4	0	1/16

Kinship Coefficient (Φ)

The **kinship coefficient (Φ)** between two individuals is the probability that a randomly chosen allele from one individual is identical by descent (IBD) to a randomly chosen allele from the same locus in the other individual. It quantifies genetic relatedness.

$$\Phi = \frac{1}{2} \times P(IBD = 2) + \frac{1}{4} \times P(IBD = 1) + P(IBD = 0) \times 0$$

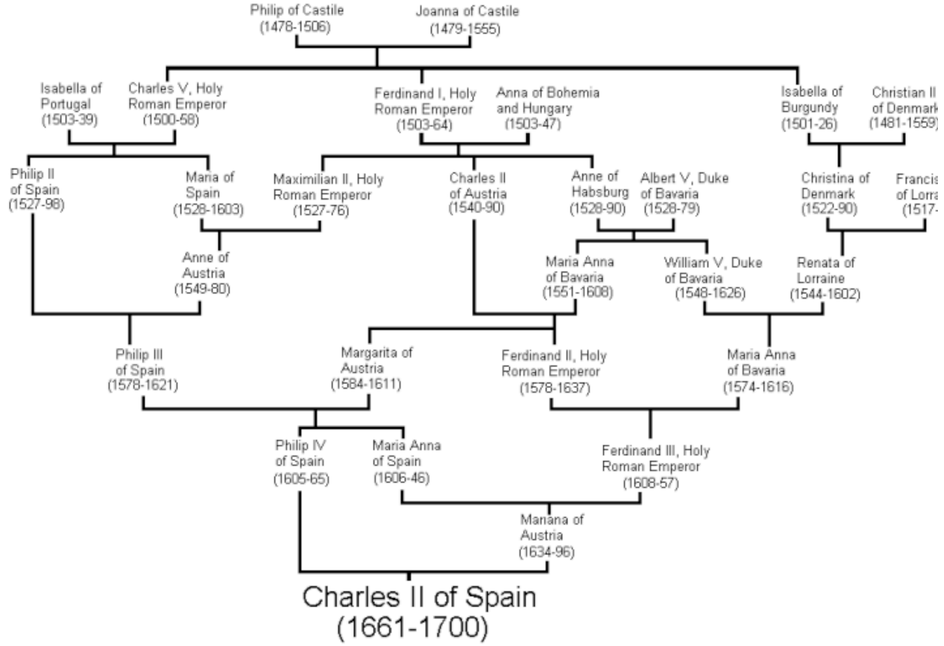
Inbreeding



- 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

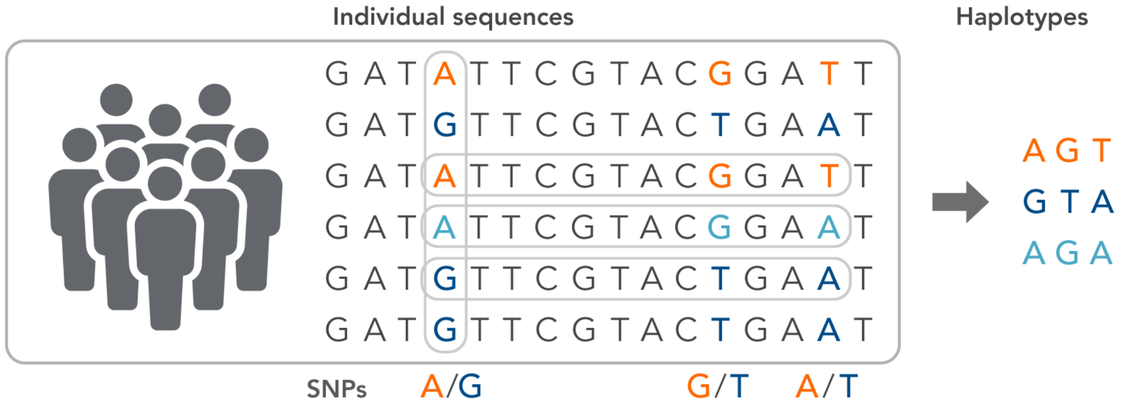
- Alvarez et al. (2009) calculated that Charles II had an inbreed- ing coefficient of 0.254, equivalent to a full-sib mating, thanks to all of the inbreeding loops in his pedigree. Therefore, he is expected to have been homozygous by descent for a full quarter of his genome. As we'll talk about later 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 $p_A; p_a; p_B; p_b$

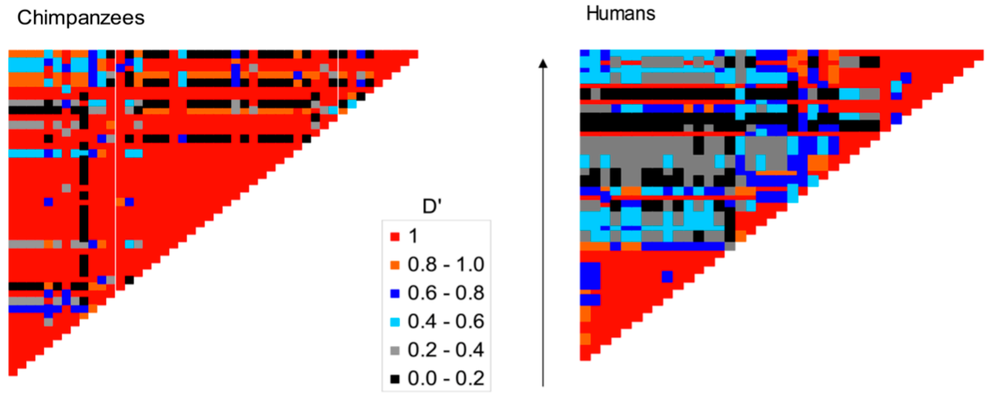
IF THEIR SEGREGATION IS INDEPENDENT: $p_{AB} = p_A\, p_B$*, OTHERWISE THE TWO LOCI ARE IN LINKAGE DISEQUILIBRIUM*

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.



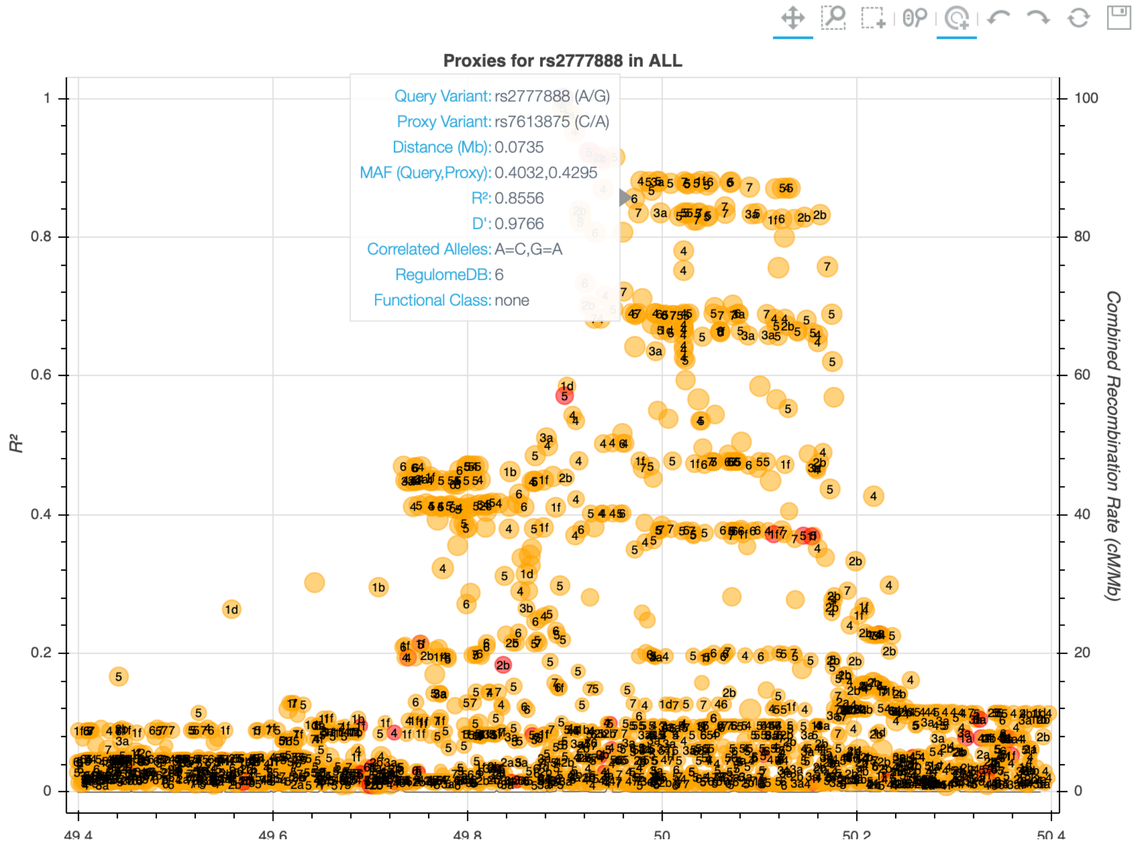
correlation coefficient

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.

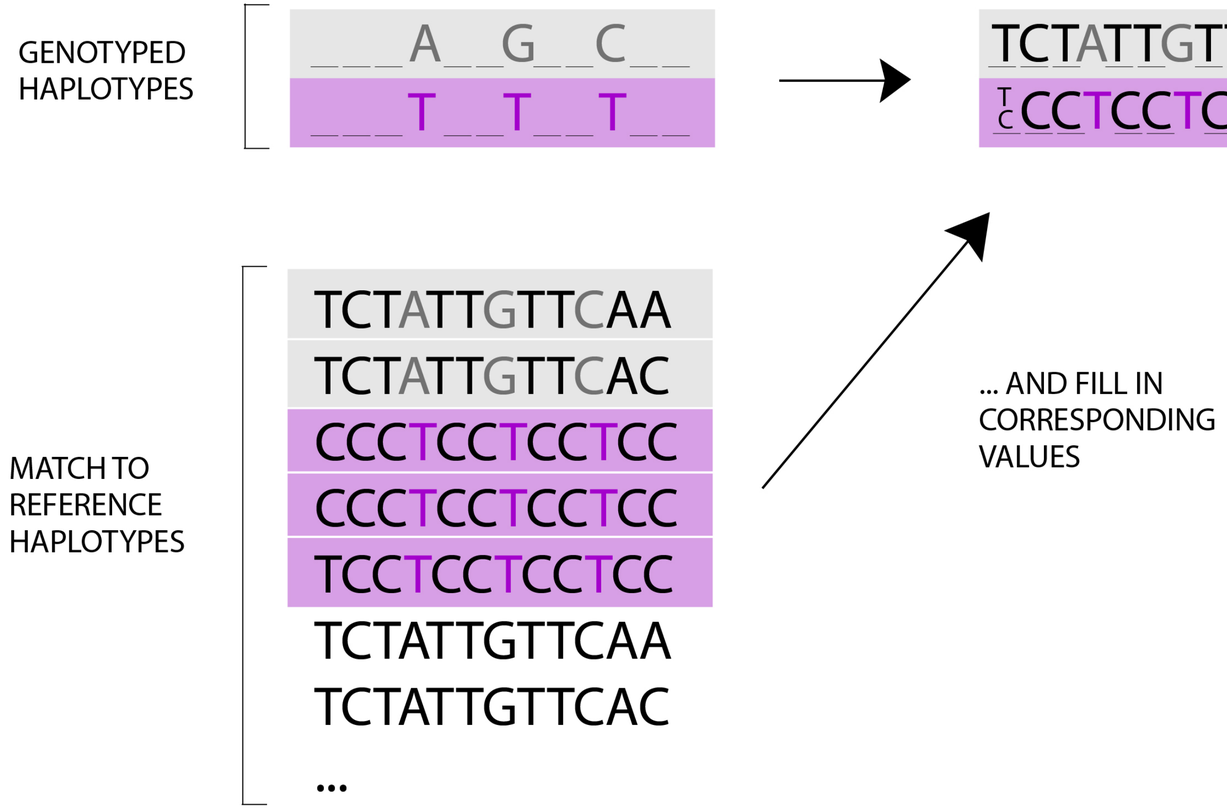
Table 1: Omni BeadChip Performance Parameters

	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

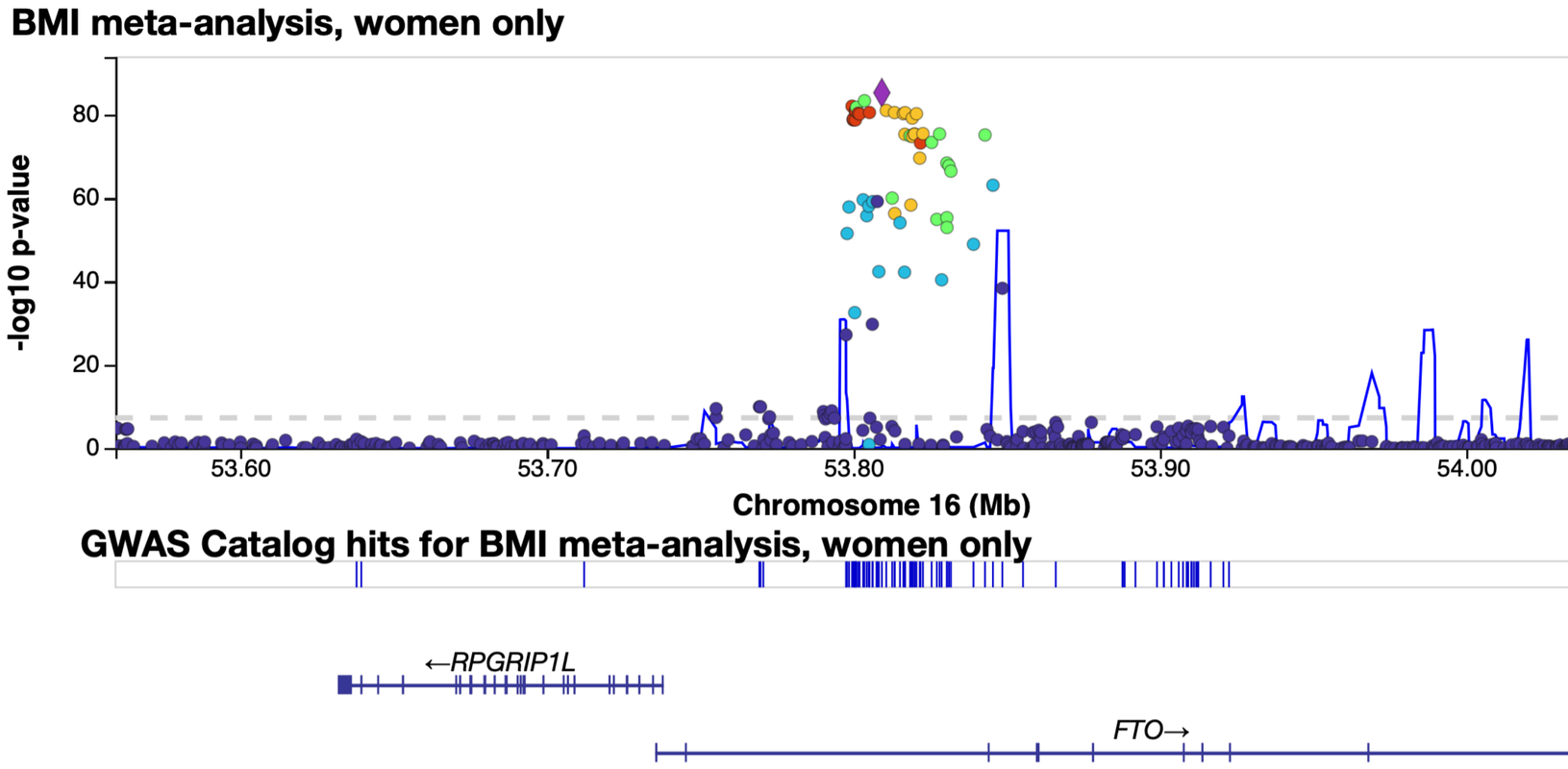


Reference Panels

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

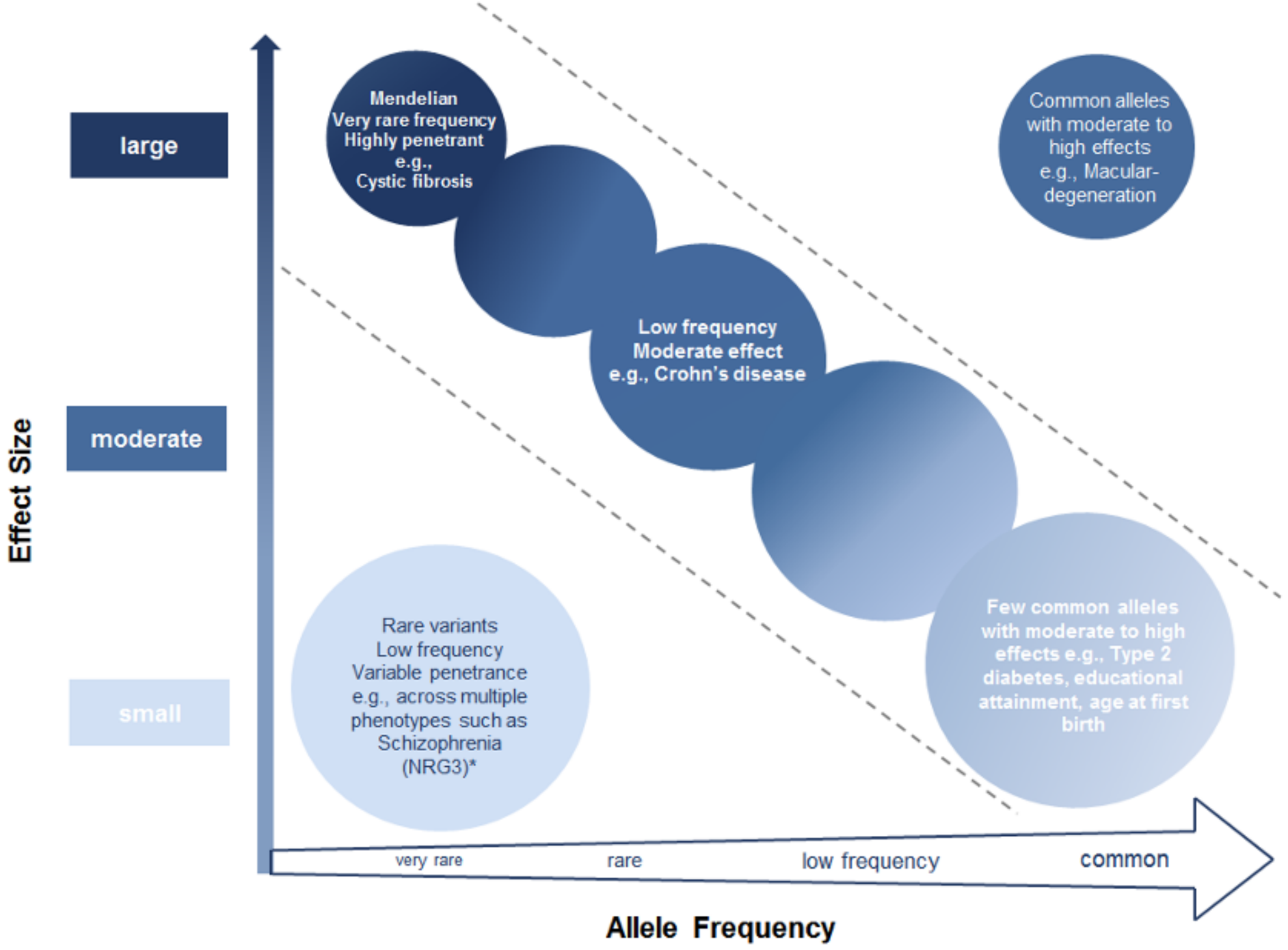
LD and genetic association

- 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