Sociogenomics

Basic concepts

Nicola Barban

Sequencing vs. Genotyping

- **Sequencing** is the process in which we “read” the whole or parts of DNA. Fundamental to characterize the genetic variation in a sample of individuals;
- Ideally you might imagine DNA sequencing providing a fully accurate end-to-end read-out of each chromosome. But no current technology can provide this directly.
- **Genotyping** refers to a variety of different experimental methods that can **determine a person’s genotype at a specific set of pre-selected SNP positions** (and nowhere else in the genome).

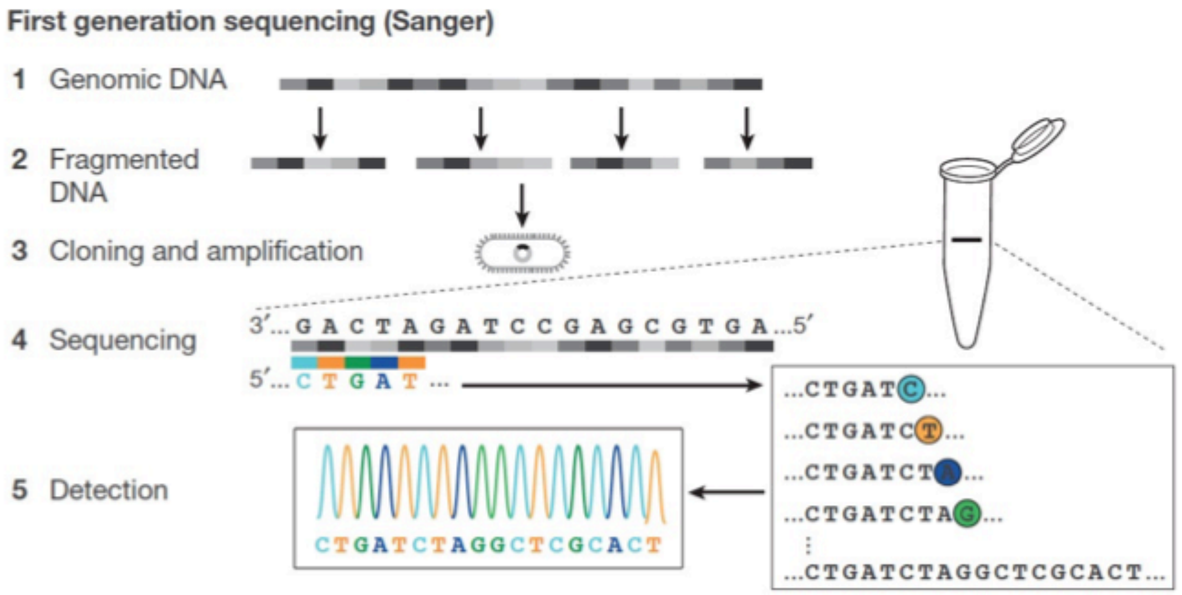
.

https://www.nature.com/immersive/d42859-020-00099-0/index.html

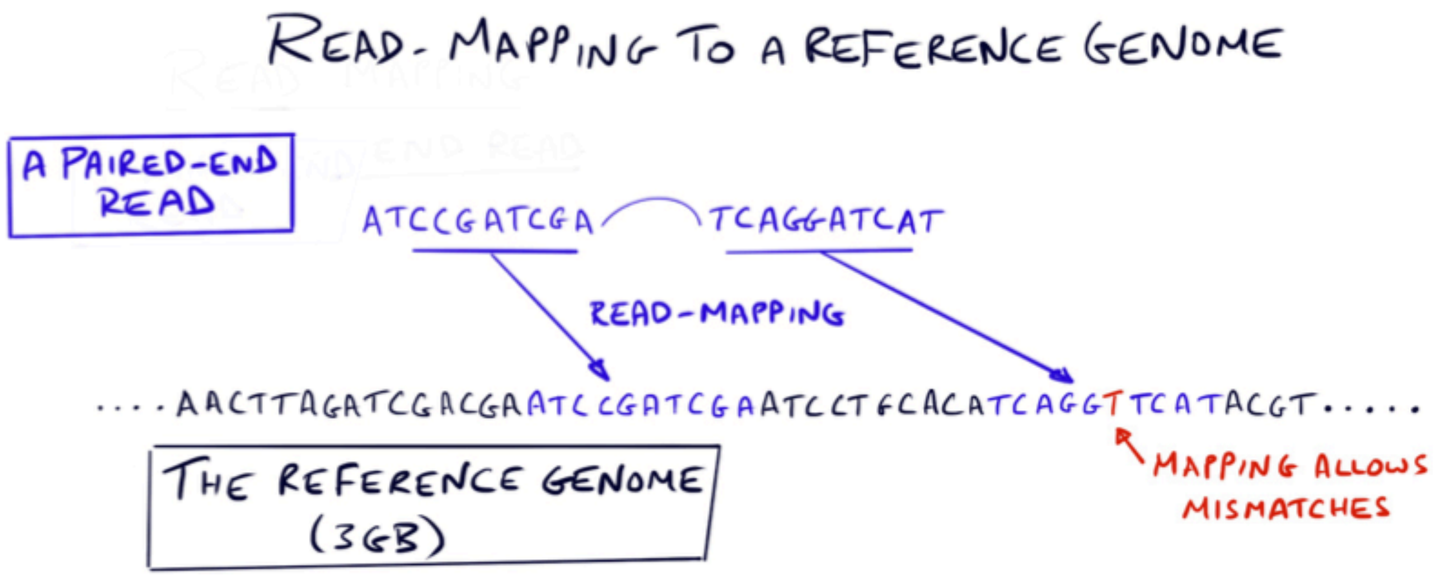
DNA sequencing

Sanger sequencing

The first practical DNA sequencing was achieved in the 1970s by two scientists (Fred Sanger and Walter Gilbert)



READ-MAPPING TO A REFERENCE GENOME



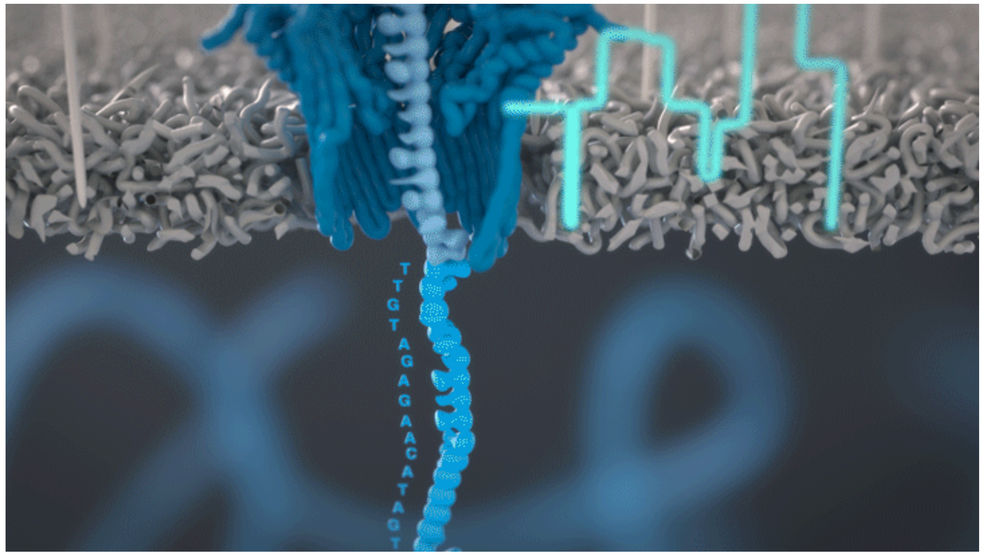
DNA sequencing

Next-Generation Sequencing

The early 2000s saw a major paradigm shift, with a group of new sequencing technologies that achieved enormous advances over Sanger sequencing. These new methods – also called **next-generation** (NSG) – were dramatically faster, and required smaller amounts of expensive reagents. These became commercially available by around 2006 and greatly reduced sequencing costs, enabling individual labs to perform genome-scale sequencing projects for the first time.

DNA sequencing

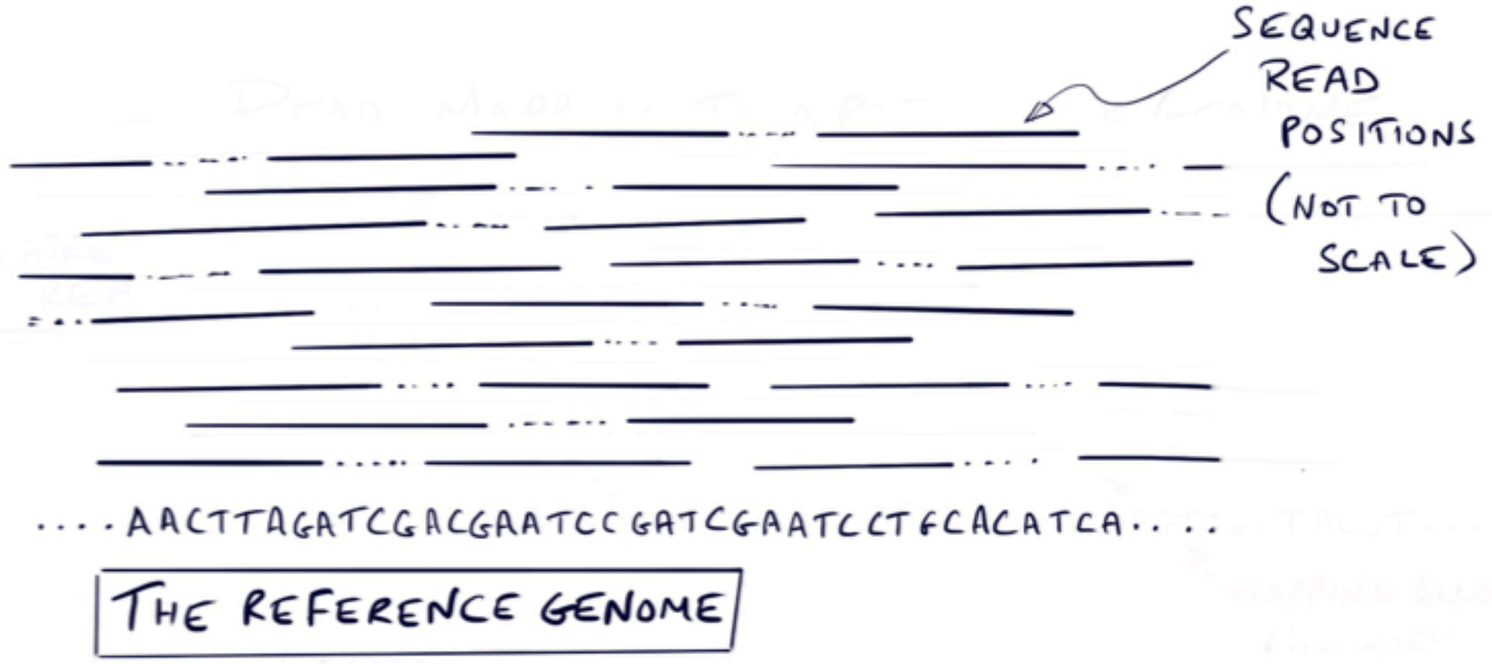
Third generation sequencing.



- https://nanoporetech.com/platform/technology

most genome sequencing uses what is called **shotgun sequencing**, in which we break the genome into many small fragments, sequence them, and rely on our ability to make sense of the sequence reads when we have them

.



human chromosomes are 50-250 Mb long, but what we get are billions of short reads of ~150bp. (bp= base pair, kb=1000 base pairs, Mb=1,000,000 bp. Human Genome approx 3.2 Gb)

Sequencing Pipeline

- **Extract DNA from a tissue sample**, e.g., from blood cells. The initial sample contains millions of cells (each with its own copy of the genome), and so the DNA fragments that we sequence are a mixture of many different copies of the same genome;

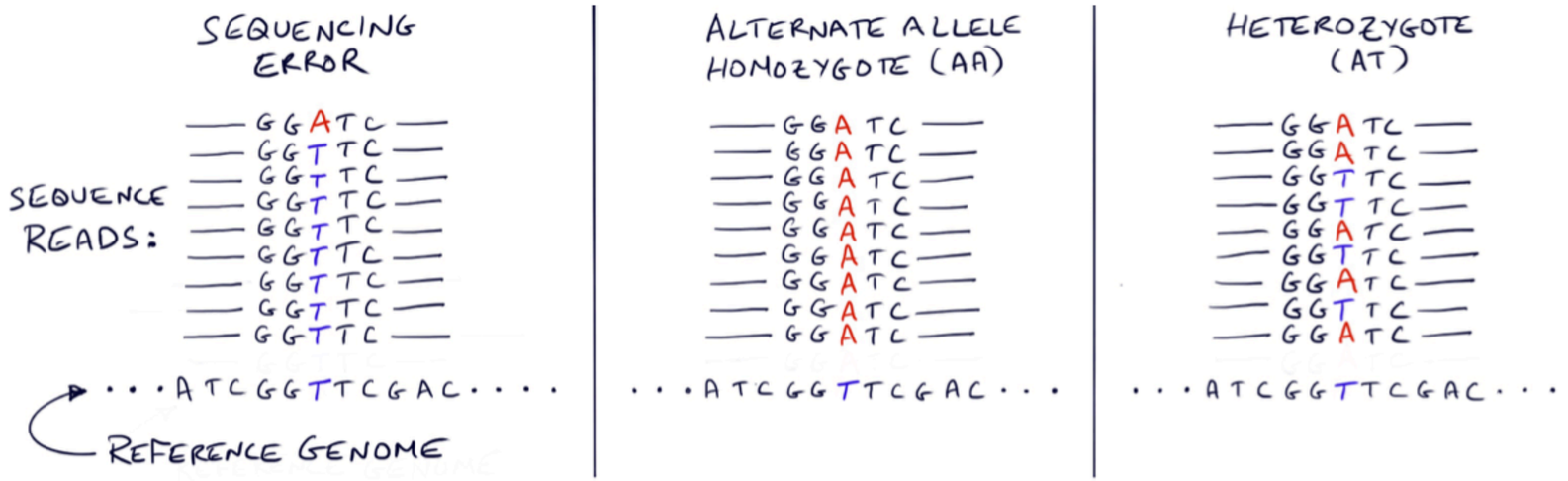
- **Smash up the genome into ~600 bp** fragments of DNA for shotgun sequencing;

- **Map reads to a standard reference human genome**

- **Infer genotype differences from the reference** (e.g., SNPs and structural variants)

SNP calling

When we see a mismatch between the sequence read and the reference genome this might indicate one of several possibilities: a homozygous difference from the reference; a heterozygous difference; a sequencing error



Genome coverage. A key experimental parameter for genome sequencing is referred to as read depth or genome coverage. These terms refer to the average sequencing depth in mappable regions of the genome. 30X coverage is a commonly-used standard for high quality genomes

Genotyping

Current commercial genotyping platforms measure between 500,000 and 2 million SNPs. Genotyping provides less information than a full genome sequence–for example it cannot tell you if carry a rare mutation in a disease gene, as that mutation is unlikely to be included on the genotyping array.

Genotyping is widely used because it can be applied to very large numbers of samples, and is accurate and relatively cheap (less than $100 per sample)

- 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

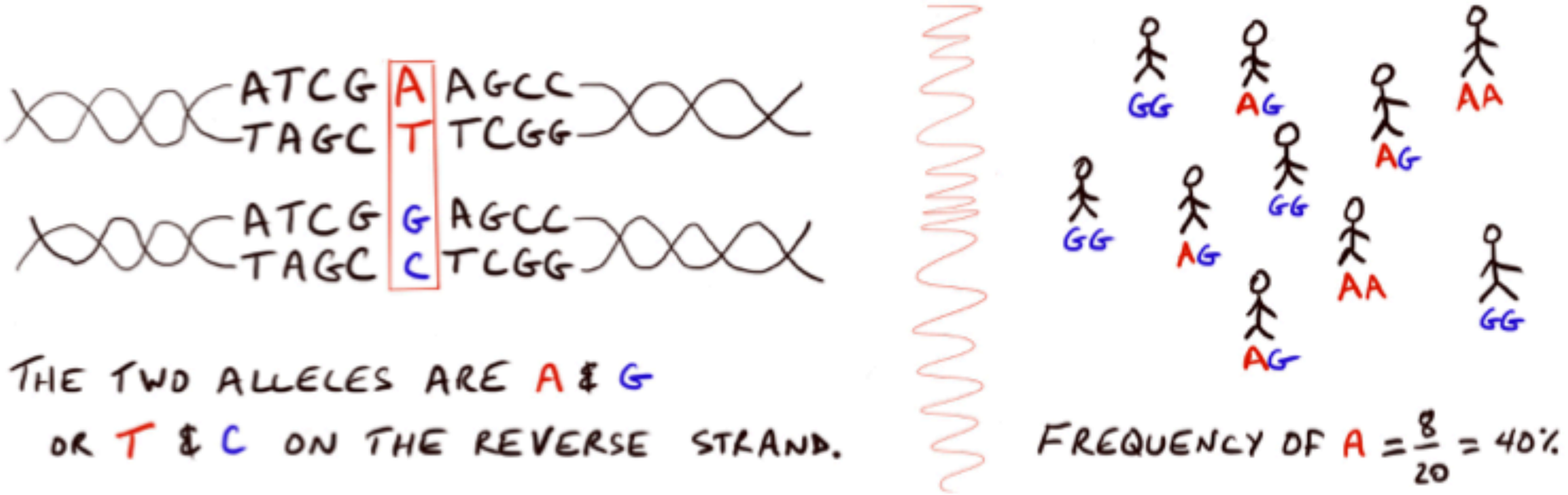
Omni arrays provide flexibility for timing and budget to help investigators effectively achieve their research goals.

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

Single Nucleotide polymorphisms (SNP)



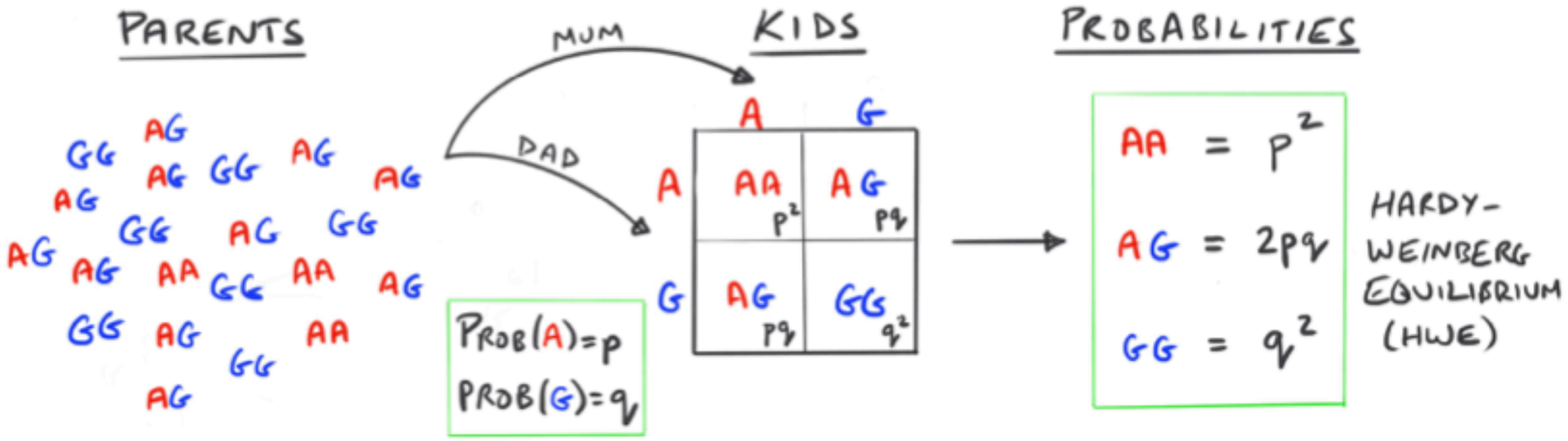
If you're analyzing data it's important to be keep track of **which strand of the DNA the SNP refers to**; in the example above we would consider this an A/G SNP if we're looking at one strand, but a T/C SNP on the other strand.

• **Reference/Alternate allele:** The reference allele is the allele listed at that position in the Human Reference Genome.

• **Minor/Major allele:** The minor allele is the less common allele in a population (frequency < 50%). MAF stands for Minor Allele Frequency.

• **Ancestral/Derived allele:** The ancestral allele is the allele that was present in the common ancestor of humans (this can be inferred if one allele matches the nucleotide found at this position in other great apes), while the derived allele is inferred to have arisen by mutation within the human population.

Hardy-Weinberg Equilibrium



How many SNPs are there?

- Since you inherited one copy of each chromosome from your mum, and one from your dad, o**ne way to measure genetic diversity is to count up how many differences you have between these two genome copies**.
- Any difference between the two genomes – for example, you got an A from mum and a G from dad at a particular position – is a **heterozygous SNP**
- how frequently would you find heterozygous SNPs between the homologous copies of your genome?
- **you can expect to find a heterozygous SNP about once every 1,000–2,000 bp, depending on your ancestry.**

Heterozygosity

The fraction of heterozygous sites is referred to as **heterozygosity**, and is a useful measure of genetic diversity

.

Region	Population	Heterozygosity × 1000
Africa	San	0.95
	Yoruba	0.96
	Maasai	0.93
	Mbuti	0.91
Near East	Palestinian	0.73
	Iranian	0.71
Europe	Spanish	0.69
	Polish	0.67
South Asia	Punjabi	0.71
	Bengali	0.72
East Asia	Thai	0.69
	Japanese	0.67
Oceania	Australian	0.63
	Papuan	0.58
Americas	Inuit	0.63
	Surui	0.50

How many SNPs in your genome?

Since your genome is about 3.2 billion basepairs, this table implies that **you have about 1.5–3.0 million heterozygous sites, depending on your ancestry**.

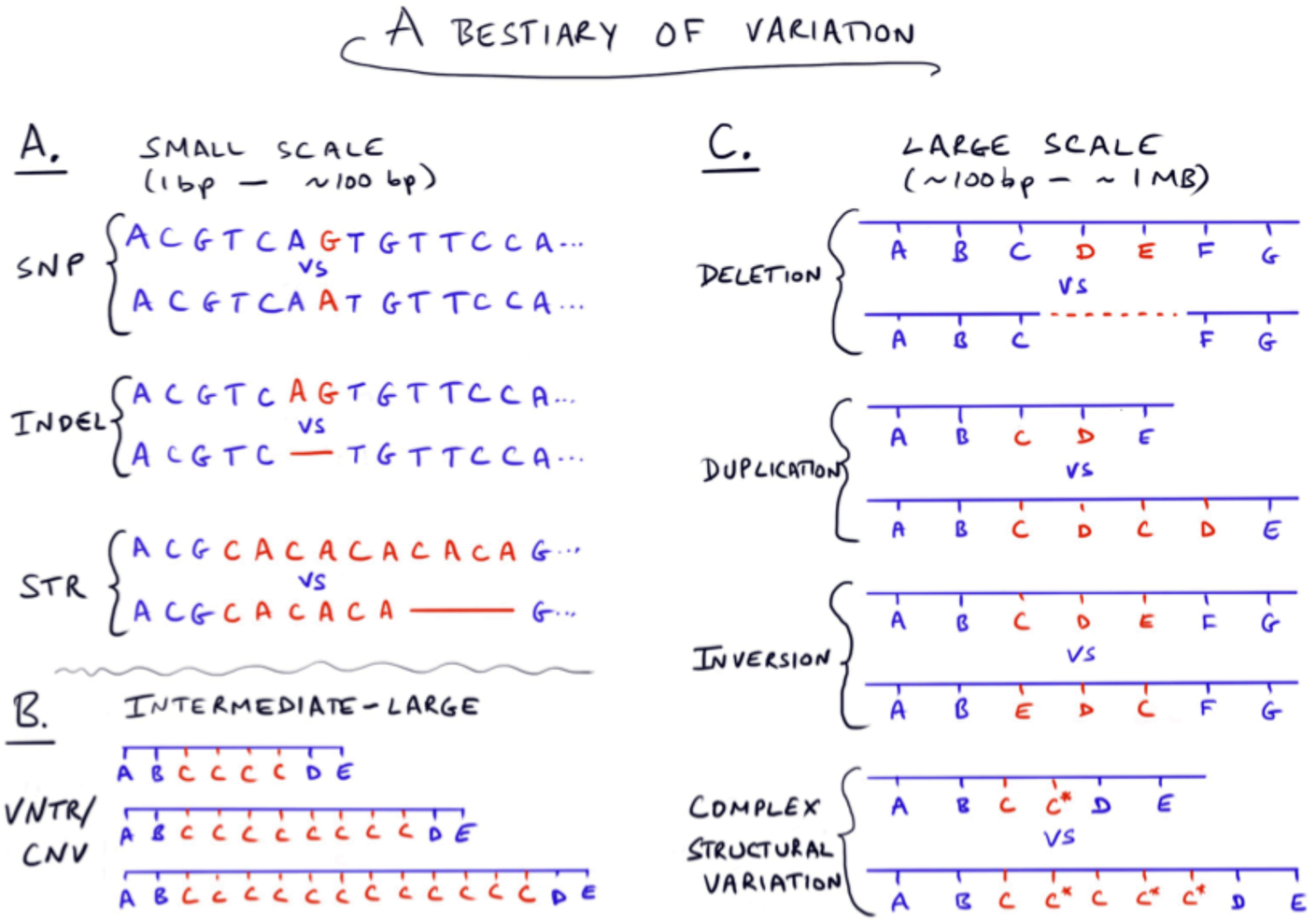
What if we look at SNPs **in a larger number of individuals?** For example, the 1000 Genomes Project sequenced the genomes of 2500 individuals from a diverse set of global populations. They reported 85 million SNPs, most of which were very rare: 65 million were below frequency 0.5%; 12 million were between 0.5%–5%, and 8 million SNPs were above 5%. In other words, **there is a common SNP with frequency>5% about once per 400 bp**.

nearly every possible SNP allele exists somewhere in the world

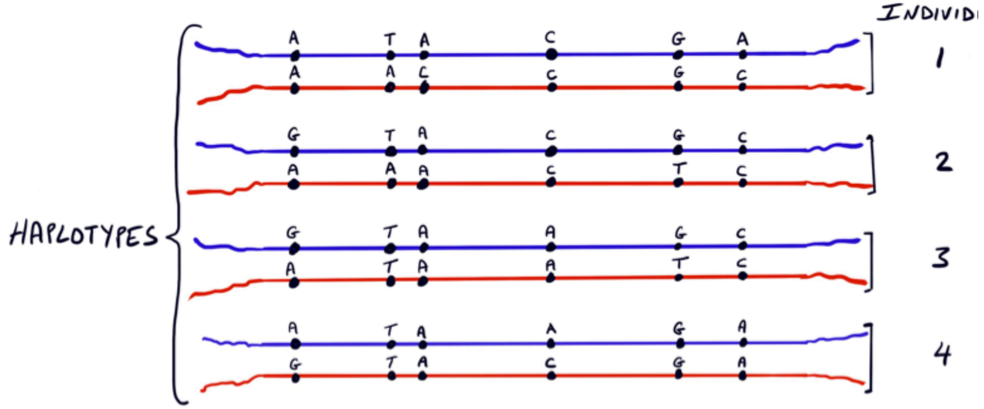
we should expect to find that nearly every possible SNP allele exists somewhere in the world. The world population is nearly 10 billion people and, as we'll see

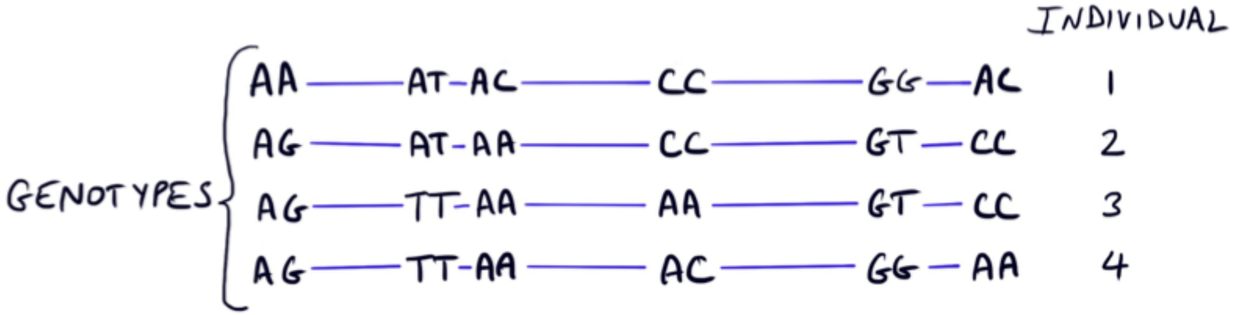
Mutation rate is around 10^{-08} per nucleotide per generation. This implies that nearly every possible single nucleotide variant must occur many times each generation, somewhere in the world.

Other types of genetic variations

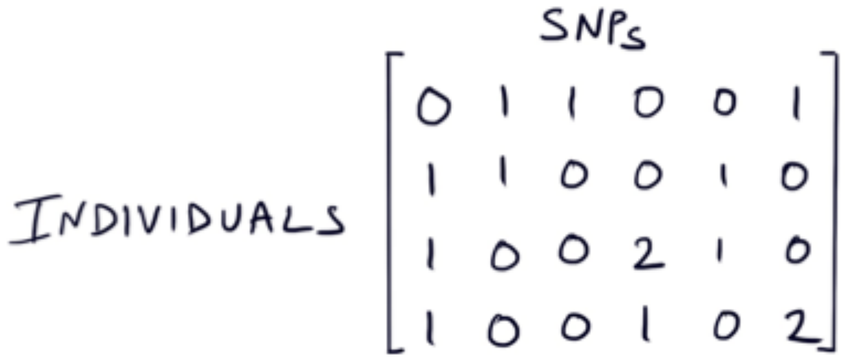


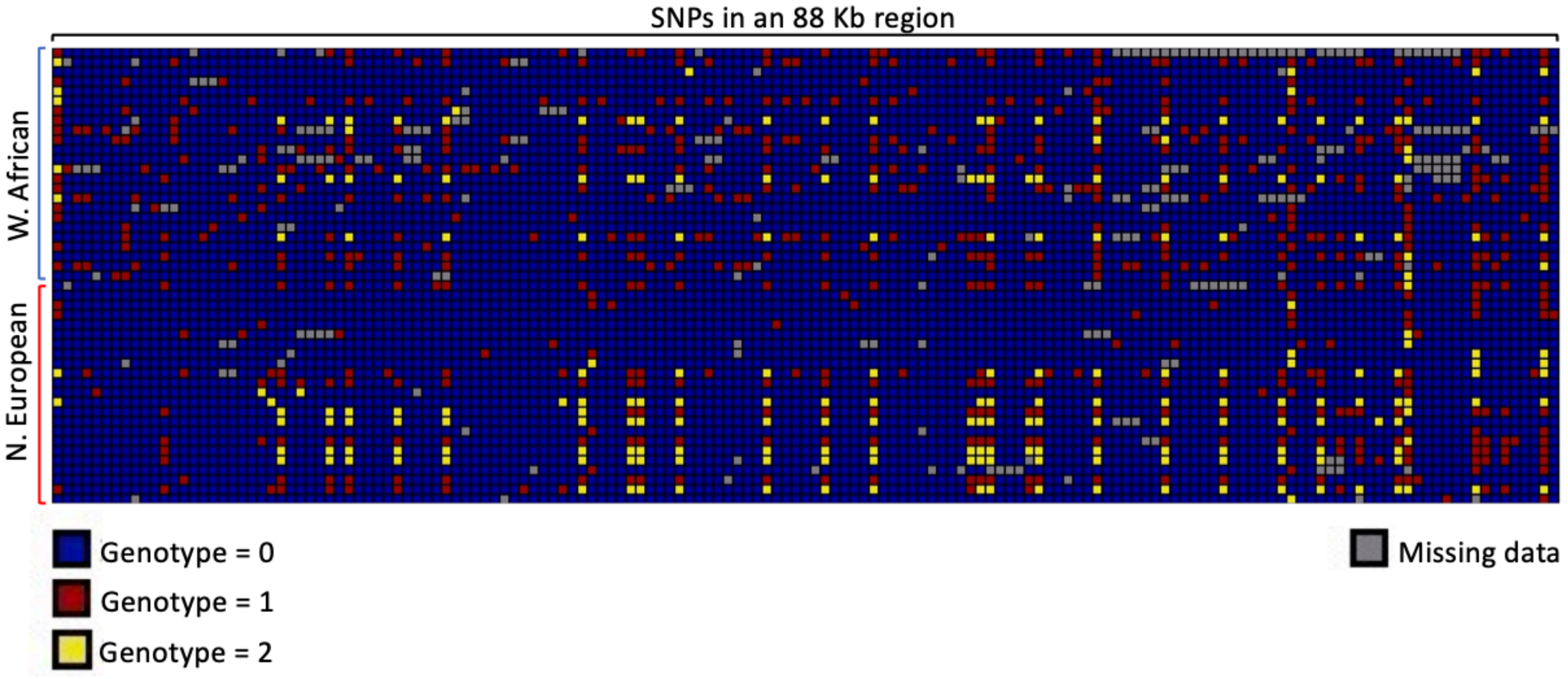
Genotypes and haplotypes





Genotype Matrix representation

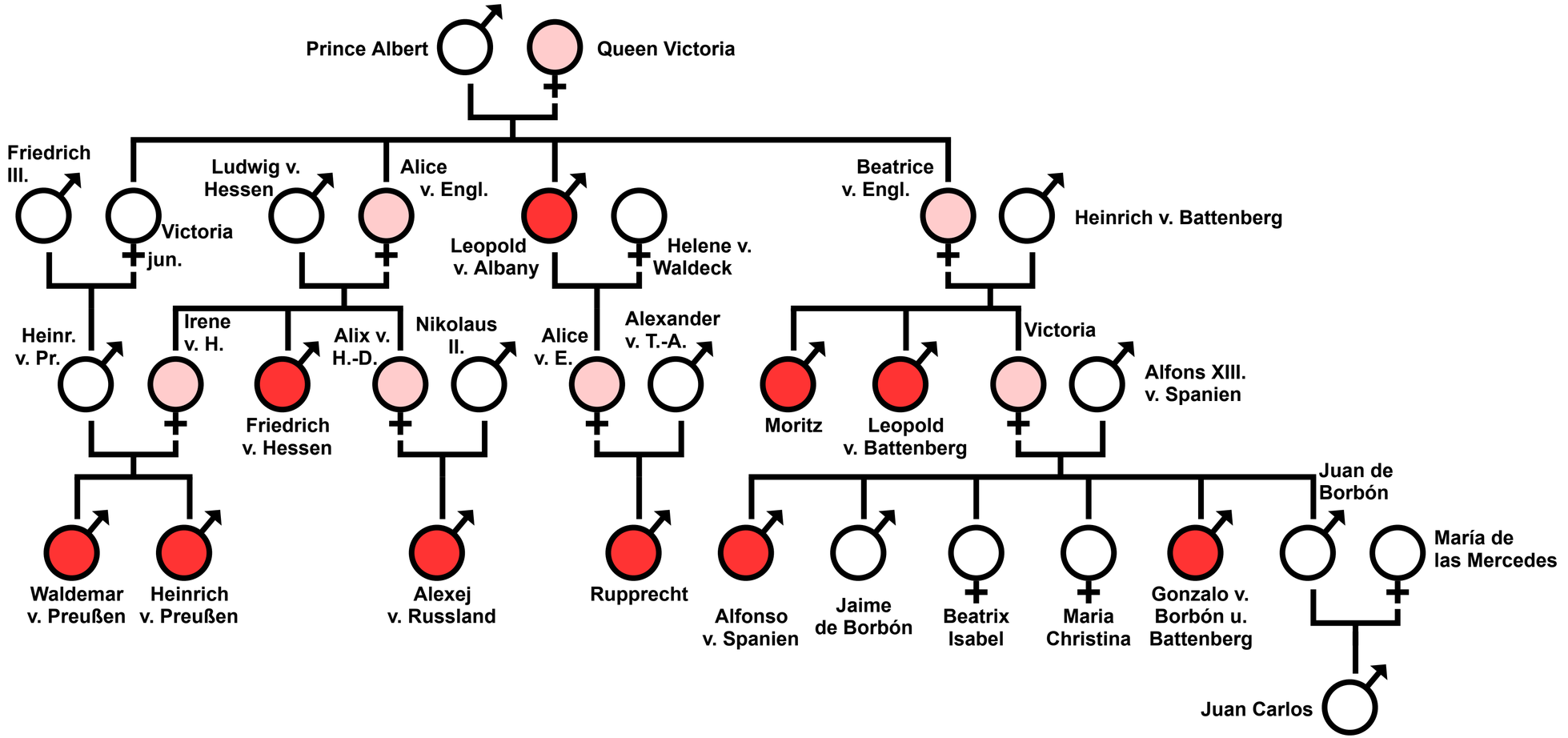



SNPs in an 88 Kb region
W. African
N. European
Genotype = 0
Genotype = 1
Genotype = 2
Missing data

Example: hemophilia in the royal families of Europe.

Hemophilia is a genetic disease, caused by mutations in either of two X chromosome genes that are essential for normal blood clotting. Since males only have one X chromosome, any male with the mutation will have the disease. In contrast, females with one copy of the mutation do not have the disease, but can transmit to their children. Prior to modern treatments, affected individuals often died at young ages.

The 19th century British queen, **Queen Victoria** (1819-1901), is the first person in her family known to have carried the mutation. One of her three sons had the disease; he and two of Victoria's daughters passed it into the royal families of Spain, Germany and Russia. Ultimately, eleven male-line descendants of Victoria had hemophilia, spread across 4 generations. Victoria's last known descendent with hemophilia died in 1945.

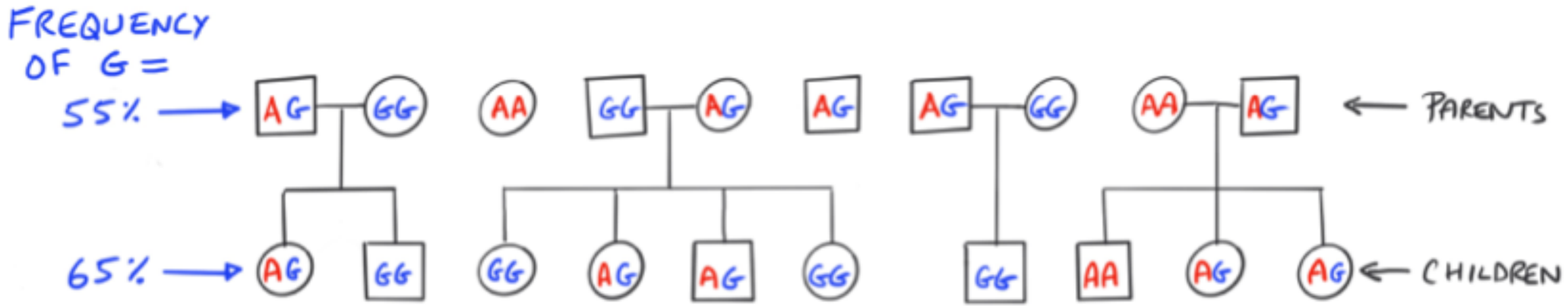


Each child of a carrier will have a 50% chance of inheriting their mother's mutation, of being a haemophiliac (sons) or carrier (daughters). The daughter of a male haemophiliac will always inherit his mutation, while a son cannot ever inherit it.

We pick up the story with one of Victoria’s great-grandsons, the Tsarevich Alexei Nikolaevich, born in 1904 as heir apparent to the Russia throne. Alexei inherited the hemophilia mutation from his mother, the Tsarina Alexandra. He almost died from blood loss at birth, and suffered throughout his life from dangerous hemorrhages resulting from the minor bumps and bruises of childhood. After the Russian Revolution of 1917, Alexei and his family were exiled to Siberia.

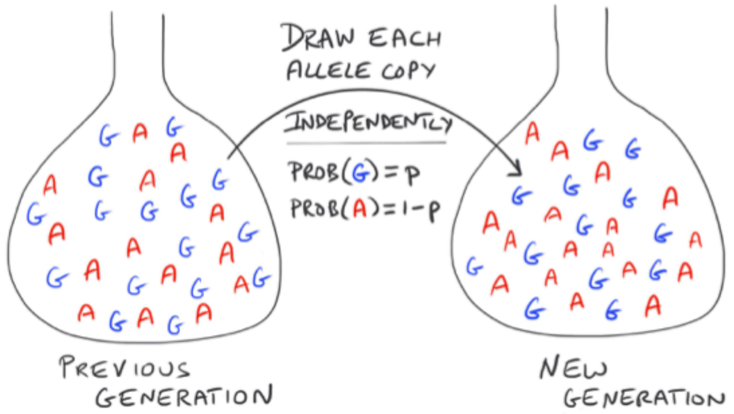
The following year the Bolsheviks executed the entire family. Much later, amid persistent rumors that Alexei and one of his sisters had escaped, the remains were exhumed and eventually subjected to genetic analysis in the mid-1990s that confirmed their identities.

Genetic drift

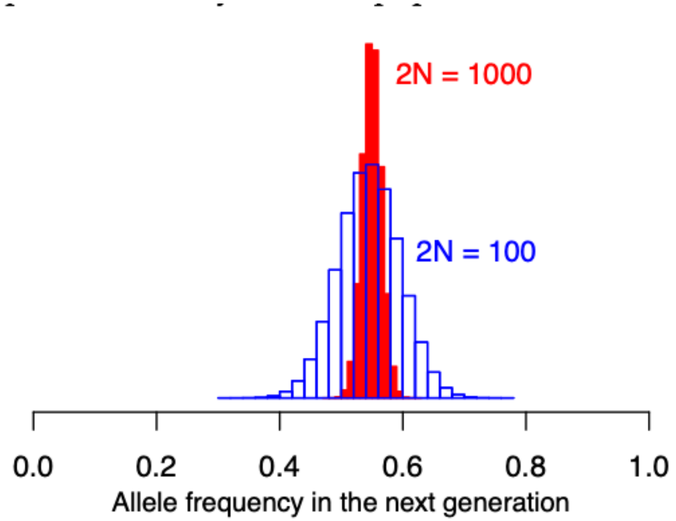


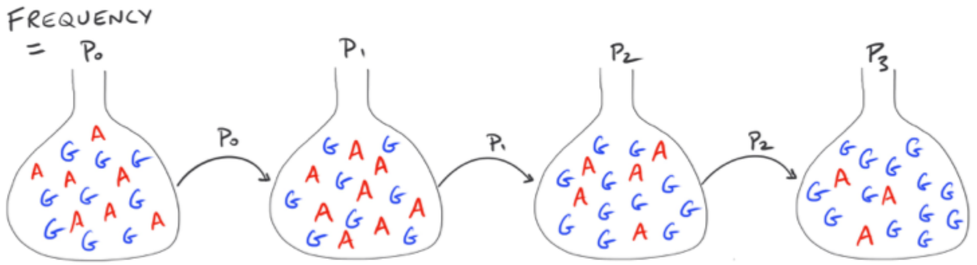
Modelling genetic drift

- The Wright Fisher model provides a framework for modeling how allele frequencies change over time
- We start by assuming a population with *N* individuals (2*N* copies of each locus). We assume that there are discrete generations, and that the *N* individuals mate at random to generate *N* individuals who form the next generation
- *sampling with replacement.*

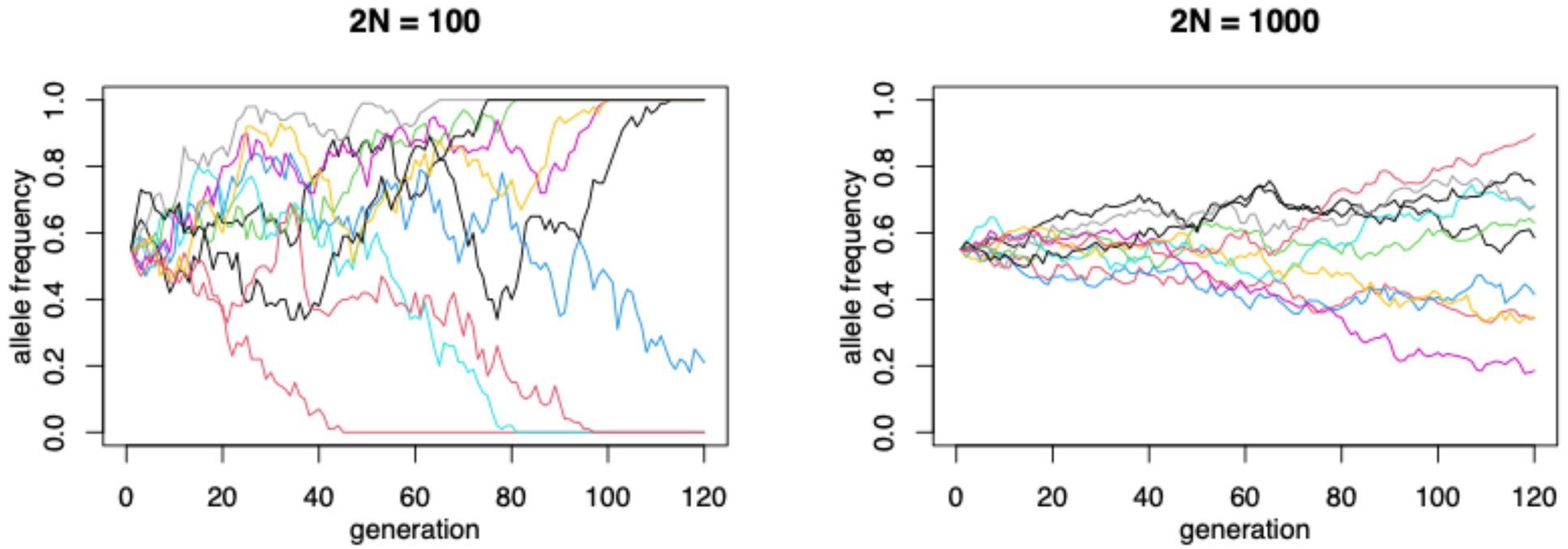


Binomial distribution



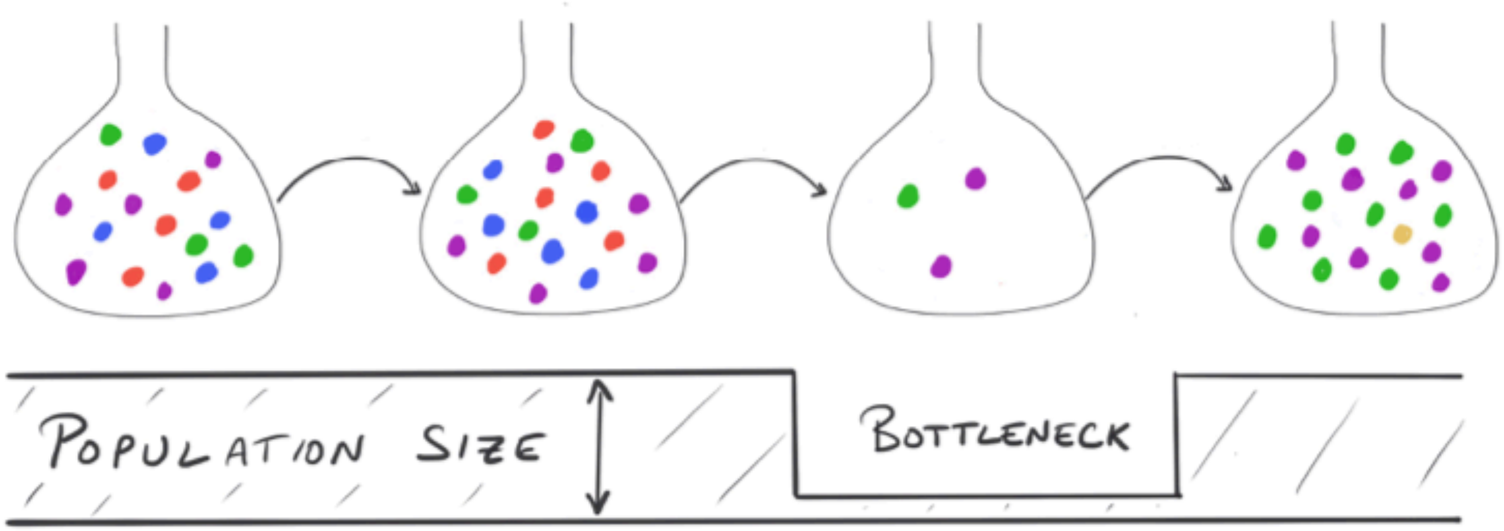


Simulation of genetic drift



Genetic drift under bottleneck

decrease:



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

An intriguing new genetic study suggests that around 900,000 years ago our ancestors dwindled to just 1,280 breeding individuals during that era, and they nearly vanished from the earth long before modern humans even appeared on the scene.

SCIENCE

The New York Times

ORIGINS

Humanity's Ancestors Nearly Died Out, Genetic Study Suggests

The population crashed following climate change about 930,000 years ago, scientists concluded. Other experts aren't convinced by the analysis.

Science

Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 381, NO. 6661 > GENOMIC INFERENCE OF A SEVERE HUMAN BOTTLENECK DURING THE EARLY TO MIDDLE PLEISTOCENE TRANSITION

RESEARCH ARTICLE | HUMAN EVOLUTION

Genomic inference of a severe human bottleneck during the Early to Middle Pleistocene transition

WANGJIE HU, ZIQIAN HAO, PENGYUAN DU, FABIO DI VINCENZO, GIORGIO MANZI, JIALONG CUI, YUN-XIN FU, YI-HSUAN PAN, AND HAIPENG LI

Out of Africa

Humans initially spread out of Africa through the Middle East, ranging further north into Europe, east across Asia and south to Australasia. Later, they eventually spread north-east over the top of Beringia into the Americas.

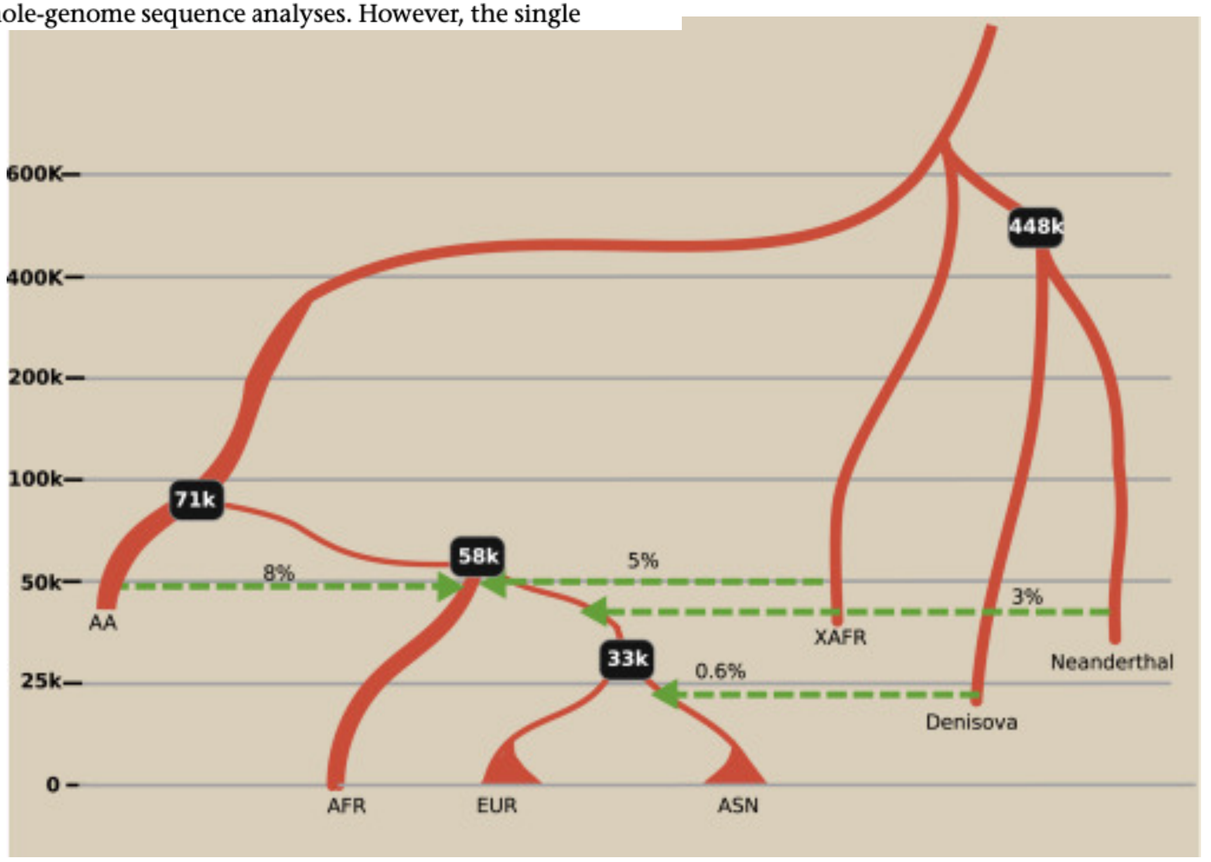
Around 60,0000 years ago a small number of Africans moved out of the continent – taking a much reduced gene-pool with them**. This genetic bottleneck, and the subsequent growth of non-African populations, meant that there was less genetic diversity to go round, and so there are fewer differences, on average, between the genomes of non-Africans compared to Africans.**

Revisiting the out of Africa event with a deep-learning approach

Francesco Montinaro,[1,2,6] Vasili Pankratov,[1,6] Burak Yelmen,[1,3,4] Luca Pagani,[1,5] and Mayukh Mondal[1,*]

Summary

Anatomically modern humans evolved around 300 thousand years ago in Africa. They started to appear in the fossil record outside of Africa as early as 100 thousand years ago, although other hominins existed throughout Eurasia much earlier. Recently, several studies argued in favor of a single out of Africa event for modern humans on the basis of whole-genome sequence analyses. However, the single



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)

Selection pressures due to diet.

- Diet has been a major driver of selection in recent human evolution. As humans spread around the globe to inhabit virtually all possible ecosystems, they were forced to learn to survive on a wide array of different foods. Further enormous shifts in diet were driven by the transition to **agriculture**, starting in the past 5,000-10,000 years, in many parts of the world.

- Several potential signals of selection have been hypothesized as relating to diet, including at the FADS locus, which is involved in metabolism of fatty acids and at Amylase1, which is involved in starch digestion

.

nature human behaviour

Explore content · About the journal · Publish with us · Subscribe

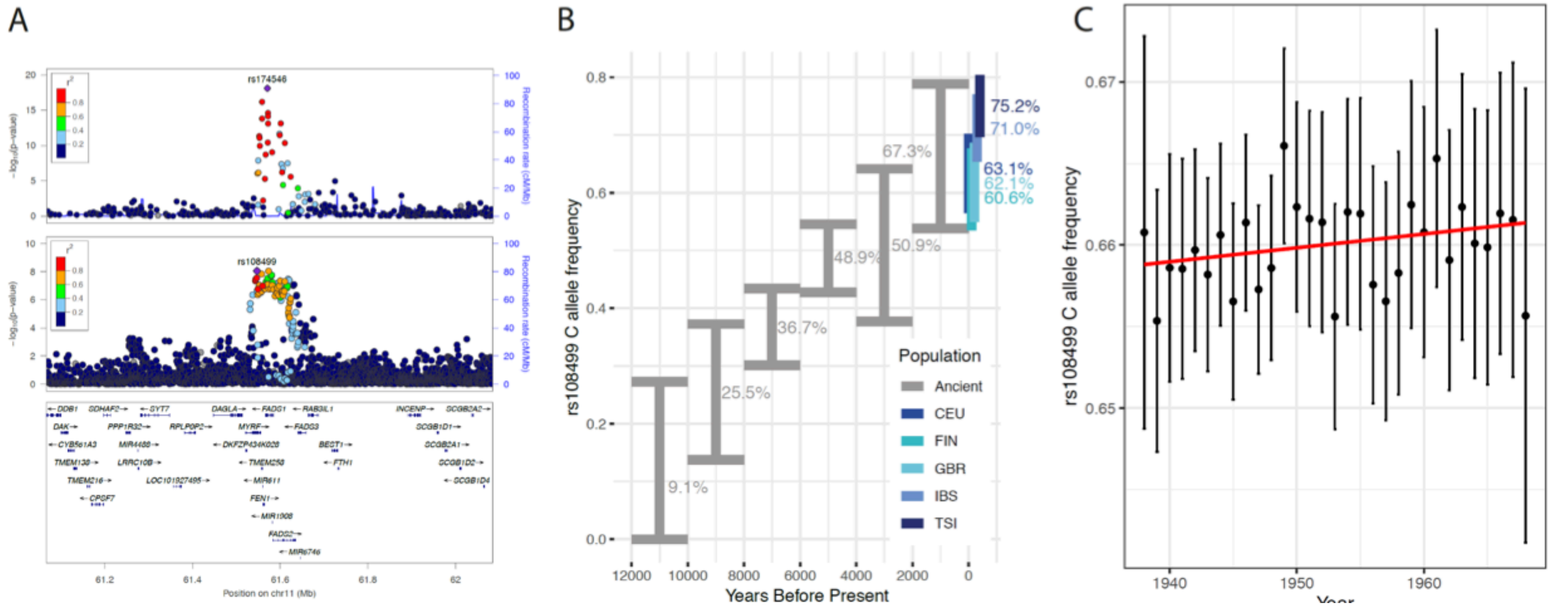
nature > nature human behaviour > articles > article

Article | Published: 02 March 2023

Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the *FADS* locus

Iain Mathieson, Felix R. Day, Nicola Barban, Felix C. Tropf, David M. Brazel, eQTLGen Consortium, BIOS Consortium, Ahmad Vaez, Natalie van Zuydam, Bárbara D. Bitarello, Eugene J. Gardner, Evelina Akimova, Ajuna Azad, Sven Bergmann, Lawrence F. Bielak, Dorret I. Boomsma, Kristina Bosak, Marco Brumat, Julie E. Buring, David Cesarini, Daniel I. Chasman, Jorge E. Chavarro, Massimiliano Cocca, Maria Pina Concas, FinnGen Study, Lifelines Cohort Study, ... John R. B. Perry + Show authors

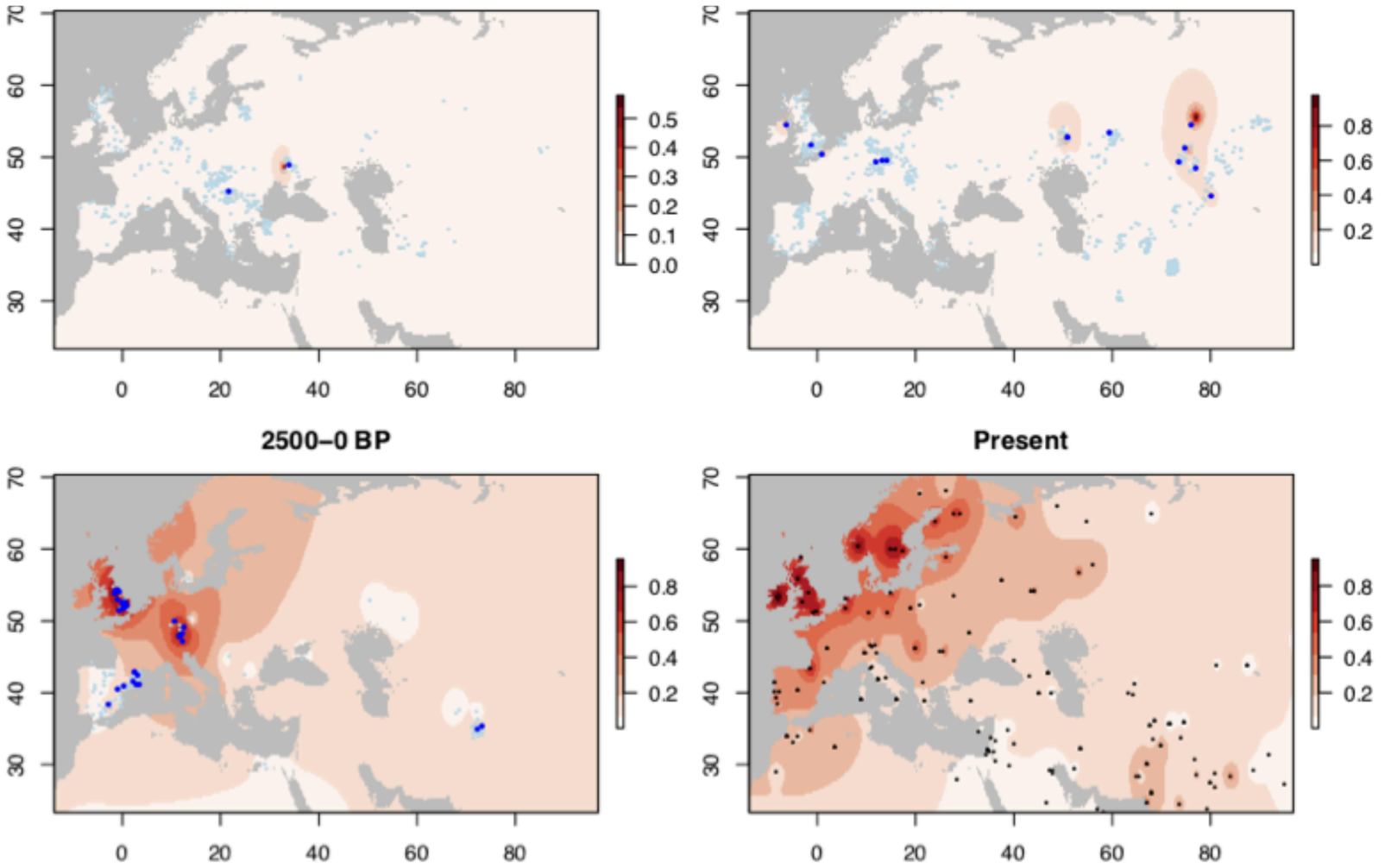
Figure 3 | Evidence for historical and ongoing selection at the FADS locus. A: Colocalization of the ancient DNA selection signal[17] (upper panel) and the NEB GWAS signal (lower panel). **B**: Estimated frequency (95% confidence intervals) for the derived *FADS* allele in Europe, based on direct evidence from ancient DNA. Present-day frequencies in 1000 Genomes European populations shown in blue. **C**: Frequency (95% confidence intervals) of the derived *FADS* allele in UK biobank as a function of birth year from 1938 to 1968.



In particular, the derived haplotype at this locus has increased from a frequency of <10% 10,000 years ago to 60-75% in present-day European populations (Figure 3B). While some of this increase is due to admixture, there is strong evidence of positive selection over the past few thousand years, even accounting for changes in ancestry

Lactase persistence

- The clearest diet-related signal is at the **lactase** locus. Lactase is the en-zyme that is responsible for digesting the sugar lactose, which is present in milk. Most mammals stop consuming milk (and lactose) after weaning, and expression of the lactase gene is generally turned off in adults.
- The first known evidence for dairy farming is in Anatolia (modern day Turkey) in the early Neolithic, about 9,000 years ago. Dairy farming sub- sequently became important in many places, including in Europe, in India and the Middle East, and in east Africa. This, in turn, provided
- a strong selective pressure for early humans to be able to digest milk throughout life. Consequently, several different **regulatory mutations** that cause the lactase gene to be expressed throughout life have spread to intermediate or high frequency in different farming populations. These regulatory mutations are often referred to as **lactase persistence** alleles as they cause lactase to persist throughout life.

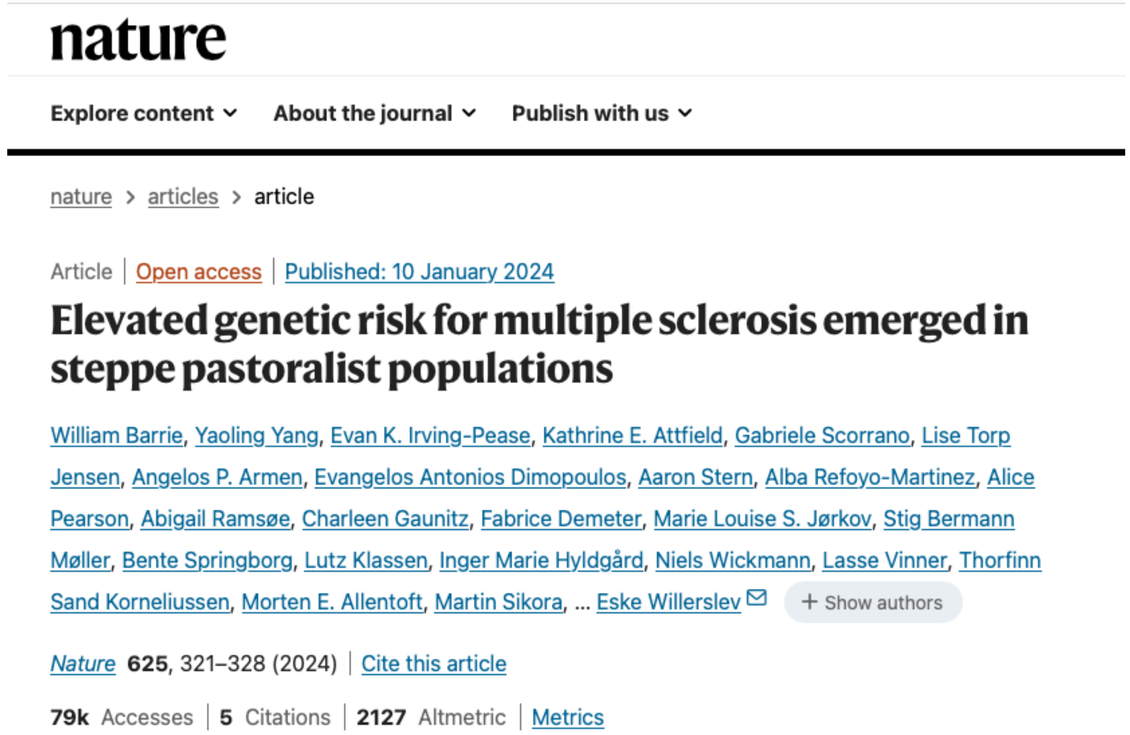


We propose that lactase non-persistent individuals consumed milk when it became available but, under conditions of famine and/or increased pathogen exposure, this was disadvantageous, driving LP selection in prehistoric Europe.

Evershed, R.P., Davey Smith, G., Roffet-Salque, M. *et al.* Dairying, diseases and the evolution of lactase persistence in Europe. *Nature* **608**, 336–345 (2022). https://doi.org/10.1038/s41586-022-05010-7

https://www.nature.com/articles/d41586-022-02068-1#MO0

Multiple sclerosis



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open access | Published: 10 January 2024

Elevated genetic risk for multiple sclerosis emerged in steppe pastoralist populations

William Barrie, Yaoling Yang, Evan K. Irving-Pease, Kathrine E. Attfield, Gabriele Scorrano, Lise Torp Jensen, Angelos P. Armen, Evangelos Antonios Dimopoulos, Aaron Stern, Alba Refoyo-Martinez, Alice Pearson, Abigail Ramsøe, Charleen Gaunitz, Fabrice Demeter, Marie Louise S. Jørkov, Stig Bermann Møller, Bente Springborg, Lutz Klassen, Inger Marie Hyldgård, Niels Wickmann, Lasse Vinner, Thorfinn Sand Korneliussen, Morten E. Allentoft, Martin Sikora, ... Eske Willerslev + Show authors

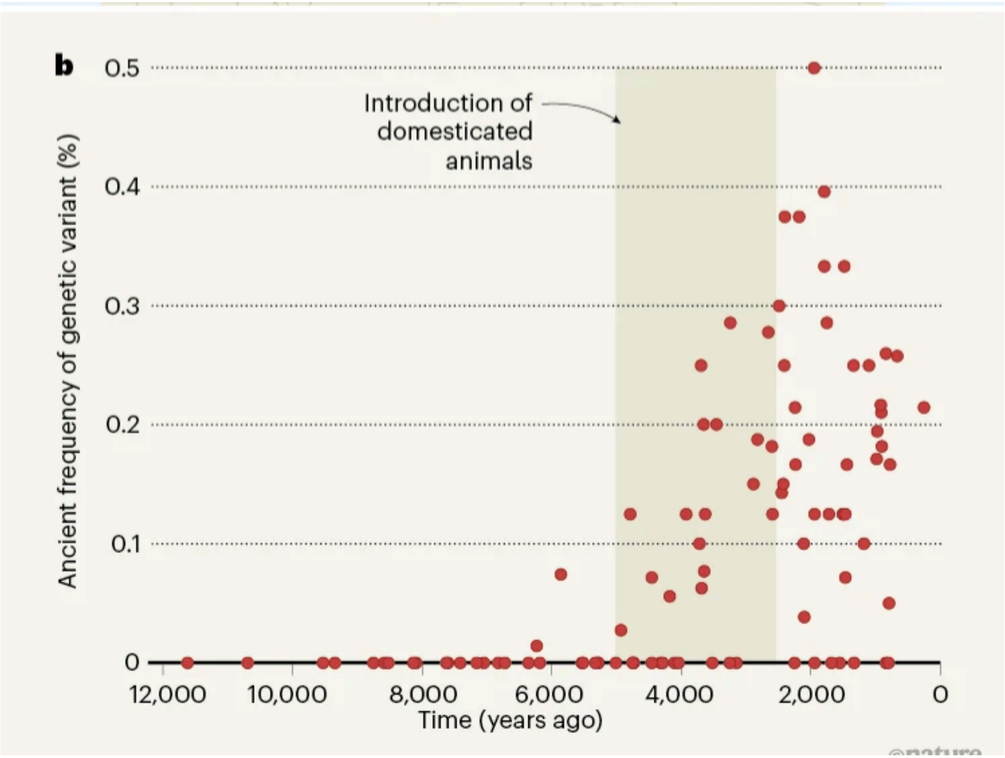
Nature **625**, 321–328 (2024) | Cite this article

79k Accesses | **5** Citations | **2127** Altmetric | Metrics

- Multiple sclerosis (MS) is a neuro-inflammatory and neurodegenerative disease that is most prevalent in Northern Europe
- a genetic variant on chromosome 6 called HLA-DRB1*15:01 is present in up to one-fifth of northern Europeans, and individuals with this variant have a threefold higher risk of developing multiple sclerosis compared with those who do not have the variant
- genetic risk for MS rose among pastoralists from the Pontic steppe and was brought into Europe by the Yamnaya-related migration approximately 5,000 years ago.

Why would a variant bearing such a health risk be favoured by evolution?

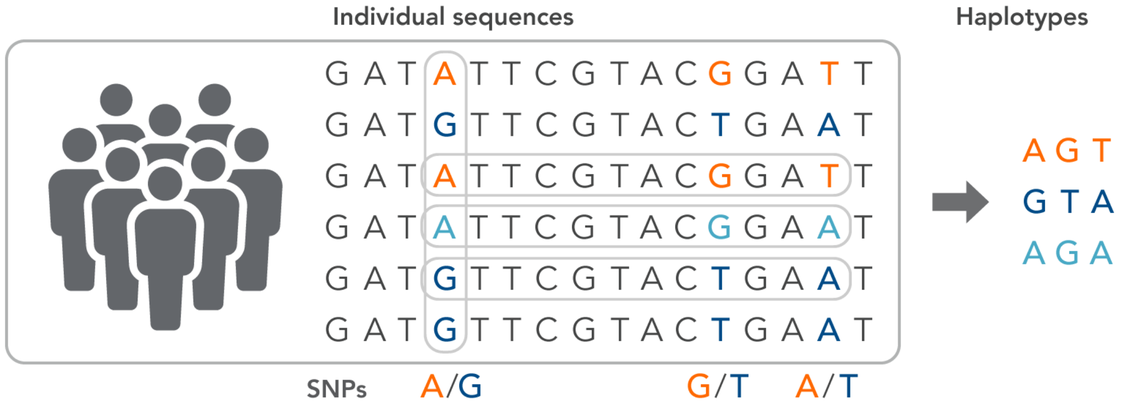
The authors suggest that HLA-DRB1*15:01 **could have protected ancient Europeans against pathogens that came with the shift from hunting and gathering to farming and pastoralism.** This could explain why the variant has been positively selected for in recent human evolution.



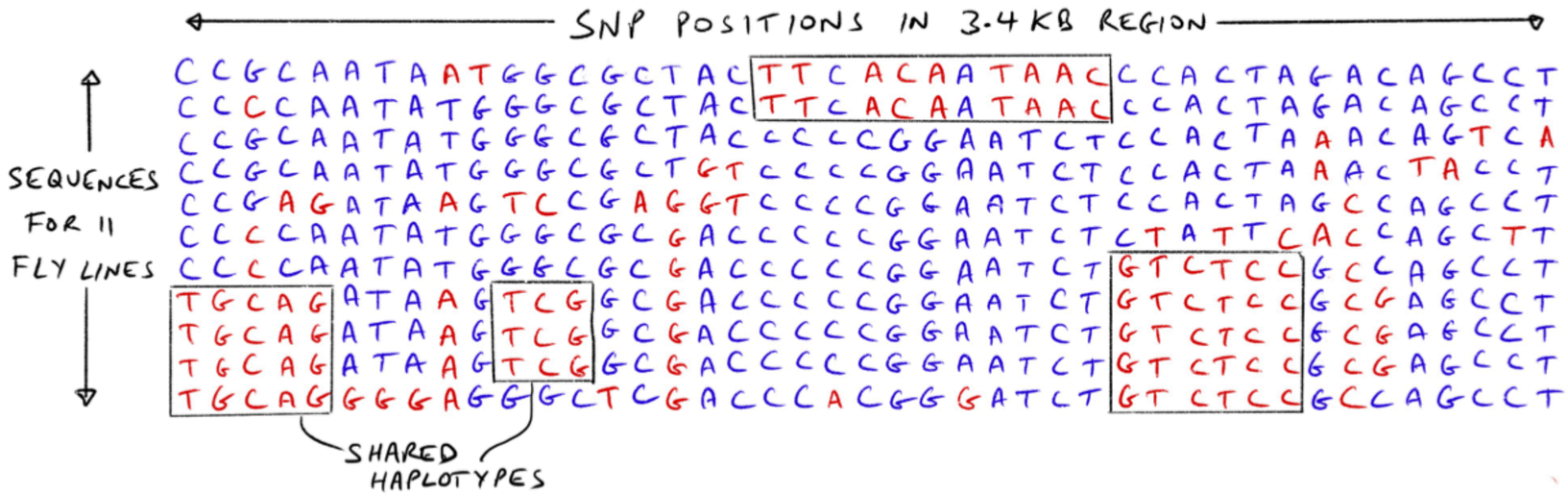
Correlations Among Loci

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.



combinations of alleles at different SNPs frequently appear together.



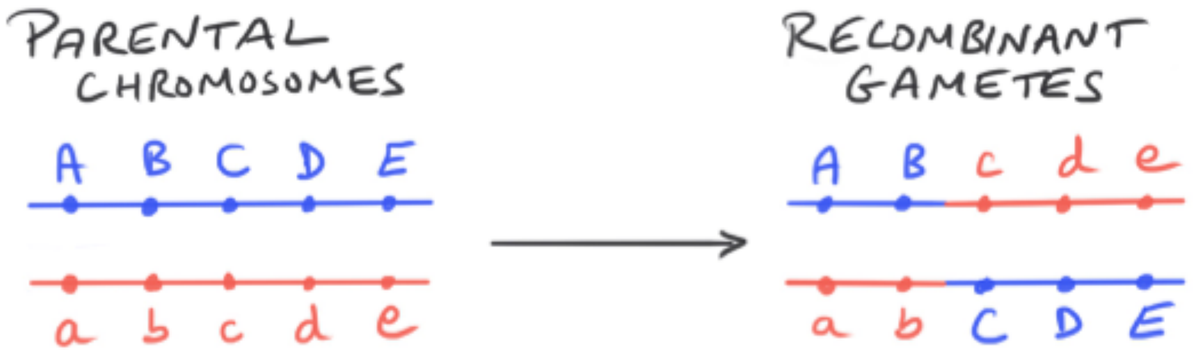
particular alleles at nearby SNPs often appear together more often than expected by chance. This nonrandom assortment of alleles at different sites is referred to as **linkage disequilibrium (LD).**

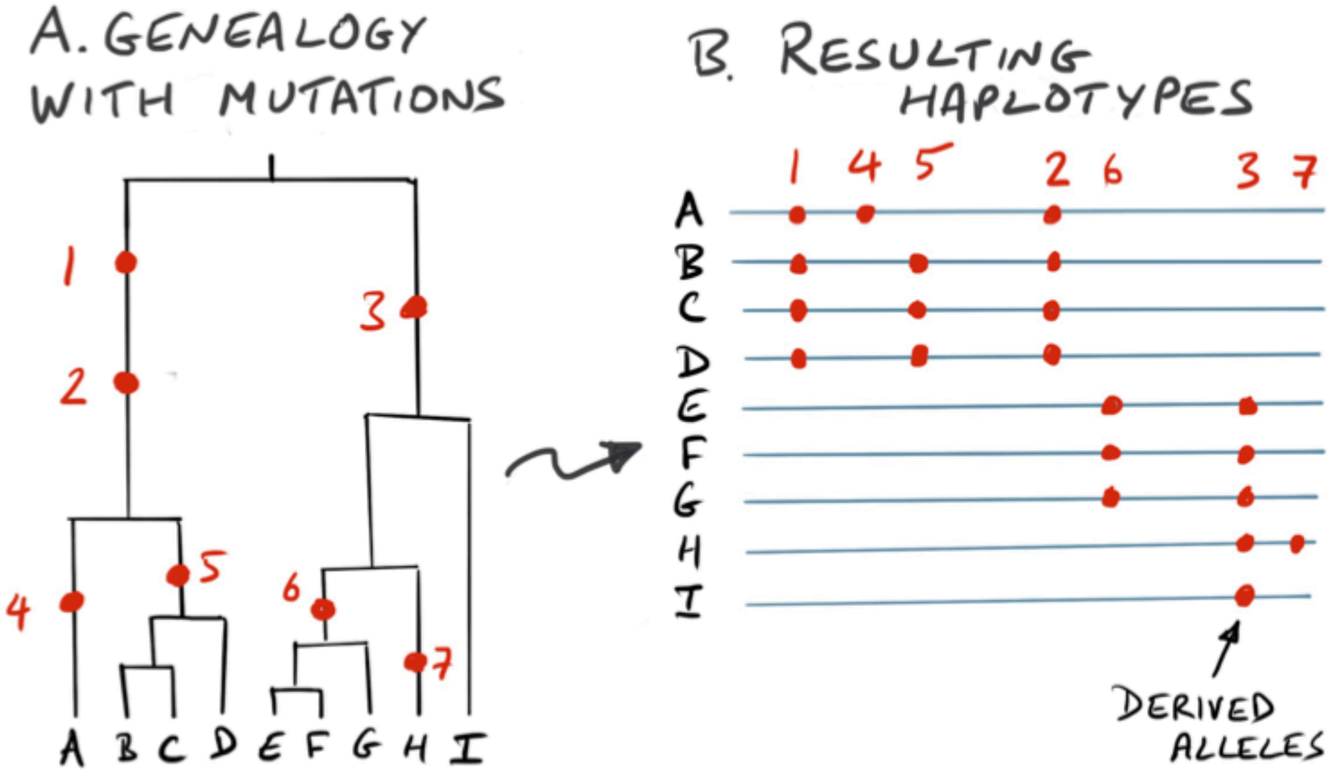
Recombination

- During the production of eggs and sperm, the chromosomes go through meiosis. In humans, this reduces the number of chromosomes from 46 to 23.
- During this process, the maternal and paternal chromosomes are broken and then joined back together so that chromosomes in the resulting gametes are mixtures of the parental chromosomes. This is called **recombination**, or **crossover**
- Crossover events are positioned more-or-less randomly across the genome with an average of 26 crossovers per sperm and 42 per egg
- **It's useful to remember that the human recombination rate is about 1% per megabase.**

.

Recombination mixes haplotypes




A. GENEALOGY WITH MUTATIONS
1
2
3
4
5
6
7
A B C D E F G H I
B. RESULTING HAPLOTYPES
1 4 5 2 6 3 7
A
B
C
D
E
F
G
H
I
DERIVED ALLELES

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 $p_A; p_a; p_B; p_b$

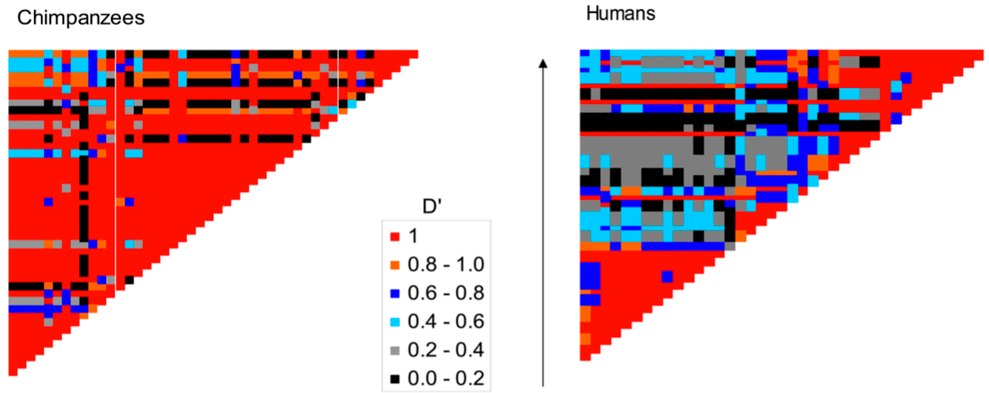
IF THEIR SEGREGATION IS INDEPENDENT: $p_{AB} = p_A\, p_B$*, OTHERWISE THE TWO LOCI ARE IN LINKAGE DISEQUILIBRIUM*

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.



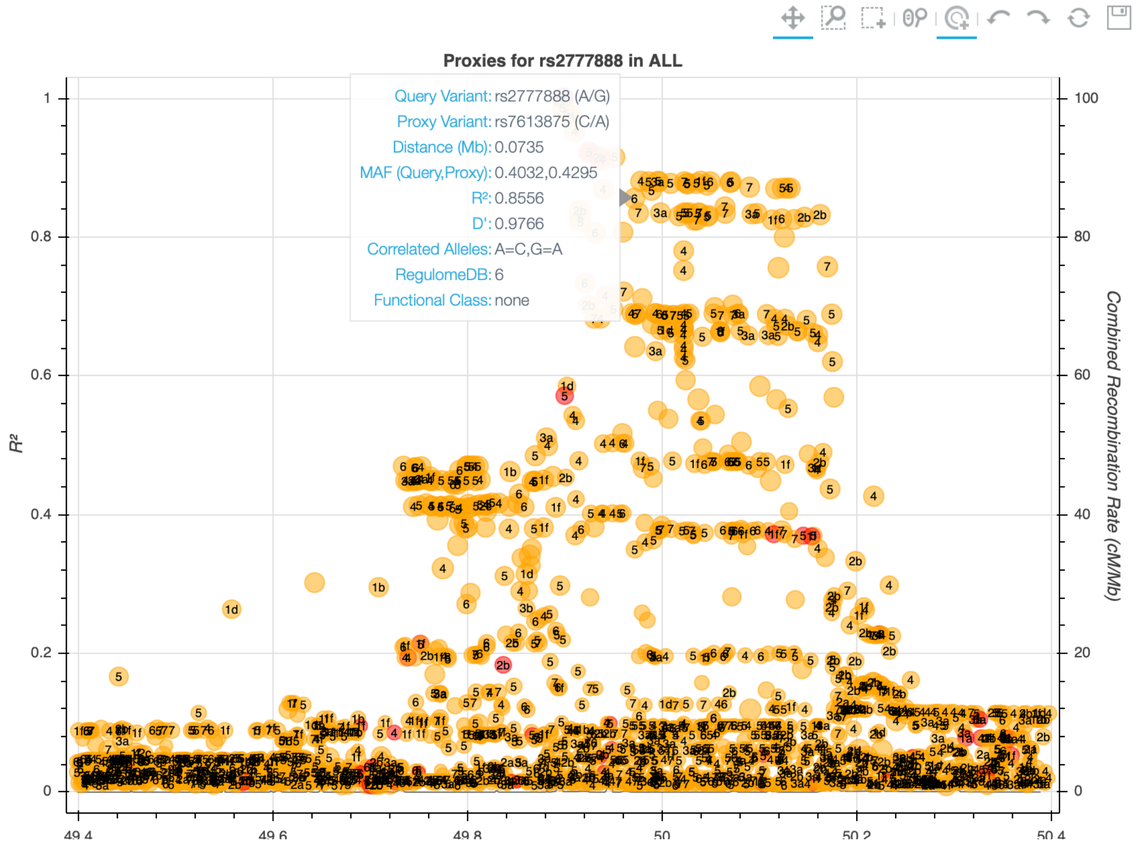
correlation coefficient

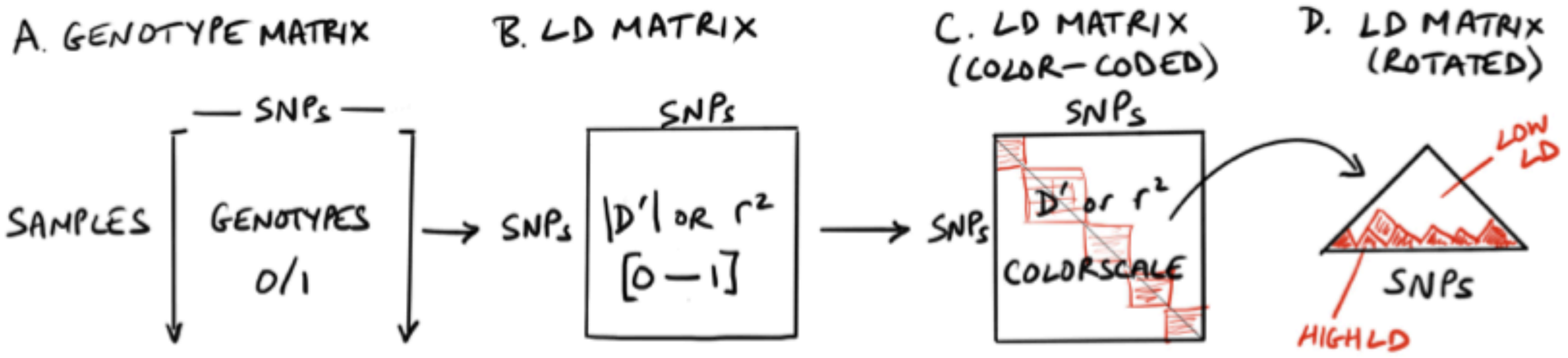
As D is a covariance, and $p_A(1 - p_A)$ is the variance of an allele drawn at random from locus A

.

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

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




A. GENOTYPE MATRIX
SNPs
SAMPLES
GENOTYPES
0/1
B. LD MATRIX
SNPs
SNPs
|D'| OR r²
[0–1]
C. LD MATRIX
(COLOR-CODED)
SNPs
SNPs
D' or r²
COLORSCALE
D. LD MATRIX
(ROTATED)
LOW
LD
SNPs
HIGH LD

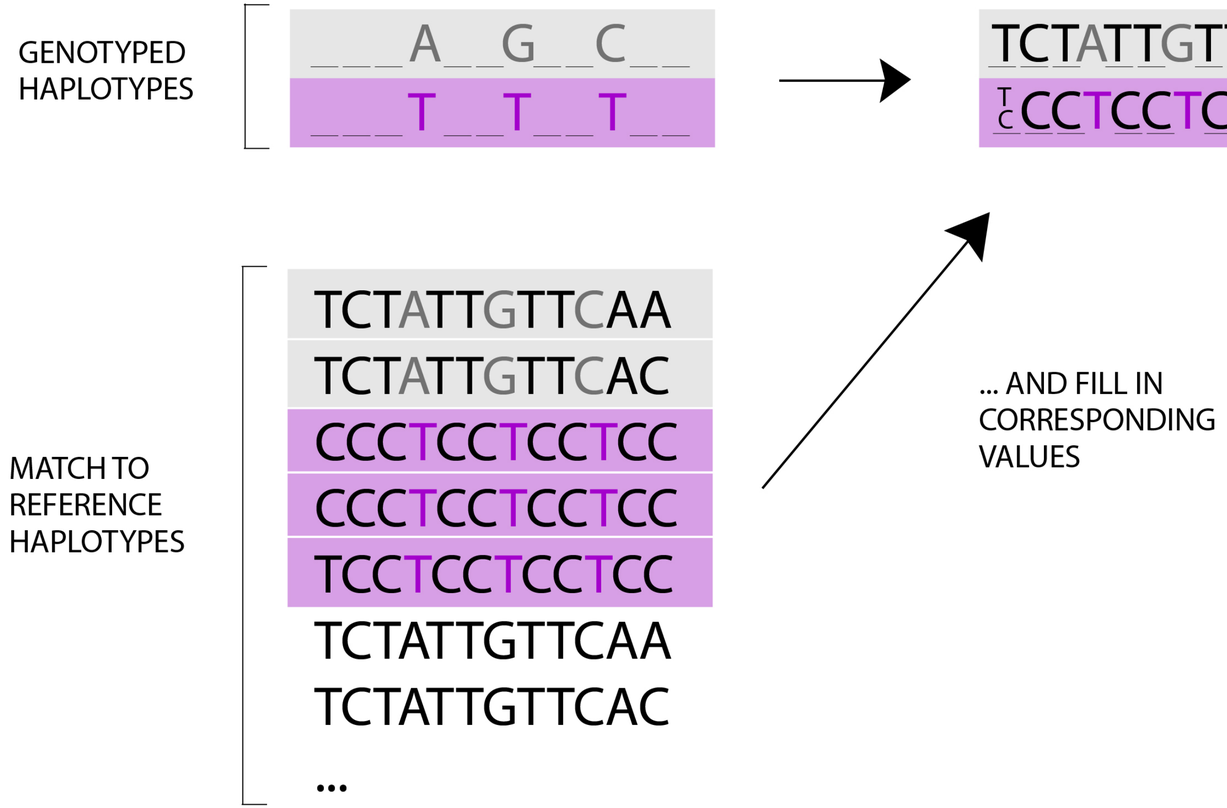
Why is it relevant?

1. Evolutionary biology

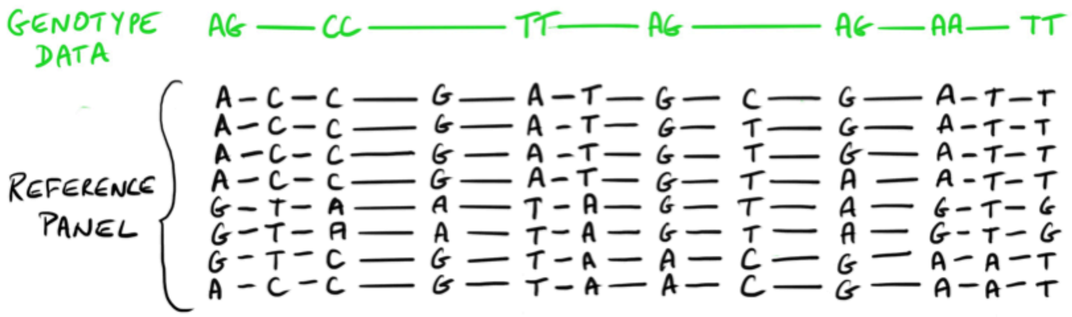
- LD is of importance in evolutionary biology provides information about past events and it constrains the potential response to both natural and artificial selection.
- LD in each genomic region reflects the history of natural selection, gene conversion, mutation and other forces that cause gene-frequency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

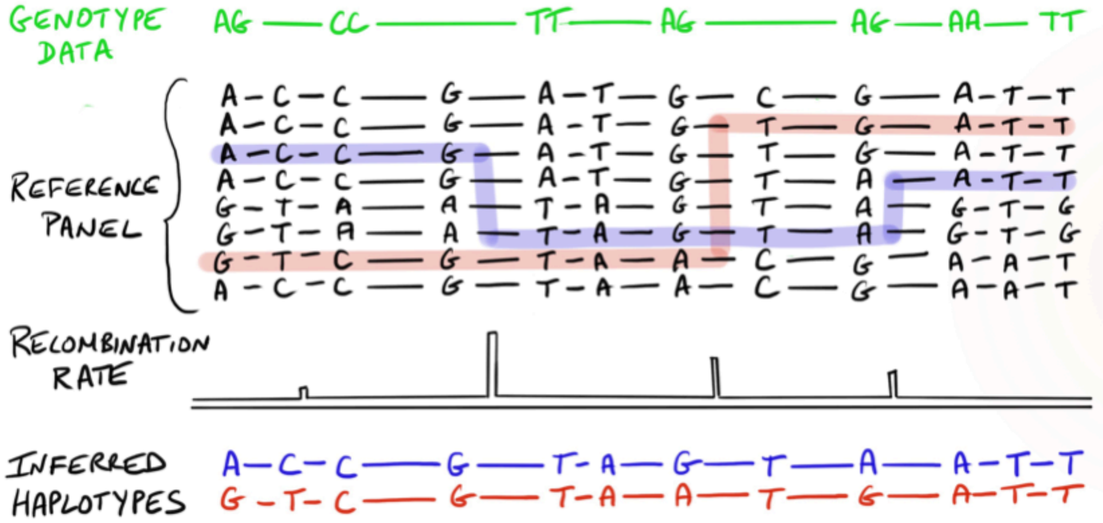
Genetic Imputation

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



Genetic imputation



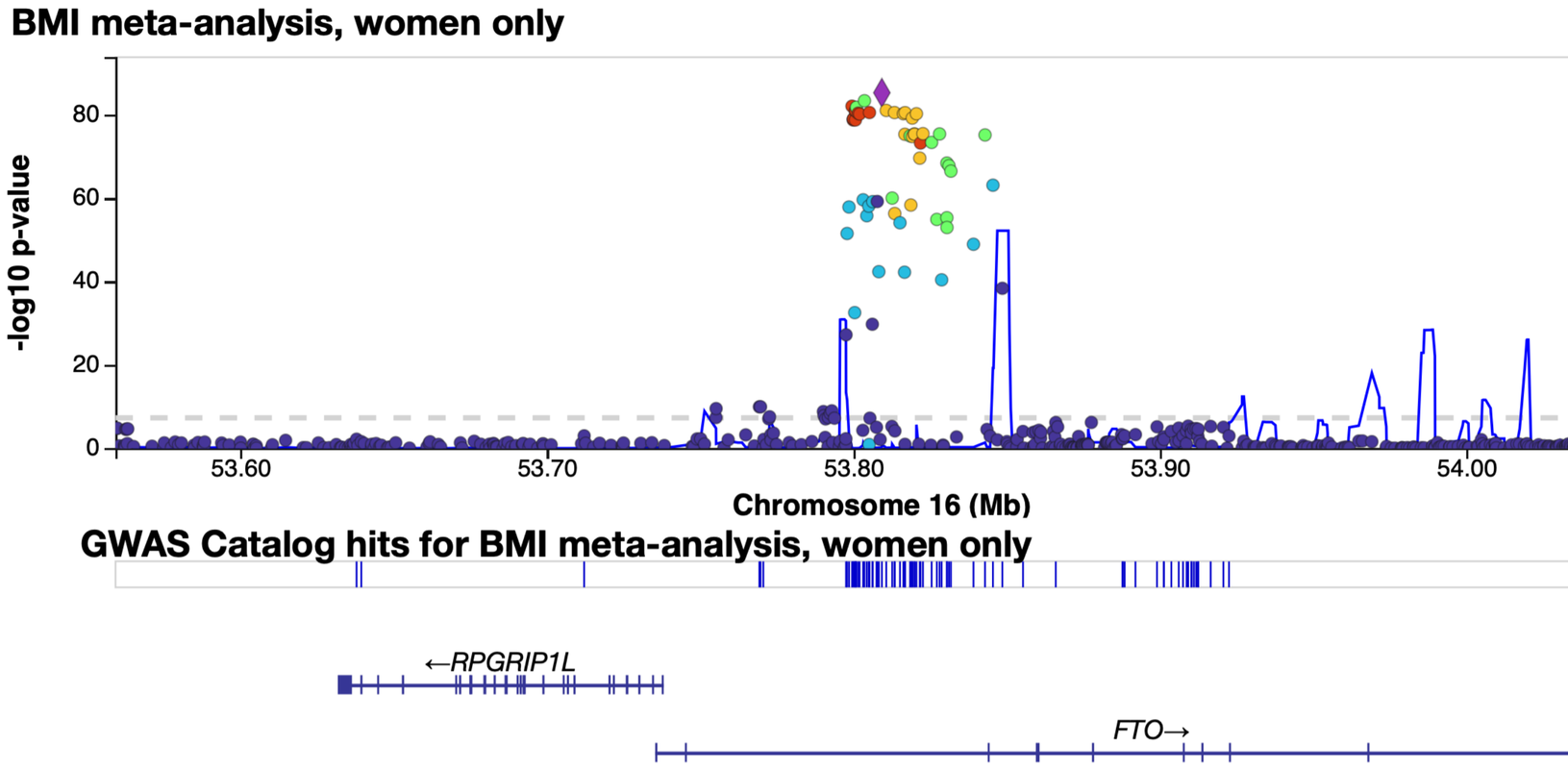


Reference Panels

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

LD and genetic association

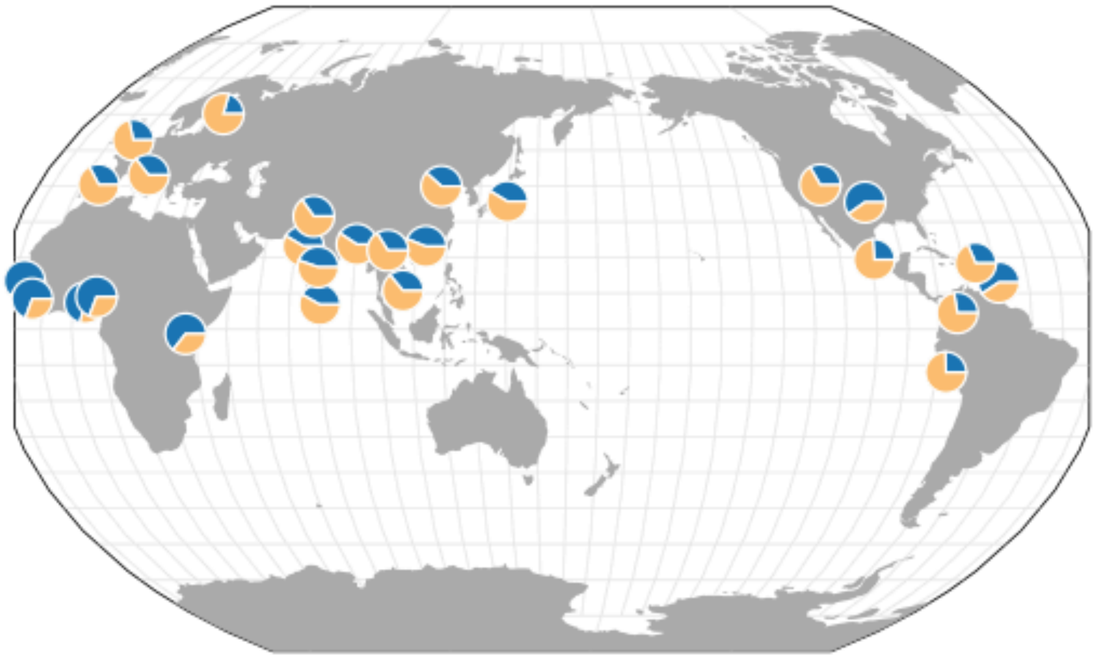
- The LD structure allows to identify “genomic regions” in association results.
- As it is not possible to genotype all genetic variants, we identify “markers” that can be in LD with the real causal variants.
-



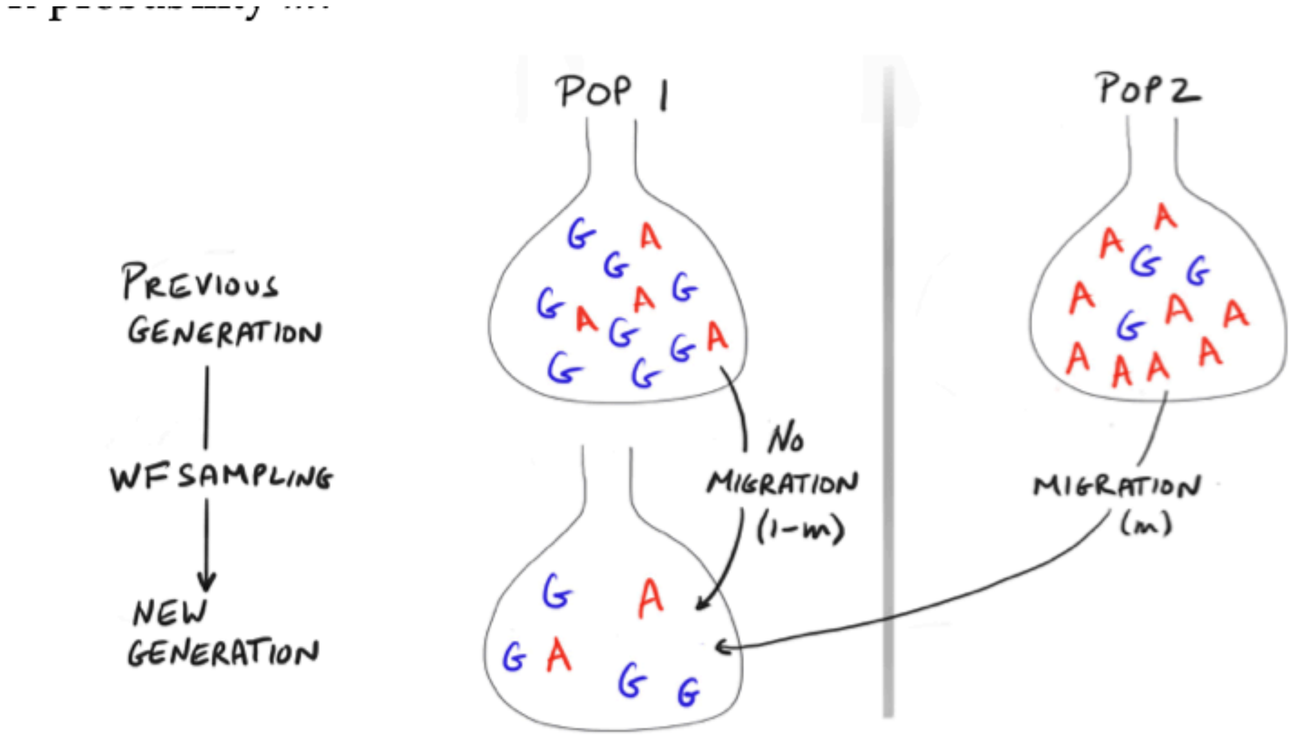
Consequences of Imputing genotypes

- 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

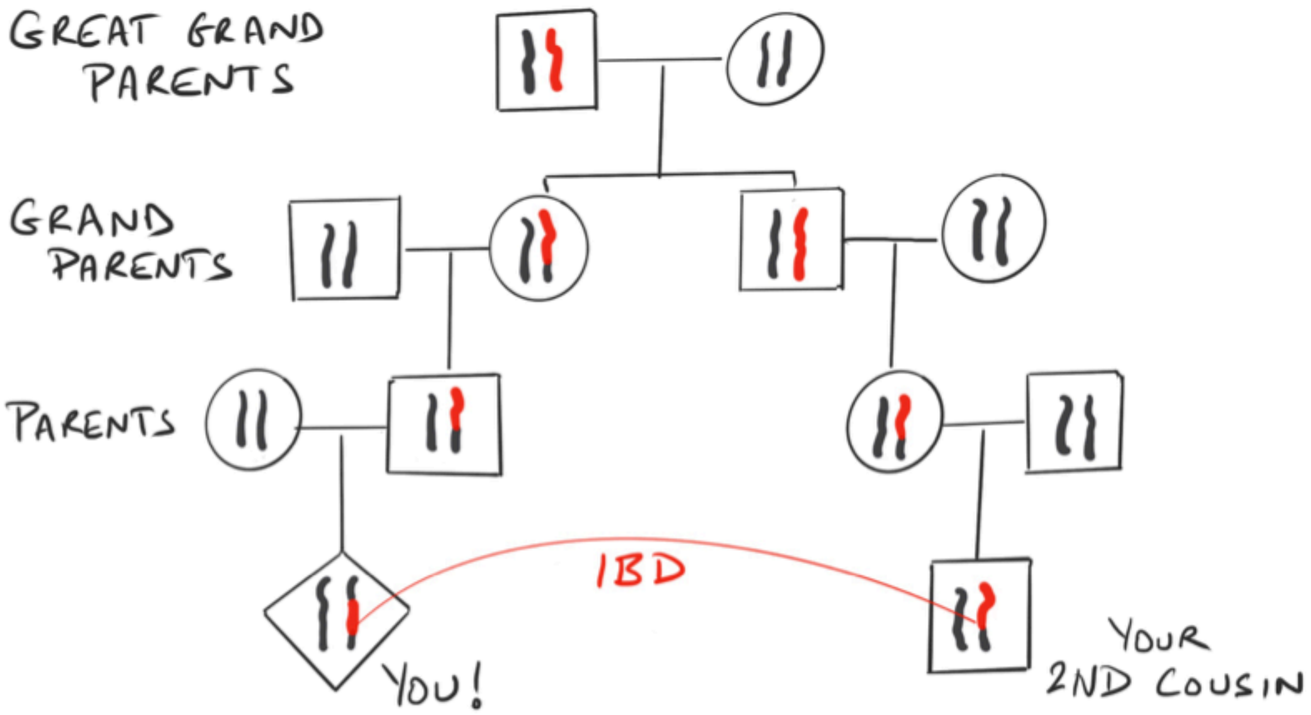
Allele frequency variation across populations.



Migration


PREVIOUS GENERATION
WF SAMPLING
NEW GENERATION
POP 1
POP 2
NO MIGRATION (1-m)
MIGRATION (m)

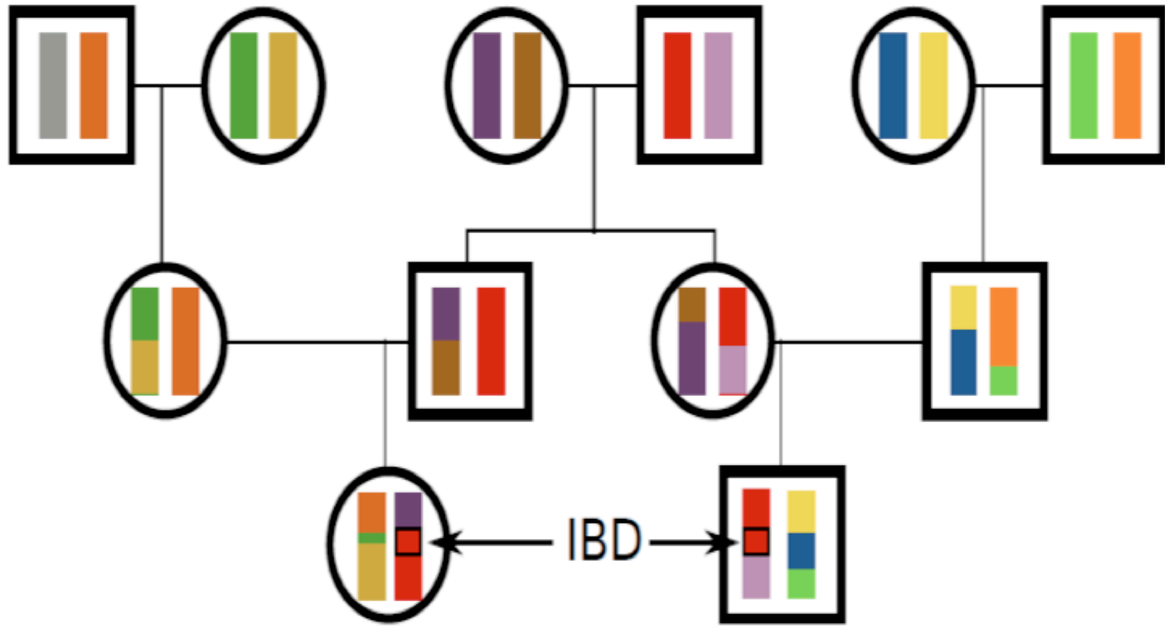
Measuring relatedness



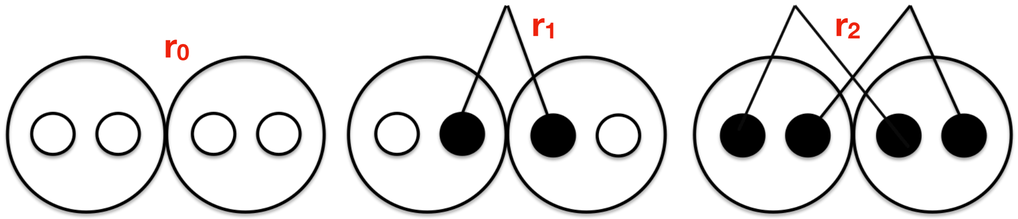
- This part of your genome **coalesced** with the corresponding part of your cousin's genome 3 generations ago
- The great-grandparent is your **common ancestor** at this locus.
- Coalescence means that this part of both your, and your cousin's genomes, are descended as copies of this ancestral genome

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

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.

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

Relationship with Coefficient of Relatedness (r):

The kinship coefficient represents the expected proportion of loci where the alleles are IBD. If Φ = 0.25, it means that, on average, 25% of randomly chosen alleles from the first individual are IBD with randomly chosen alleles from the second individual.

$$r = 2\Phi$$

Relationship	Kinship Coefficient (Φ)	Coefficient of Relatedness (r)
Identical Twins	0.5	1.0
Parent-Offspring	0.25	0.5
Full Siblings	0.25	0.5
Half Siblings	0.125	0.25
First Cousins	0.0625	125
Unrelated	0	0