BIOL 350: Bioinformatics

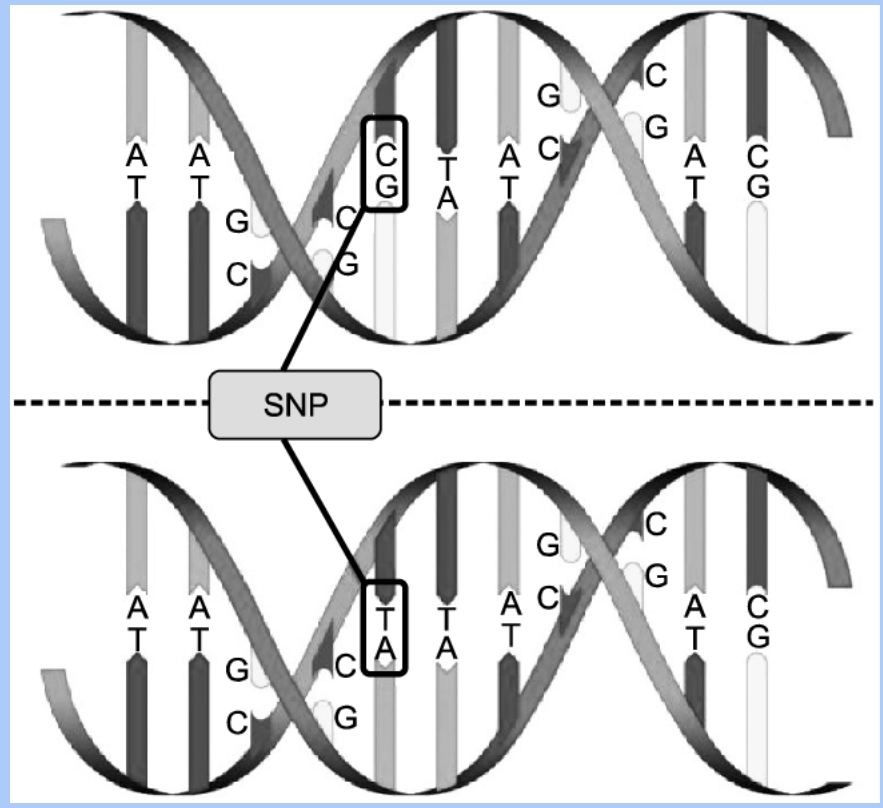
Introduction to genetic association studies

What is a SNP?

Polymorphisms and their role in genetics

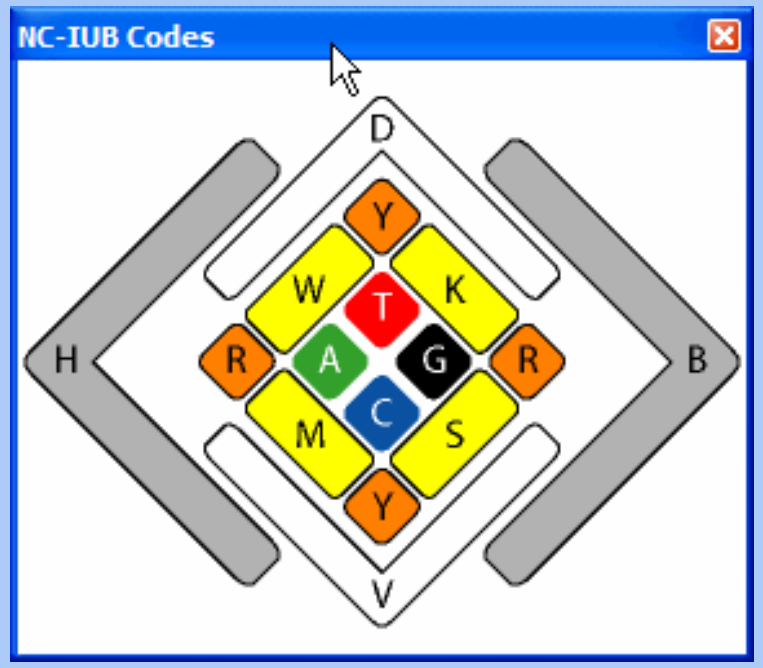
Single-nucleotide polymorphisms

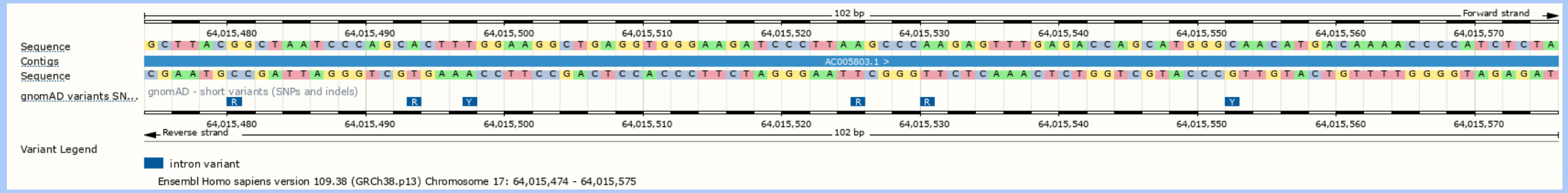
- **Polymorphism** is the tendency of DNA to admit different nucleotide pairs at a single locus
- Of 3.2 billion bases, any individual is polymorphic at 4–5 million sites
- The more common allele is called the **major allele**; the less common allele is called the **minor allele**



IUPAC-IUB SNP codes

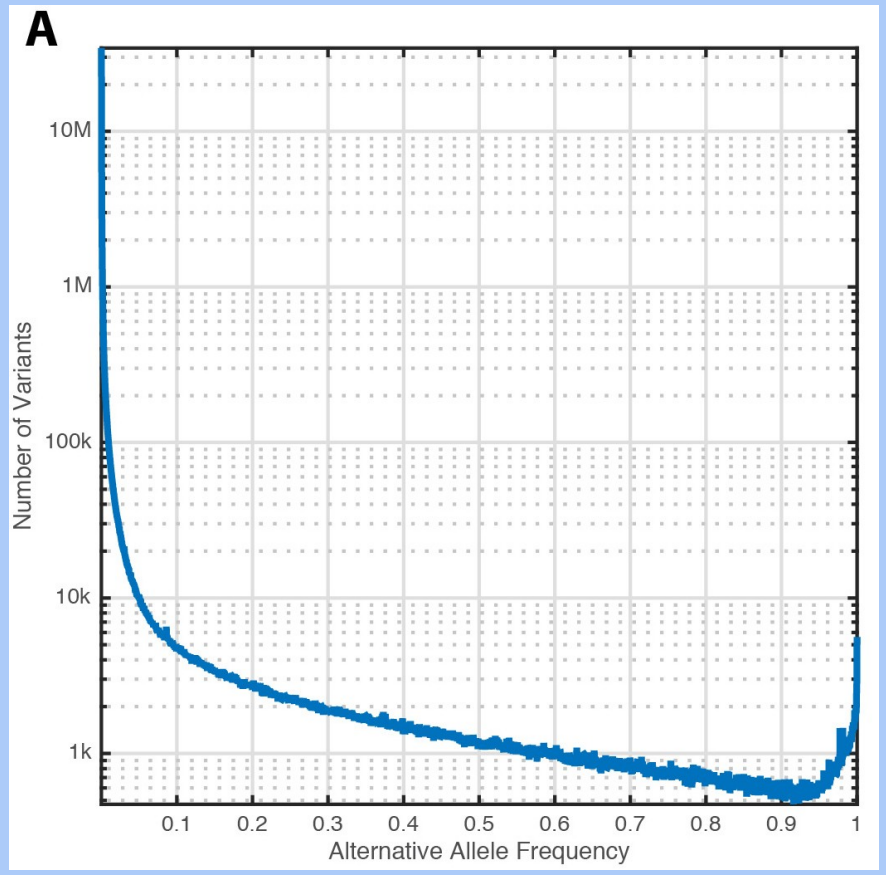
- More than just A, T, G, and C?
- Each polymorphism is coded by its possible alleles





Many rare SNPs

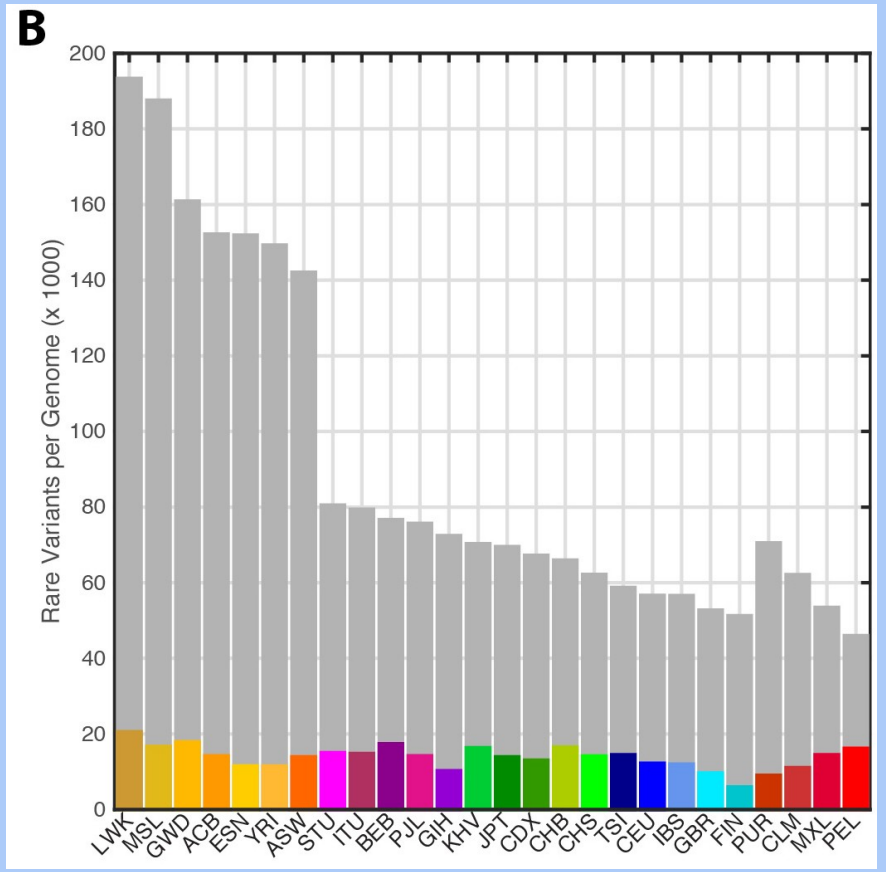
- Most SNPs of the >600 million known SNPs are very rare (frequency < 0.5%), but <5% of an individual's genome consists of rare SNPs



https://www.nature.com/articles/nature15393

Few rare SNPs per genome

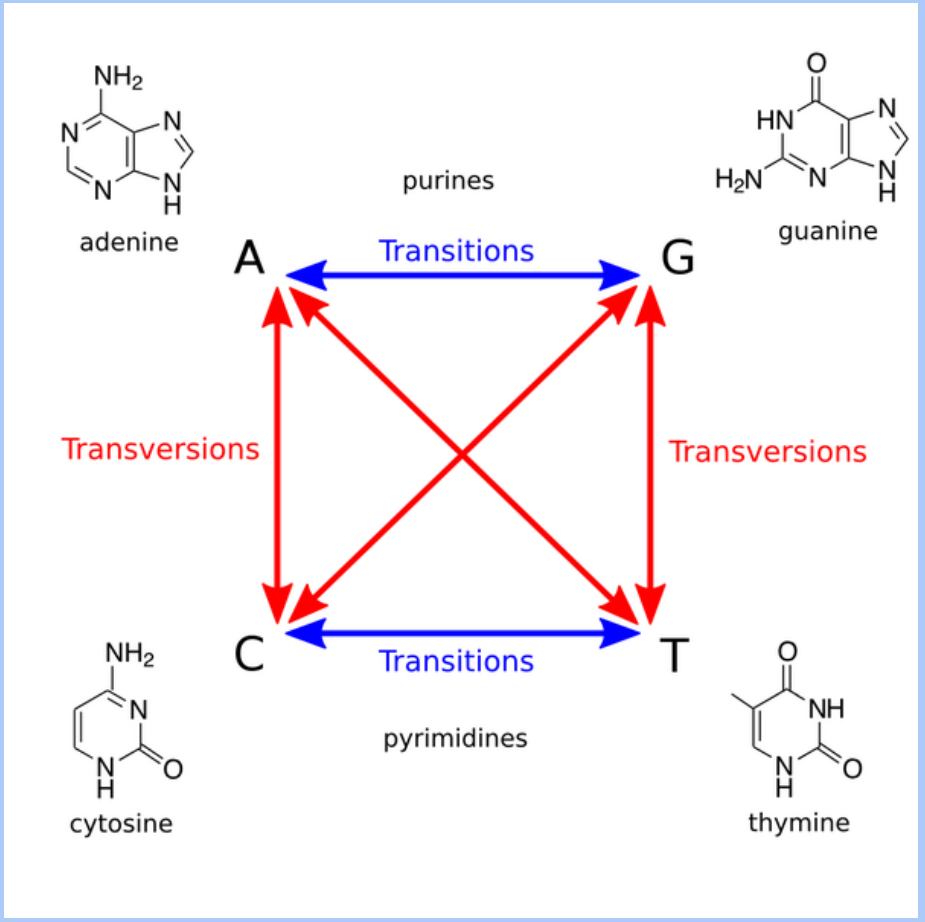
- Most SNPs of the >600 million known SNPs are very rare (frequency < 0.5%), but <5% of an individual's genome consists of rare SNPs
- Common SNPs have **minor allele frequency (MAF)** >5%



https://www.nature.com/articles/nature15393

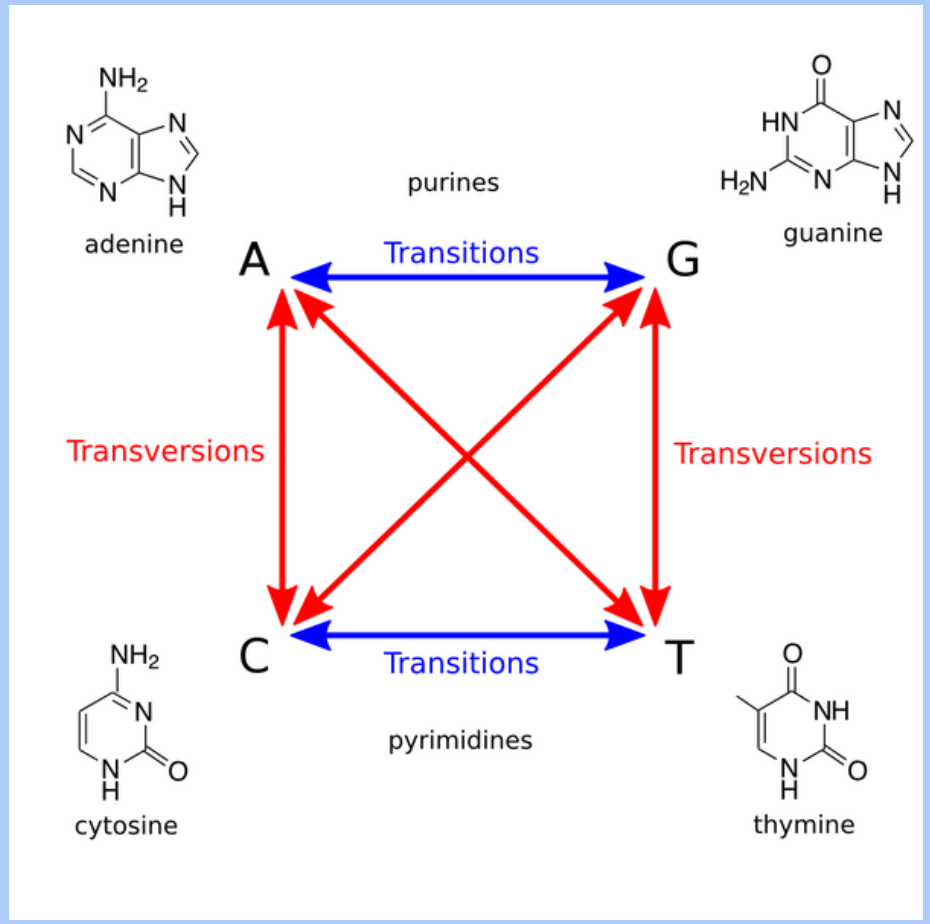
Transitions and transversions

- **Transitions** occur between nucleotides of the same type (purines or pyrimidines)
- **Transversions** occur between nucleotides of opposite type (between purines and pyrimidine)



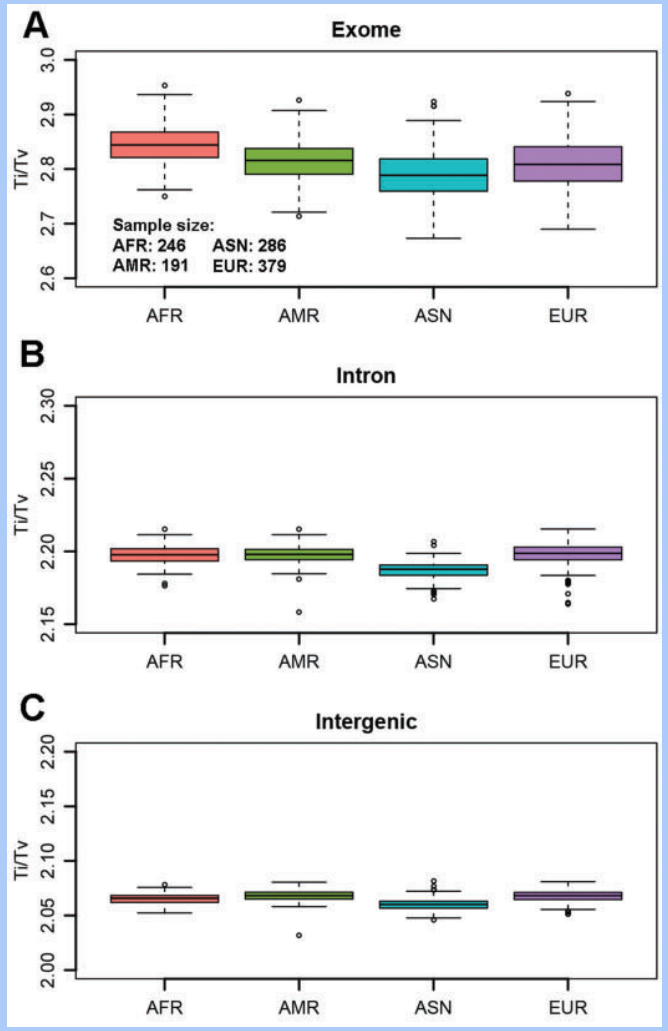
How many polymorphisms are there?

- If there are n nucleotide pairs, there are n symmetric conversions: A/T → T/A and C/G → G/C transitions
- If there are n nucleotide pairs, there are $n(n-1)$ asymmetric conversions: A/T → C/G transversions and A/T → G/C transitions
- A total of $n + n(n-1) = n^2$ polymorphims



Transition-transversion ratio

- Even though there are twice as many transversions possible as transitions, in humans the ratio of transitions to transversions is approximately 2 genome-wide
- In coding regions, the ratio is as high as 3



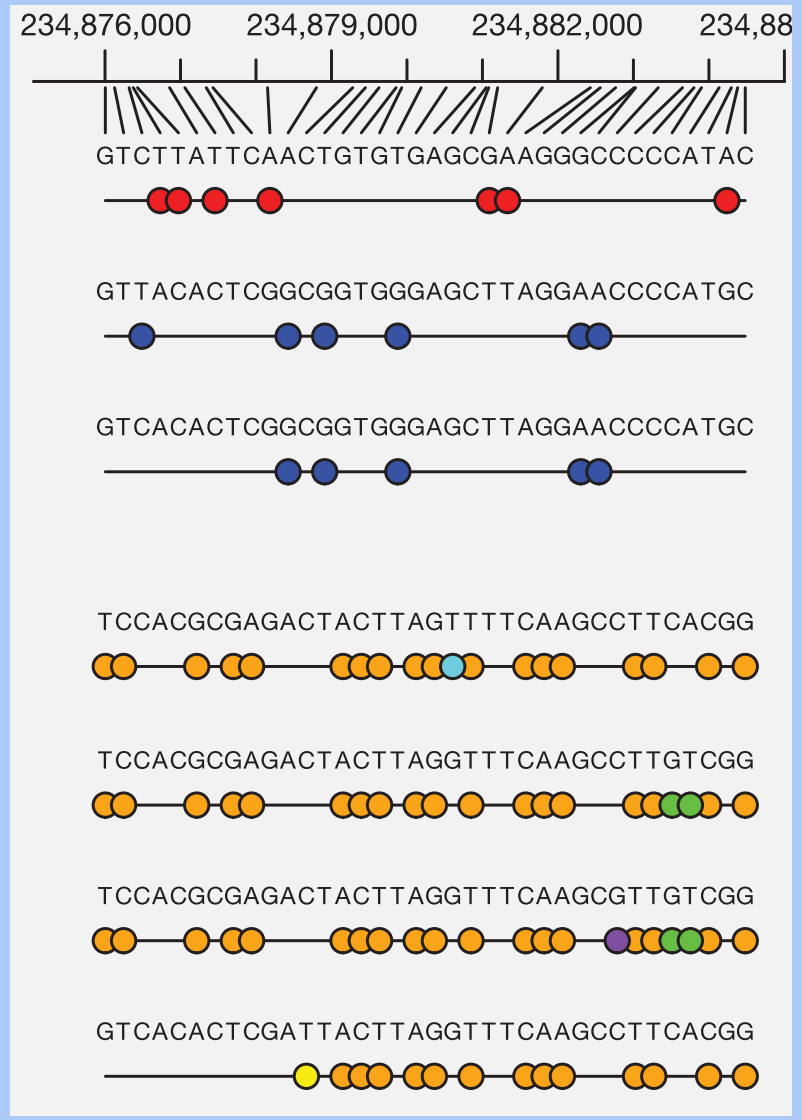
https://pubmed.ncbi.nlm.nih.gov/25297068/

Generation of sequencing data

Sequencing projects and technologies

The HapMap Project

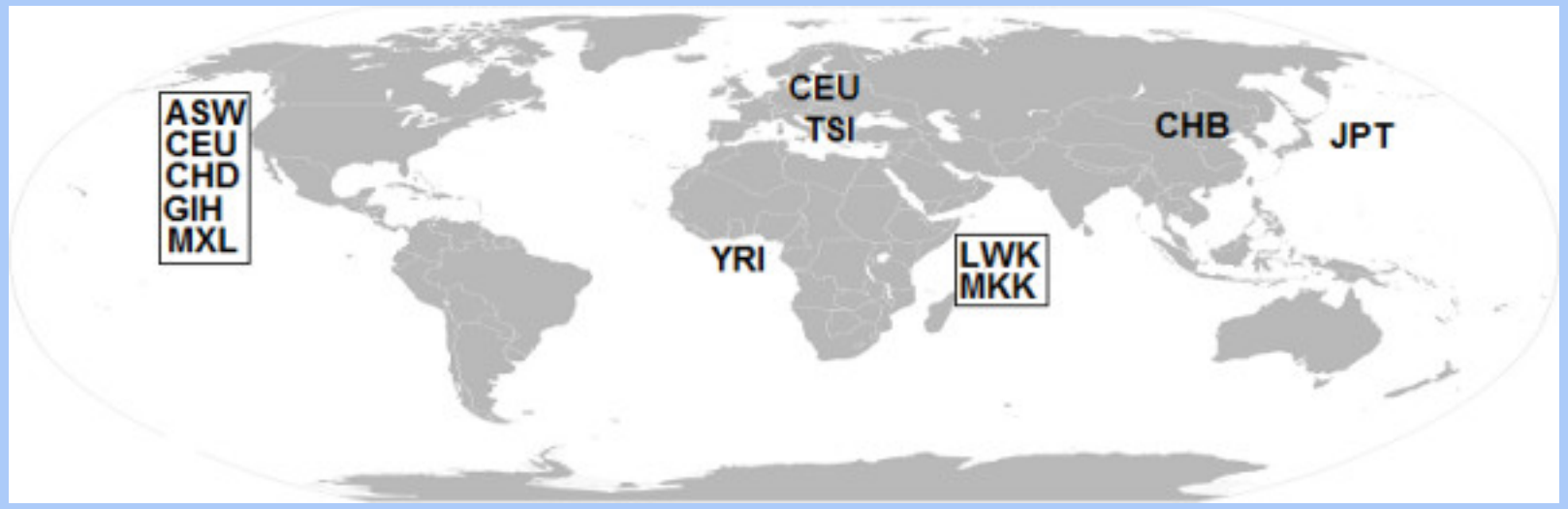
- International genotyping consortium launched in 2002 to find common polymorphisms linked to rare disease loci
- Variants occur together on a small number of **haplotypes**



https://pubmed.ncbi.nlm.nih.gov/16255080/

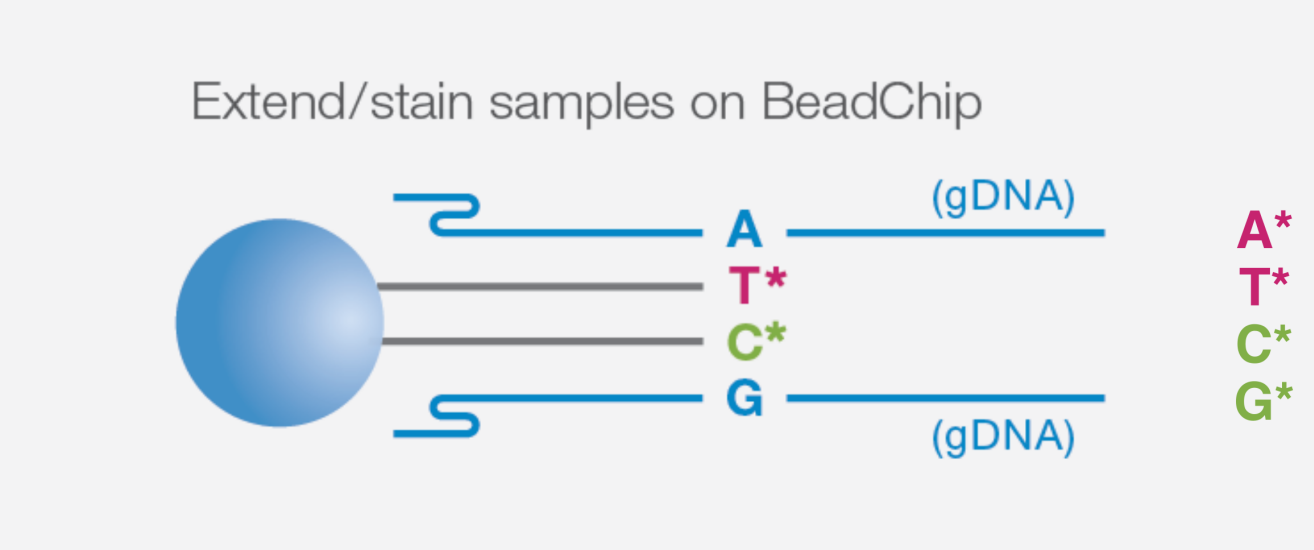
The HapMap Project

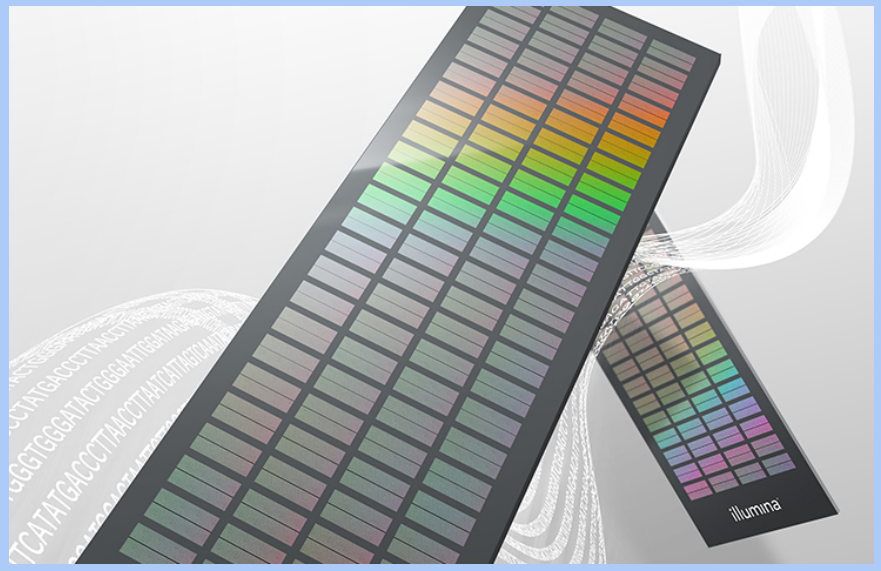
- Phase 3 (2010): genotyping and PCR resequencing of 1.6 million SNPs from 1,184 human samples from different parts of the world



SNP Genotyping

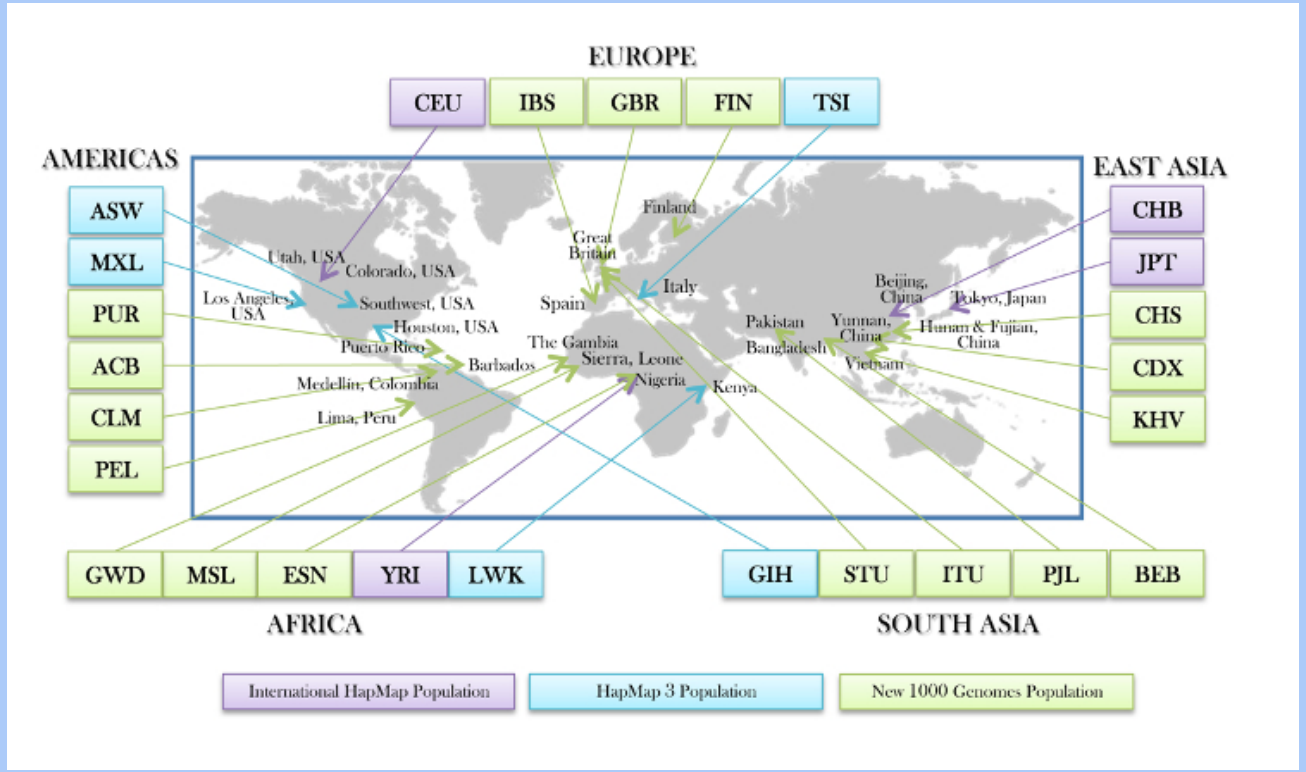
- Genomic DNA with binds to a complementary sequence and incorporates a fluorescently labelled nucleotide
- The ratio of red to green at a spot identifies the sample allele





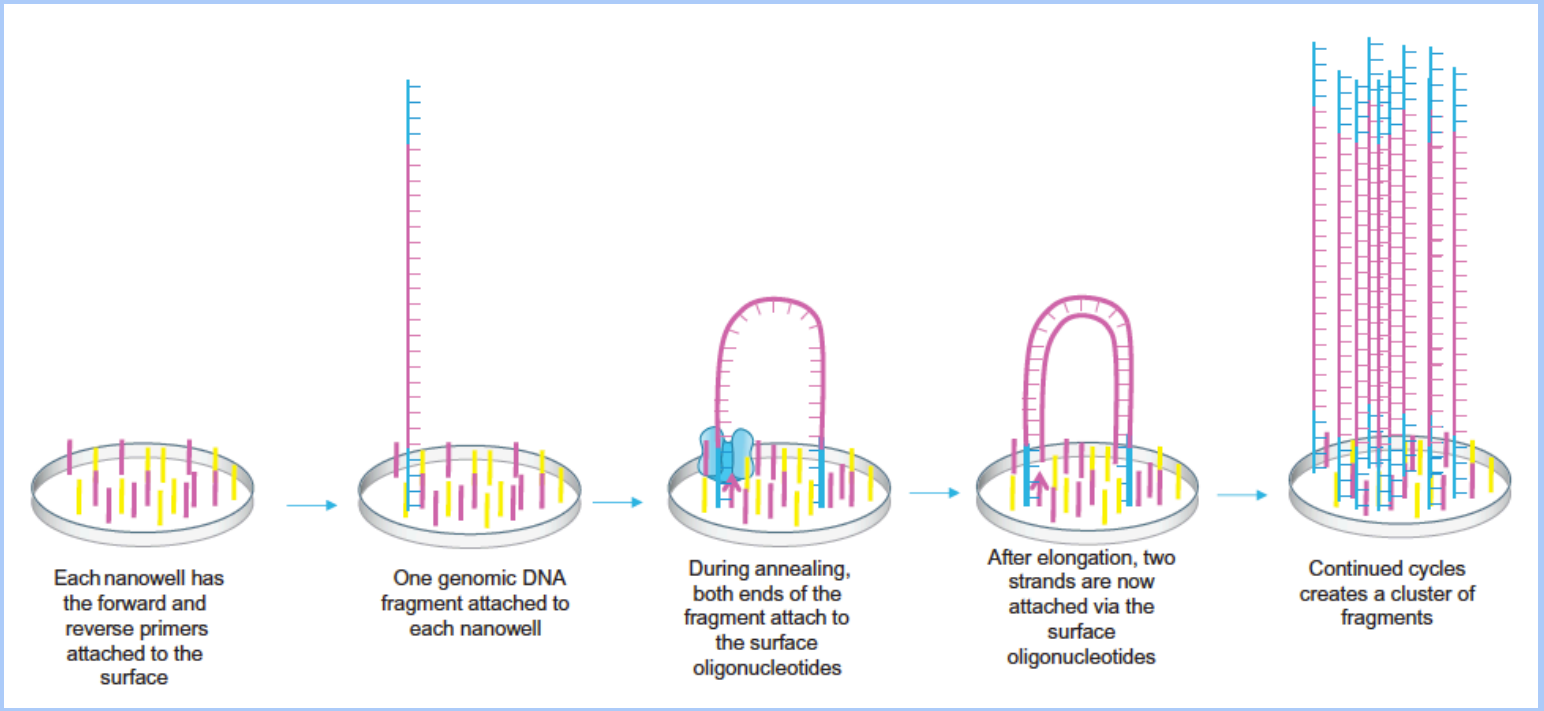
The 1000 Genomes Project

- An international consortium launched in 2008 to catalog rarer variants (frequency < 1%) taking advantage of new sequencing technologies
- Phase 3 release (2015) contained data from 2,504 individuals representing 26 populations across the globe and identified 85 million new SNPs



Whole-genome sequencing (WGS)

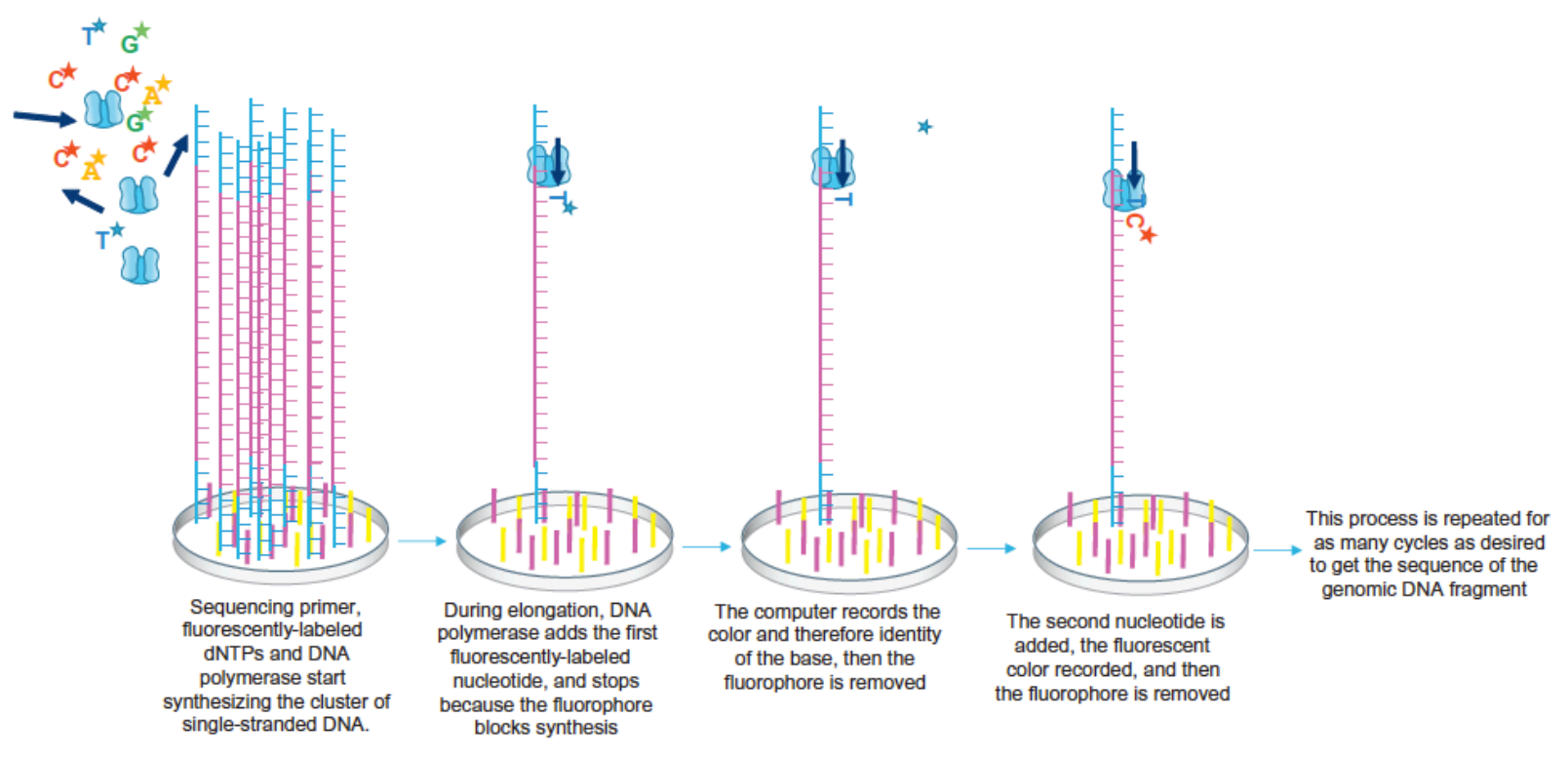
- DNA fragments from a sample are attached to a flow cell and amplified



Clark et al. *Molecular Biology (3rd Edition)*. Ch. 8: DNA Sequencing, 240-269 (2019)

Whole-genome sequencing (WGS)

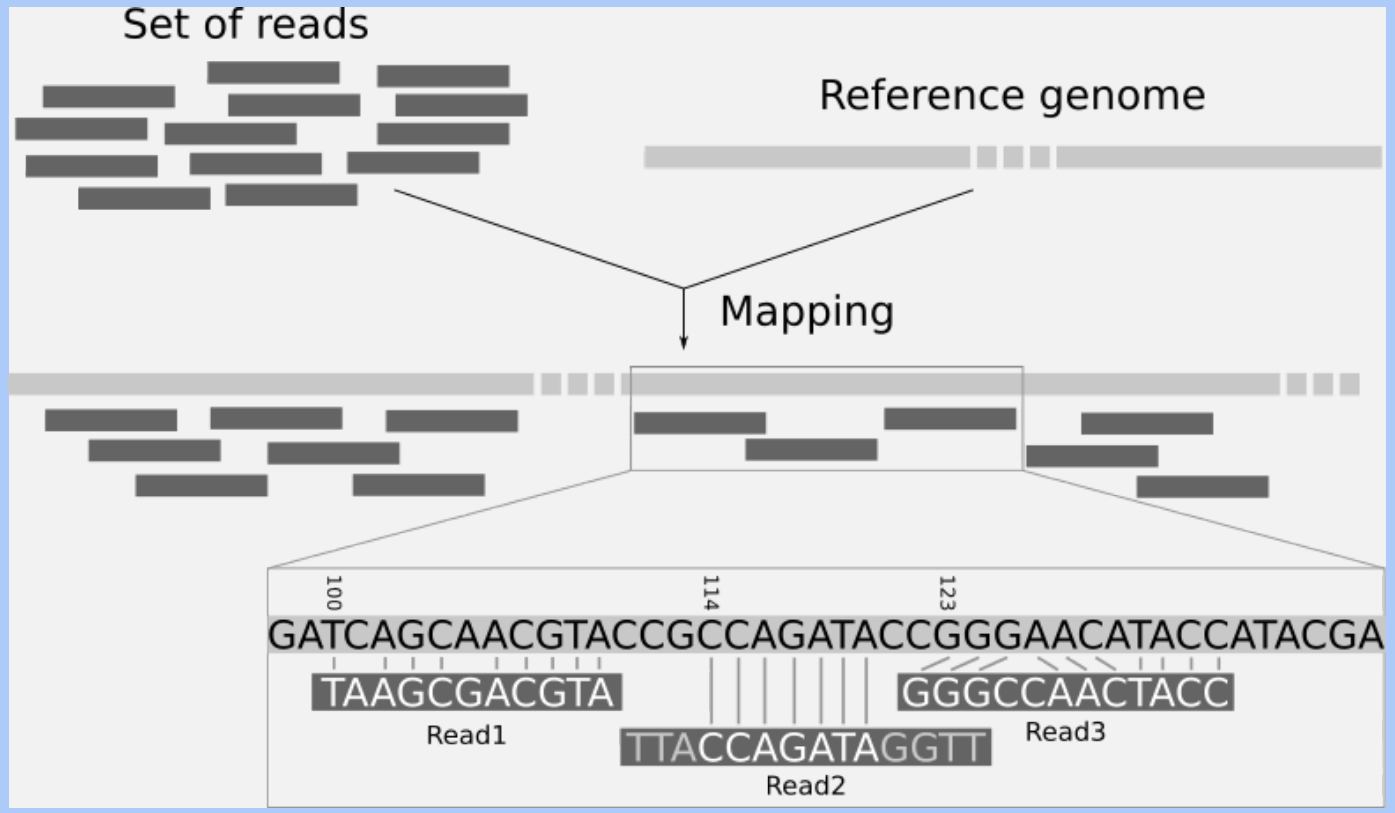
- Sequencing by synthesis: Short reads are produced as fluorescent nucleotides are incorporated one base at a time



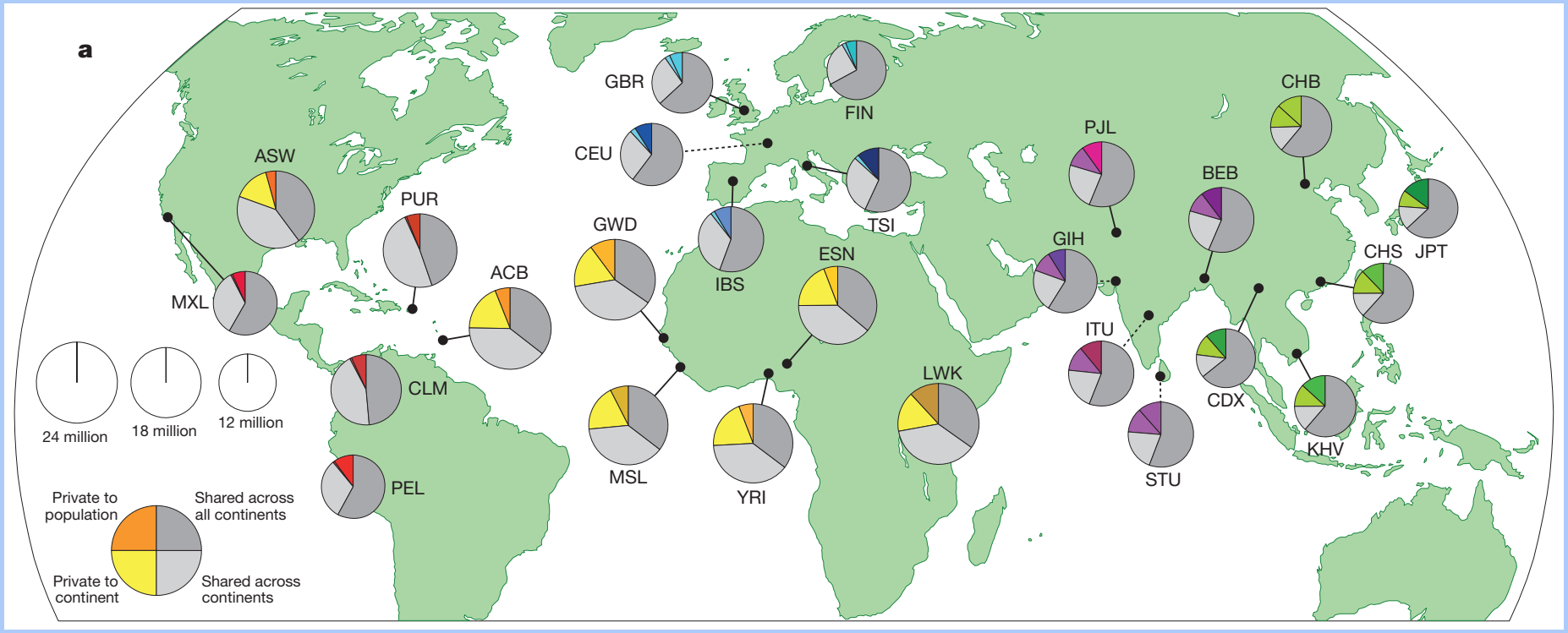
Clark et al. *Molecular Biology (3rd Edition)*. Ch. 8: DNA Sequencing, 240-269 (2019)

Mapping to the reference genome

- Locate from where in the genome the reads came
- Detect single-nucleotide differences from the reference sequence



Global genetic variation



- Most SNPs are shared across continents, and the majority of variation is within rather than between populations

Statistical variation of an allele

- Variation of the counts x_i of an allele about the group mean $\overline{x}_j$ and the population mean $\overline{x}$

$$\sum_i (x_i - \overline{x})^2 = \sum_i \left(x_i - \overline{x}_{j(i)}\right)^2 + \sum_j (\overline{x}_j - \overline{x})^2$$

Total variation | Within-population variation | Between-population variation

- Most SNPs are shared across continents, and the majority of variation is within rather than between populations

Common data formats

How genotype data are stored

Principal components analysis

The concept of genetic ancestry

The same yet different?

- Most variation is within-populations rather than between-populations
- Yet regional differences in allele frequencies lead to noticeable differences in phenotypes

Example: lactase nonpersistence (lactose intolerance)

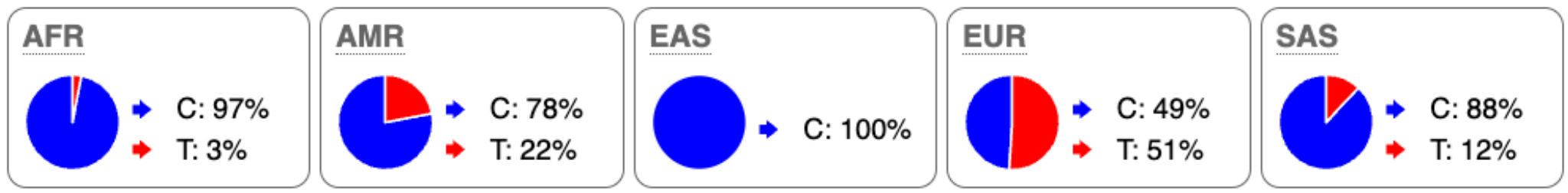
- The T allele of rs182549 is completely associated with the ability to digest lactose in Europeans

	CC	CT	TT
Non-persistence	59	0	0
Persistence	0	63	74

https://pubmed.ncbi.nlm.nih.gov/11788828/

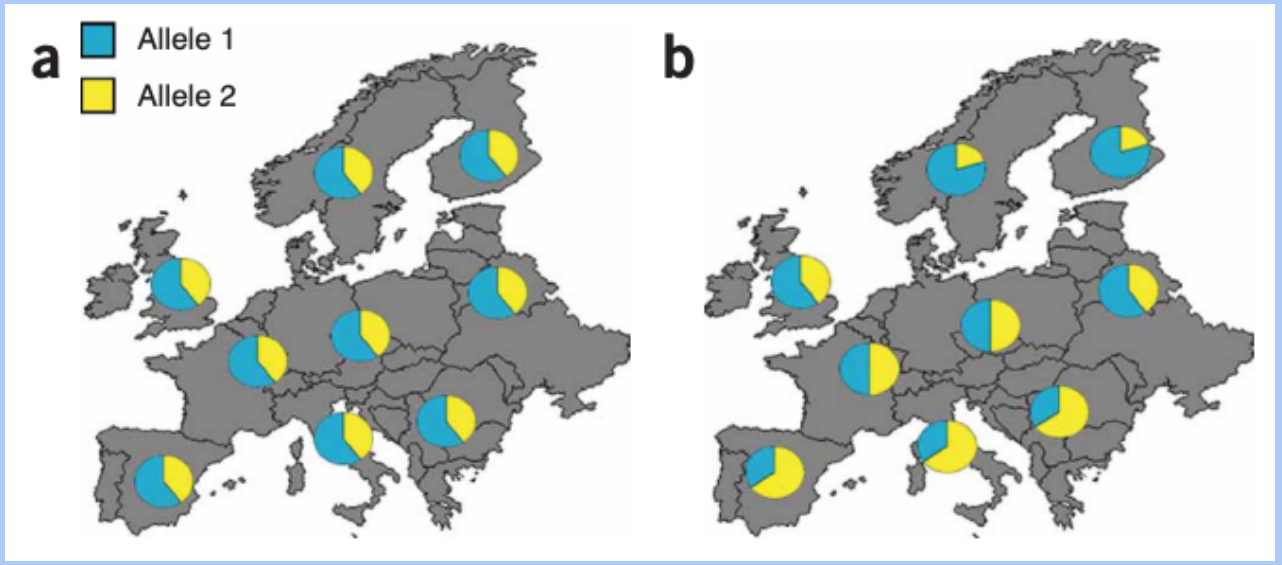
Example: lactase nonpersistence (lactose intolerance)

- Yet the polymorphism is almost absent in the African population, despite the presence of lactase persistence https://pubmed.ncbi.nlm.nih.gov/15106124/



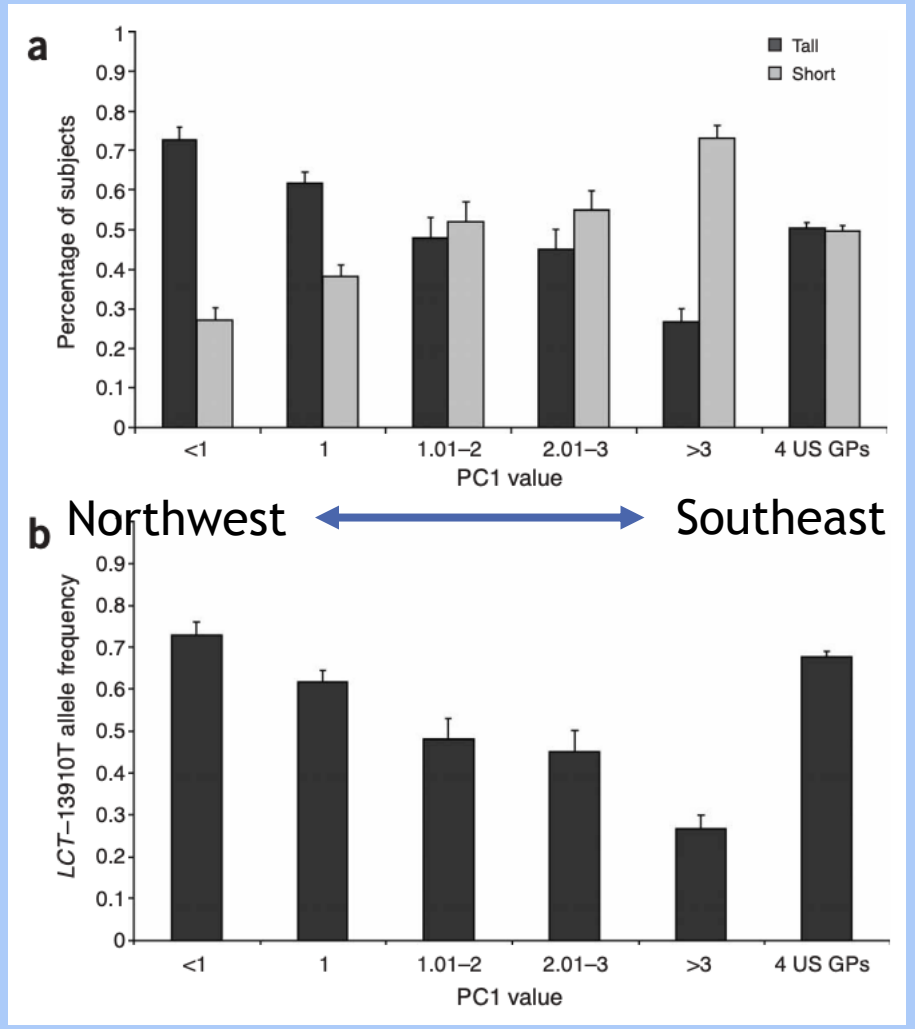
Population stratification

- An allele may appear associated with a phenotype when in fact it is associated with geographic origin (genetic ancestry)



Spurious association

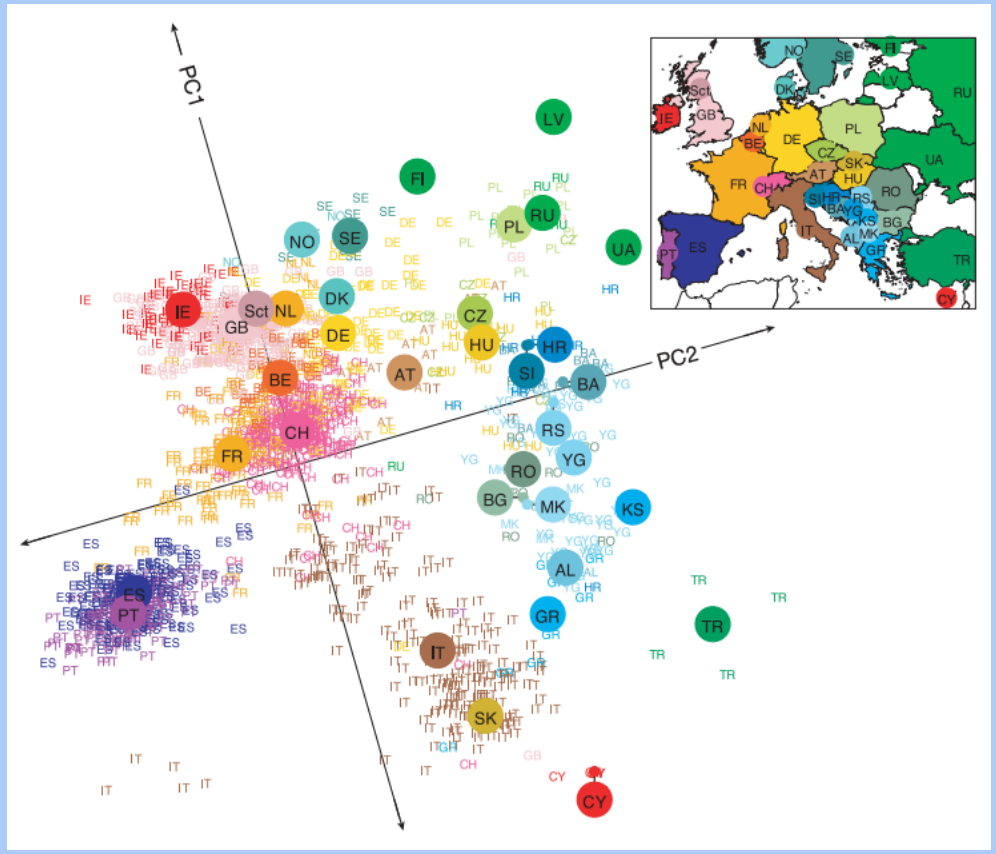
- An allele of the lactase-persistence SNP is spuriously associated with height, as its frequency is higher in individuals with Northern European ancestry vs. Southern



https://pubmed.ncbi.nlm.nih.gov/16041375/

Principal components analysis

- Genotypes can distinguish population groups
- Looking a **which variants segregate together** can tell us about an individual's likely genetic ancestry



https://pubmed.ncbi.nlm.nih.gov/18758442/

Genotype matrix

- n individuals are genotyped at m SNPs
- The number of alternate alleles is 0, 1, or 2
- “Standardize” each genotype by subtracting the mean allele frequency and dividing by its standard error

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

“Idealized” individuals

- An “idealized” subject of a particular genetic ancestry has genotypes v at m SNPs
- The position of individual 1 on PC1 is the “amount” of idealized person 1 in individual 1

$$\mathbf{XV}^T = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ \vdots & \vdots & \vdots \\ v_{1m} & v_{2m} & v_{3m} \end{pmatrix}$$

Genomic relationship matrix (GRM)

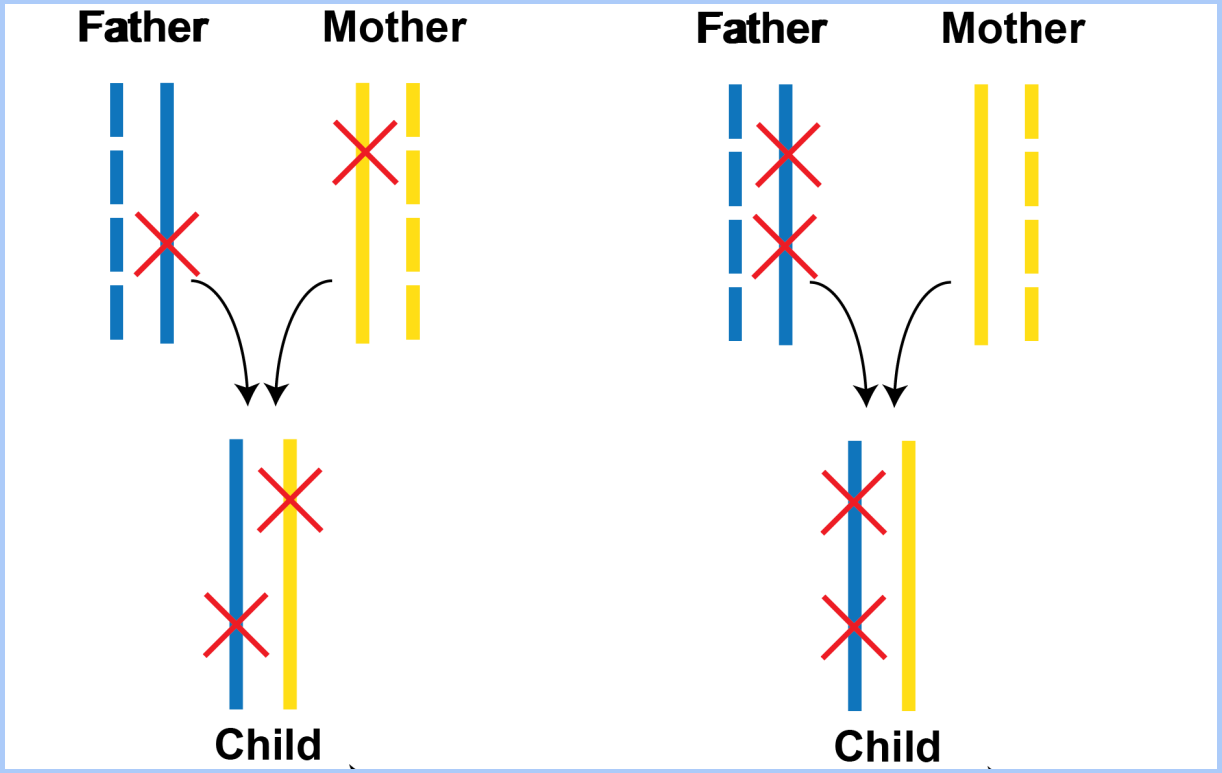
- The idea of PCA is to find the amount of each idealized individual in each actual individual
- The **eigenvectors** of the GRM contain the ancestry components
- The GRM is computed by comparing how similar any subject is to any other

$$\mathbf{X}\mathbf{X}^T = \begin{pmatrix} \mathbf{x}_1 \cdot \mathbf{x}_1 & \cdots & \mathbf{x}_1 \cdot \mathbf{x}_n \\ \mathbf{x}_2 \cdot \mathbf{x}_1 & \cdots & \mathbf{x}_2 \cdot \mathbf{x}_n \\ \vdots & & \vdots \\ \mathbf{x}_n \cdot \mathbf{x}_1 & \cdots & \mathbf{x}_n \cdot \mathbf{x}_n \end{pmatrix}$$

Linkage disequilibrium

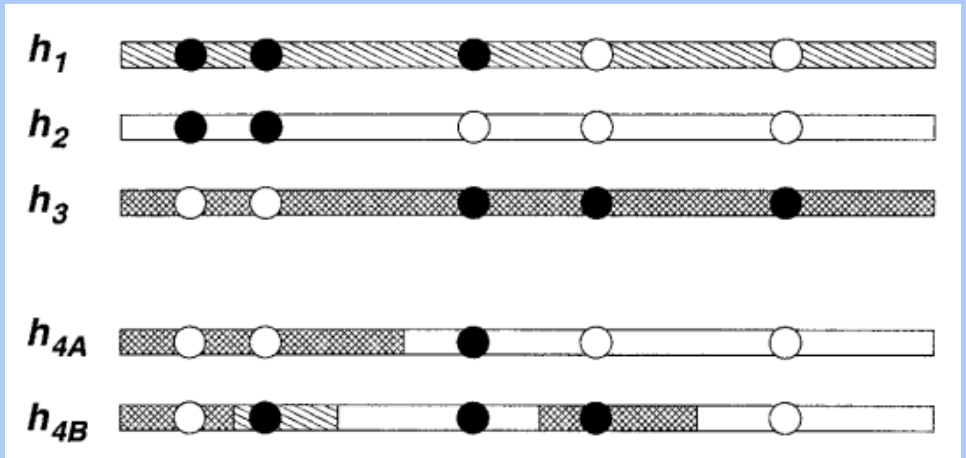
Determining a set of independent SNPs

SNPs can occur on either of two chromosomes



- Genotype data do not tell us which chromosomes carry the polymorphism
- When at least one parent is homozygous at each SNP, **haplotype phase** can be unambiguously assigned

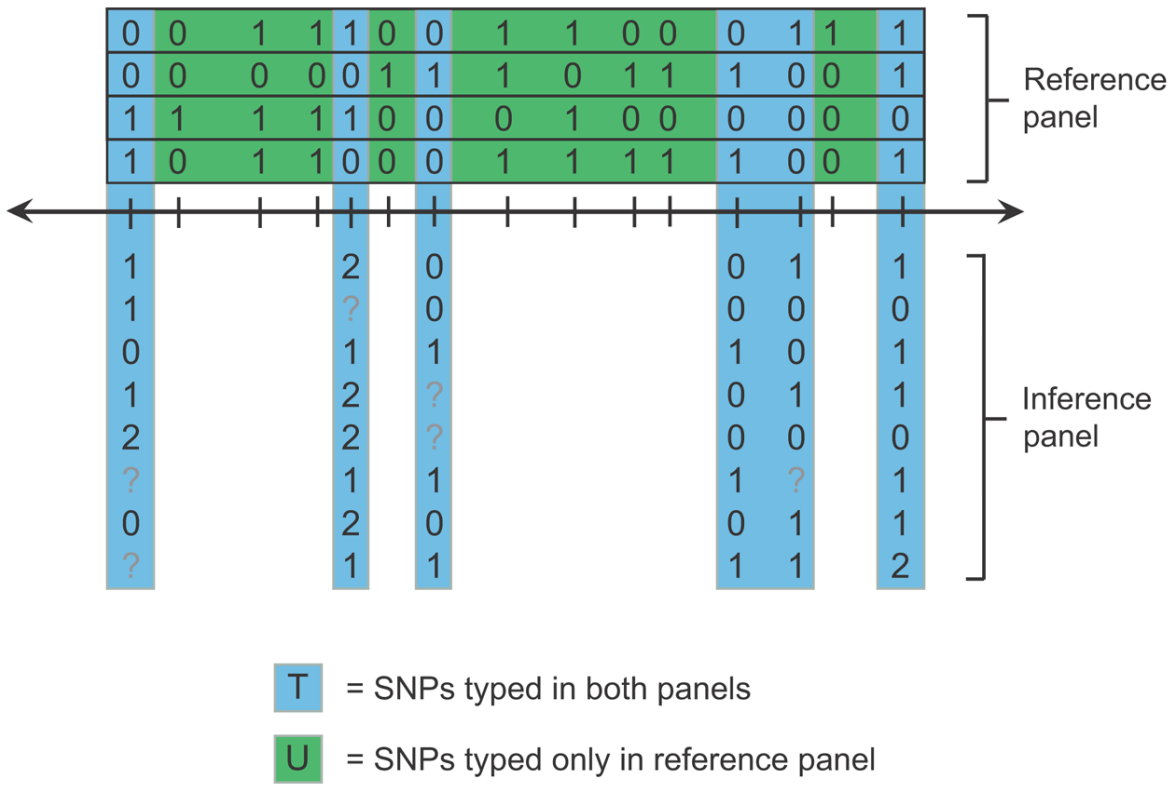
Statistical phasing and imputation



https://pubmed.ncbi.nlm.nih.gov/14704198/

- Genotyped individuals can be computationally **phased** by modelling each chromosome as an imperfect **mosaic** of chromosomes in a reference panel

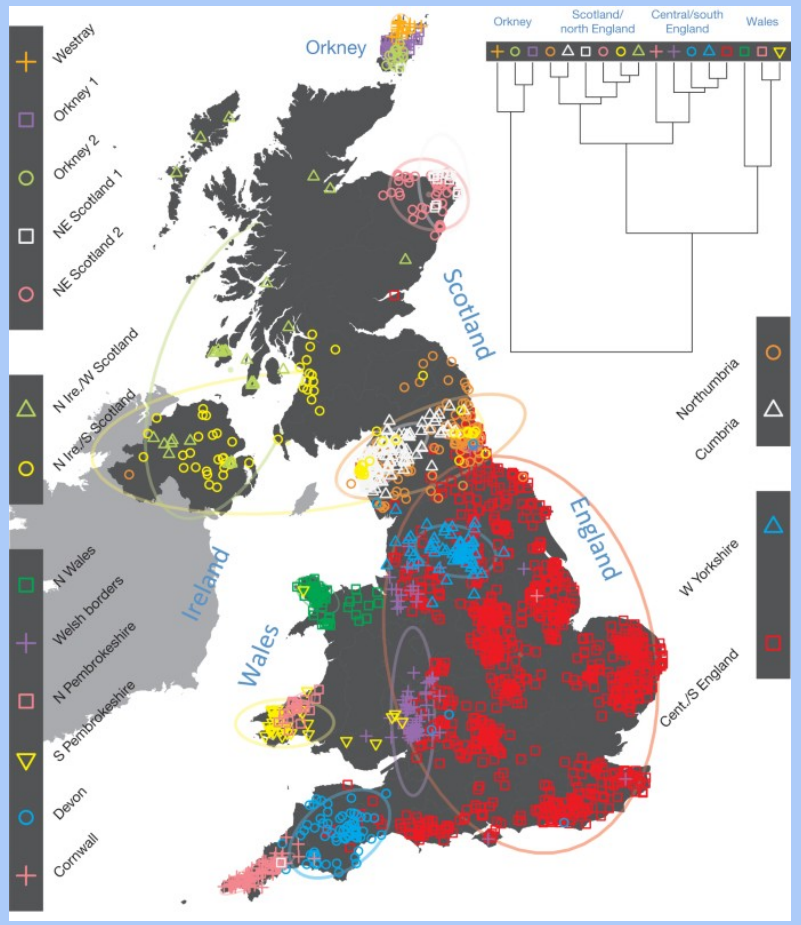
Statistical phasing and imputation



https://pubmed.ncbi.nlm.nih.gov/19543373/

- Variants that have not been typed can be **imputed** into the inference sample
- Imputation accuracy depends on the inference and reference samples being of similar genetic ancestry

Different haplotypes distinguish different populations



https://pubmed.ncbi.nlm.nih.gov/25788095/

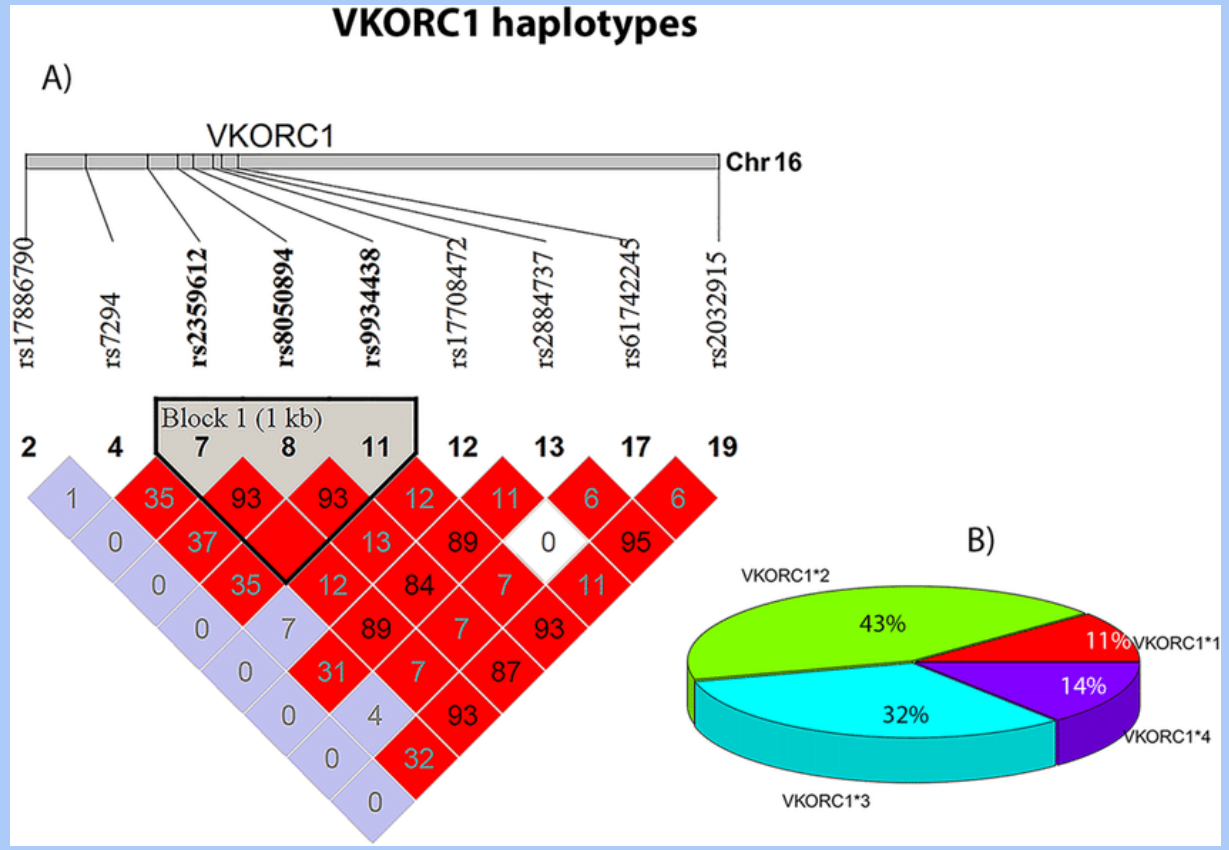
- Individuals can be grouped into populations with which they have the most haplotype-sharing

Linkage disequilibrium

- Linkage disequilibrium is the population tendency of alleles to be inherited on a single chromosomes and is measure using a correlation coefficient between the alleles of different SNPs

$$r_{A,B} = \frac{p_{A,B} - p_A p_B}{\sqrt{p_A (1 - p_A) p_B (1 - p_B)}}$$

LD blocks and haplotype structure



https://pubmed.ncbi.nlm.nih.gov/32221414/

- Plots of pairwise r^2 values show which SNPs are inherited together in the population as common haplotypes

Haplotype simulation using LD

- Simulate a haplotype by ensuring that the frequencies and LD between two alleles match the reference data
- No assumptions about phylogenies or knowledge of evolutionary theory required

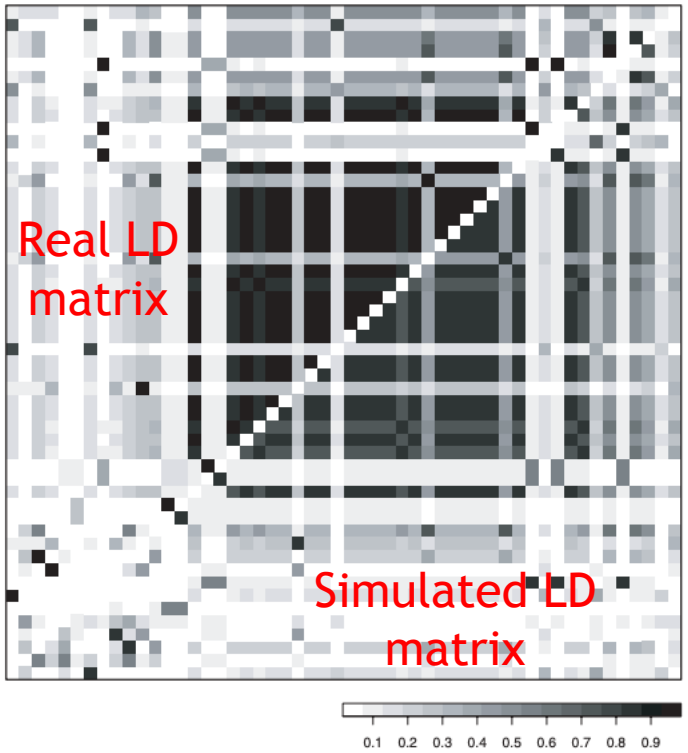


Fig. 2. ACE data: hybrid LD matrix obtained from real and artificial samples of 11 individuals each.

https://pubmed.ncbi.nlm.nih.gov/16188927/

sim1000G: simulate haplotypes from an input vcf

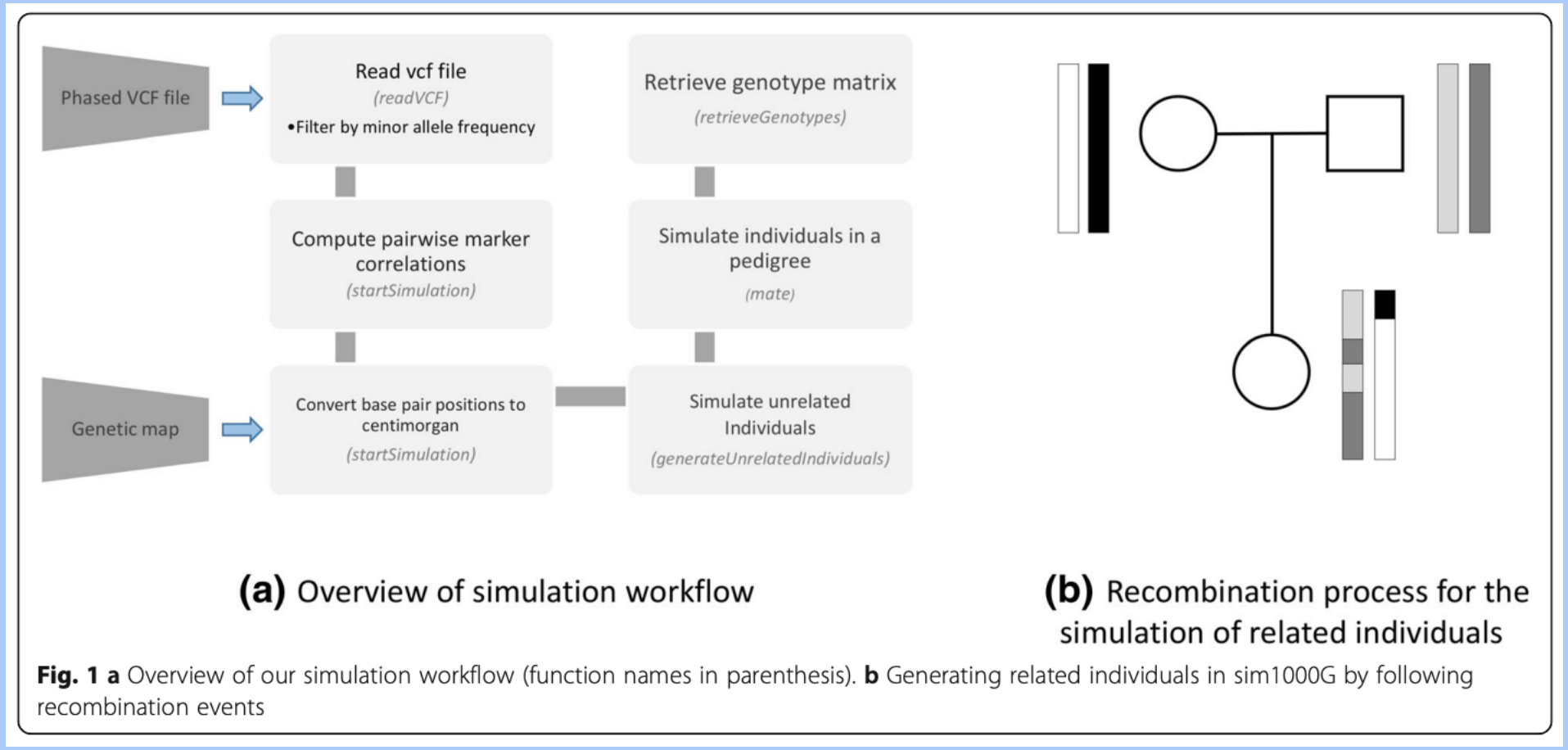


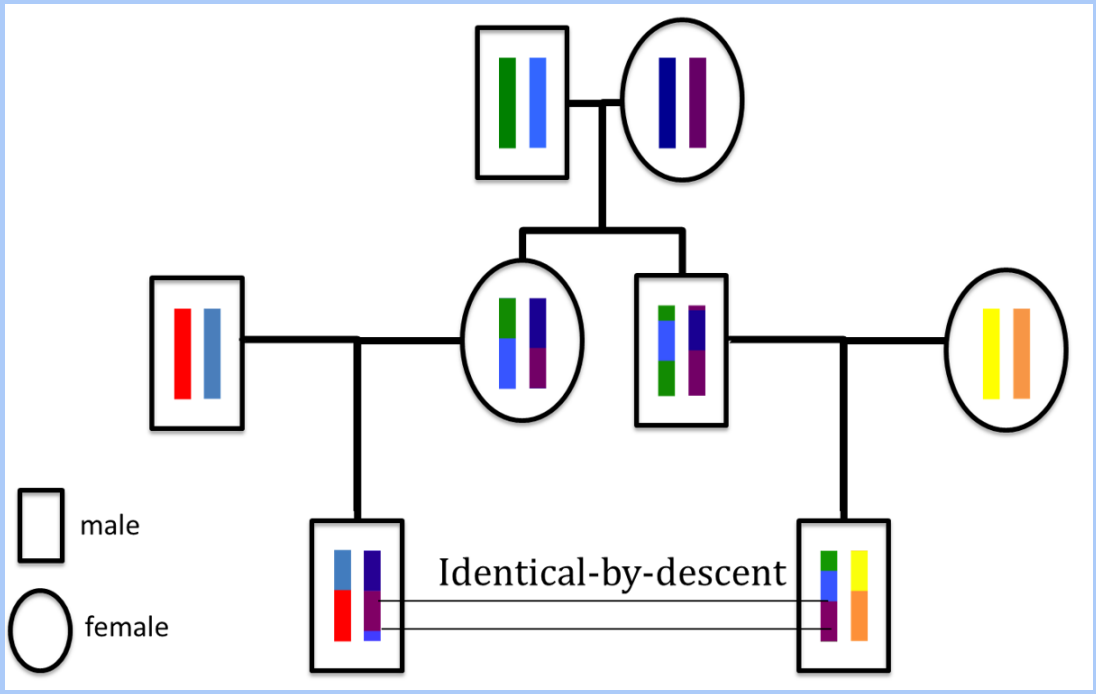
Fig. 1 a Overview of our simulation workflow (function names in parenthesis). **b** Generating related individuals in sim1000G by following recombination events

https://pubmed.ncbi.nlm.nih.gov/30646839/

Kinship analysis

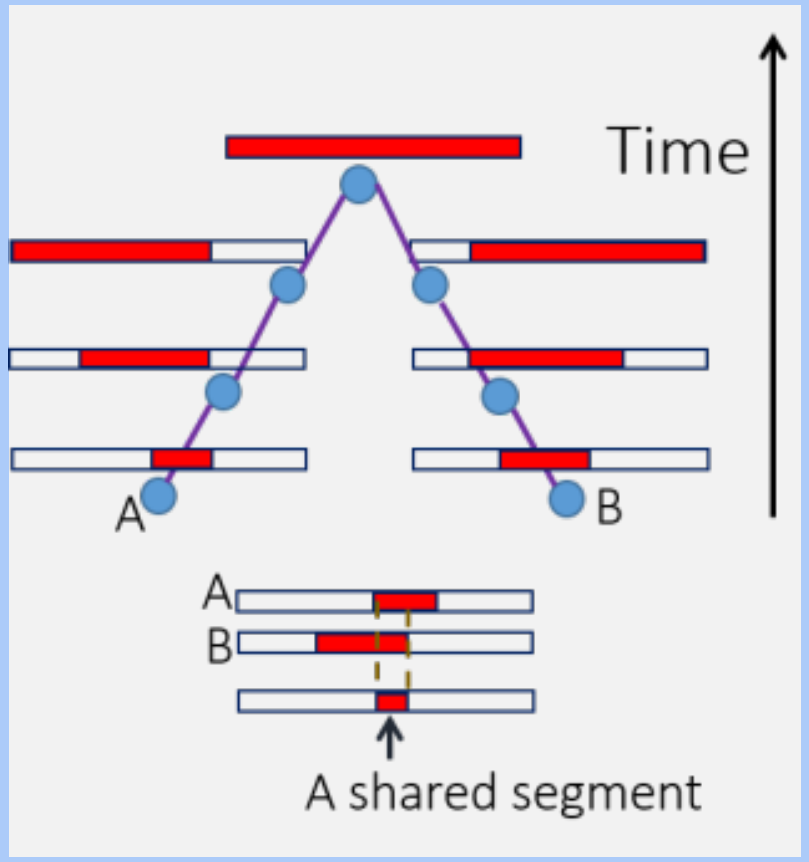
The concept of genetic relatedness

Relatives share haplotypes IBD



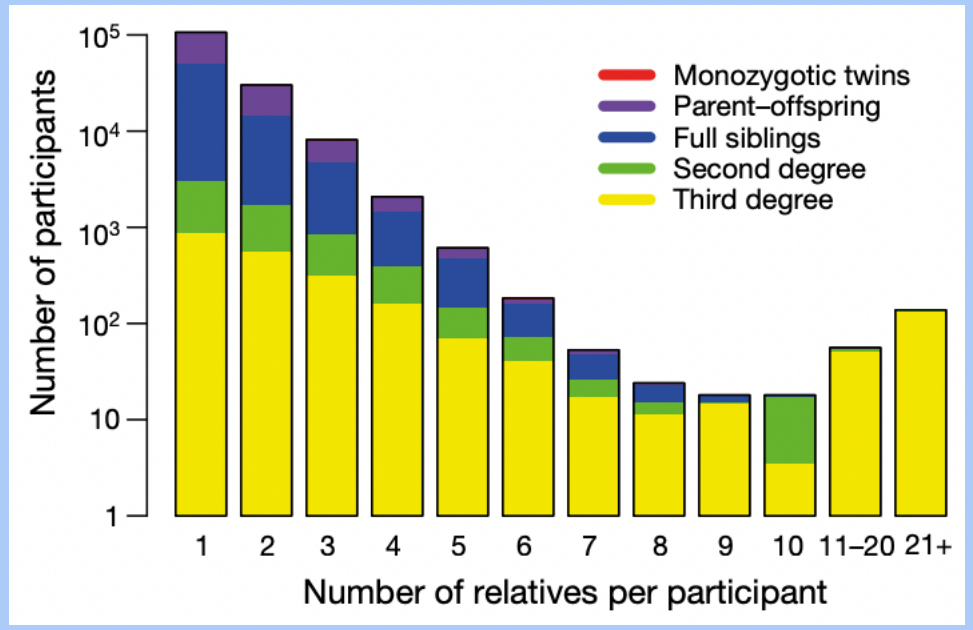
- Segments of DNA inherited from a common ancestor are said to be **identical by descent**
- DNA that just happens to be the same is **identical by state**

Haplotype sharing decays over time



- The longer the IBD segment, the more closely related are the individuals

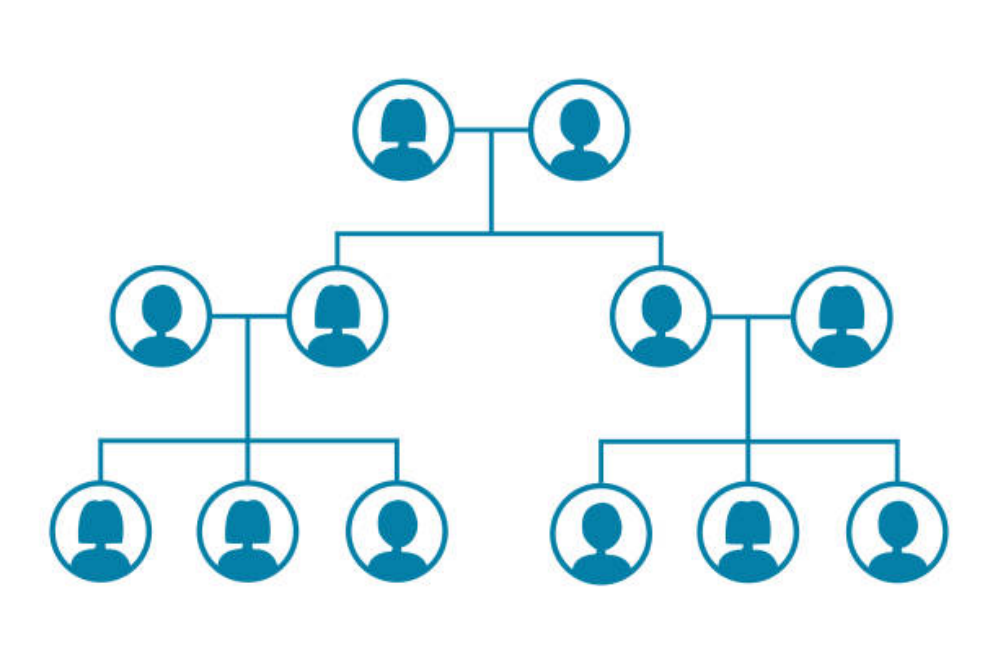
Kinship in genetic association studies



https://pubmed.ncbi.nlm.nih.gov/30305743/

- Genomic datasets, such as the UK Biobank, contain related individuals
- Sometimes there is even "cryptic" relatedness
- Because of IBD sharing, not all the observations are independent, and genotype-phenotype associations may be **confounded**

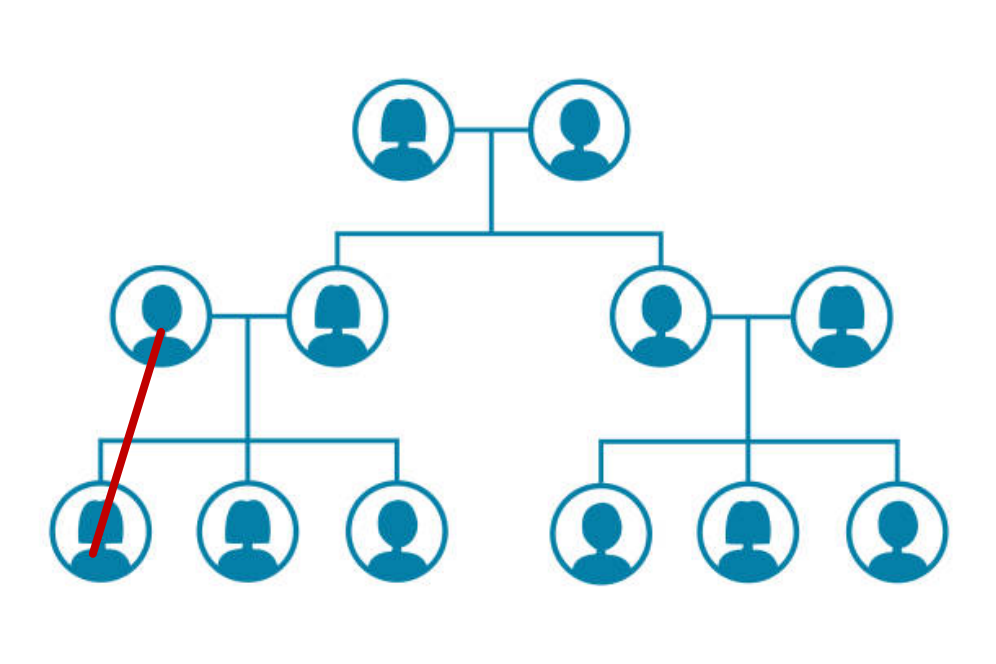
Degree of relatedness



- R is the effective number of meioses separating two individuals through their two parents
- R → ∞ for unrelated individuals

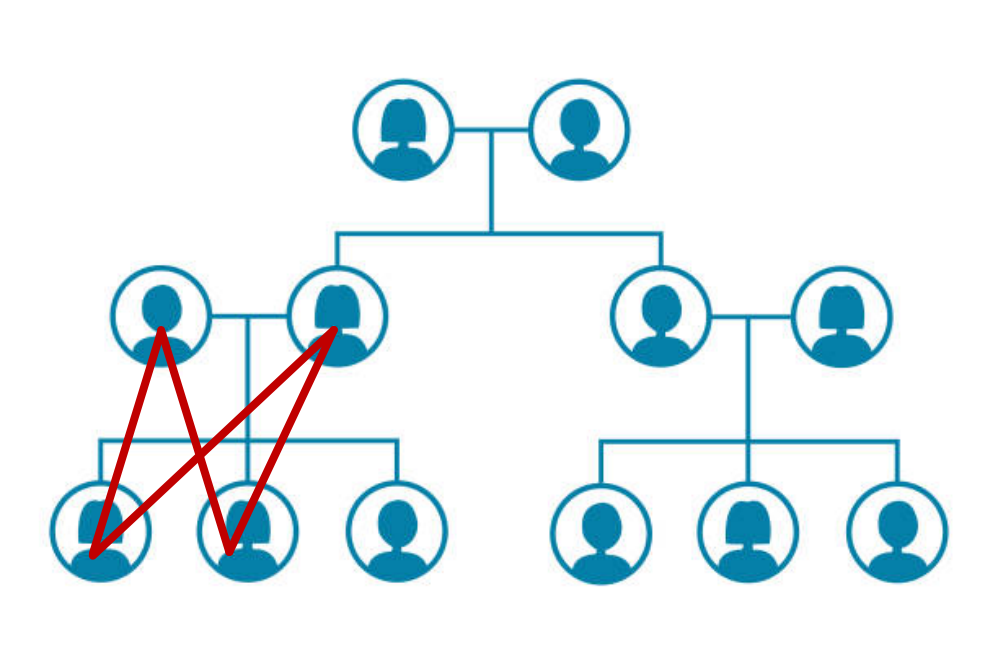
$$\frac{1}{2^R} = \frac{1}{2^{R_1}} + \frac{1}{2^{R_2}}$$

Degree of relatedness



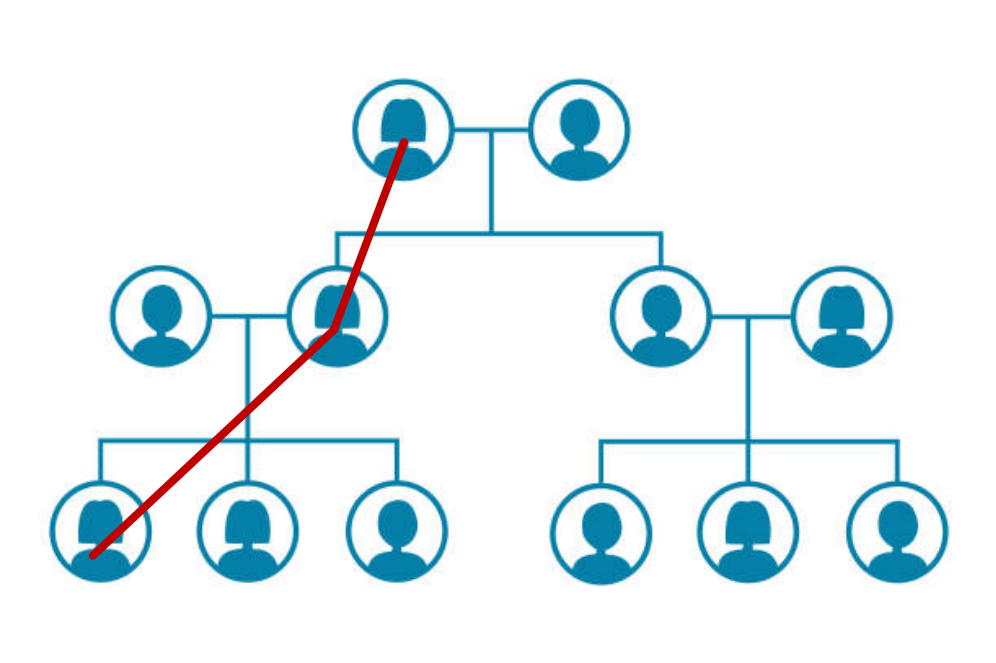
- Parent-child: R = 1 meiosis

Degree of relatedness



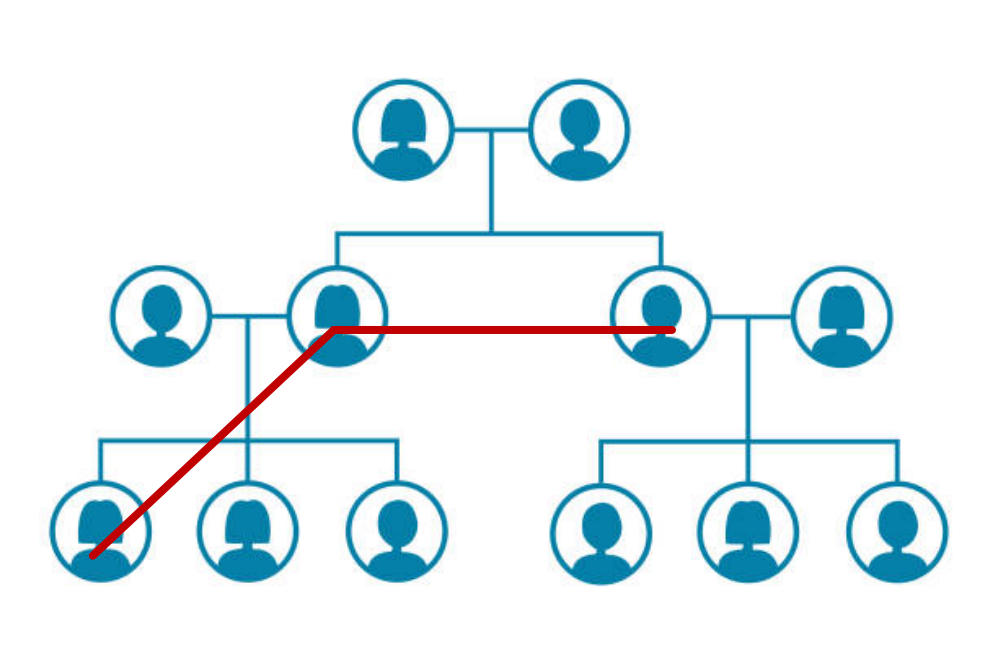
- Siblings: R = 1 “effective” meiosis: ½ + ½

Degree of relatedness



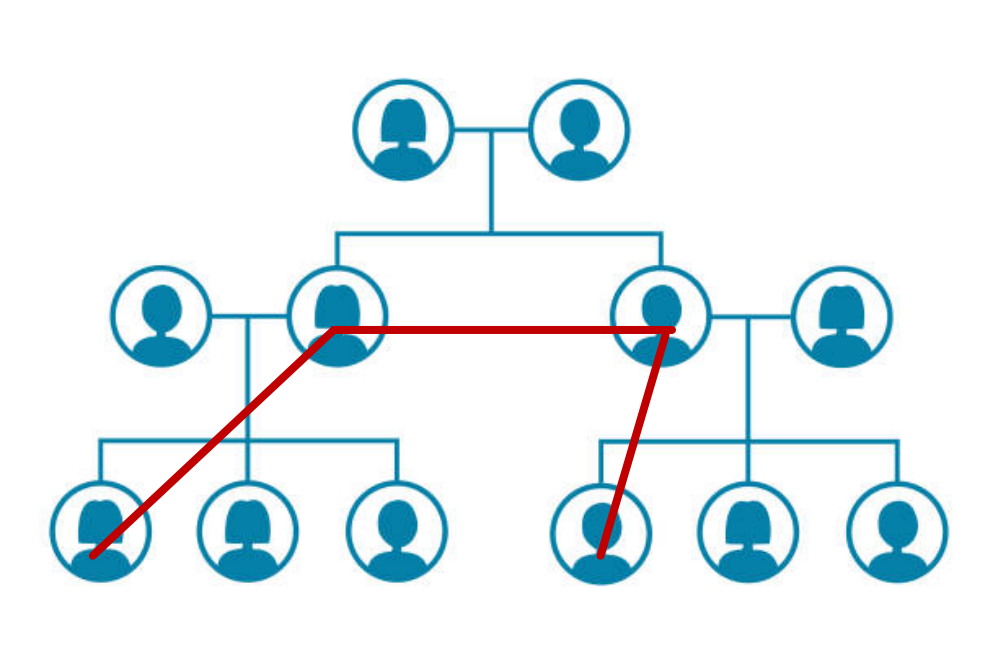
- Grandparent-grandchild: R = 2 meioses

Degree of relatedness



- Avuncular: R = 2 meioses

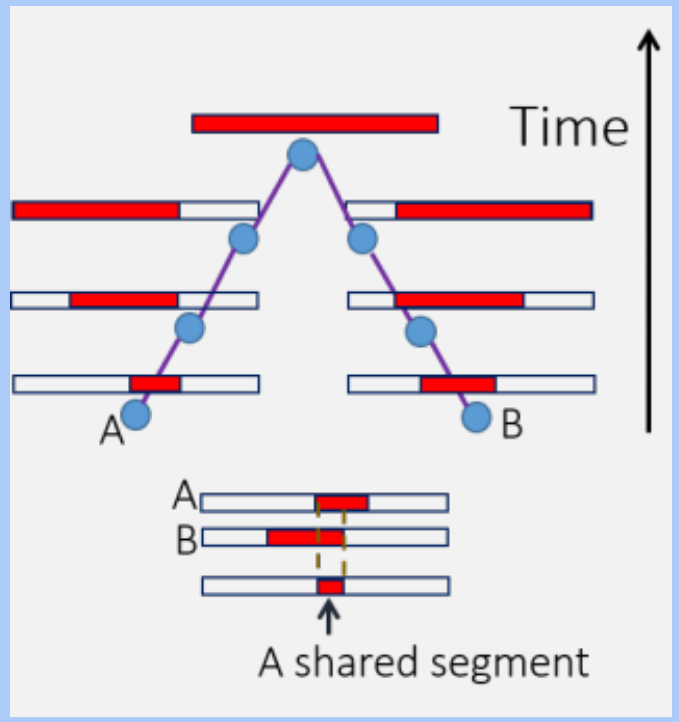
Degree of relatedness



- Cousins: R = 3 meioses

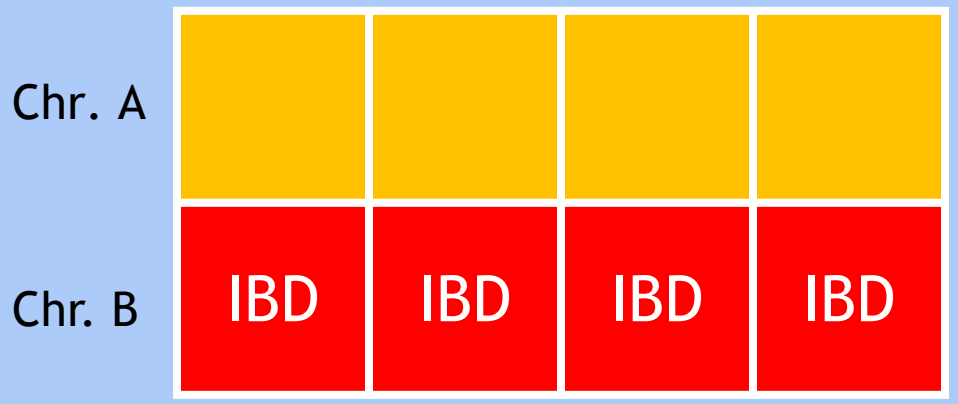
Degree of relatedness and the fraction of the genome shared IBD

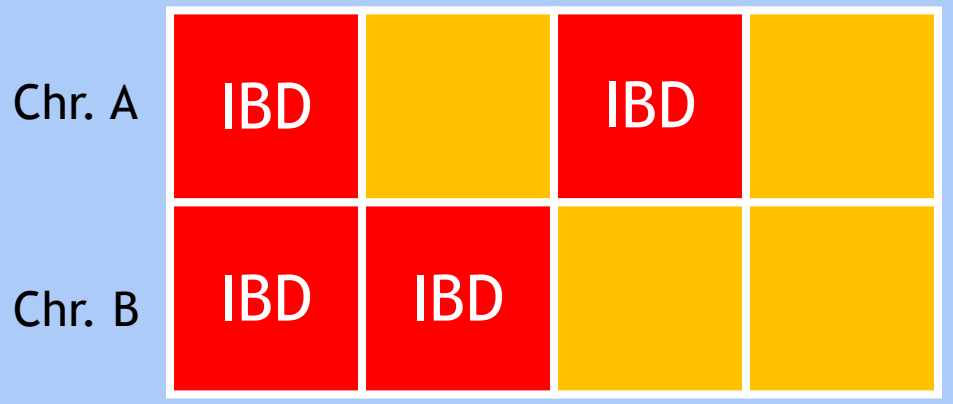
- $r = 1 / 2^R$ is the fraction of the genome shared IBD, because there is a ½ probability that the gene is passed on in each of R meioses



Degree of relatedness and the fraction of the genome shared IBD

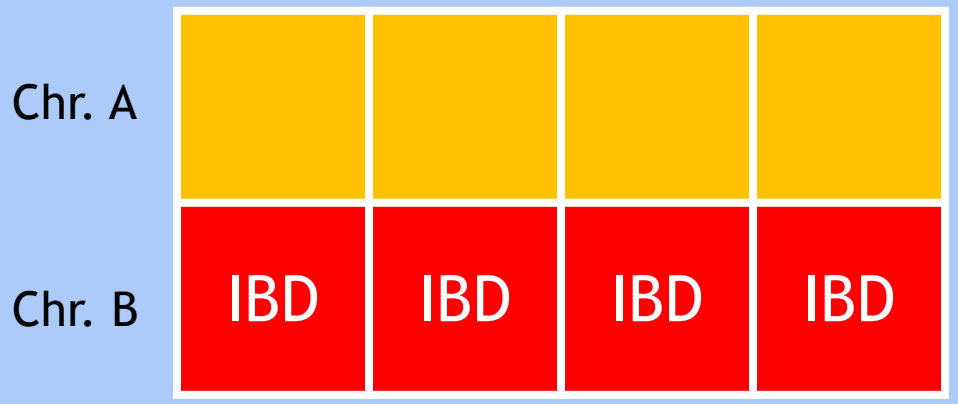
- A child shares **half** of its DNA with its parent
- A child shares (a different) **half** its DNA with its full sib

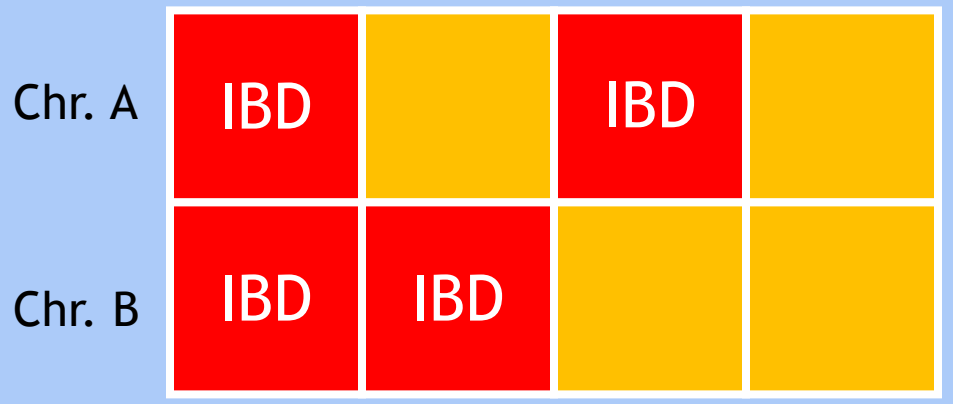




Degree of relatedness and the fraction of the genome shared IBD

- A child has 0 probability of IBD = 0 with its parent
- A child has 0.25 probability of IBD = 0 with its sib





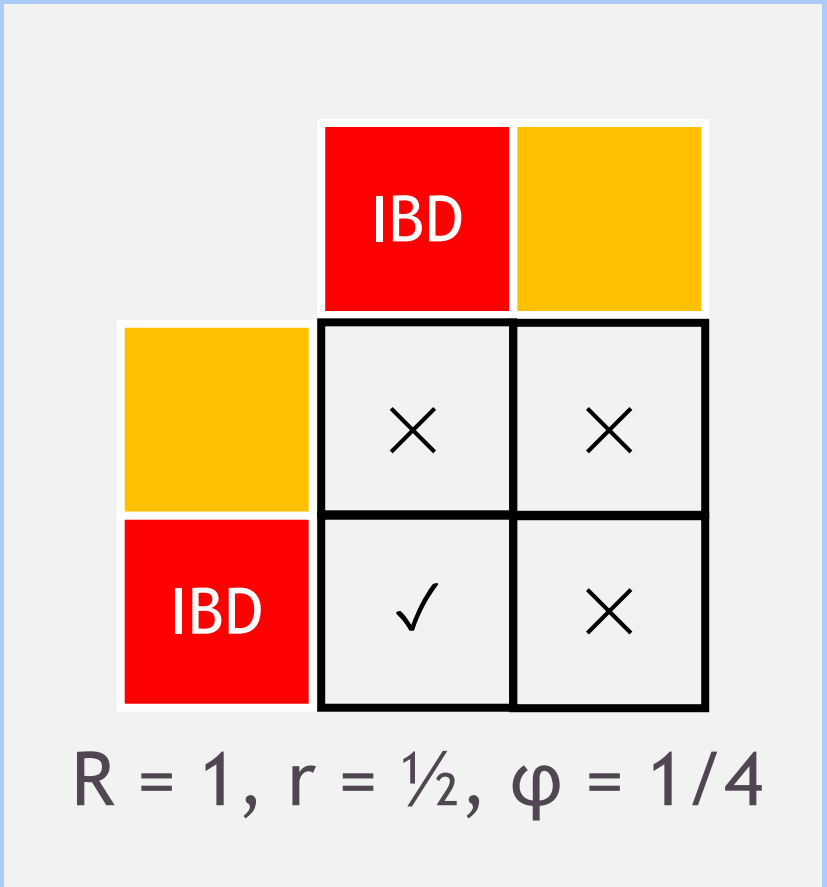
Coefficient of relatedness and IBD = 0

- φ decreases as the probability that a pair of individuals are IBD = 0 increases

Relationship	R	φ	IBS = 0
Monozygotic twins	0	0.5	0
Parent-child	1	0.25	0
Full sibs	1	0.25	0.25
2nd degree	2	0.125	0.5
3rd degree	3	0.0625	0.75
Unrelated	∞	0	1

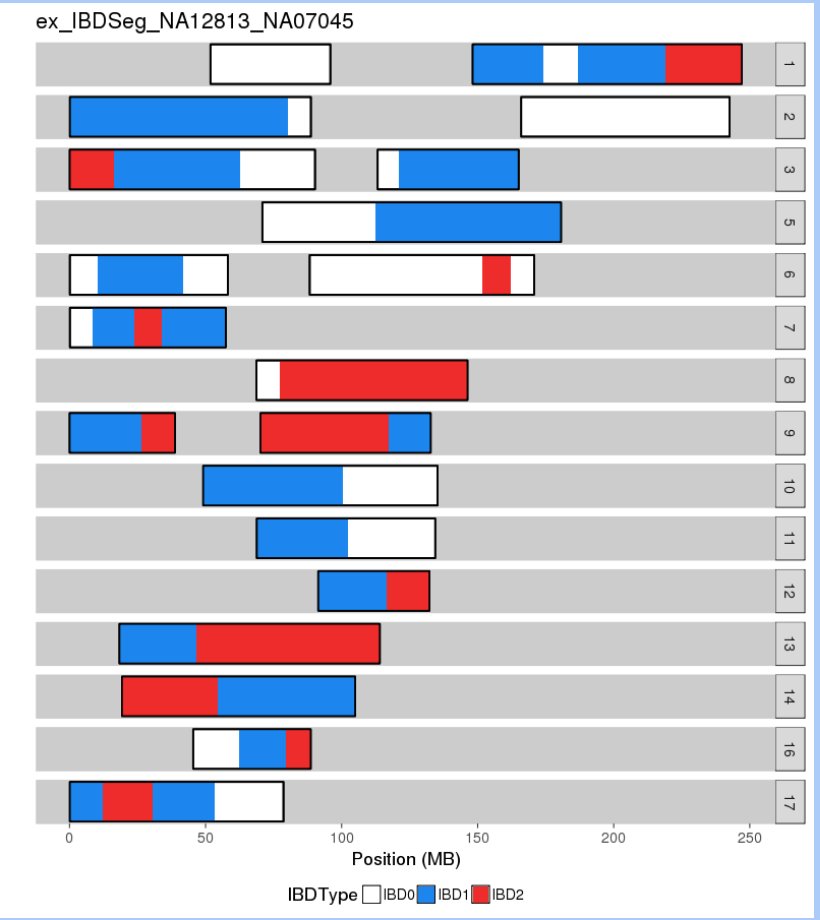
The coefficient of relatedness φ

- φ is the probability that any two alleles at a single locus chosen from two individuals are shared IBD
- φ is equal to half of $r = 1 / 2^R$



Kinship-based Inference for GWAS (KING)

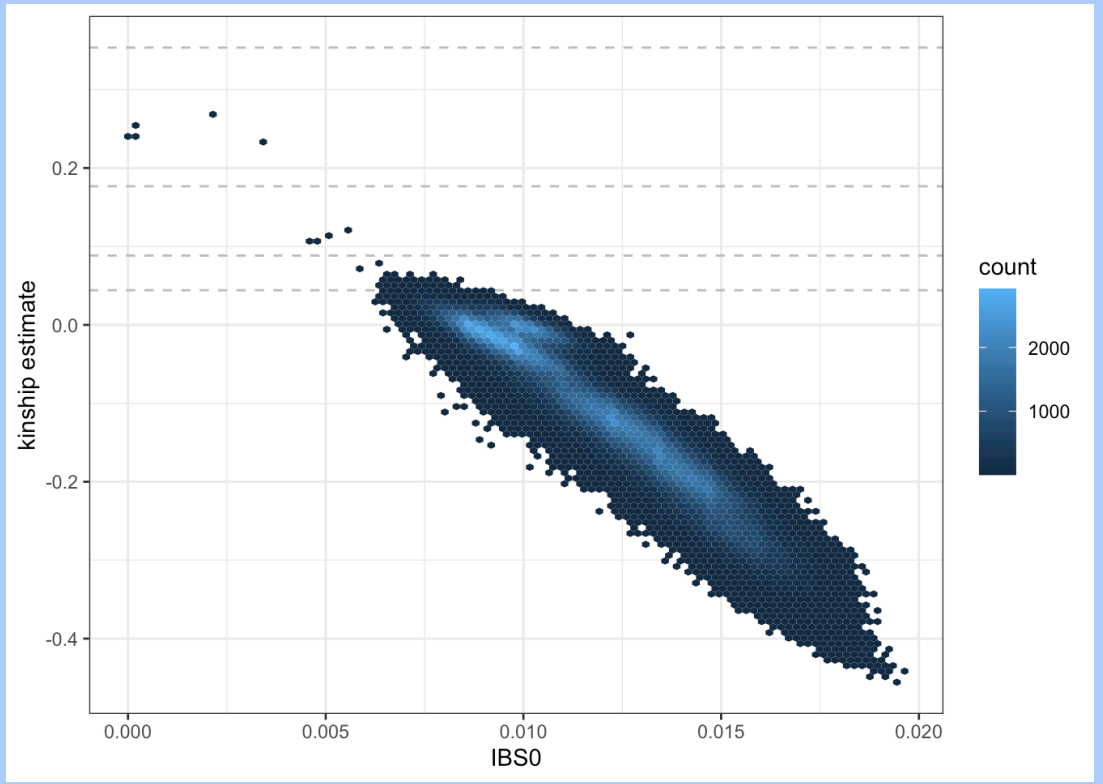
- Estimate φ and IBD sharing from the number of sites two individuals are both heterozygotes (Aa,Aa) or opposite homozygotes (AA,aa)



https://www.kingrelatedness.com/manual.shtml

Kinship-based Inference for GWAS (KING)

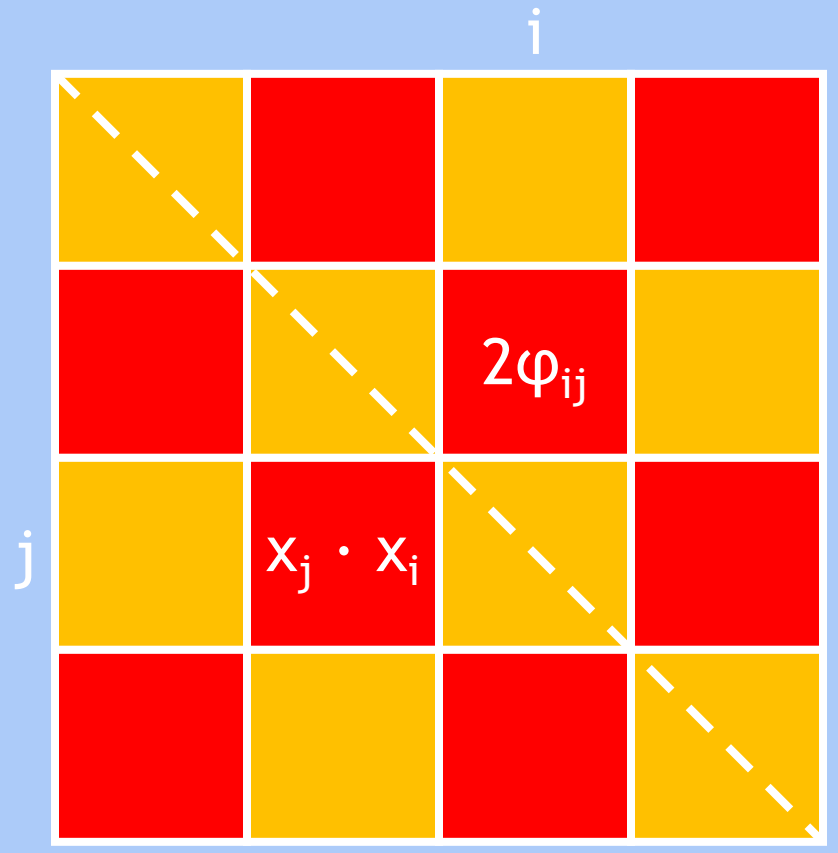
- φ is plotted vs. the fraction of IBS = 0 sites (AA,aa)
- Negative estimates indicate unrelated individuals from different populations



https://uw-gac.github.io/SISG_2021/ancestry-and-relatedness-inference.html

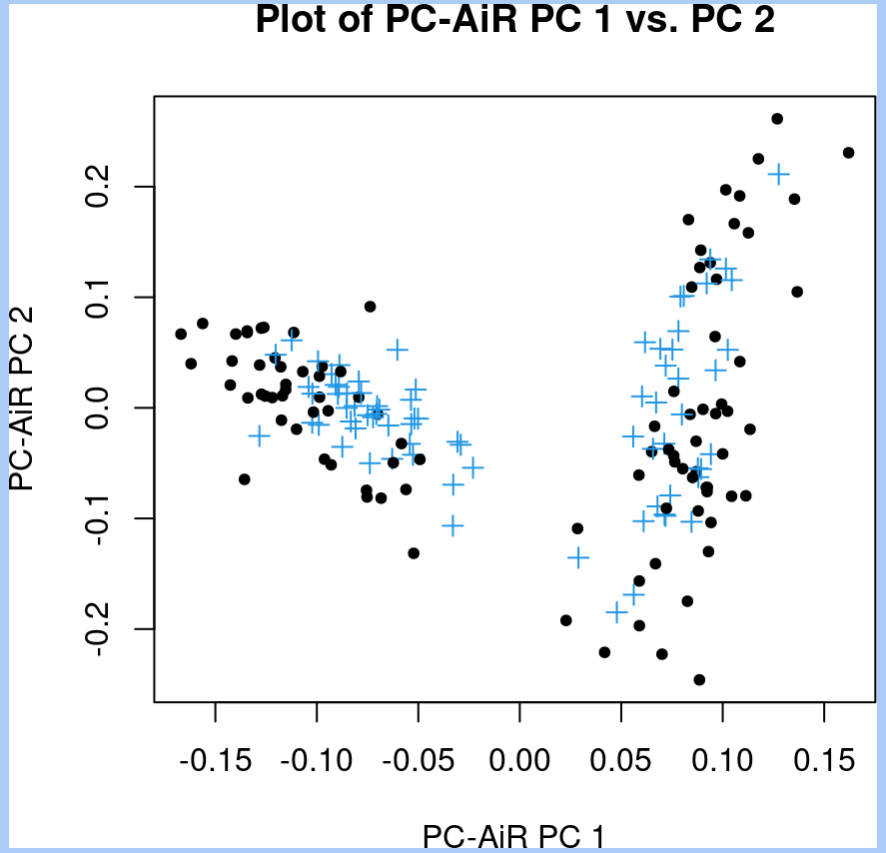
Updating the GRM

- The KING kinship coefficients 2φ are approximately equal to the GRM, but the estimate may be biased by population structure



PC-AiR: PCA in Related Samples

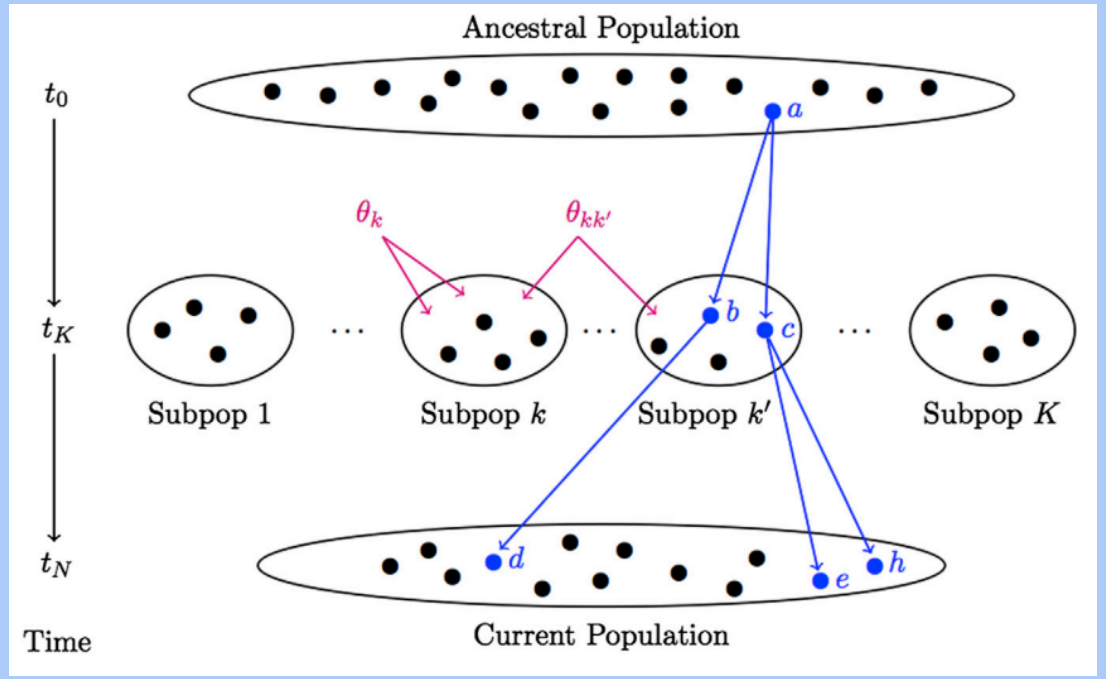
- Based on the KING estimates, PC-AiR computes PCs for a set of unrelated individuals (black)
- PCs for the remaining samples (blue) are estimated from their similarity to the unrelated subset



https://bioconductor.org/packages/devel/bioc/vignettes/GENESIS/inst/doc/pcair.html

PC-Relate

- PC-Relate uses the updated PCs to correct the GRM for population structure
- The updated GRM reflects just recent kinship



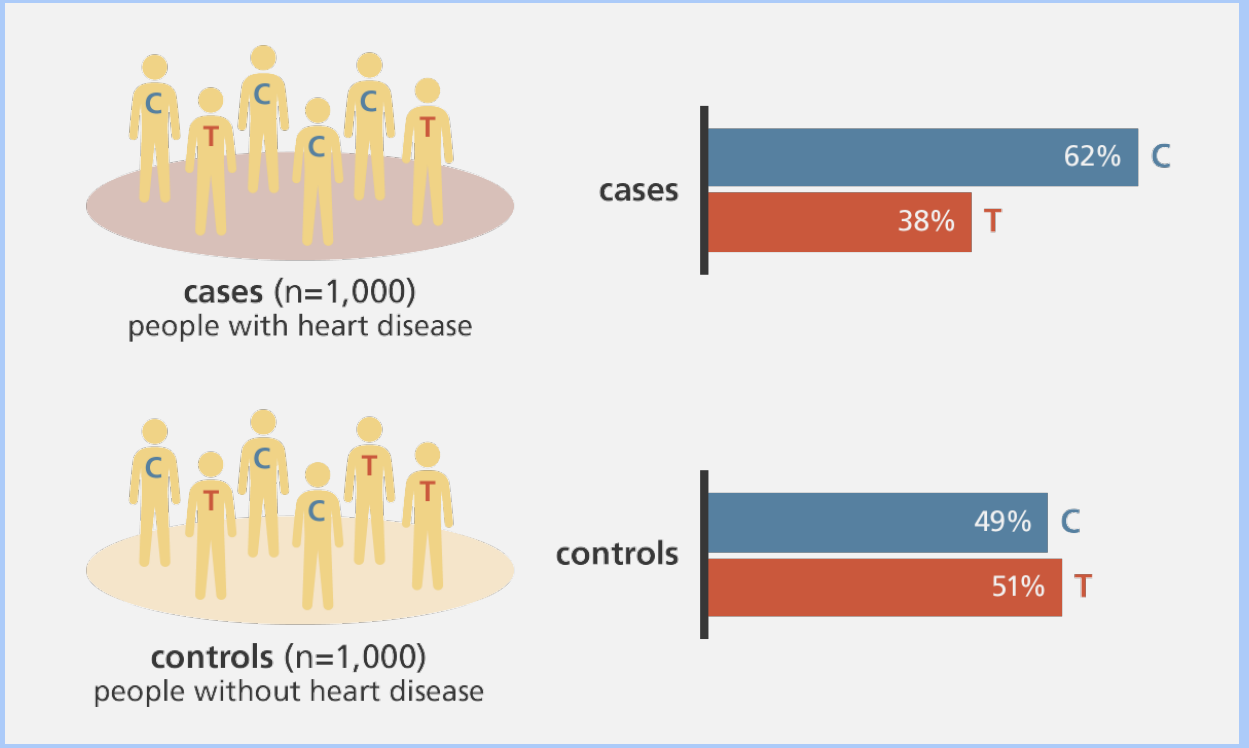
https://pubmed.ncbi.nlm.nih.gov/26748516/

Association testing

Logistic regression and linear mixed models

Case-control studies

- Is a genetic variant associated with disease?
- Is a genetic variant enriched in people with disease compared to people without?
- To find out, collect many people with disease (Cases) and many healthy individuals (Controls) from the same population



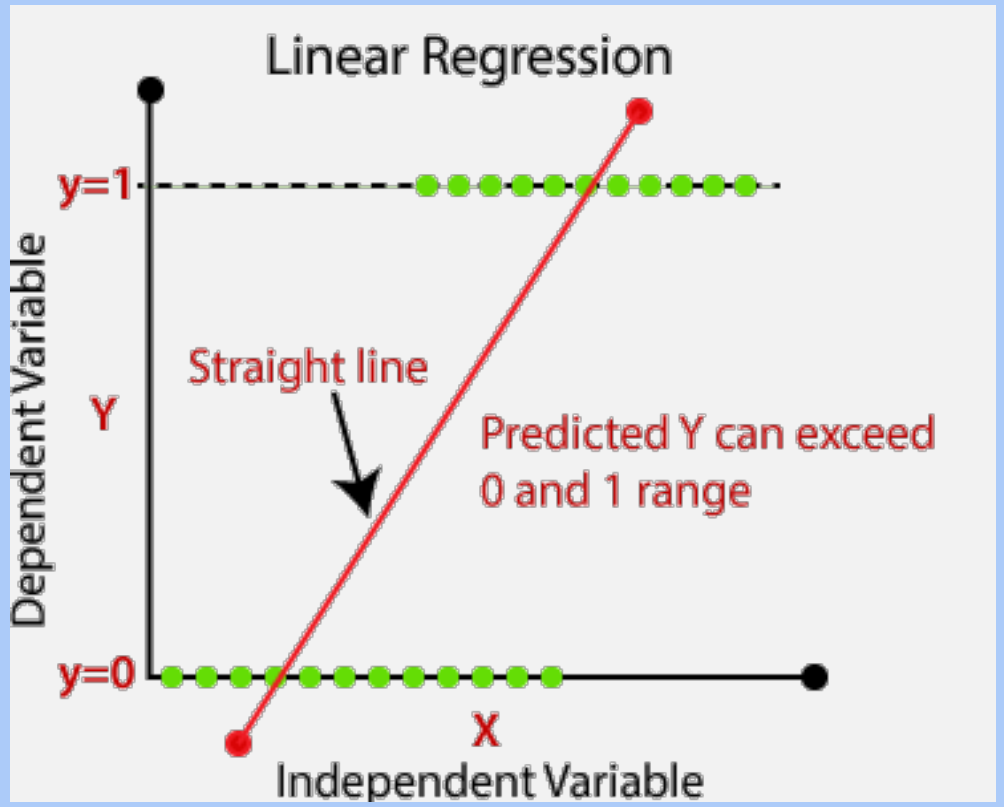
The odds ratio

- The OR is the ratio of the odds that Cases have the risk allele (620 / 380) to the odds that Controls have the risk allele (490 / 510)
- The OR is a **crude** measure of association that is not **adjusted** for other covariates (age, sex, ethnicity, etc.) that may also be associated with disease

	Cases	Controls
C	620	490
T	380	510

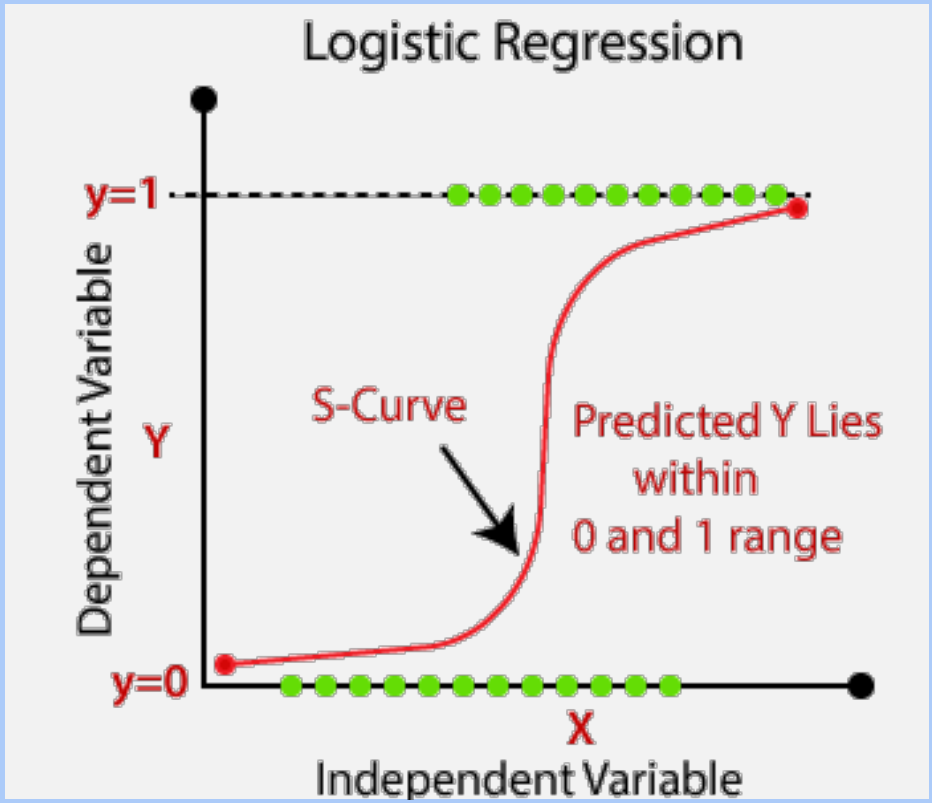
OR = (620 × 510) / (490 × 380) = 1.70

Linear vs. logistic regression



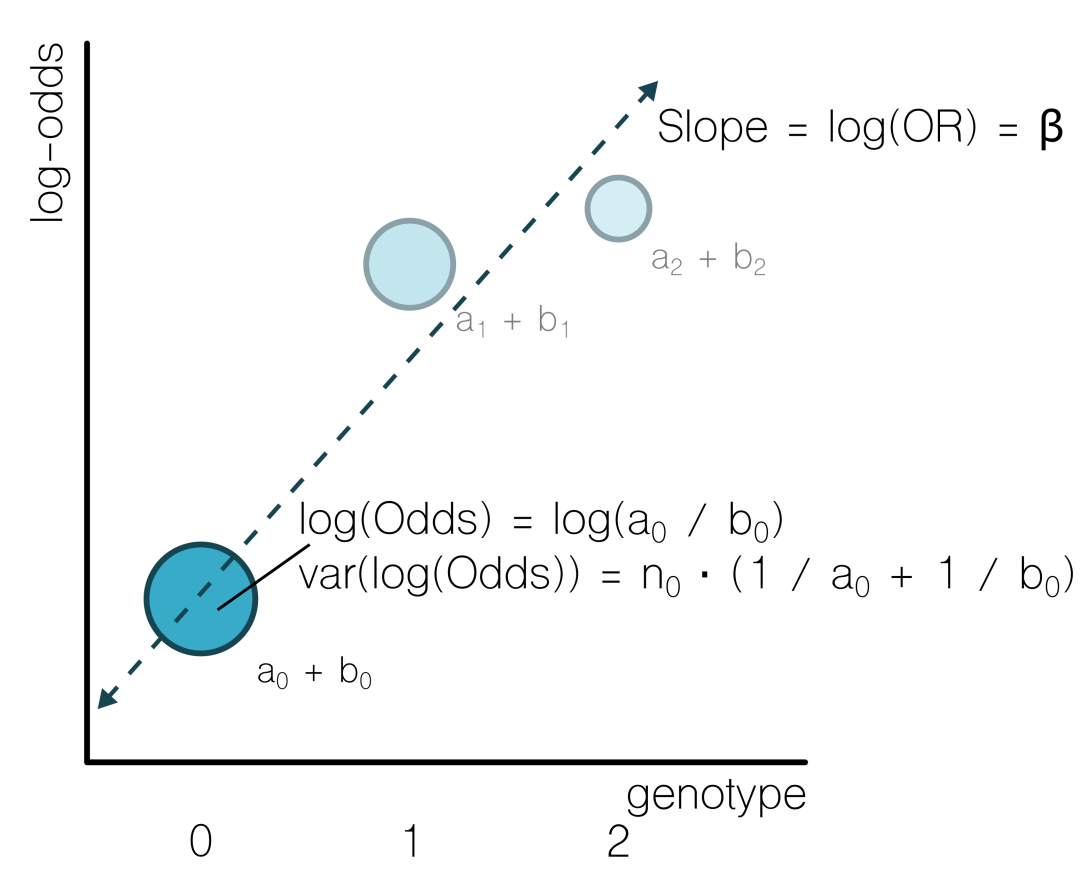
- In linear regression, we can find the association of a **continuous variate** Y with a predictor X_1 and other covariates X_2, X_3, etc.

Linear vs. logistic regression



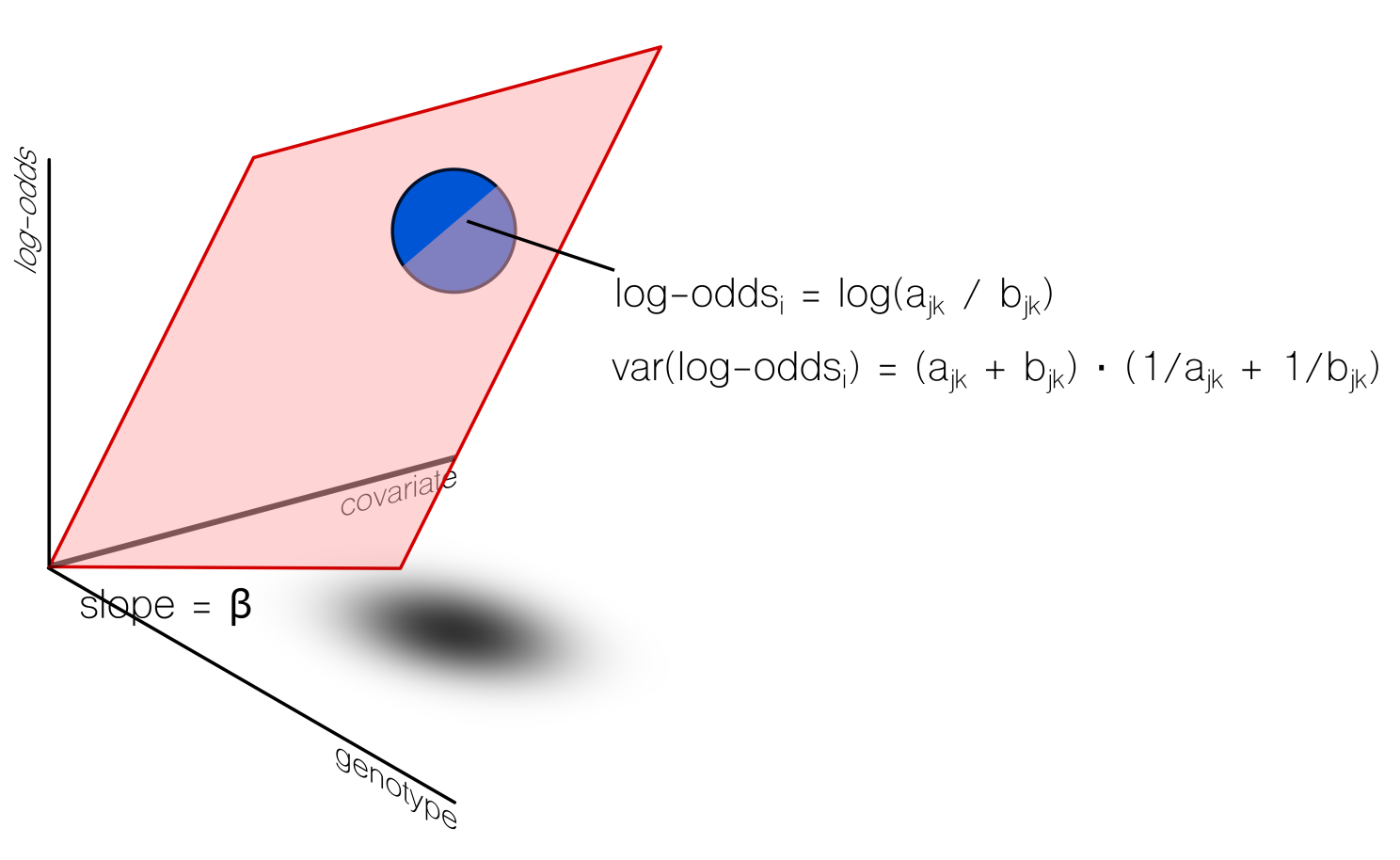
- In logistic regression, we can find the association of a **binary variate** Y with a predictor X_1 and other covariates X_2, X_3, etc.
- The sigmoid curve is an individual's probability of developing disease

Linear vs. logistic regression



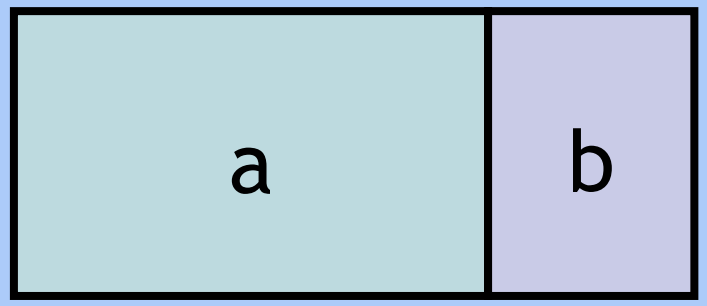
- Logistic regression can be thought of like linear regression is we transform the OR into the log(OR) and regress vs. SNP genotype

Linear vs. logistic regression



- Other covariates can be accounted for as additional independent variables
- The model is actually fit using the principle of **maximum-likelihood**

Simulating a binary phenotype

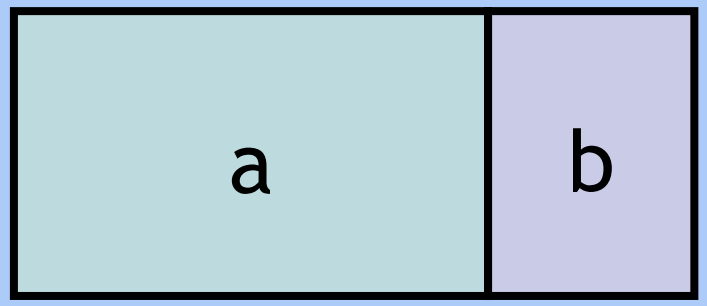


- If odds = a / b, then prob = a / (a + b) = odds / (1 + odds)

Simulating a binary phenotype

$$\log(\text{odds}) = \beta_0 + X_1\beta_1$$

- β_0 is the baseline odds
- β_1 is the log-OR
- X_1 is the SNP genotype

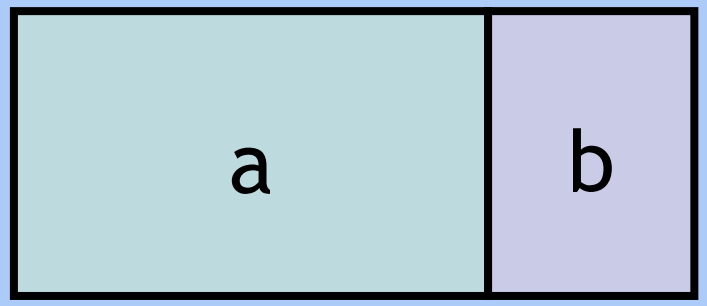


- If odds = a / b, then prob = a / (a + b) = odds / (1 + odds)

Simulating a binary phenotype

$$\text{prob} = \frac{e^{\beta_0 + X_1\beta_1}}{1 + e^{\beta_0 + X_1\beta_1}}$$

- prob is the probability of developing disease (being a Case in the study)

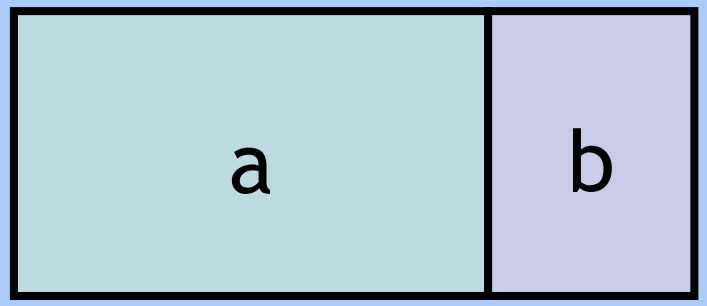


- If odds = a / b, then prob = a / (a + b) = odds / (1 + odds)

Simulating a binary phenotype

$$\text{prob} = \frac{e^{(X_1 - \overline{X}_1)\beta_1}}{1 + e^{(X_1 - \overline{X}_1)\beta_1}}$$

- β_0 becomes the mean log-odds so that the mean odds of disease is 1 (50% Cases, 50% Controls)

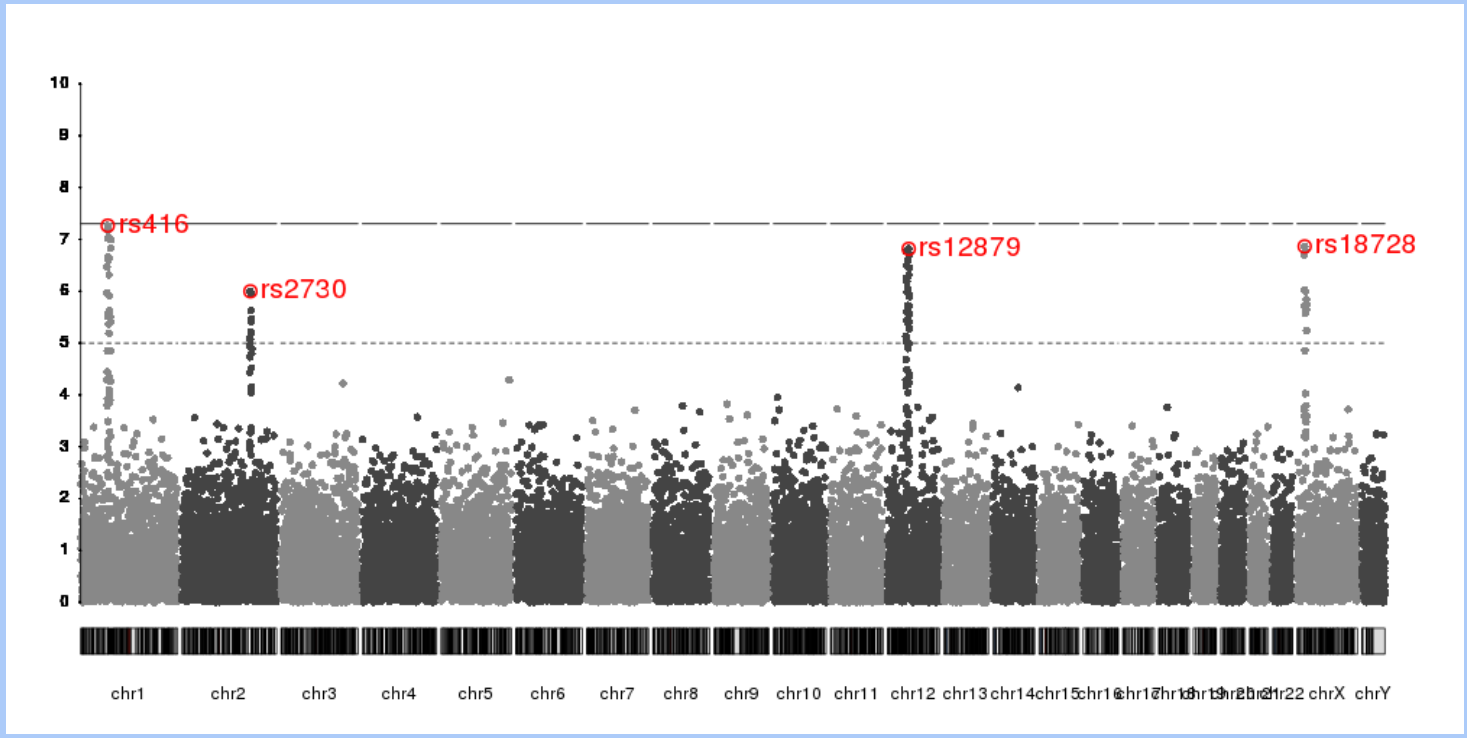


- If odds = a / b, then prob = a / (a + b) = odds / (1 + odds)

Estimating the SNP effect

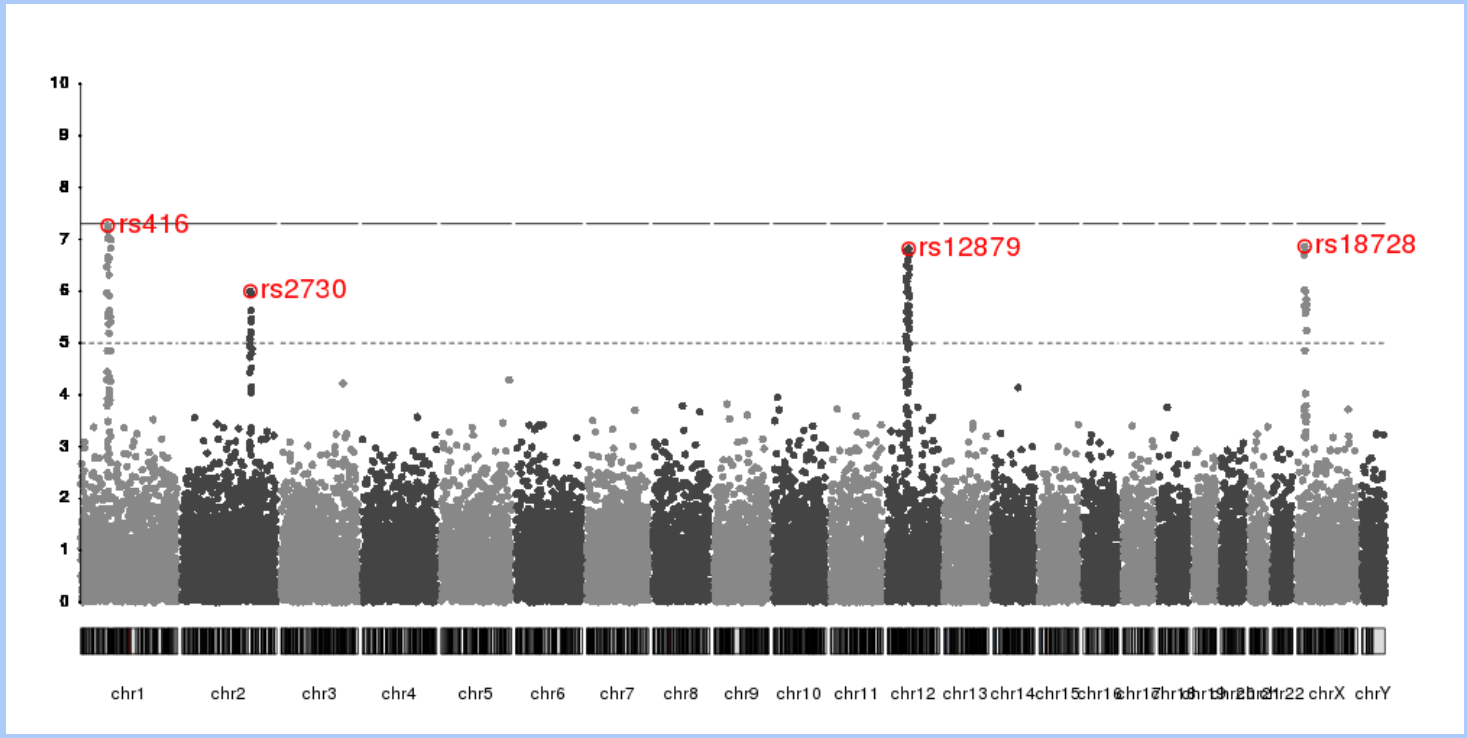
- We want to be able to **detect** the association of one SNP with disease by fitting the model $Y = \beta_0 + \beta_1 X_1 + \cdots$ and finding a slope β_1 significantly different from 0

Estimating the SNP effect



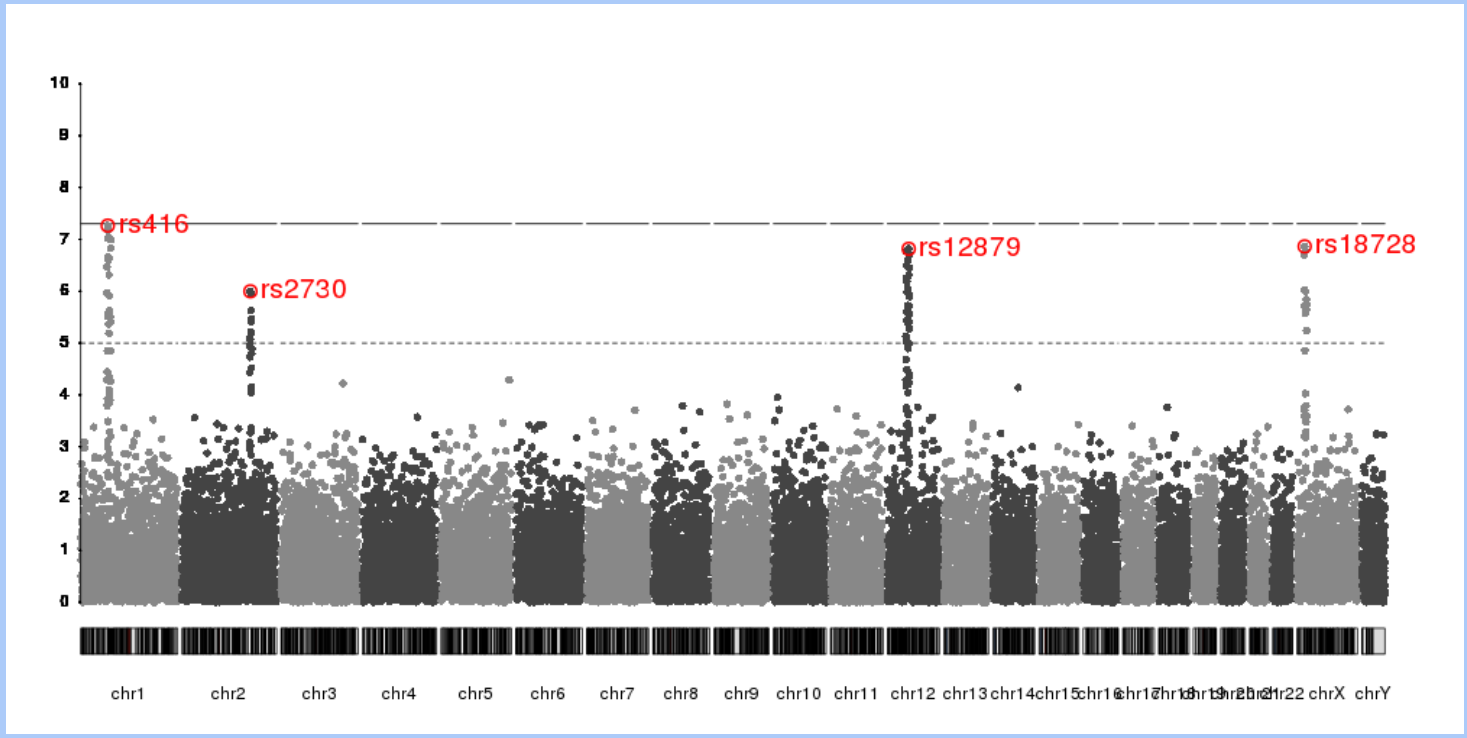
- A **Manhattan plot** gives the p-value of the log-OR estimate for each SNP

Estimating the SNP effect



- Because there are more SNPs than subjects, we cannot fit all SNPs at once

Estimating the SNP effect

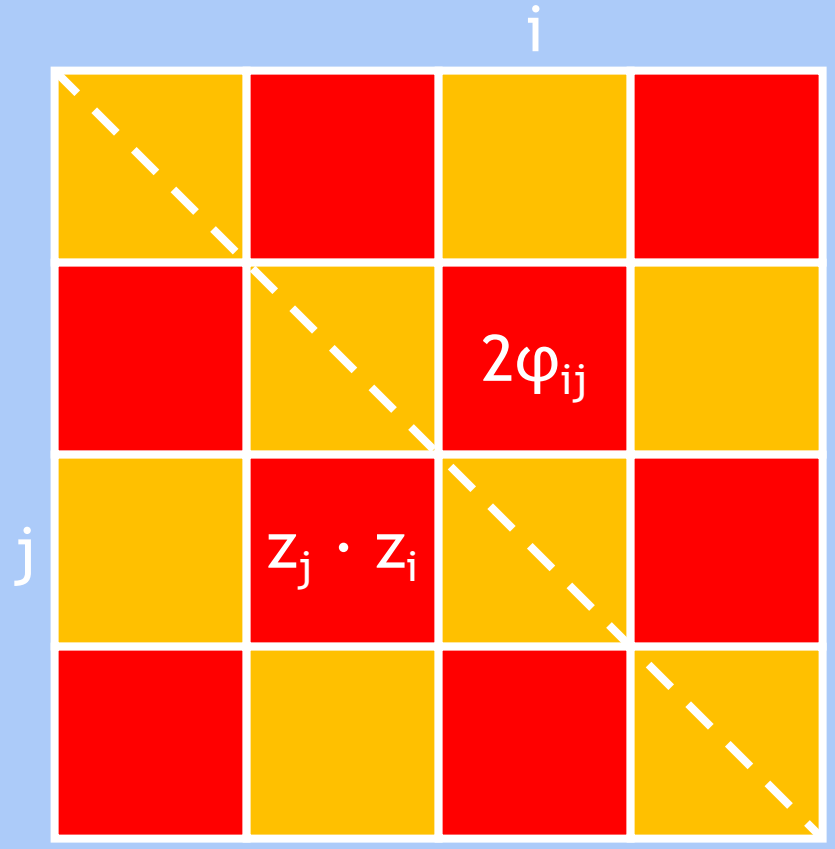


- But we can fit one SNP plus the “average” effect of all the remaining SNPs

Linear mixed models

- The solution for the **best estimate** of the SNP effect β_1 in the presence of all the remaining SNPs involves the GRM ZZ^T (from PC-Relate)

$$\mathbf{X}^T \left(\mathbf{I} + \mathbf{Z}\mathbf{Z}^T\right)^{-1} \hat{\beta} = \mathbf{X}^T \left(\mathbf{I} + \mathbf{Z}\mathbf{Z}^T\right)^{-1} \mathbf{Y}$$



Linear mixed models

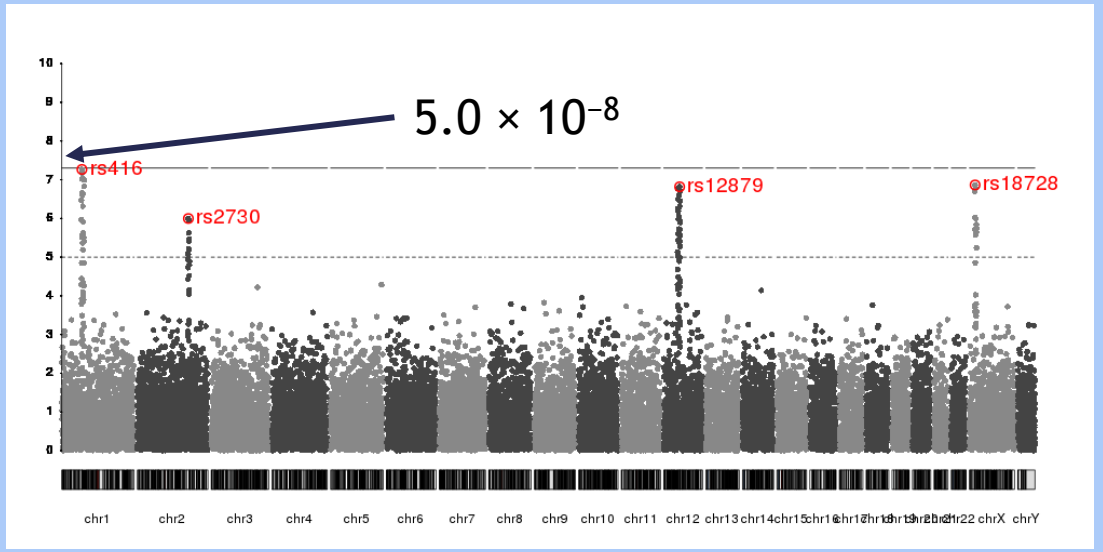
- Other covariates commonly included in the model are age, sex, and the **first few genotype principal components** (from PC-AiR)

Linear mixed models

- If the model including the SNP represents a significant improvement over the **null model** (the model without the SNP), we can reject the null hypothesis that the OR = 1

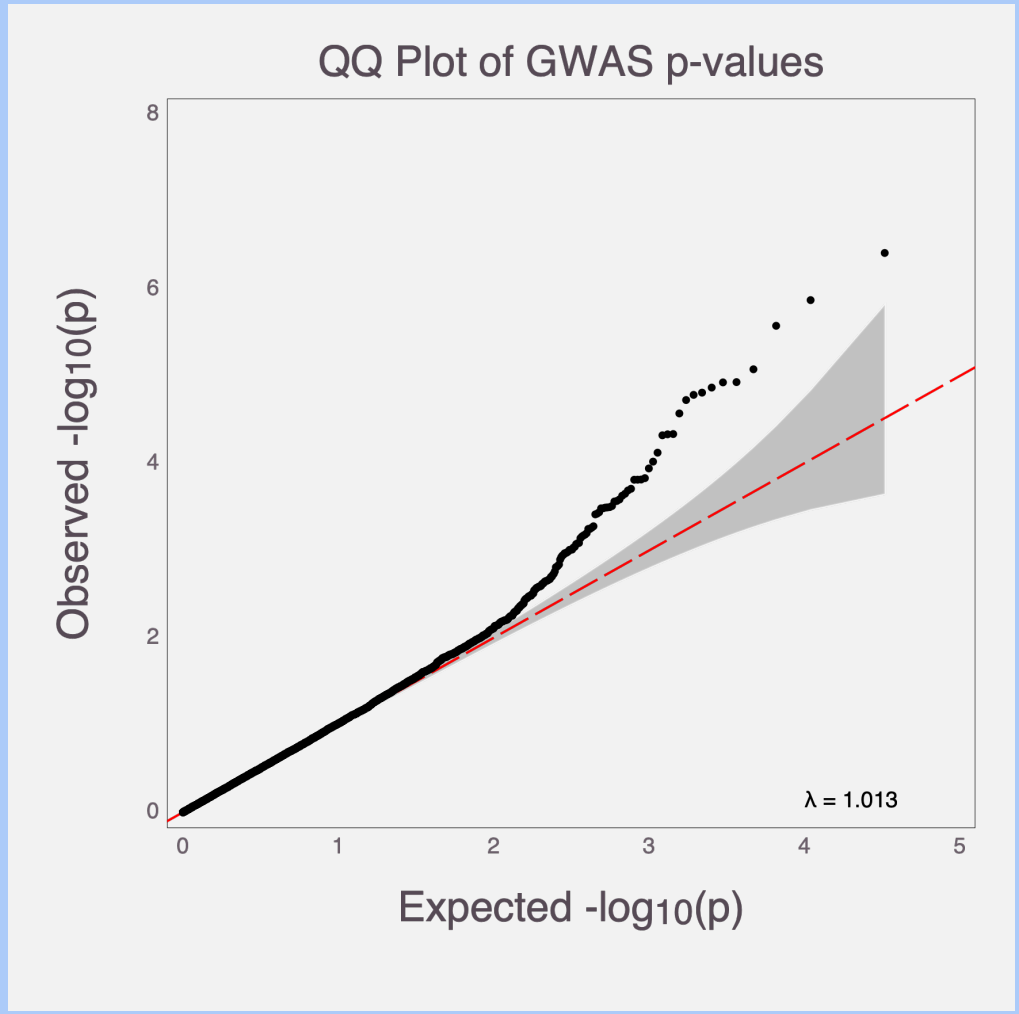
Linear mixed models

- But because of **multiple-testing**, our p-value threshold is 0.05 / 10^6 (i.e., you perform the same test 10^6 times)
- SNPs with $p < 5.0 \times 10^{-8}$ are said to achieve **genome-wide significance**



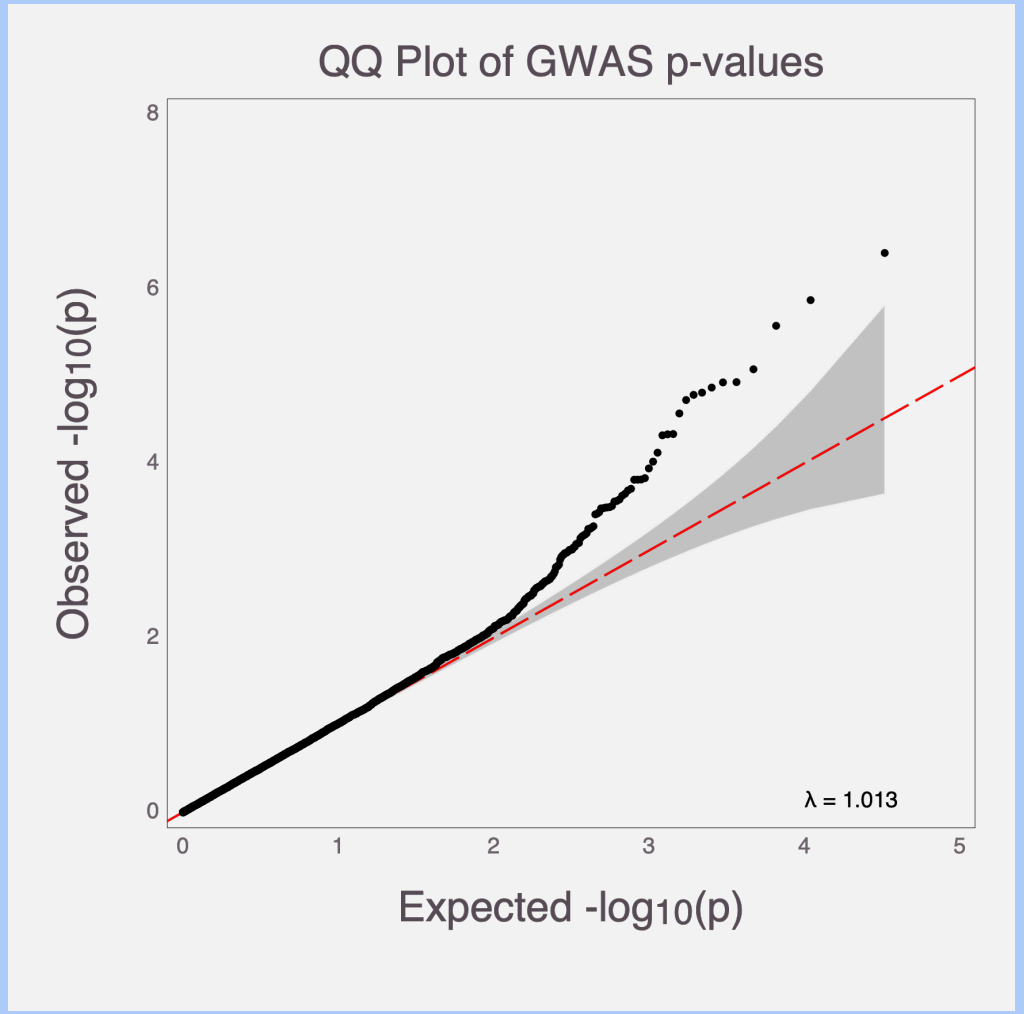
QQ plots

- To assess if the distribution of SNP effects is significantly different from that expected by chance, we make a quantile or **QQ plot**



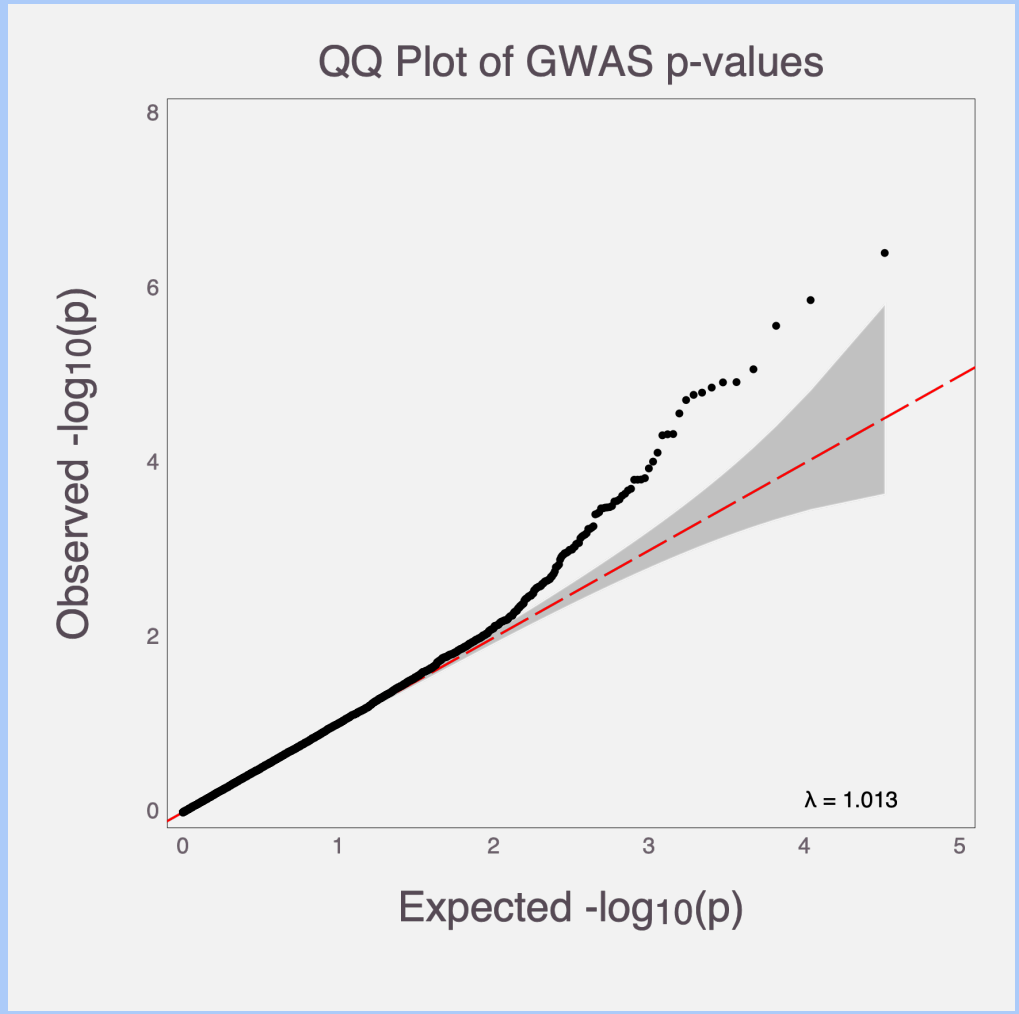
QQ plots

- Put the **observed** p-value (negative log-10) in order from smallest to biggest



QQ plots

- The **expected** p-values for the quantiles of m SNPs, are 1/m, 2/m,..., 1
- Take the negative log-10 and put in order from smallest to biggest



QQ plots

- SNPs falling above the line of identity indicate an excess of quantiles (ß's) with small probabilities