BIOL 350: Bioinformatics

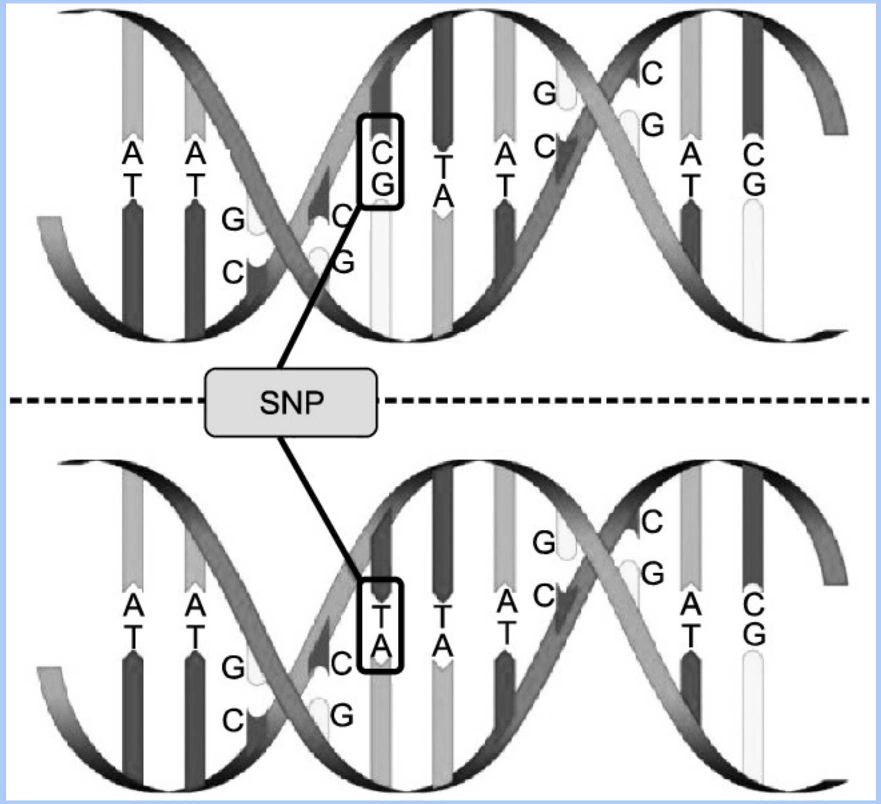
Introduction to genetic association studies

What is a SNP?

Polymorphisms and their role in genetics

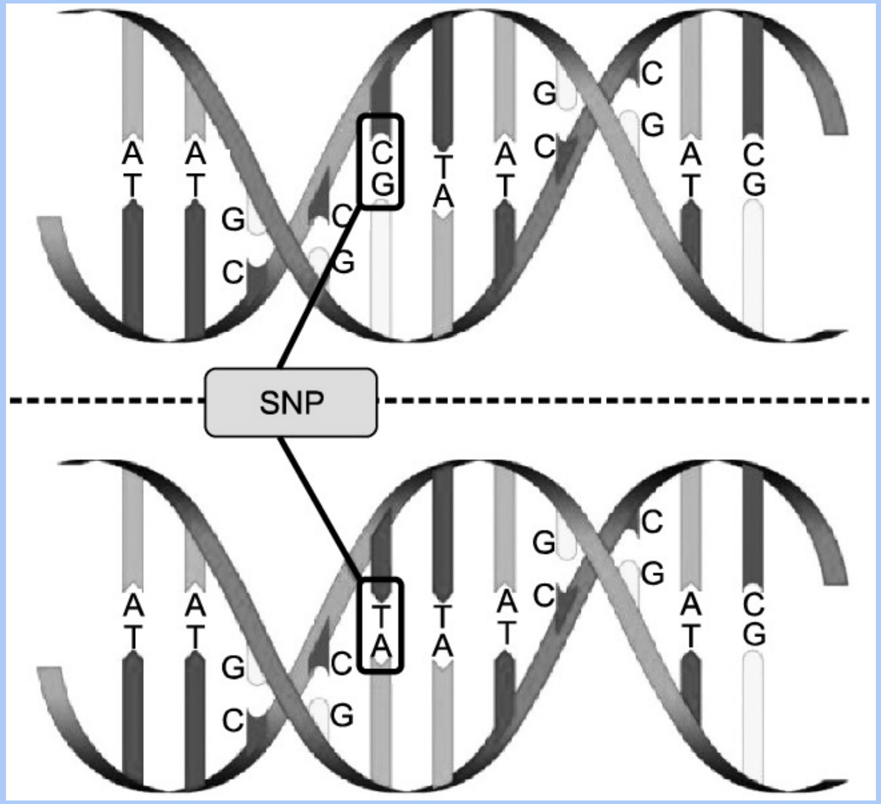
Single-nucleotide polymorphisms

- **Polymorphism** is the tendency of DNA to admit of different nucleotide pairs at a single locus



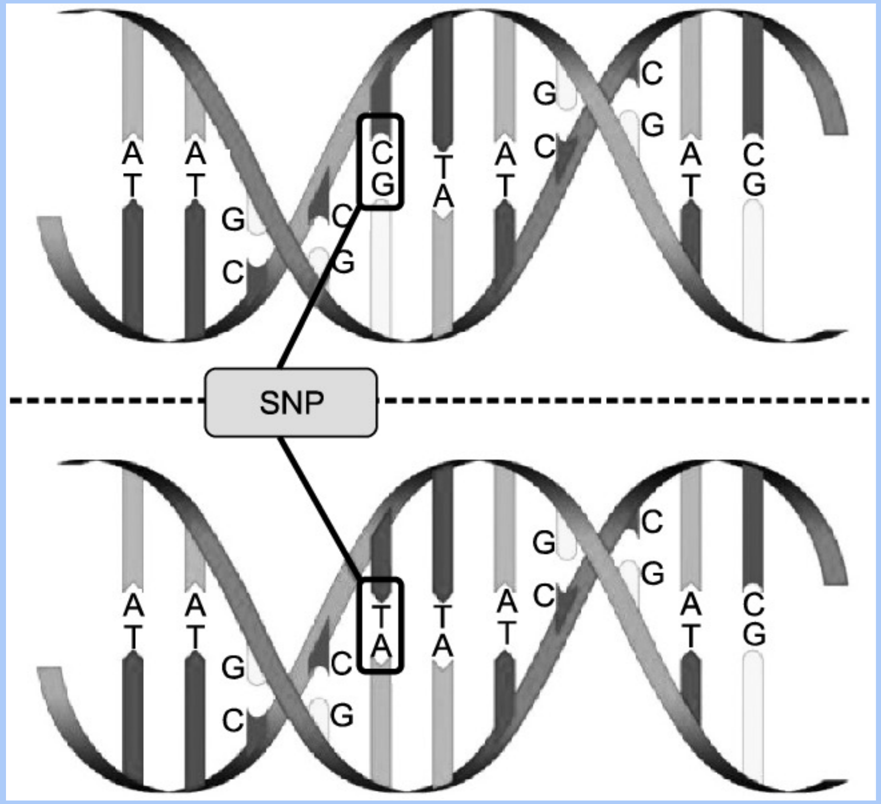
Single-nucleotide polymorphisms

- Of 3.2 billion bases, any individual is polymorphic at 4–5 million sites



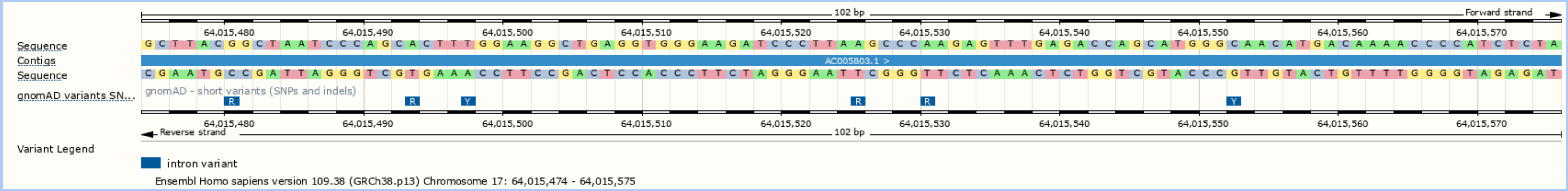
Single-nucleotide polymorphisms

- The more common allele is called the **major allele**
- The less common allele is called the **minor allele**



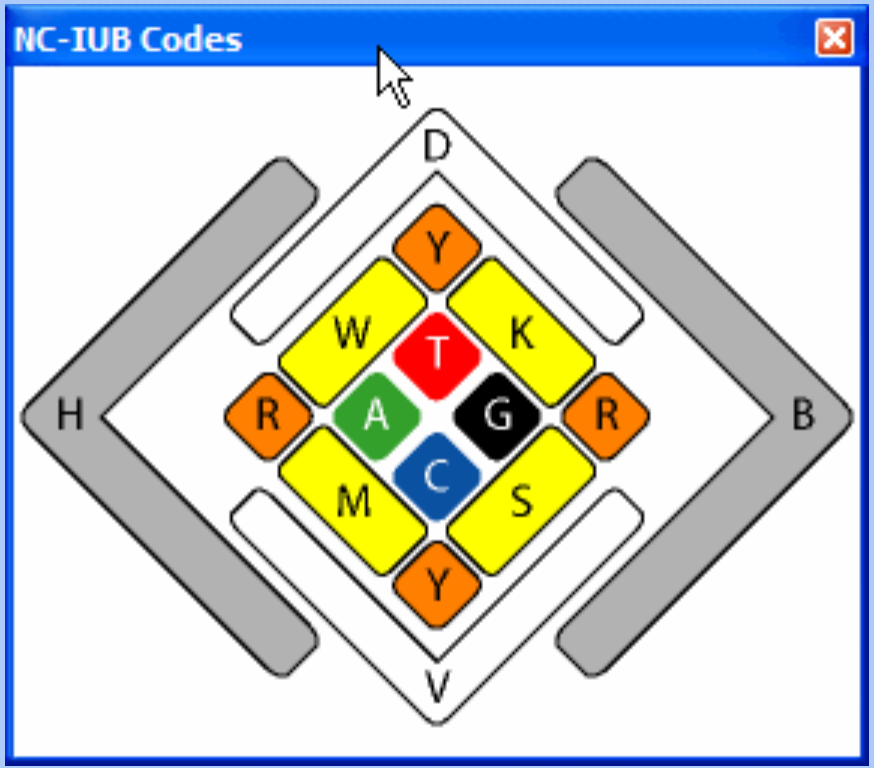
IUPAC-IUB SNP codes

- More than just A, T, G, and C?



IUPAC-IUB SNP codes

- Each polymorphism is coded by its possible alleles



https://www.gendx.com/SBTengine/Help_220/hs310.htm

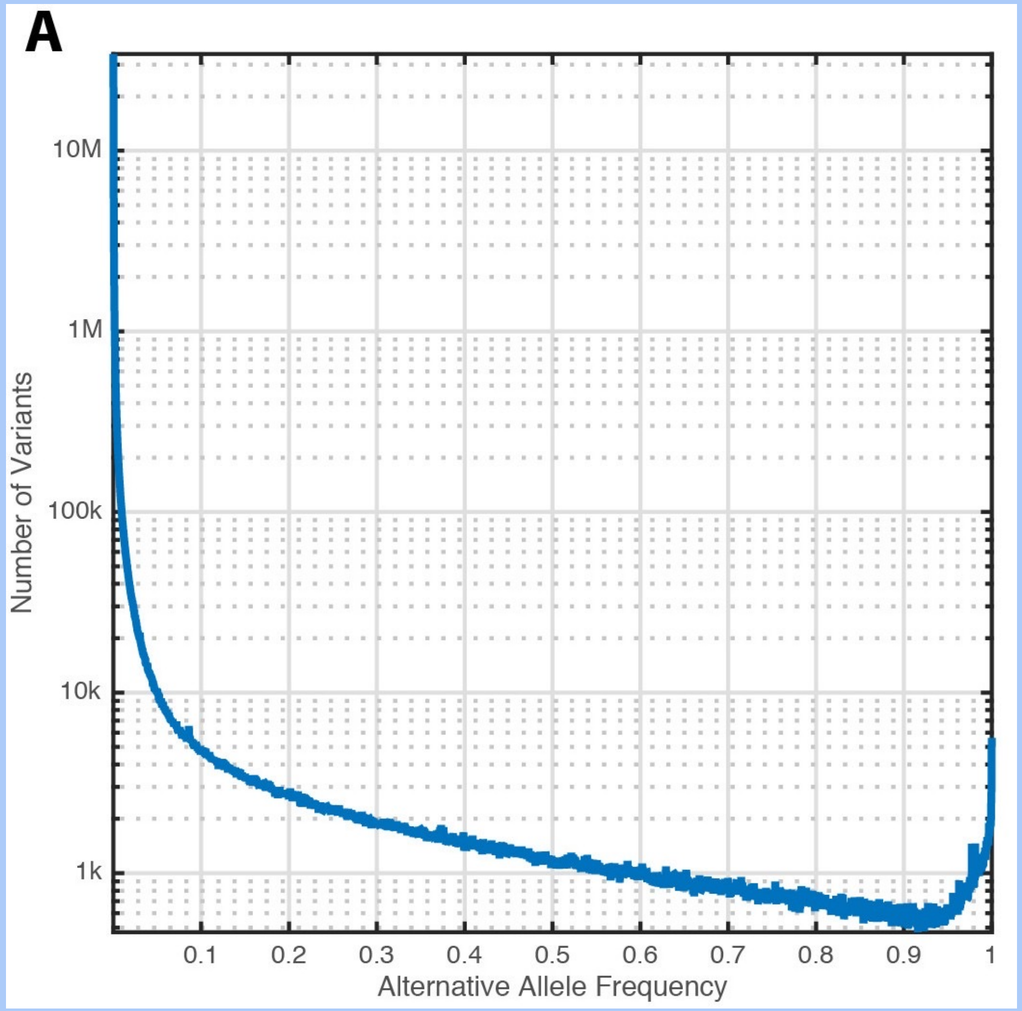
IUPAC-IUB SNP codes

- Each polymorphism is coded by its possible alleles

Code	Meaning	Explanation
R	A or G	Pu**R**rine
Y	C or T	P**Y**rimidine
S	G or C	**S**trong H-bonding
W	A or T	**W**eak H-bonging
K	G or T	**K**eto bases
M	A or C	a**M**ino bases
B	C or G or T	not A
D	A or G or T	not C
H	A or C or T	not G
V	A or C or G	not T
N	A or C or G or T	A**N**y

Many rare SNPs

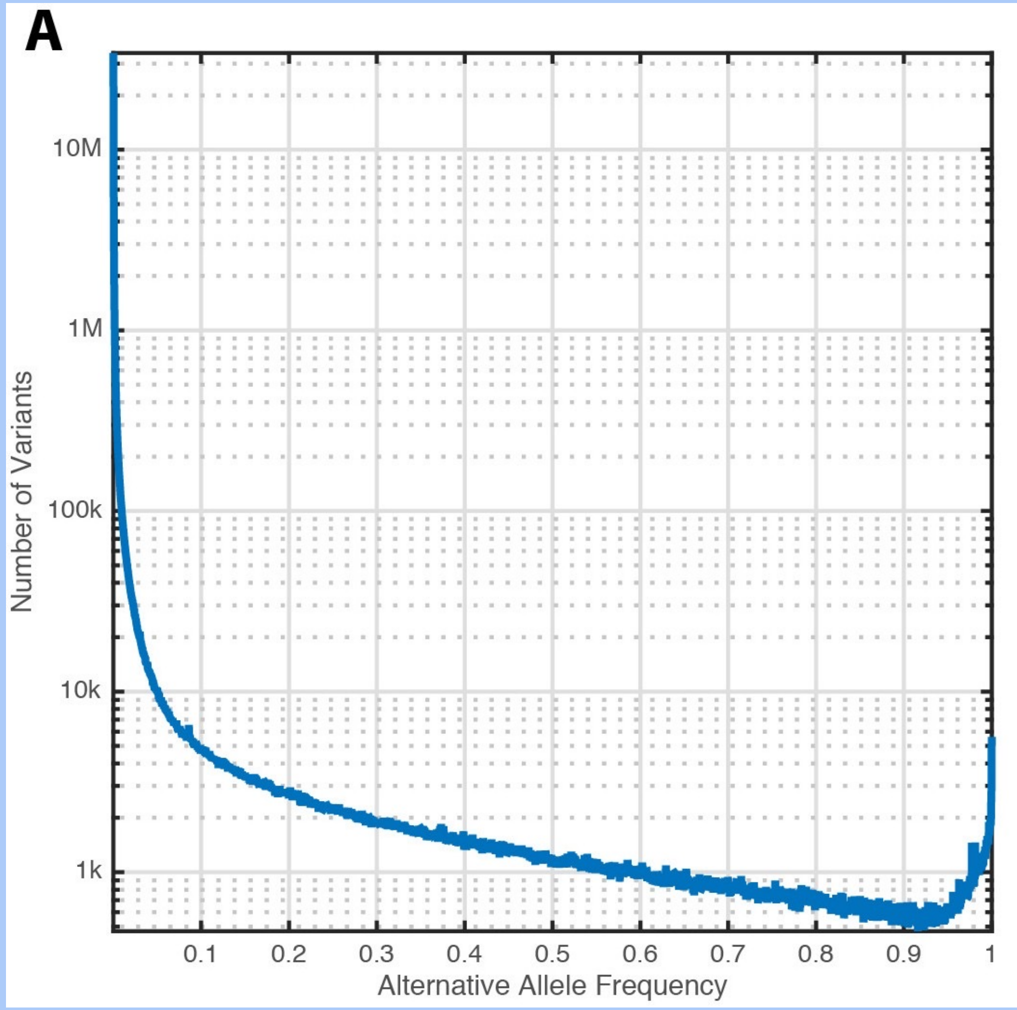
- Common SNPs have **minor allele frequency (MAF)** >5%



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

Many rare SNPs

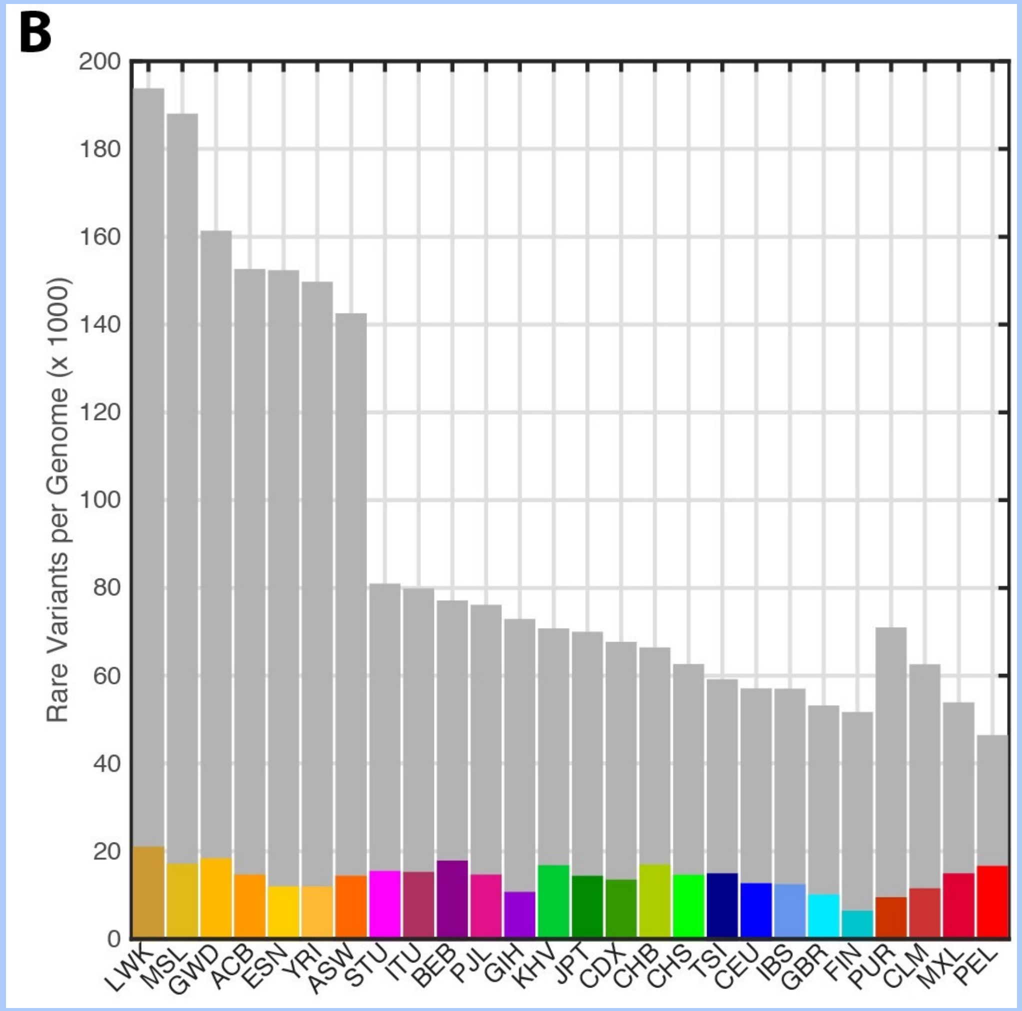
- Most SNPs of the >600 million known SNPs are very rare (frequency < 0.5%)



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

Many rare SNPs

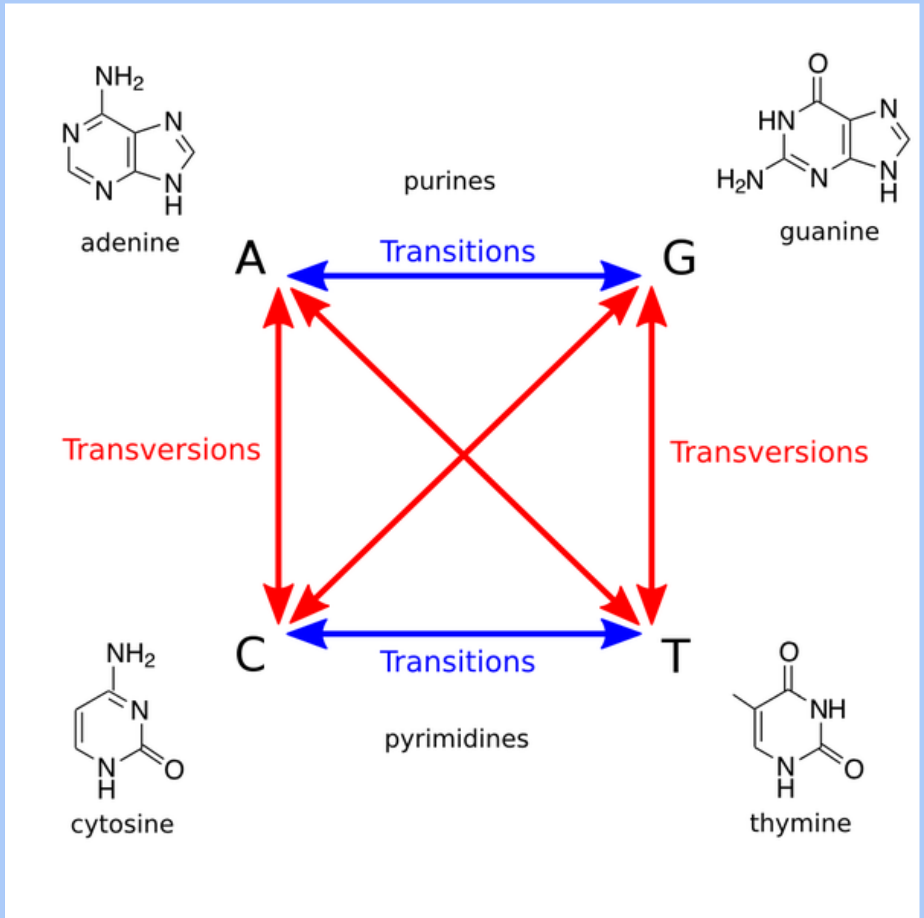
- But only <5% of an individual's genome consists of rare SNPs



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

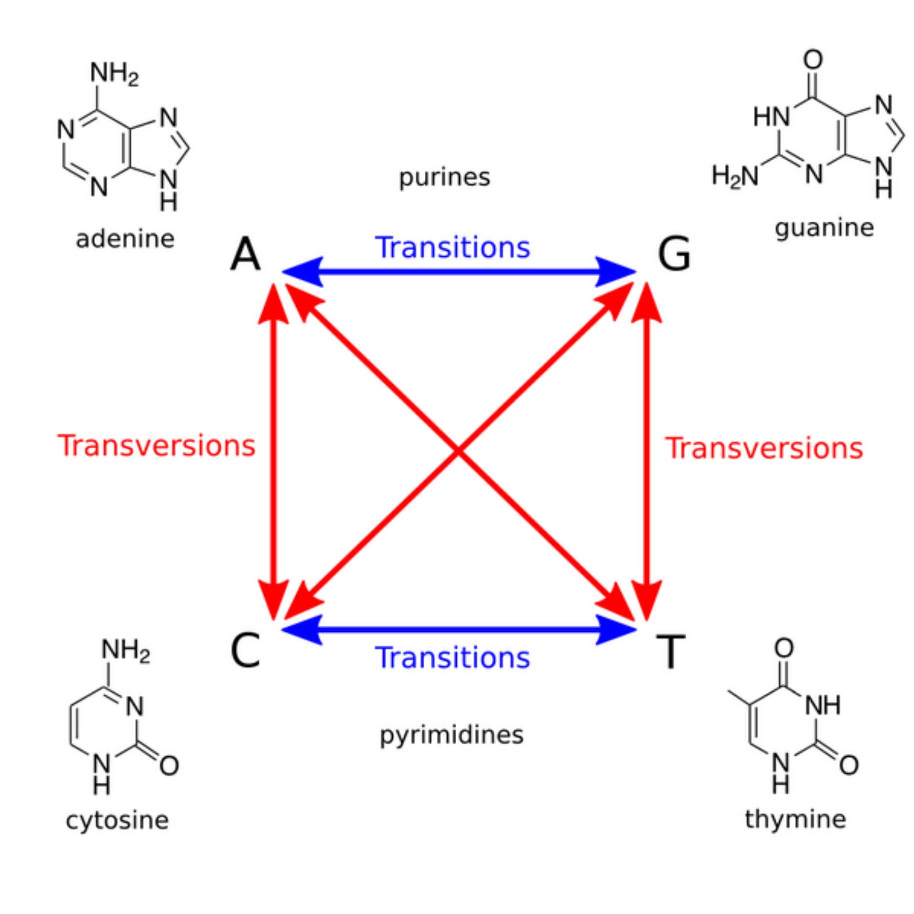
Transitions and transversions

- **Transitions** occur between nucleotides of the same type (purines or pyrimidines)



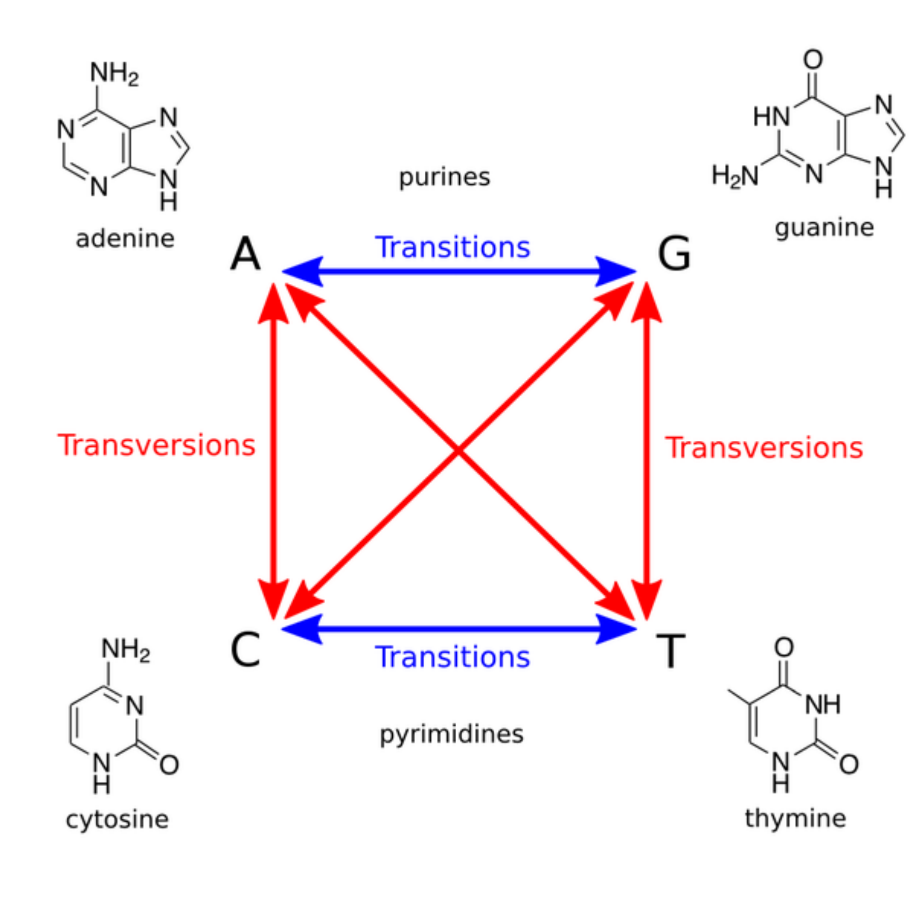
Transitions and transversions

- **Transversions** occur between nucleotides of opposite type (between purines and pyrimidine)



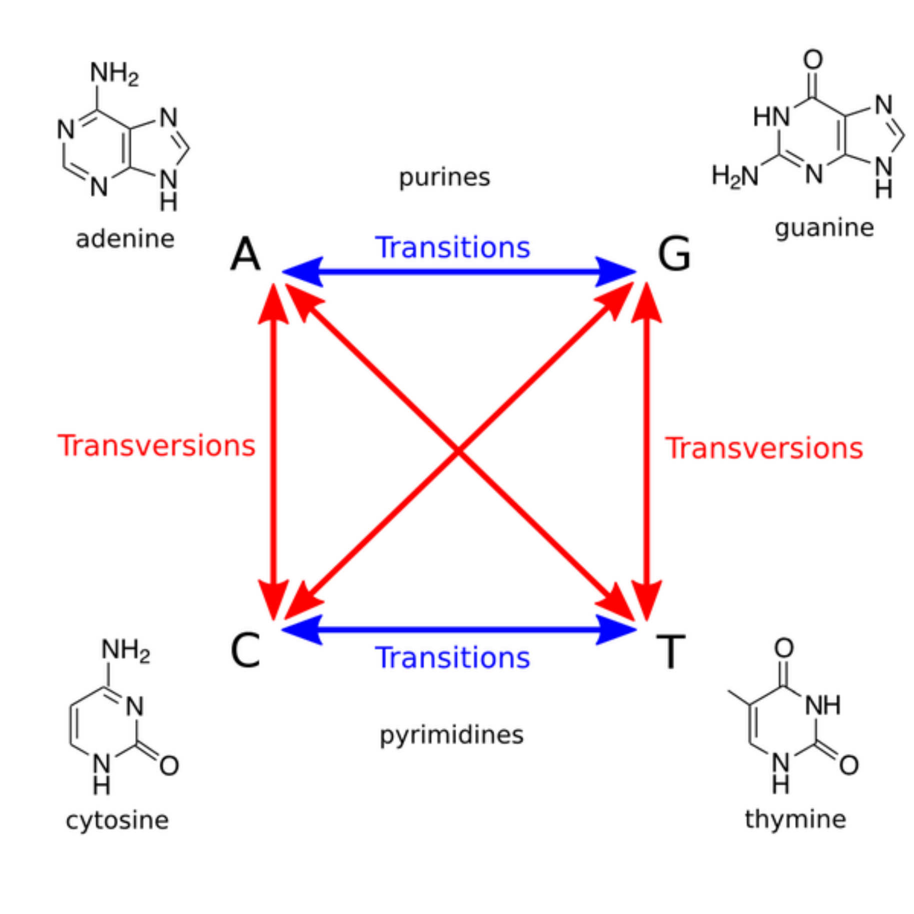
How many polymorphisms are there?

- If there are n nucleotide pairs, there are n **symmetric** conversions:
  - A/T → T/A transversion
  - C/G → G/C transversion



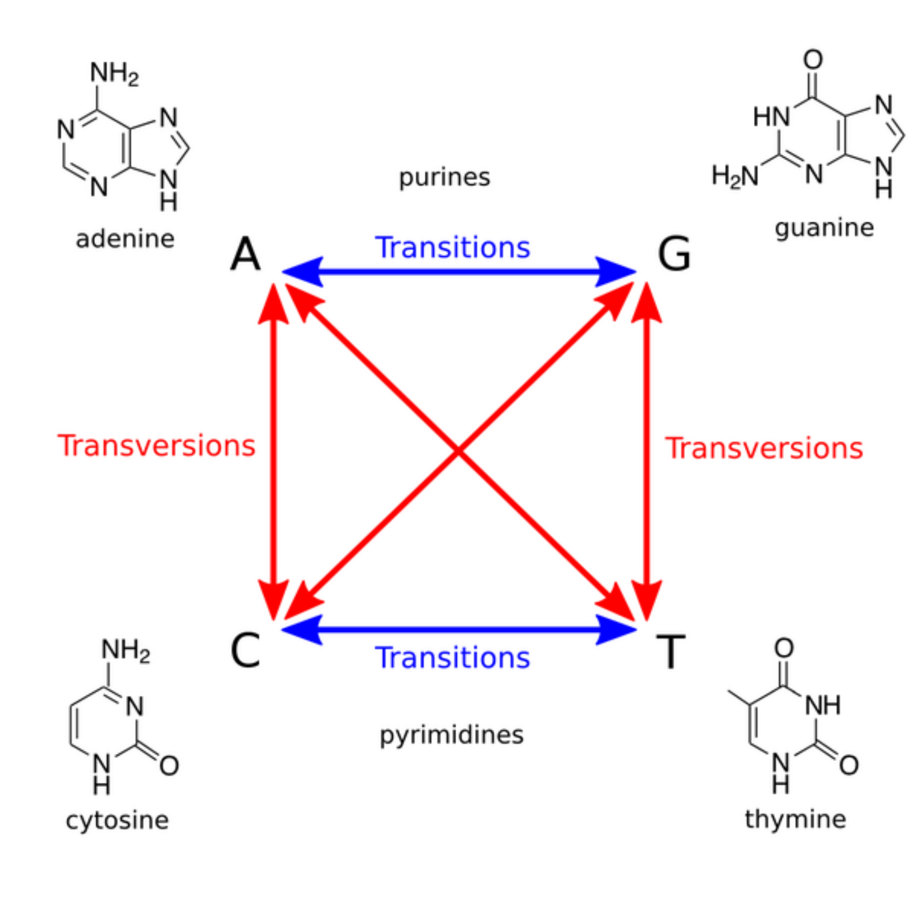
How many polymorphisms are there?

- If there are n nucleotide pairs, there are $n(n-1)$ **asymmetric** conversions:
  - A/T → C/G transversion
  - A/T → G/C transition



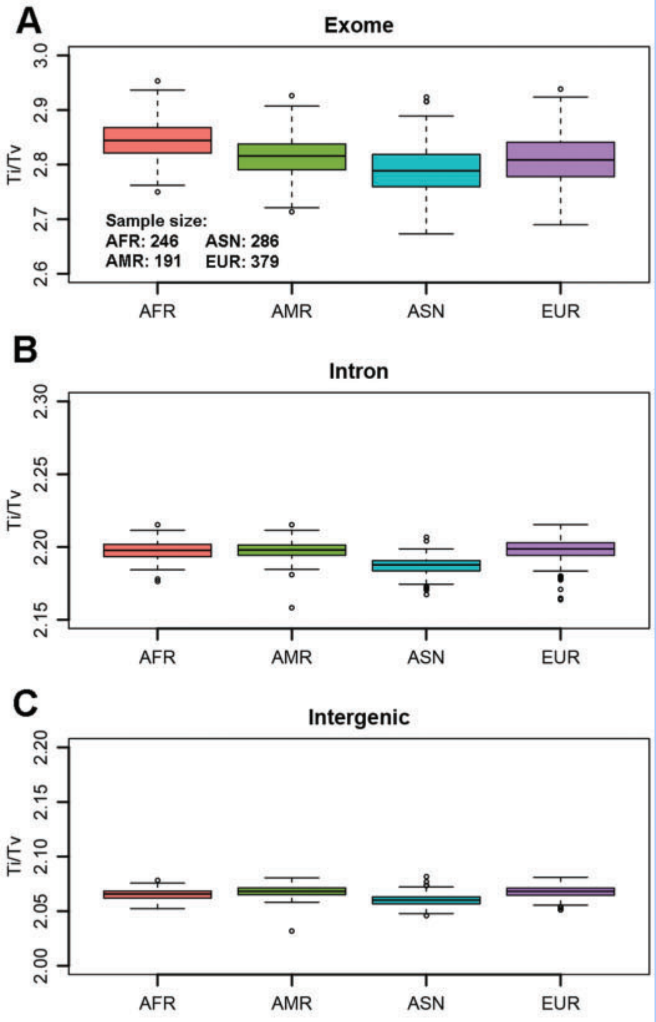
How many polymorphisms are there?

- A total of $n + n(n-1) = n^2$ polymorphims



Transition-transversion ratio

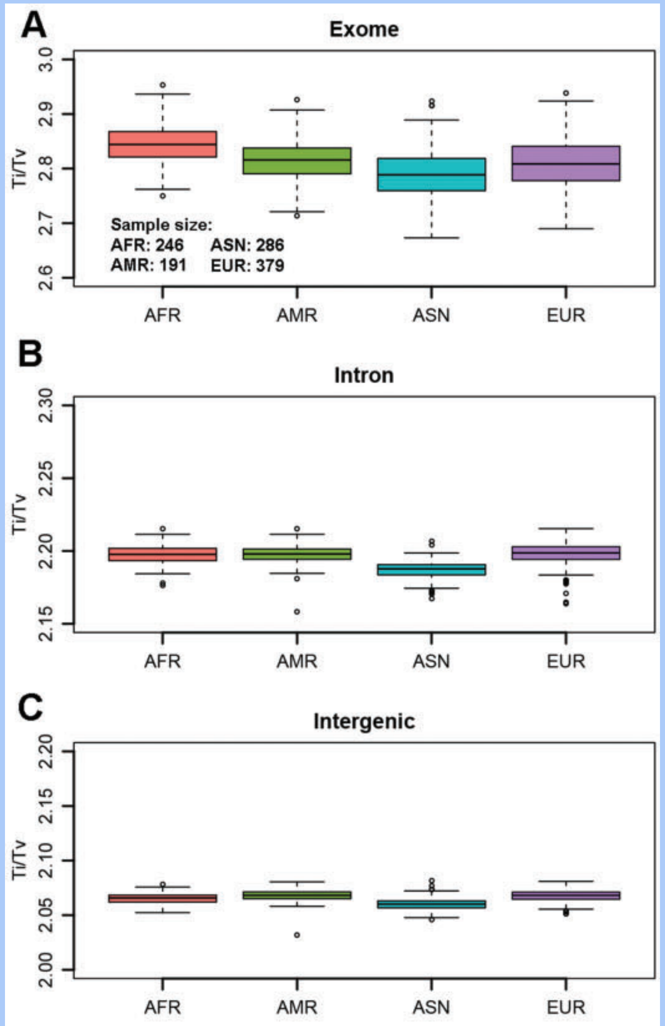
- Even though there are three times as many transversions possible as transitions, in humans the ratio of transitions to transversions is approximately 2, genome-wide



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

Transition-transversion ratio

- In coding regions, the Ti:Tv ratio is as high as 3



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

Generation of sequencing data

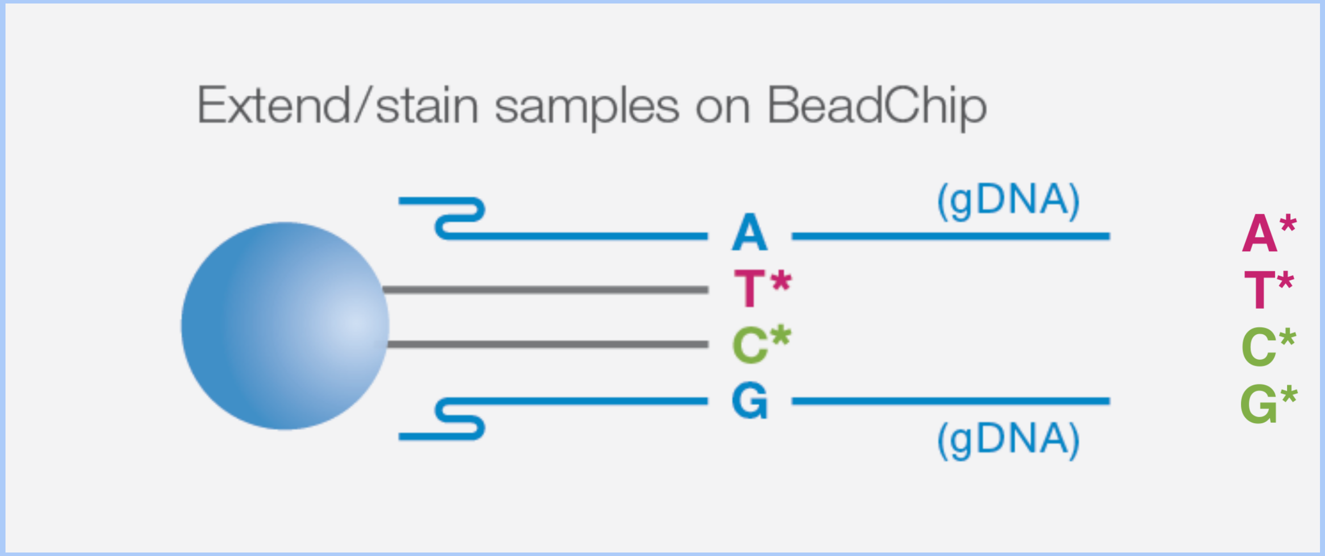
Sequencing technologies and data formats

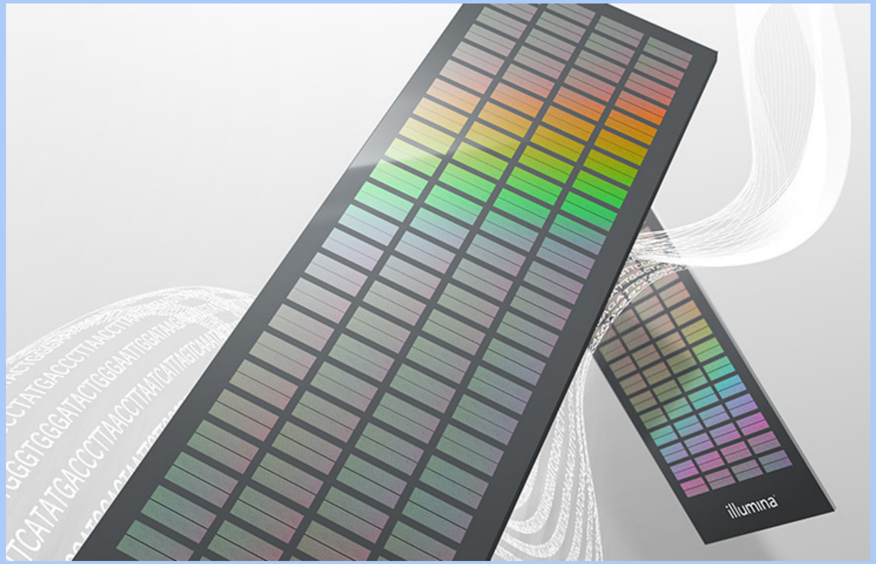
How do we get human genotypes?

- SNP Chips
- Whole-genome sequencing

SNP Chips

- Genomic DNA binds to a complementary sequence and incorporates a fluorescently labelled nucleotide



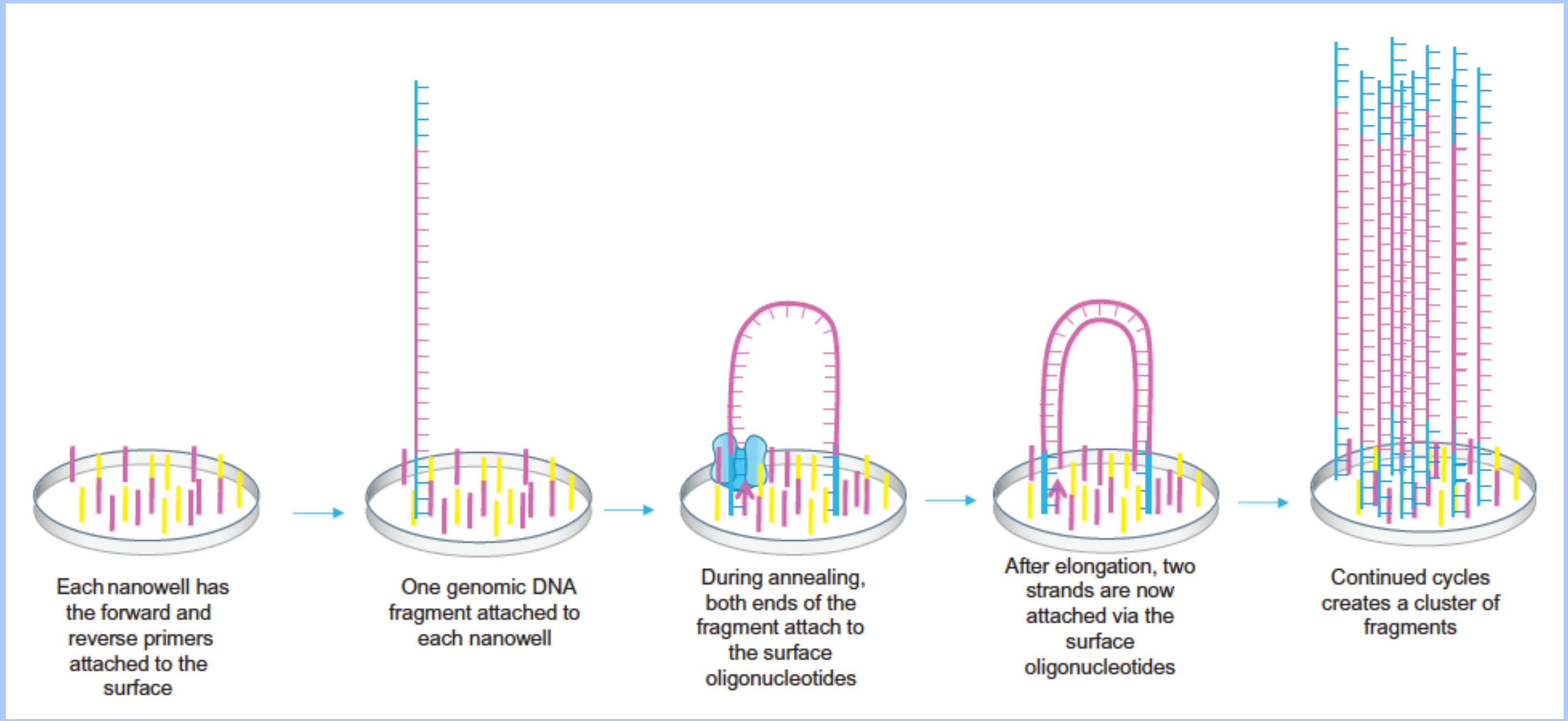


SNP Chips

- The ratio of red to green at a spot identifies the sample allele

Whole-genome sequencing (WGS)

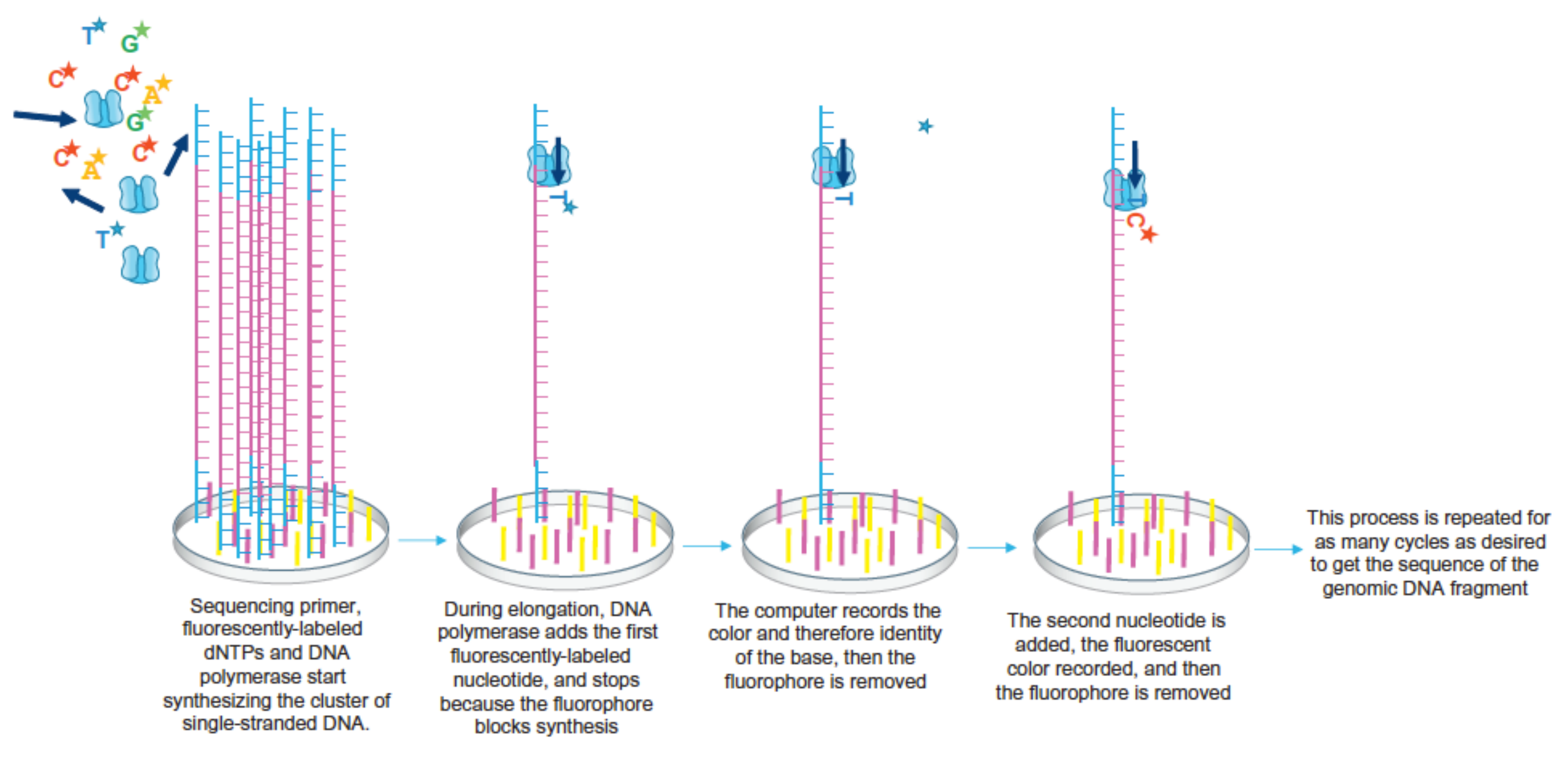
- DNA fragments from a sample are attached to a flow cell and amplified



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

Whole-genome sequencing (WGS)

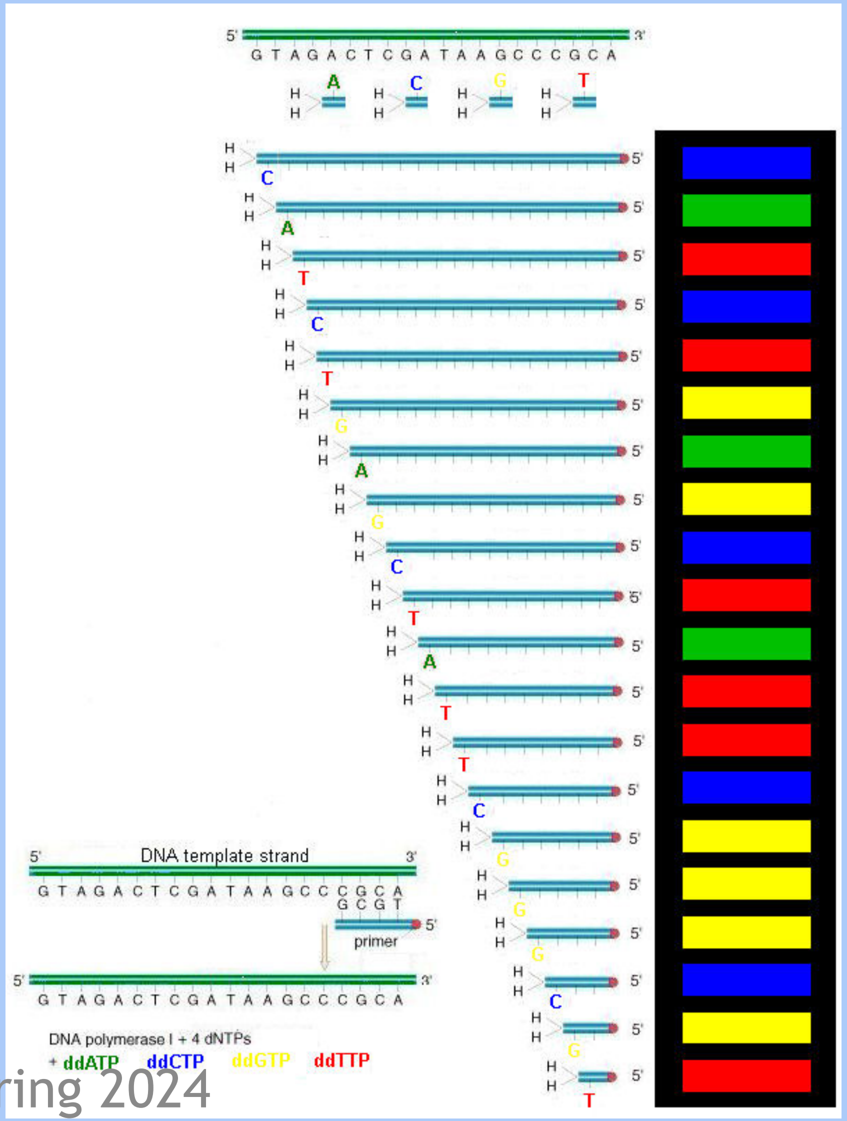
- **Sequencing by synthesis:** Short reads are produced as fluorescent nucleotides are incorporated one base at a time



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

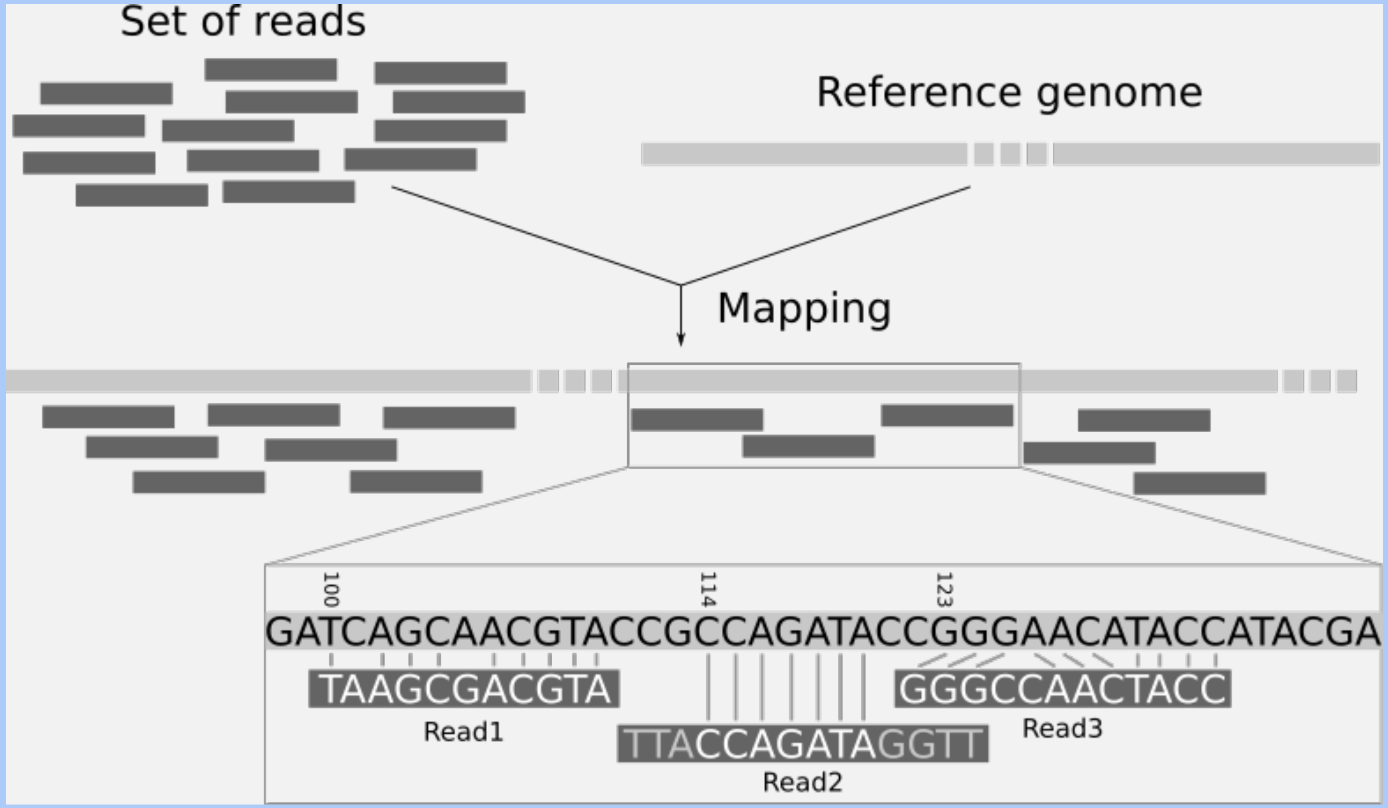
Whole-genome sequencing (WGS)

- The DNA sequence is inferred from the sequence of fluorescence images



Mapping to the reference genome

- Locate where in the genome the reads came from, and detect single-nucleotide differences from the reference sequence



Data-processing pipeline

- Generate raw reads
- Align to a reference genome
- Detect variant sites

FASTQ

- Contains raw sequence reads and their quality scores to be aligned to a reference genome (FASTA)

```
@A00178:71:HGT77DSXX:1:2171:17707:8077 2:N:0:ACAGCAAC+GTTGCTGT
GAAGAAAAGAAGGACACAGAGGAGGGAAAGGTTGAGGAAATTGATGAAGAGAAGGAAGAGAAAGAGAAGAAAAAGAAGACGATCAAGGAGGTTT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:1507:30291:23422 1:N:0:ACAGCAAC+GTTGCTGT
ACATAGAGCTTGATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGCCTCTTGGGGACAAAAAGGACAGCCTTGAACTCAAGCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:1507:30291:23422 2:N:0:ACAGCAAC+GTTGCTGT
CTGGATGAGGAAGCCTGAGGAGATCACCAAGGAGGAGTATGCTGCTTTCTATAAAAGCTTGACAAATGACTGGGAAGAGCATCTGGCTGTCAAG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:2413:22806:35790 1:N:0:ACAGCAAC+GTTGCTGT
GCTTGATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGCCTCTTGGGGACAAAAAGGACAGCCTTGAACTCAAGCTGCCCCTC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:2413:22806:35790 2:N:0:ACAGCAAC+GTTGCTGT
GAGAAGAAAAAGAAGACGATCAAGGAGGTTTCTCATGAATGGTCCTTGATCAACAAGCAGAAACCTATCTGGATGAGGAAGCCTGAGGAGATCA
+
F:FF:FFFFFFFF,:FFFFFFFF:FFFFFFFFFFFFF:F:FFFFFFF:FFFFFFFFFFFFFFFF,:FFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:2354:5620:8876 1:N:0:ACAGCAAC+GTTGCTGT
ATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGCCTCTTGGGGACAAAAAGGACAGCCTTGAACTCAAGCTGCCCCTCTACAG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:2354:5620:8876 2:N:0:ACAGCAAC+GTTGCTGT
AGAAGGAAGAGAAAGAGAAGAAAAAGAAGACGATCAAGGAGGTTTCTCATGAATGGTCCTTGATCAACAAGCAGAAACCTATCTGGATGAGGAA
+
FFFFFFF,FFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00178:71:HGT77DSXX:1:1560:6741:9815 1:N:0:ACAGCAAC+GTTGCTGT
GCAGGATTTTACCATGATCGACTACTTTTTGTCATGCCCAGAGAAGCTAGATTTTGCCAATGATGTTTATAGACCATTTAACGTTTCGCCAAGC
+
FFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

SAM (BAM)

- Paired-end reads are aligned to either the forward or reverse strand of the reference genome

```
5'  ACATAGACAGGGACCACCTGCAGGACACACACGCAGGTTTACTAAGGGTTTACTCAACACAGTGAACAGCATATACCAGA  3'
               5'  ACCTGCAGGACACACACGCAGGTTTACTAAGGGTTTACTCAACACAGTGA  3'
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
               3'  TGGACGTCCTGTGTGTGCGTCCAAATGATTCCCAAATGAGTTGTGTCACT  5'
```

https://eriqande.github.io/eca-bioinf-handbook/bioinformatic-file-formats.html#sambamfiles

SAM (BAM)

- Paired-end reads are aligned to either the forward or reverse strand of the reference genome

```
       Read 1: 5'  ACCTGCAGGA  3'
5'  ACATAGACAGGGACCACCTGCAGGACACACACGCAGGTTTACTAAGGGTTTACTCAACACAGTGAACAGCATATACCAGA  3'
forward-strand
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
reverse-strand
3'  TGTATCTGTCCCTGGTGGACGTCCTGTGTGTGCGTCCAAATGATTCCCAAATGAGTTGTGTCACTTGTCGTATATGGTCT  5'
                                               Read 2: 3'  TTGTGTCACT  5'
```

https://eriqande.github.io/eca-bioinf-handbook/bioinformatic-file-formats.html#sambamfiles

SAM (BAM)

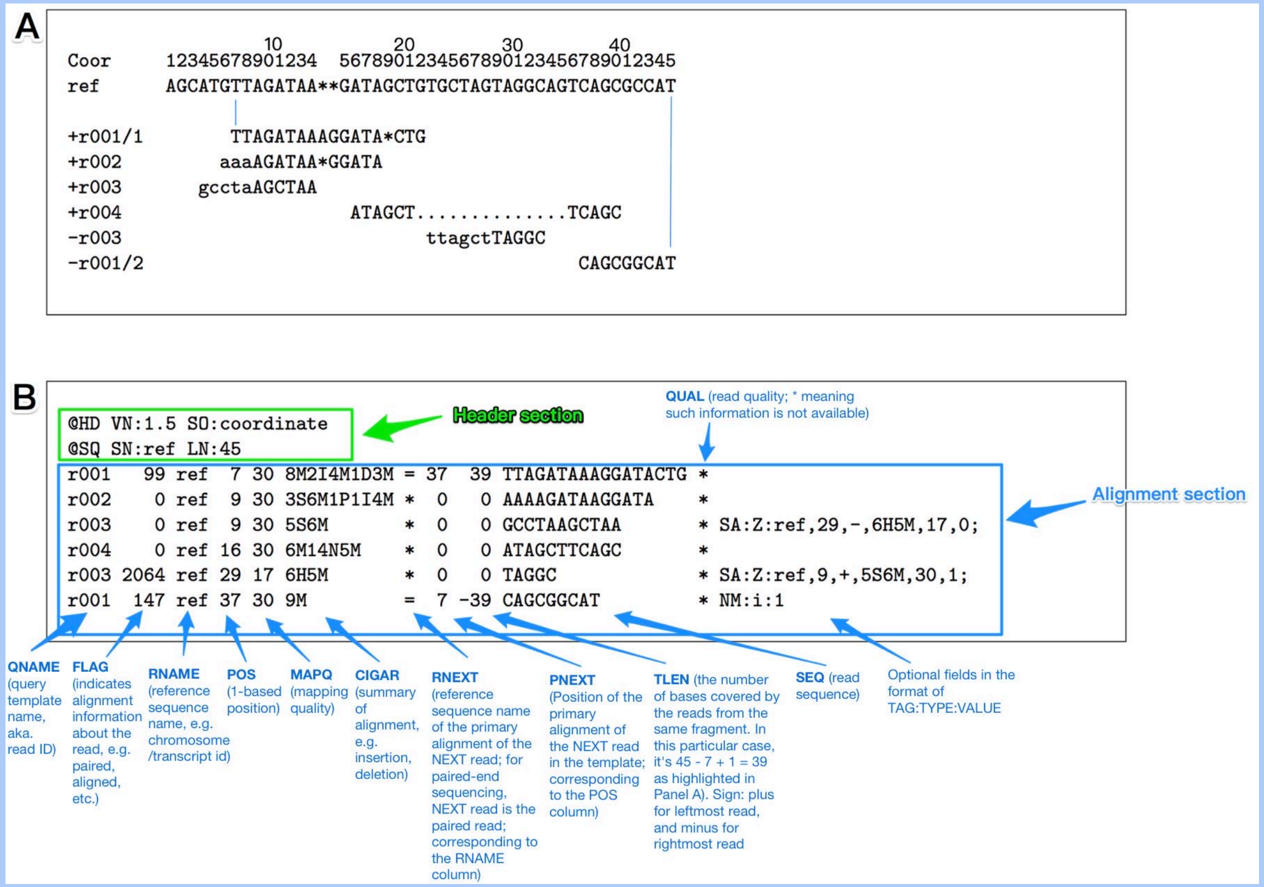
- Paired-end reads are aligned to either the forward or reverse strand of the reference genome

```
      Read 2:  5'  ACCTGCAGGA  3'
5'  ACATAGACAGGGACCACCTGCAGGACACACACGCAGGTTTACTAAGGGTTTACTCAACACAGTGAACAGCATATACCAGA  3'
forward-strand
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
reverse-strand
3'  TGTATCTGTCCCTGGTGGACGTCCTGTGTGTGCGTCCAAATGATTCCCAAATGAGTTGTGTCACTTGTCGTATATGGTCT  5'
                                              Read 1:  3'   TTGTGTCACT  5'
```

https://eriqande.github.io/eca-bioinf-handbook/bioinformatic-file-formats.html#sambamfiles

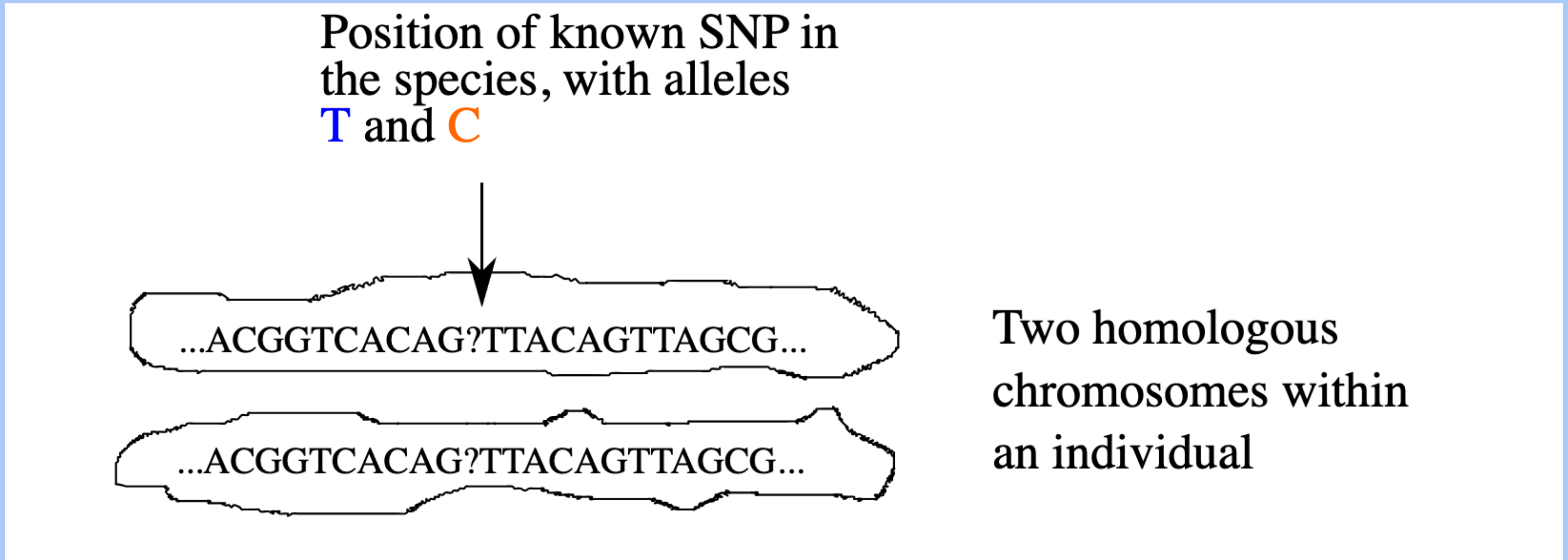
SAM (BAM)

- A Sequence alignment map (SAM) or binary alignment map (BAM) file contains the alignments to the reference genome



Variant calling (mpileup)

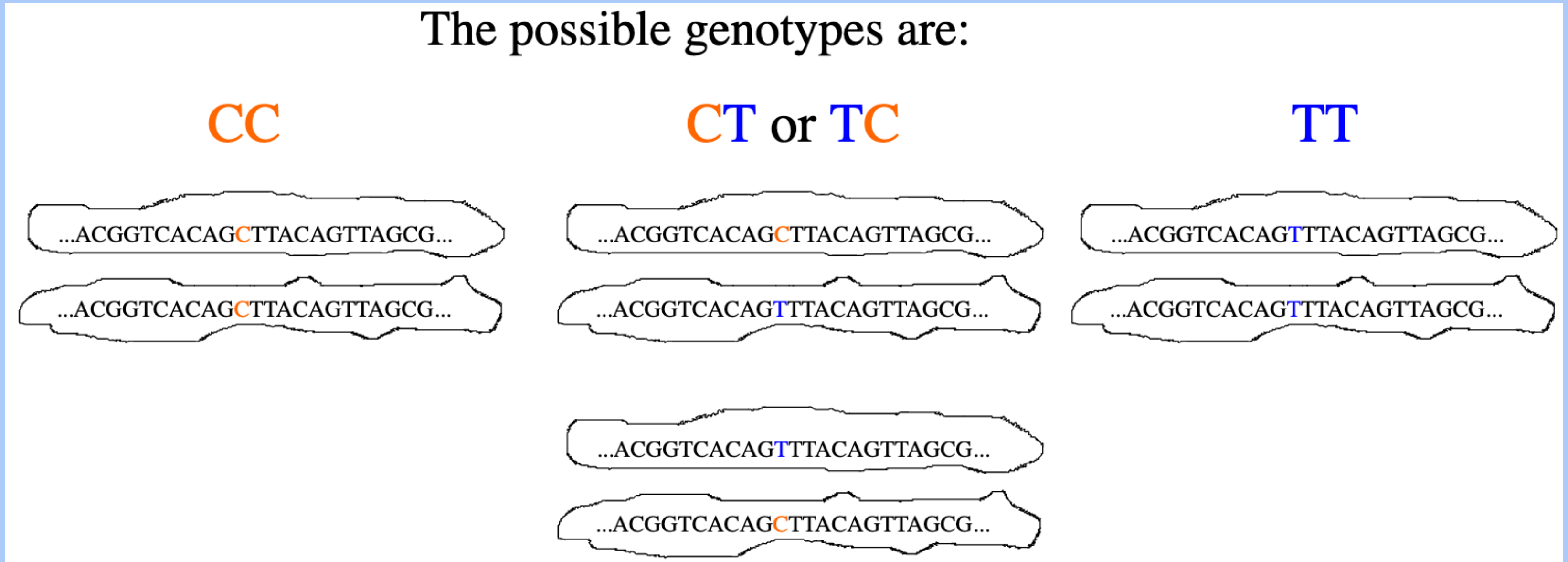
- How certain can we be of an individual's genotype?



https://eriqande.github.io/eca-bioinf-handbook/bioinformatic-file-formats.html#sambamfiles

Variant calling (mpileup)

- How certain can we be of an individual's genotype?



https://eriqande.github.io/eca-bioinf-handbook/bioinformatic-file-formats.html#sambamfiles

Variant calling (mpileup)

- How certain can we be of an individual's genotype?

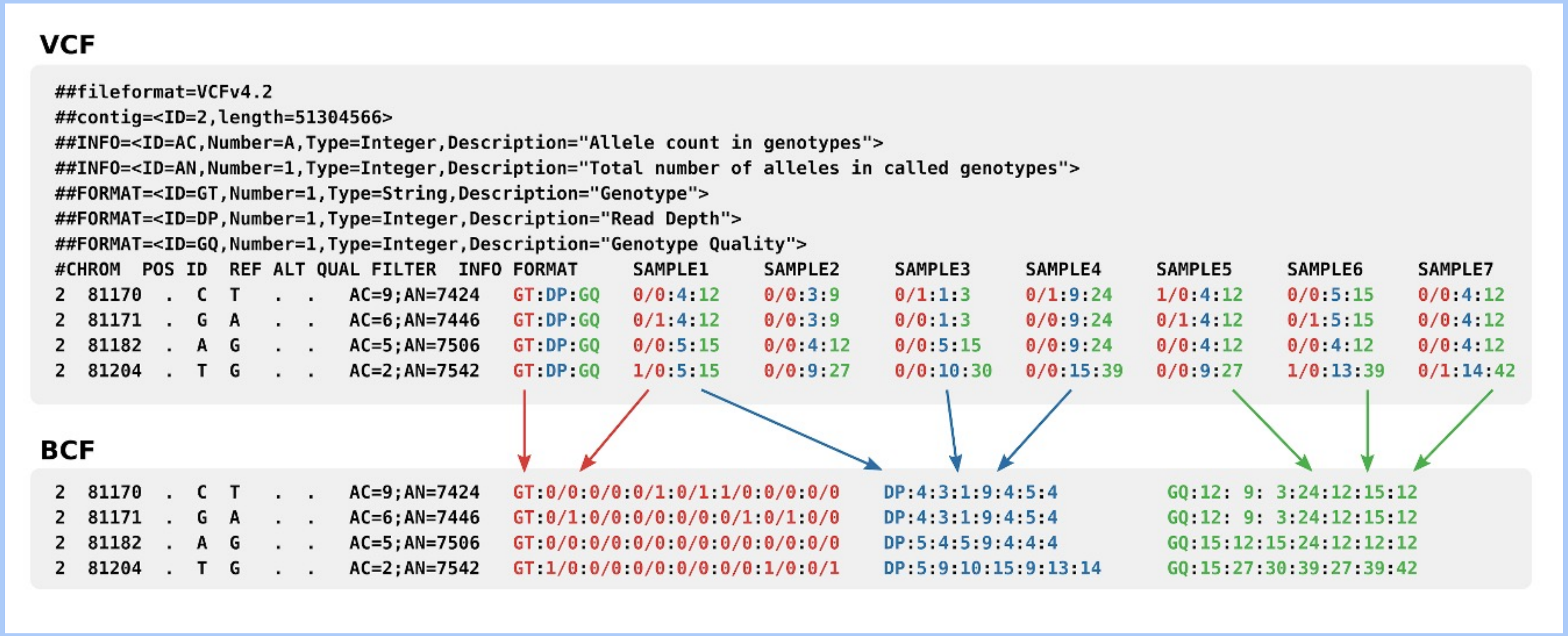
The data are: 4 reads covering that site,
and
the associated base quality scores

Read #	*Read*	*Observed Base*	*PHRED-scaled base quality score*
1	CAGCTTACA	C	32 (A)
2	ACAGCT	C	37 (F)
3	GTTTA	T	35 (D)
4	AGCTTACAG	C	33 (B)

https://eriqande.github.io/eca-bioinf-handbook/bioinformatic-file-formats.html#sambamfiles

VCF

- The results of genotype-calling are stored in a variant call format (VCF) file



VCF

- The VCF file has one row for each variant, and one column for each sequenced individual

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS    ID         REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1        Sample2        Sample3
2      4370   rs6057     G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330   .          T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696 rs6055     A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237 .          T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2      134567 microsat1  GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

VCF

- Codes such as GT, DP, GP give the genotype, read depth, and genotype probabilities for each individual

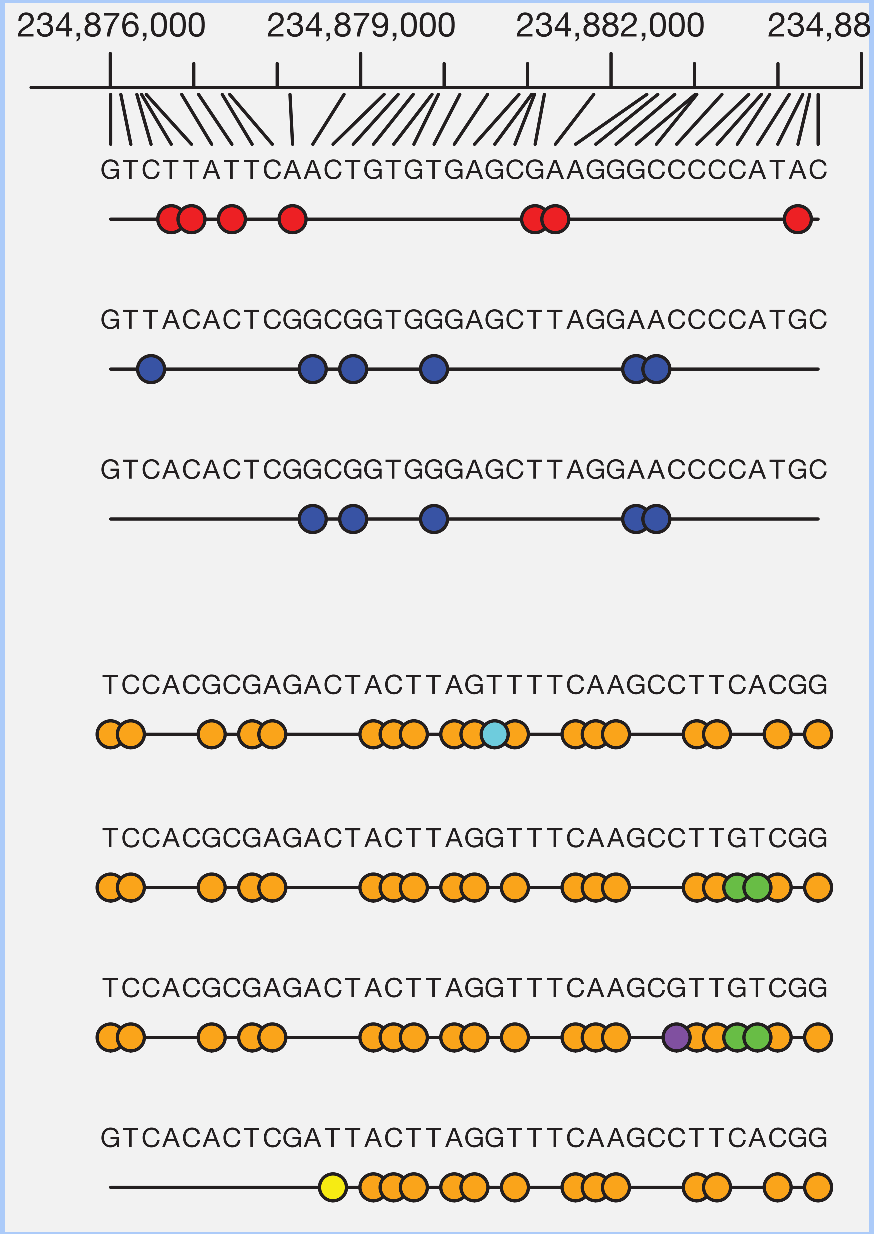
```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1        Sample2        Sample3
2      4370    rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330    .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696  rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237  .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2      134567  microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Human genetic variation

Sequencing projects and implications for association studies

The HapMap Project

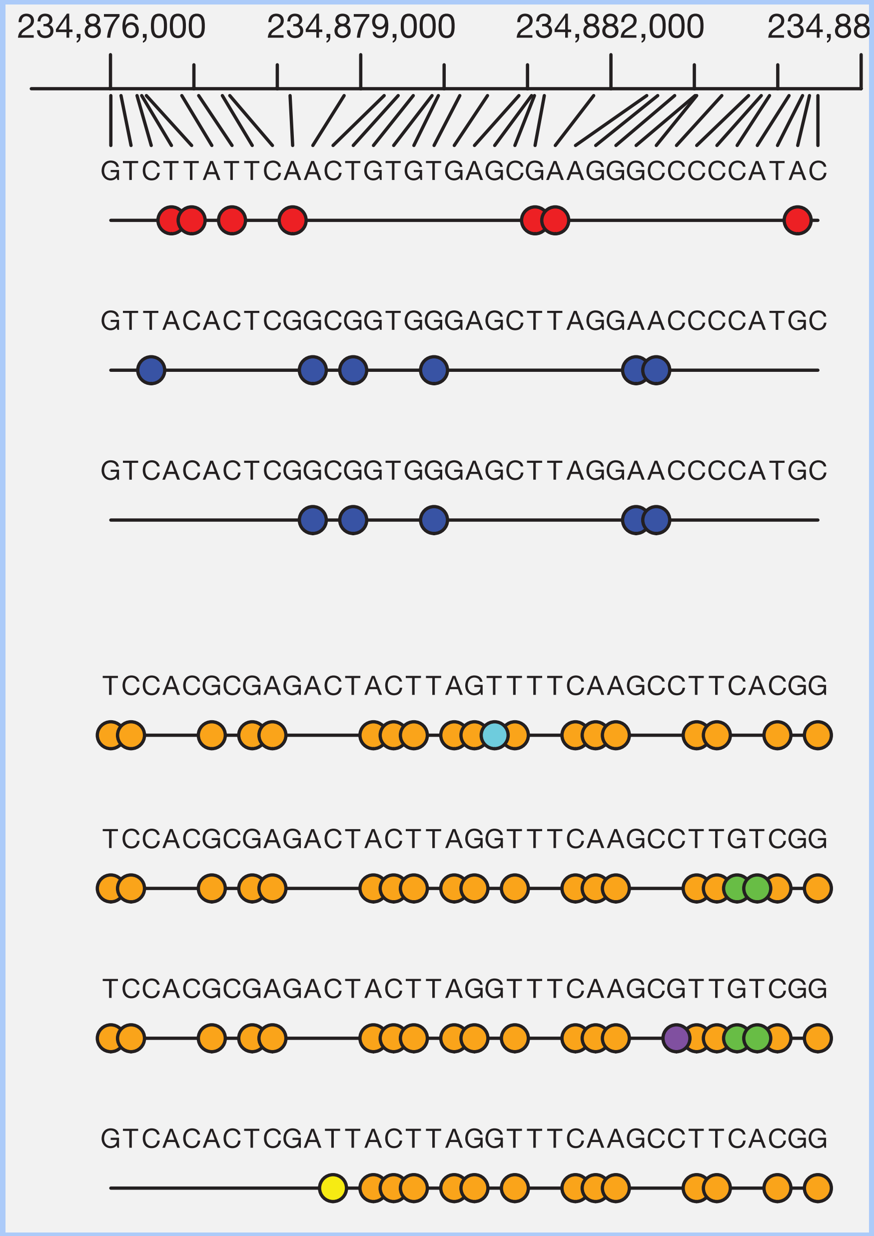
- International genotyping consortium launched in 2002 to find common polymorphisms linked to rare disease loci



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

The HapMap Project

- Variants occur together on a small number of **haplotypes**



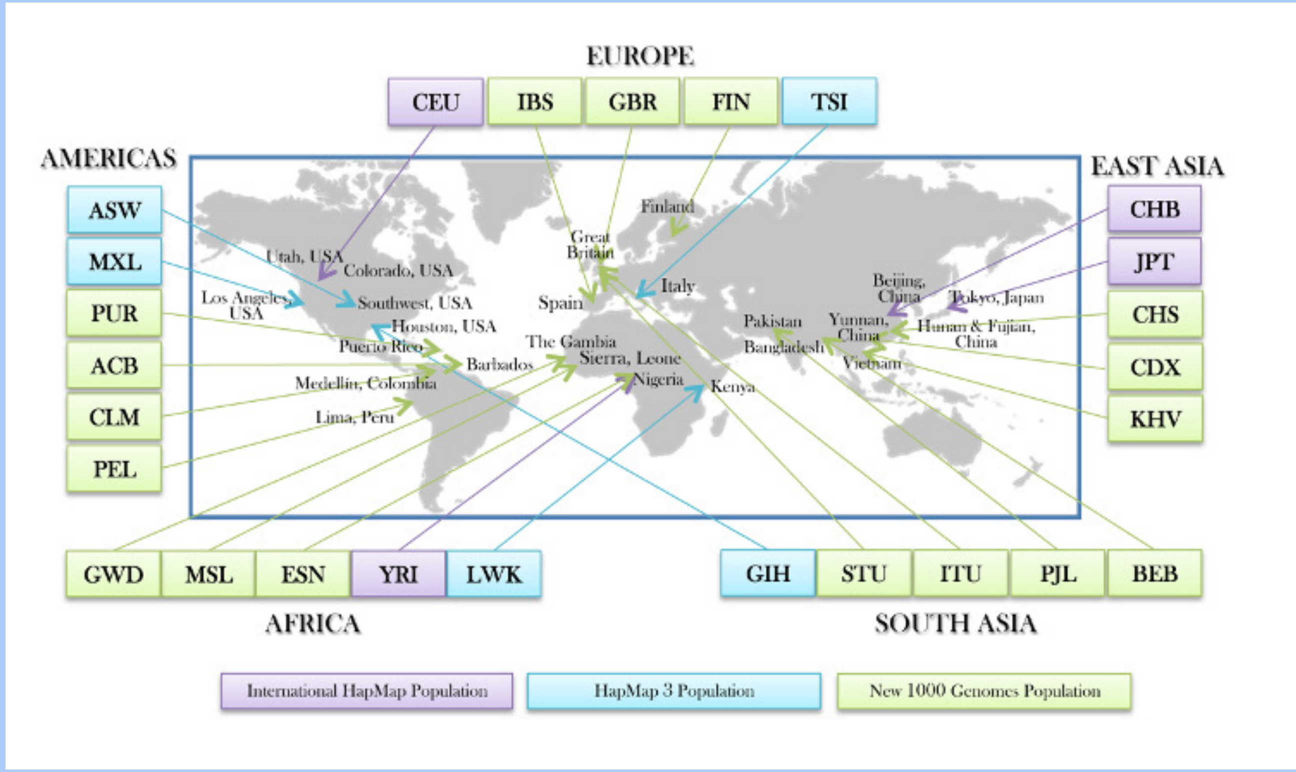
https://pubmed.ncbi.nlm.nih.gov/16255080/

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

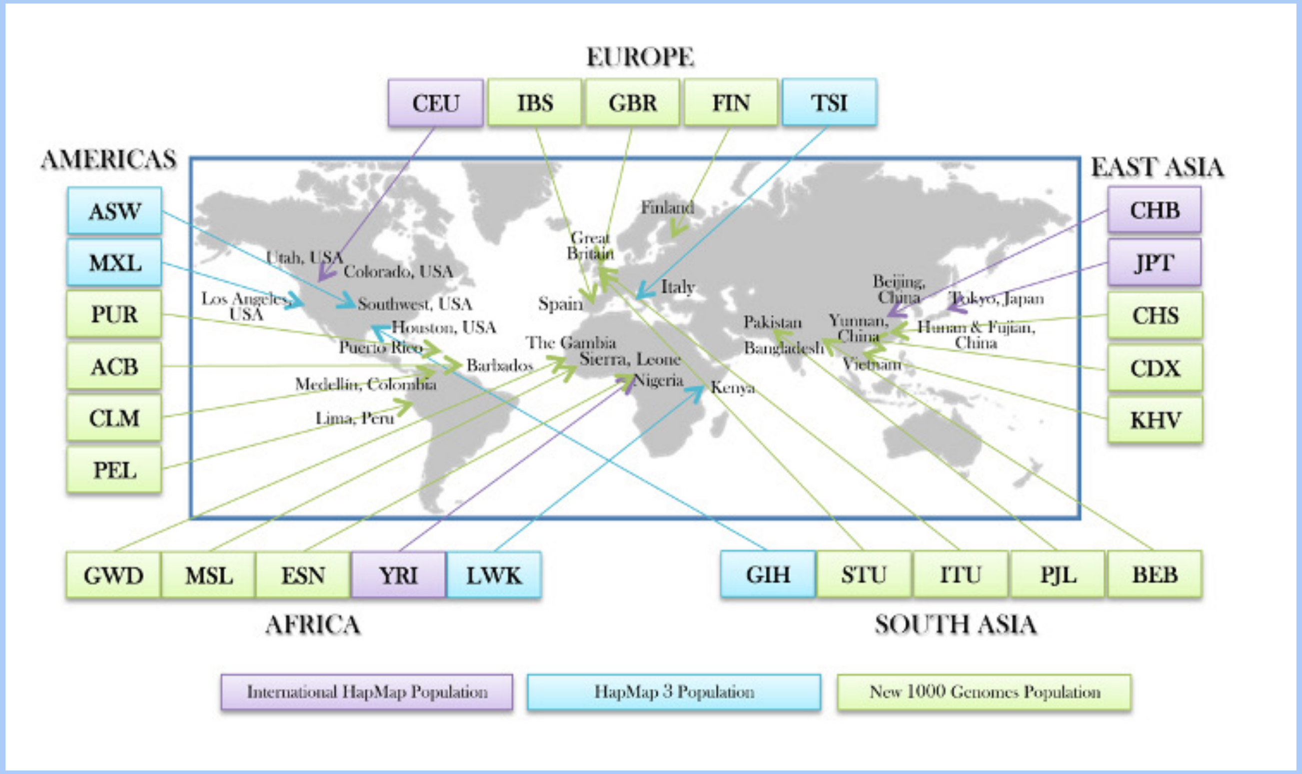
The 1000 Genomes Project

- An international consortium launched in 2008 to catalog rare variants (frequency < 1%) taking advantage of new sequencing technologies

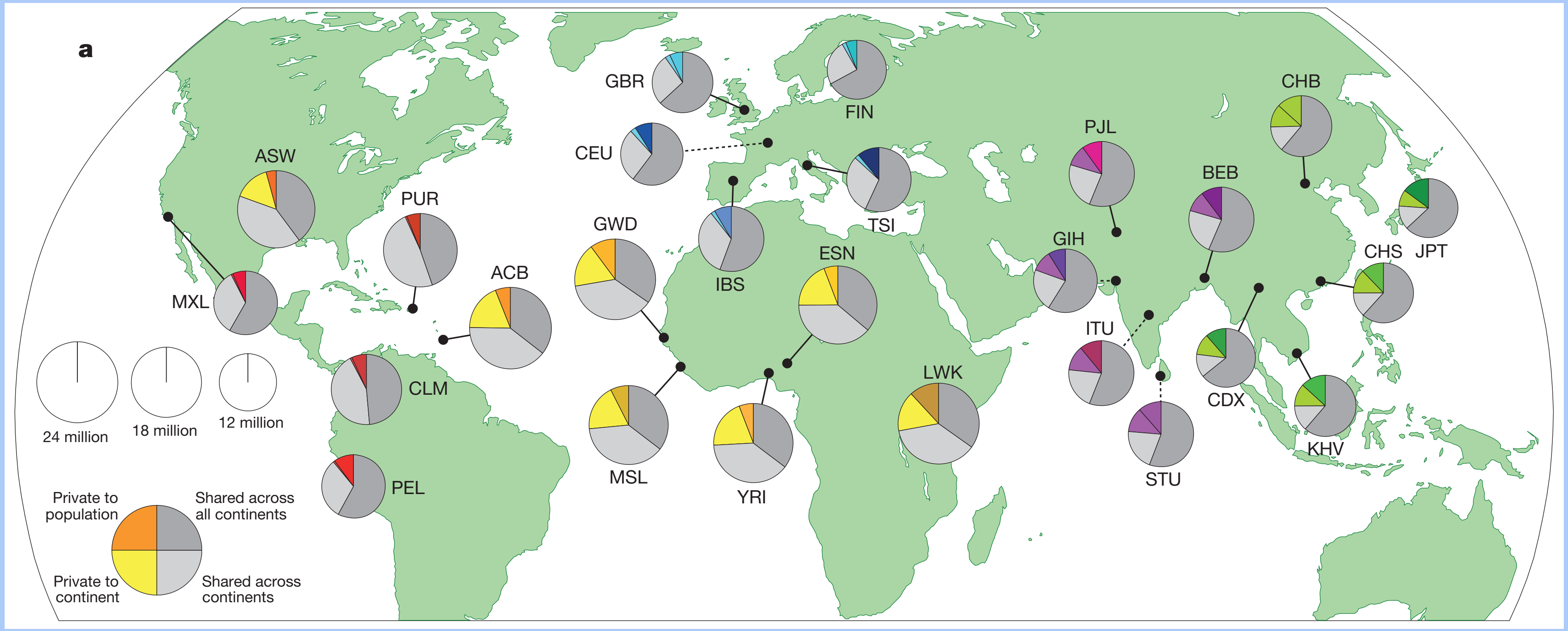


The 1000 Genomes Project

- Phase 3 release (2015) contained data from 2,504 individuals representing 26 populations across the globe, and identified 85 million new SNPs



Global genetic variation



- Most SNPs are shared across continents, and the majority of variation (~85%) is within rather than between populations

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

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 $\overline{x}$

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

Total variation

Within-population variation

Between-population variation

Pitfalls of not accounting for genetic ancestry

- Because of allele-frequency differences in global populations, **spurious associations** with disease risk can show up that may be entirely explained by genetic ancestry

Example: lactase nonpersistence (lactose intolerance)

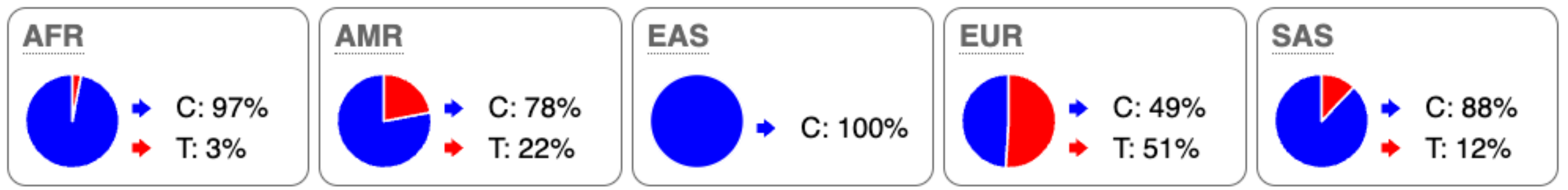
- The T allele of rs182549 is completely associated with the ability to digest lactose in Europeans

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

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

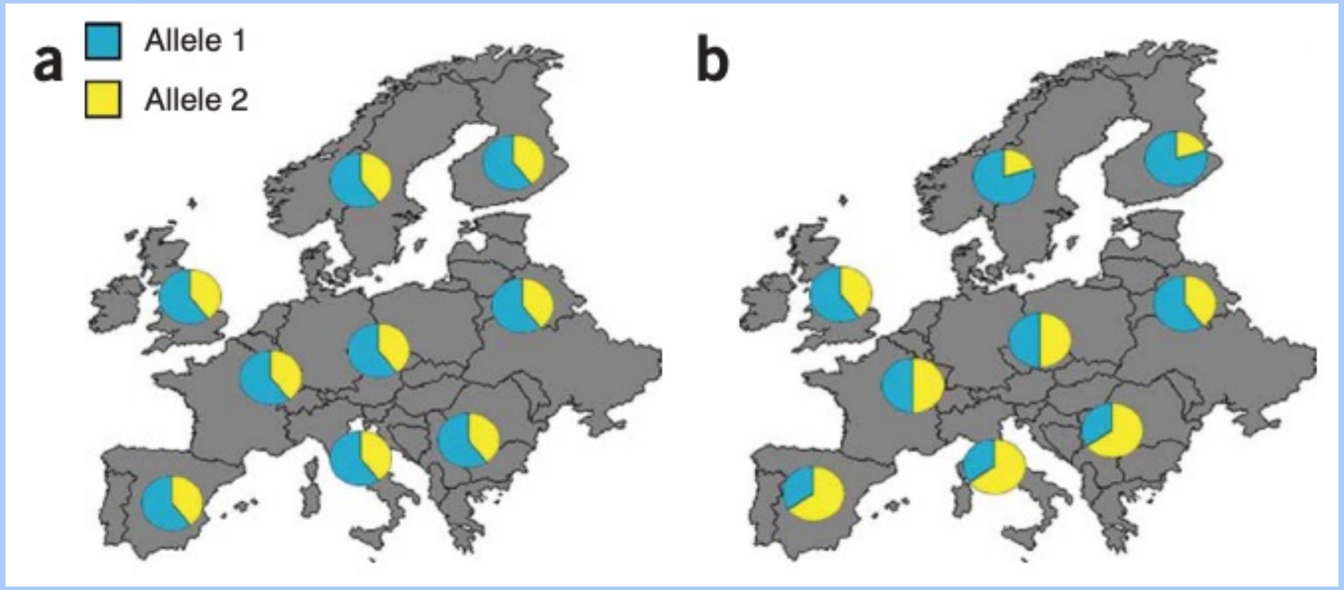
Example: lactase nonpersistence (lactose intolerance)

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



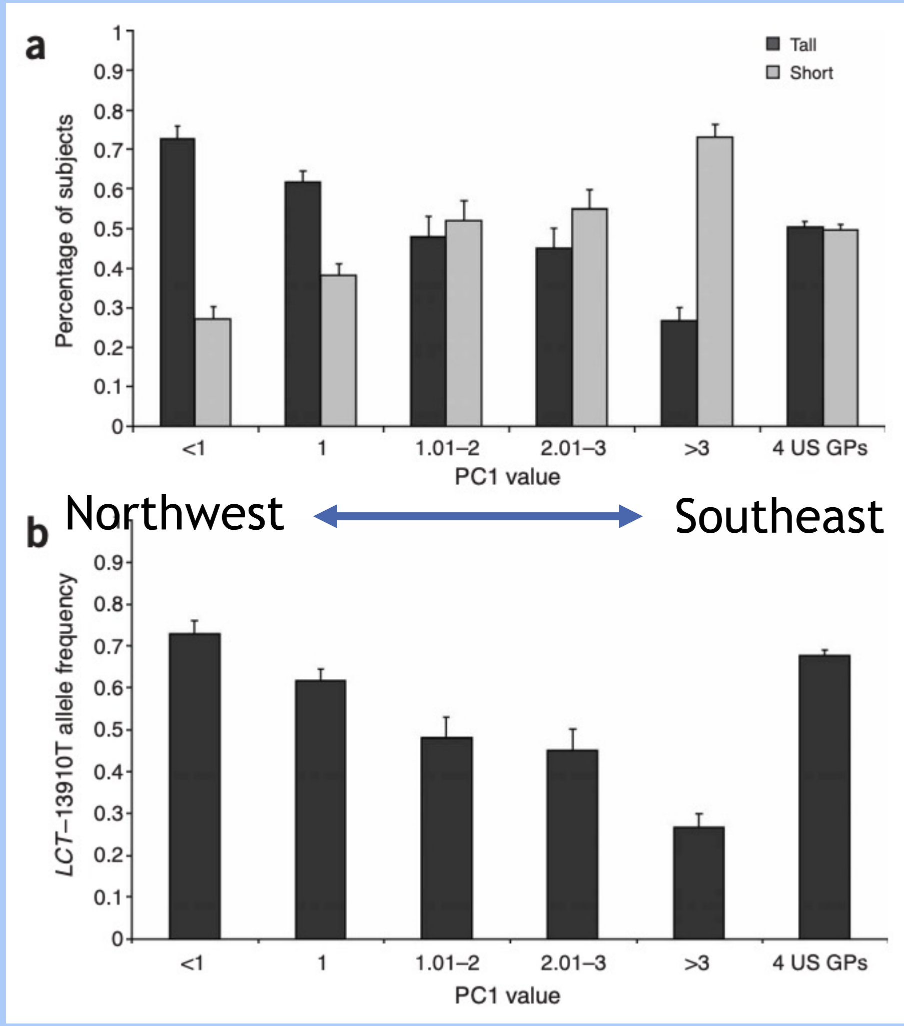
Population stratification

- 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



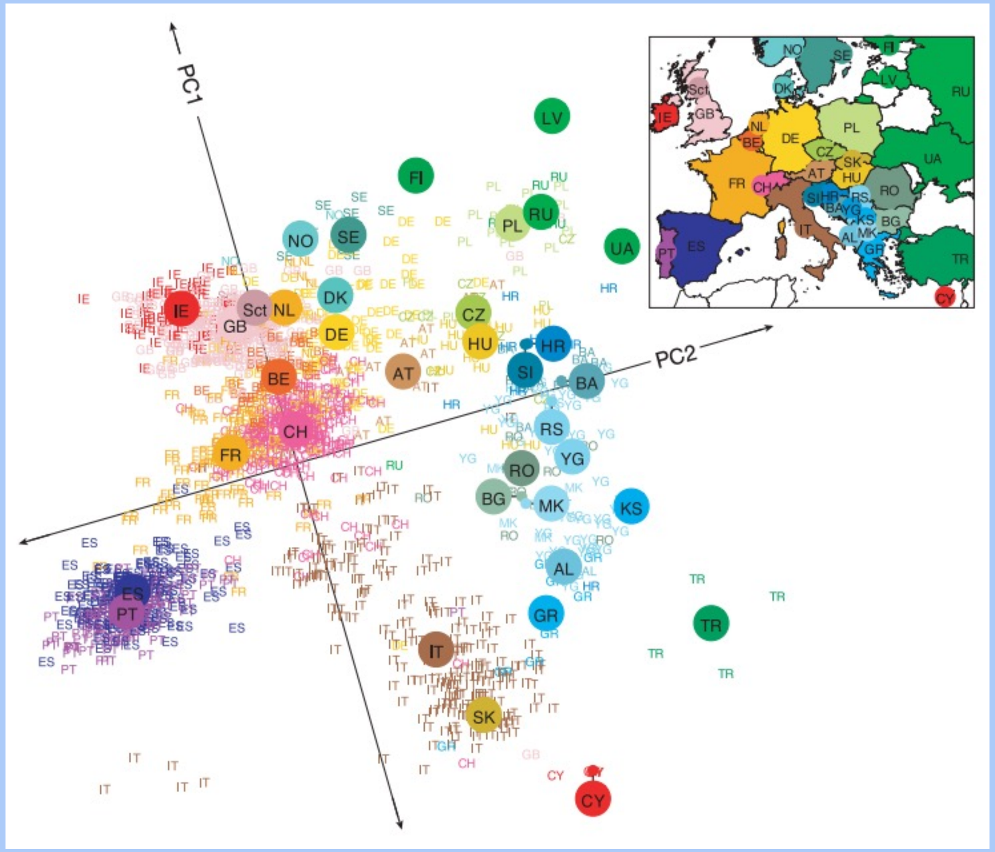
https://pubmed.ncbi.nlm.nih.gov/16041375/

Principal components analysis

The concept of genetic ancestry

Principal components analysis

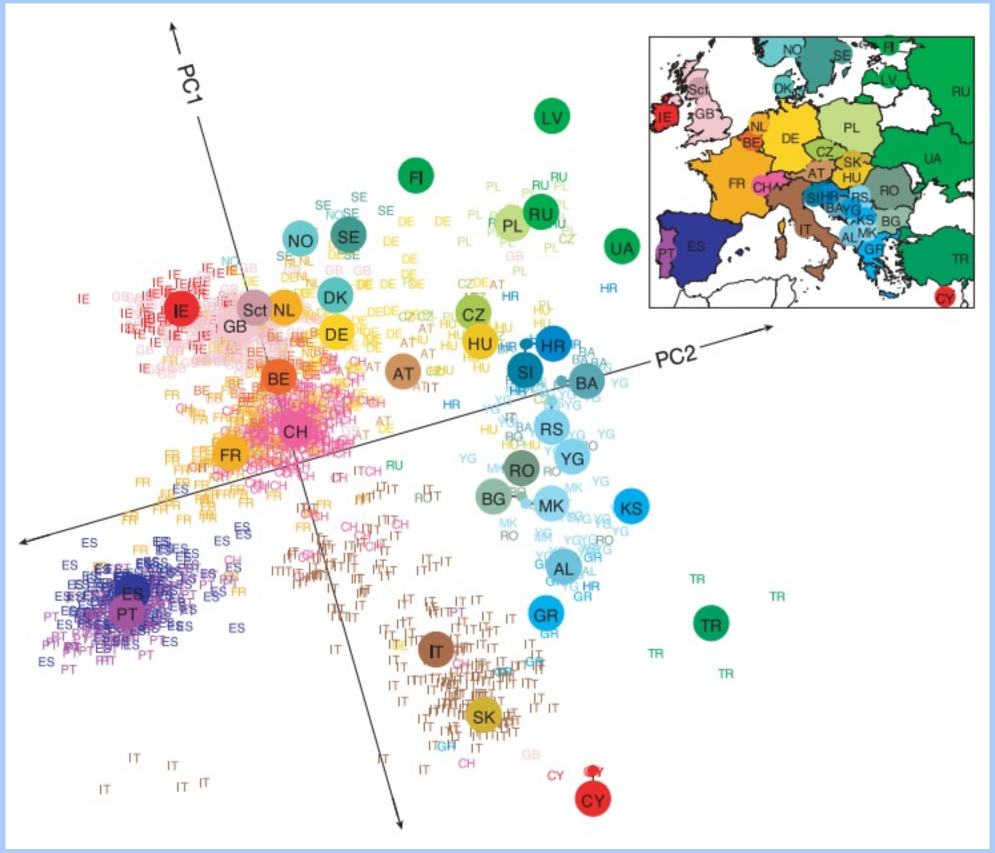
- Genotypes can distinguish population groups



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

Principal components analysis

- Looking at **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

$$\begin{array}{cc} & \text{m SNPs} \longrightarrow \\ \text{n subjects} \downarrow & \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \end{array}$$

Genotype matrix

- The number of alternate alleles is 0, 1, or 2

m SNPs

n subjects

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

Genotype matrix

- "Standardize" each genotype by subtracting the mean allele (column) frequency and dividing by its standard error

m SNPs

n subjects

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

"Idealized" individuals

- An "idealized" subject of a particular genetic ancestry has genotypes v at m SNPs

$$\mathbf{X}\mathbf{V}^T = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2n} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ \vdots & \vdots & & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{pmatrix}$$

“Idealized” individuals

- The position $u_{11}\lambda_{11}$ of individual 1 on PC1 is the “amount” of idealized person 1 in individual 1

$$\begin{pmatrix} u_{11} & \cdots & u_{1n} \\ u_{21} & \cdots & u_{2n} \\ \vdots & & \vdots \\ u_{n1} & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} \lambda_{11} & & & \\ & \lambda_{22} & & \\ & & \ddots & \\ & & & \lambda_{nn} \end{pmatrix} = \mathbf{U\Sigma}$$

“Idealized” individuals

- The position $u_{ij}\lambda_{jj}$ of individual i on PCj is the “amount” of idealized person j in individual i

$$\begin{pmatrix} u_{11} & \cdots & u_{1n} \\ u_{21} & \cdots & u_{2n} \\ \vdots & & \vdots \\ u_{n1} & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} \lambda_{11} & & & \\ & \lambda_{22} & & \\ & & \ddots & \\ & & & \lambda_{nn} \end{pmatrix} = \mathbf{U\Sigma}$$

"Idealized" individuals

- The idea of PCA is to find the amount of each idealized individual in each actual individual using the decomposition of the n × m genotype matrix **X** into n × n, n × n, and n × m matrices **U**, **Σ**, and **V**

$$\mathbf{X}\mathbf{V}^T = \mathbf{U}\mathbf{\Sigma}$$

Genomic relationship matrix (GRM)

- The GRM is computed by comparing how similar any subject is to any other

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

Genomic relationship matrix (GRM)

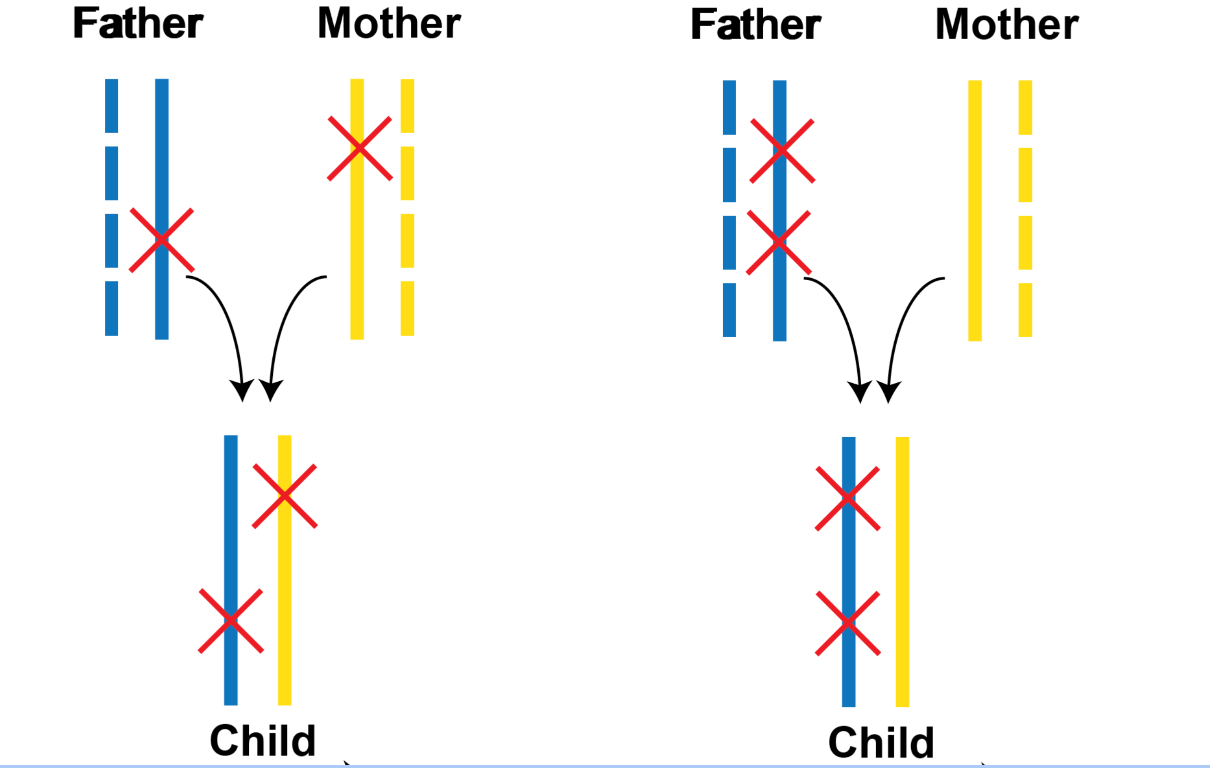
- The **eigenvectors** (columns of U) of the GRM contain the ancestry components

$$\mathbf{X}\mathbf{X}^T\mathbf{U} = \mathbf{U}\boldsymbol{\Sigma}^2$$

Linkage disequilibrium

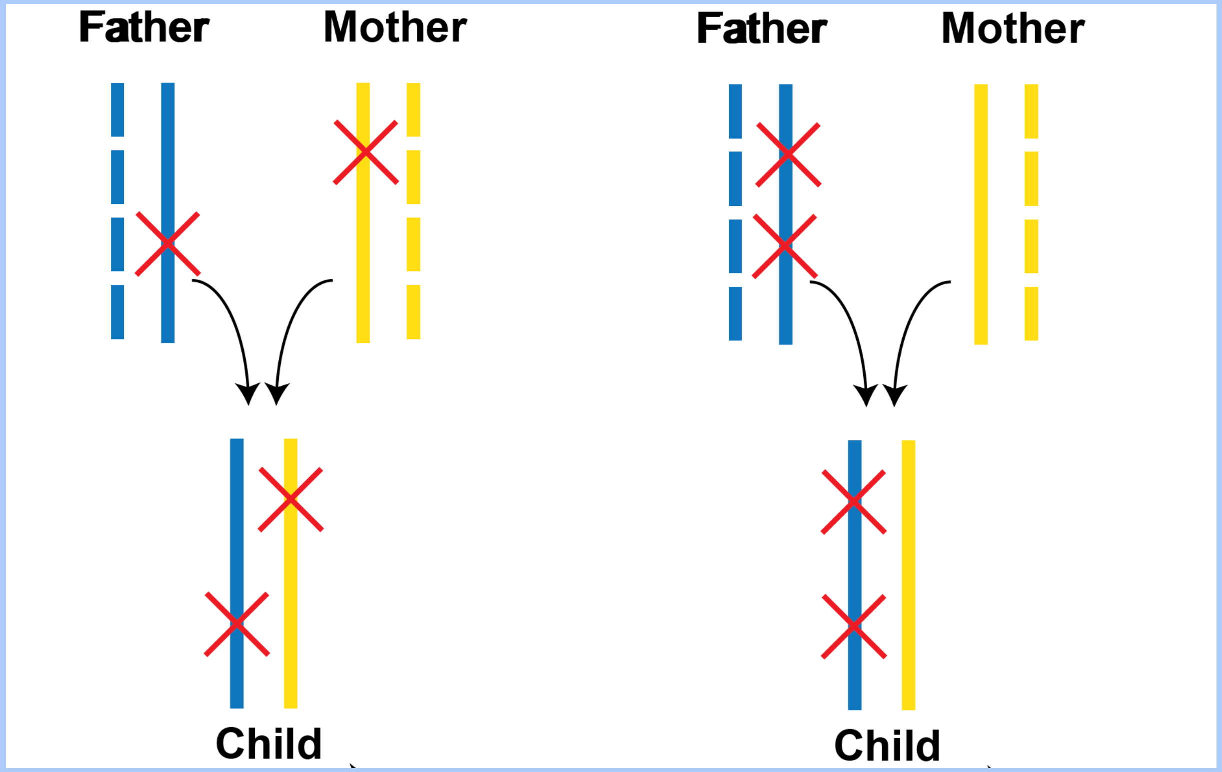
Determining a set of independent SNPs

SNPs can occur on either of two chromosomes



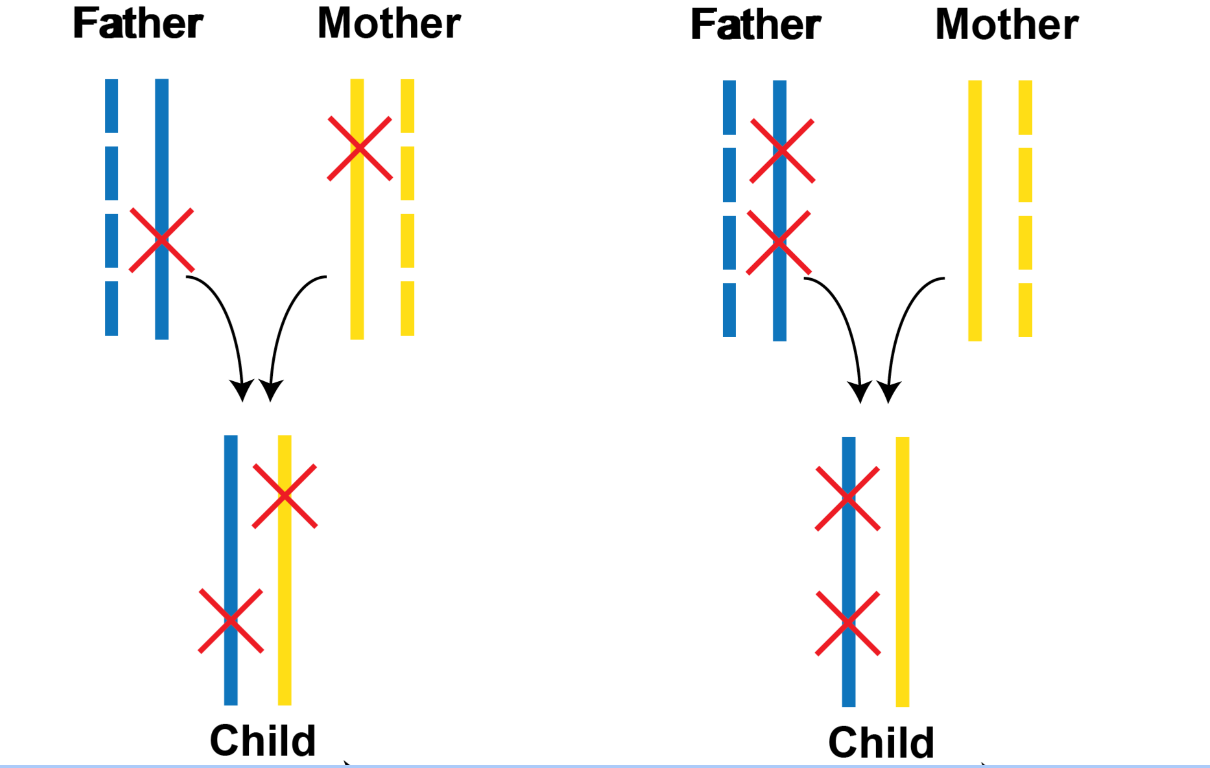
- Genotype data do not tell us which chromosomes carry the polymorphism

SNPs can occur on either of two chromosomes



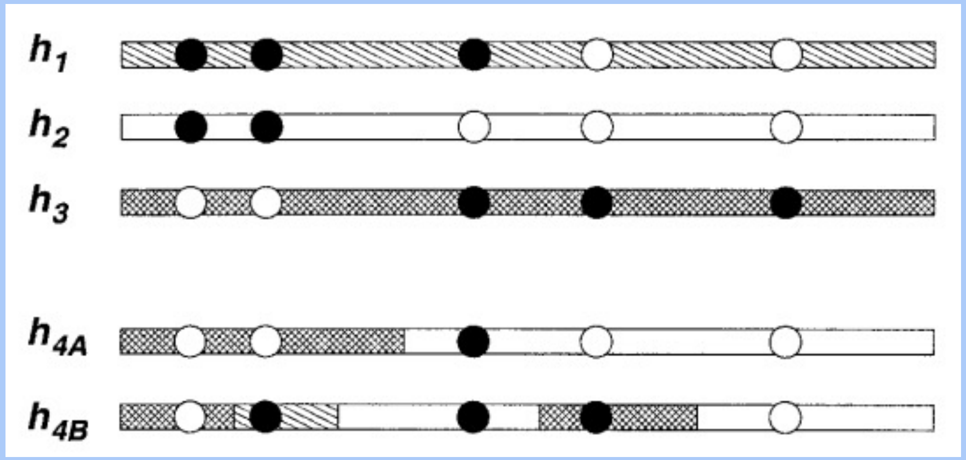
- When at least one parent is homozygous at each SNP, **haplotype phase** can be unambiguously assigned

SNPs can occur on either of two chromosomes



- and we can distinguish AB/ab from Ab/aB

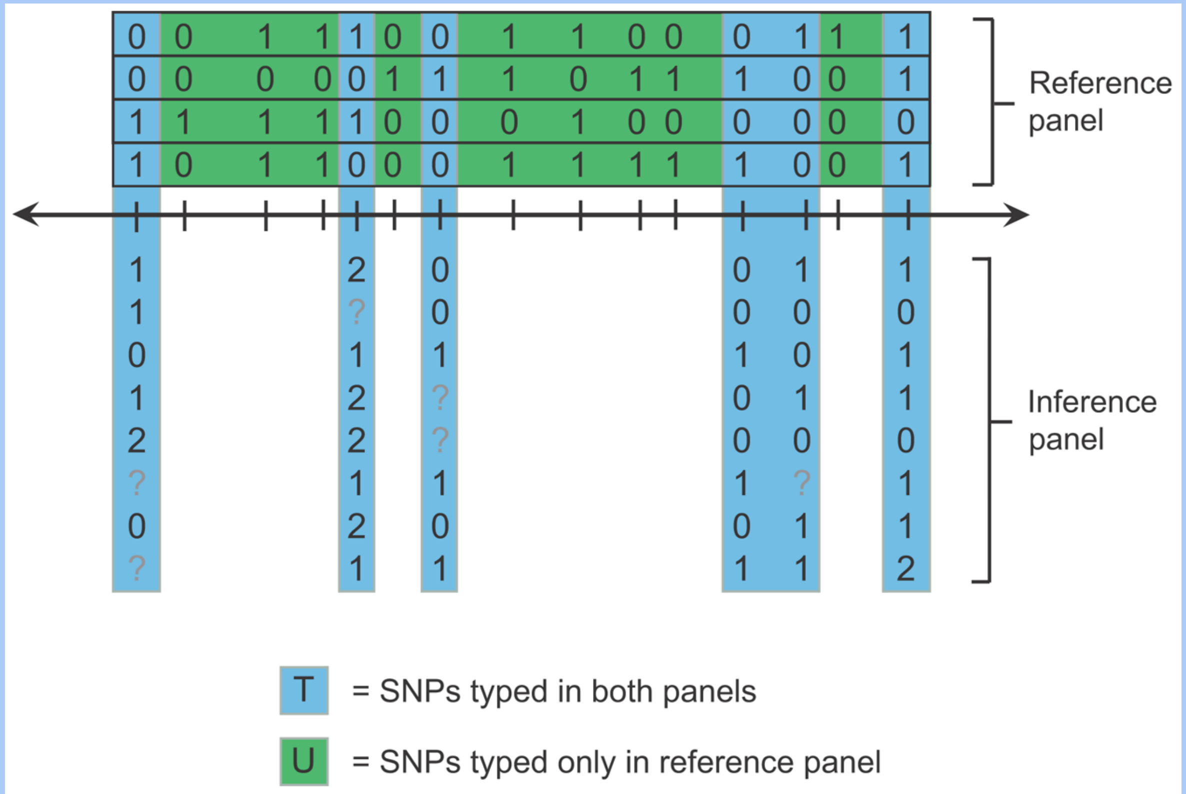
Statistical phasing and imputation



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

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

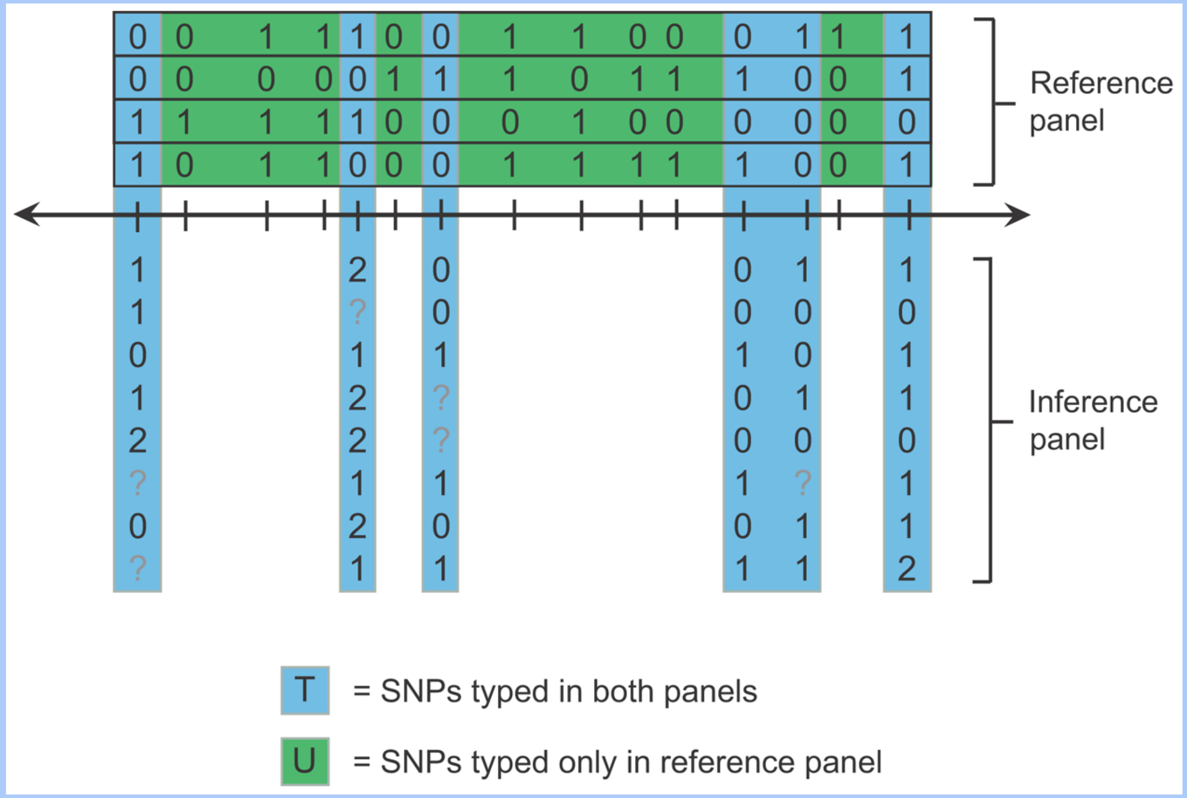
Statistical phasing and imputation



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

- Variants that have not been typed can be **imputed** into the inference sample

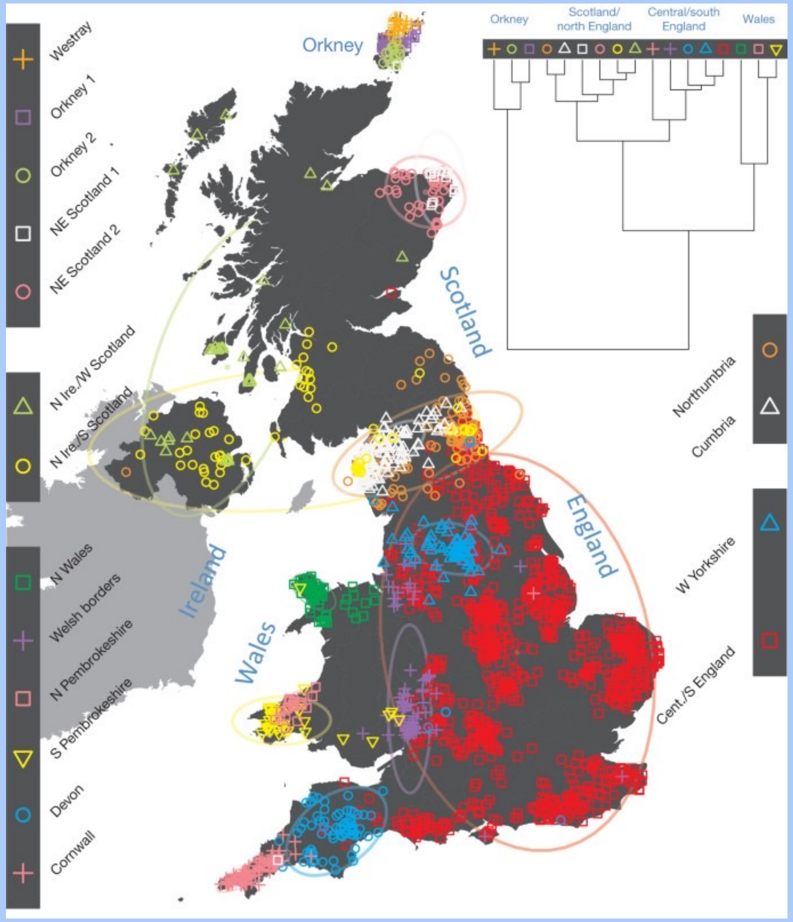
Statistical phasing and imputation



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

- 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 chromosome**

Linkage disequilibrium

- LD is measured as the correlation coefficient between the alleles of different SNPs

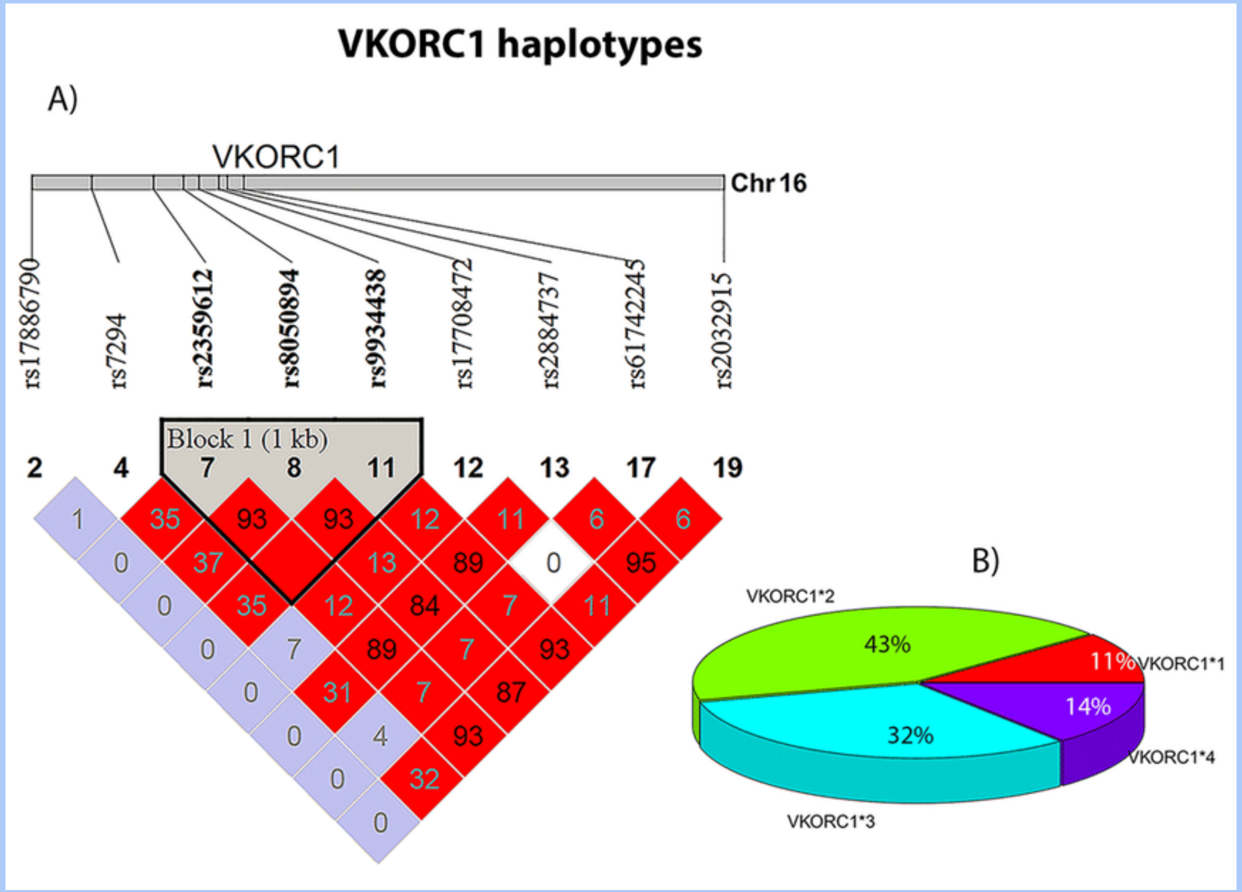
$$r_{A,B} = \frac{p_{A,B} - p_A p_B}{\sqrt{p_A (1 - p_A) p_B (1 - p_B)}}$$

Linkage disequilibrium

- p_A = fraction of chromosomes with A
- p_{AB} = fraction of chromosomes with A and B

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

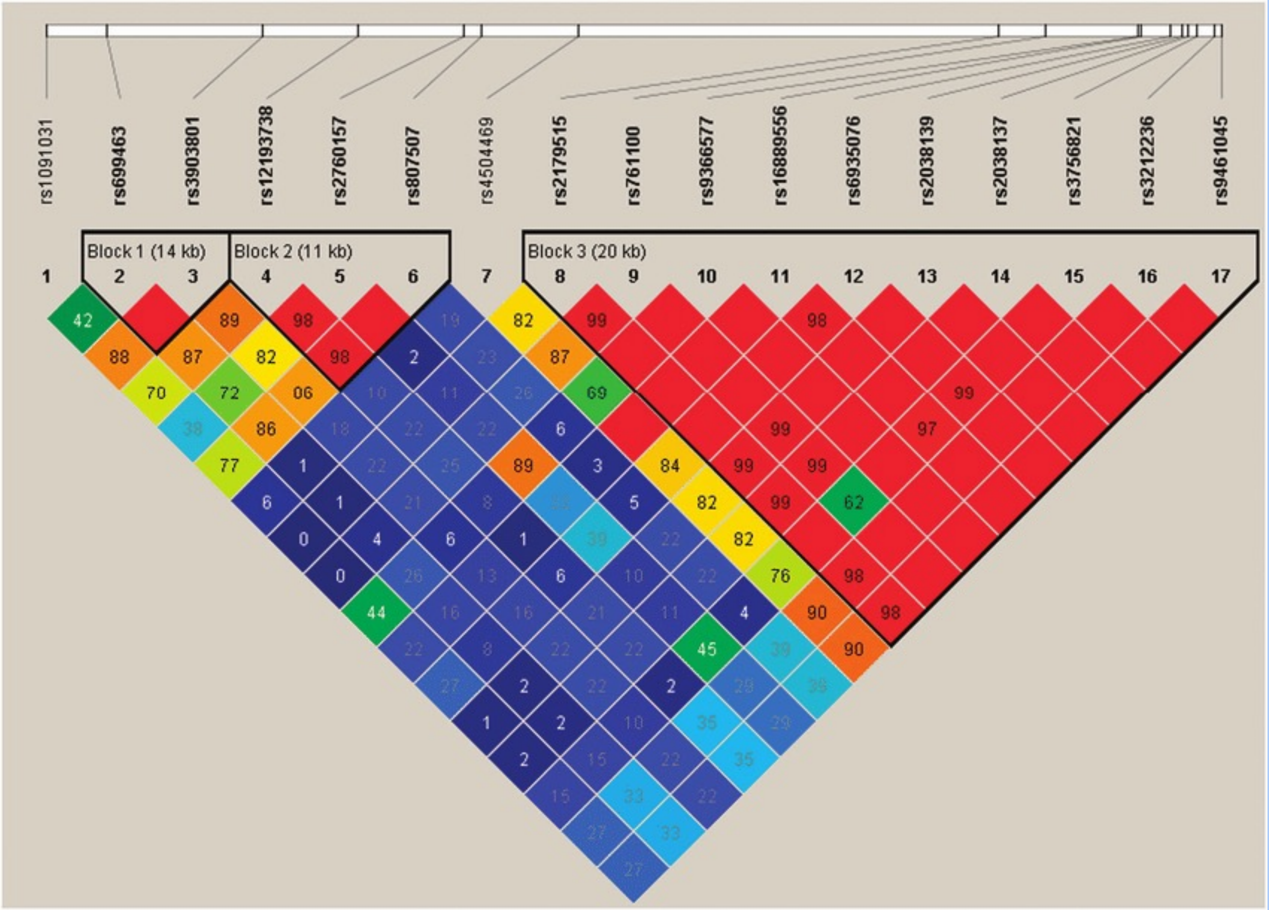
LD blocks and haplotype structure



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

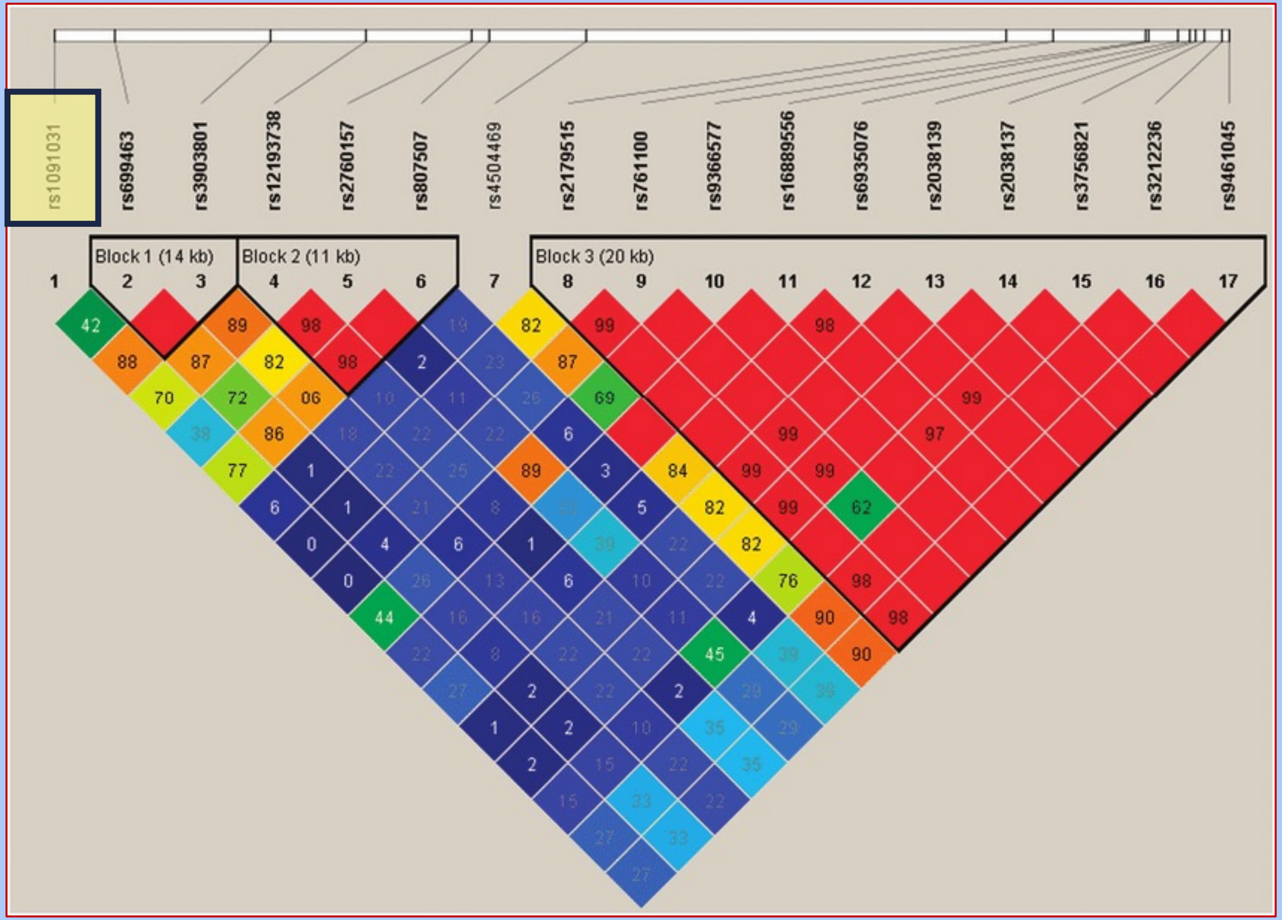
- Plots of pairwise r^2 values show which SNPs are inherited together in the population as common haplotypes

An algorithm for computing an independent subset of alleles



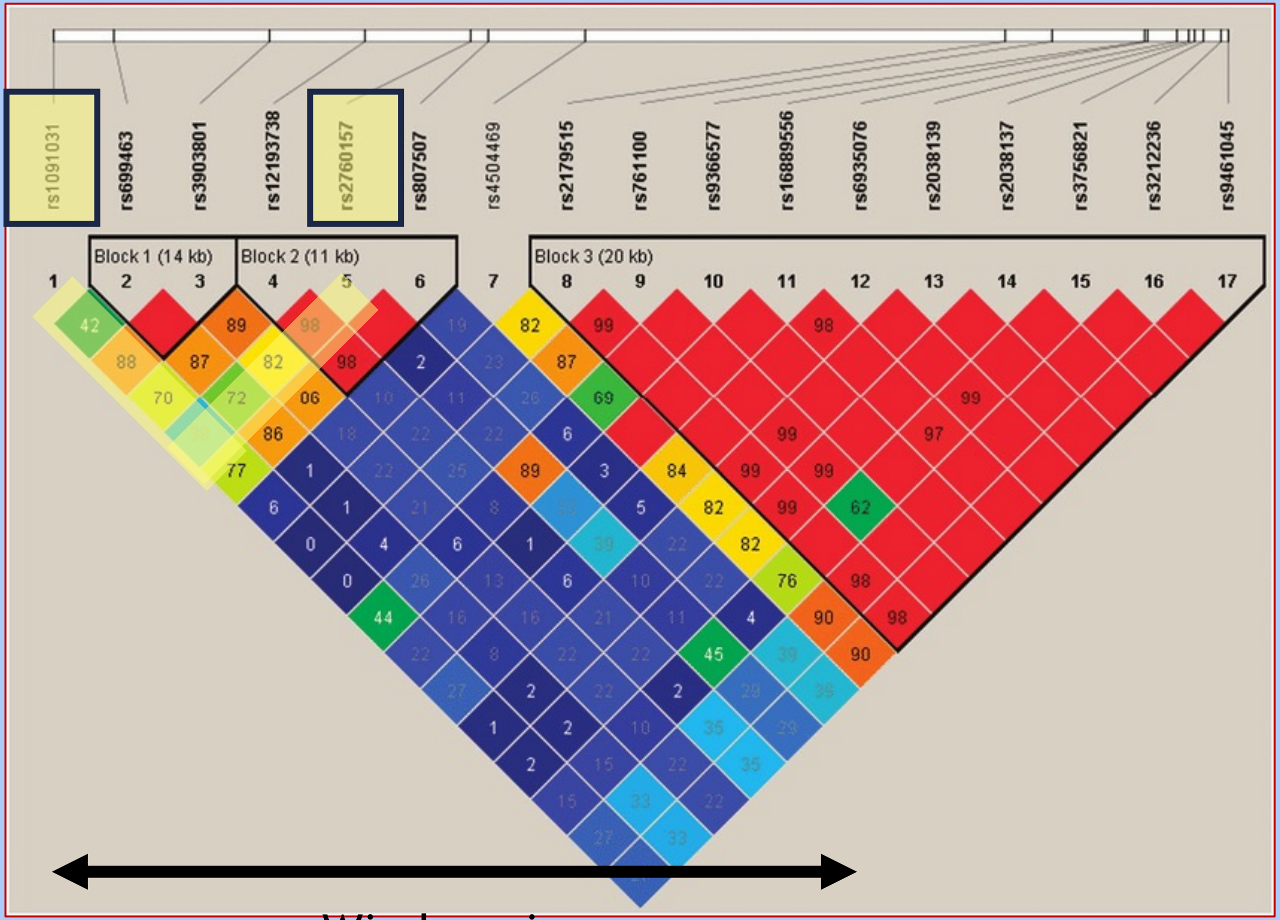
- From the SNPRelate package https://rdrr.io/bioc/SNPRelate/man/snpgdsLDpruning.html

An algorithm for computing an independent subset of alleles



- Pick a random SNP

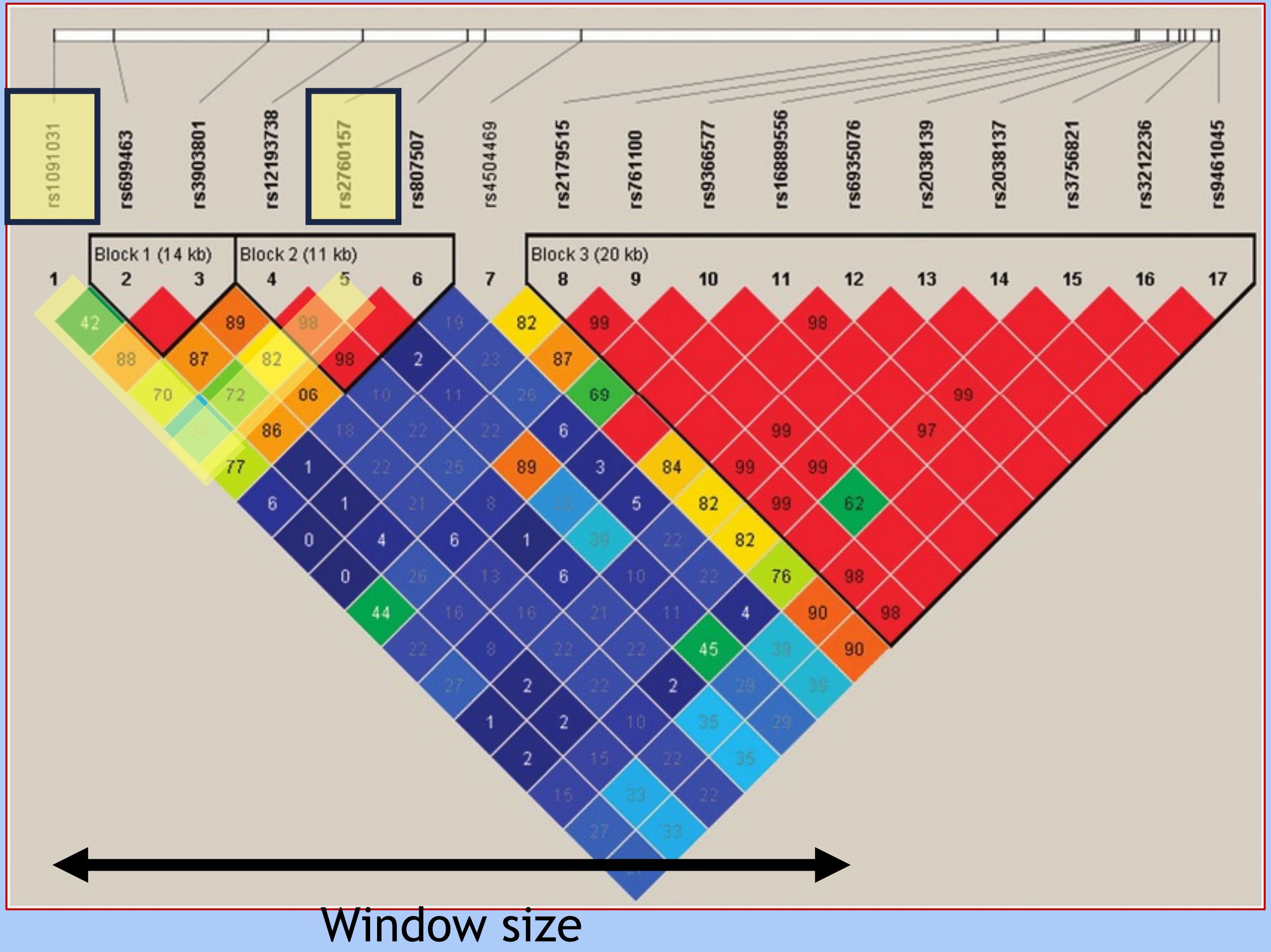
An algorithm for computing an independent subset of alleles



Window size

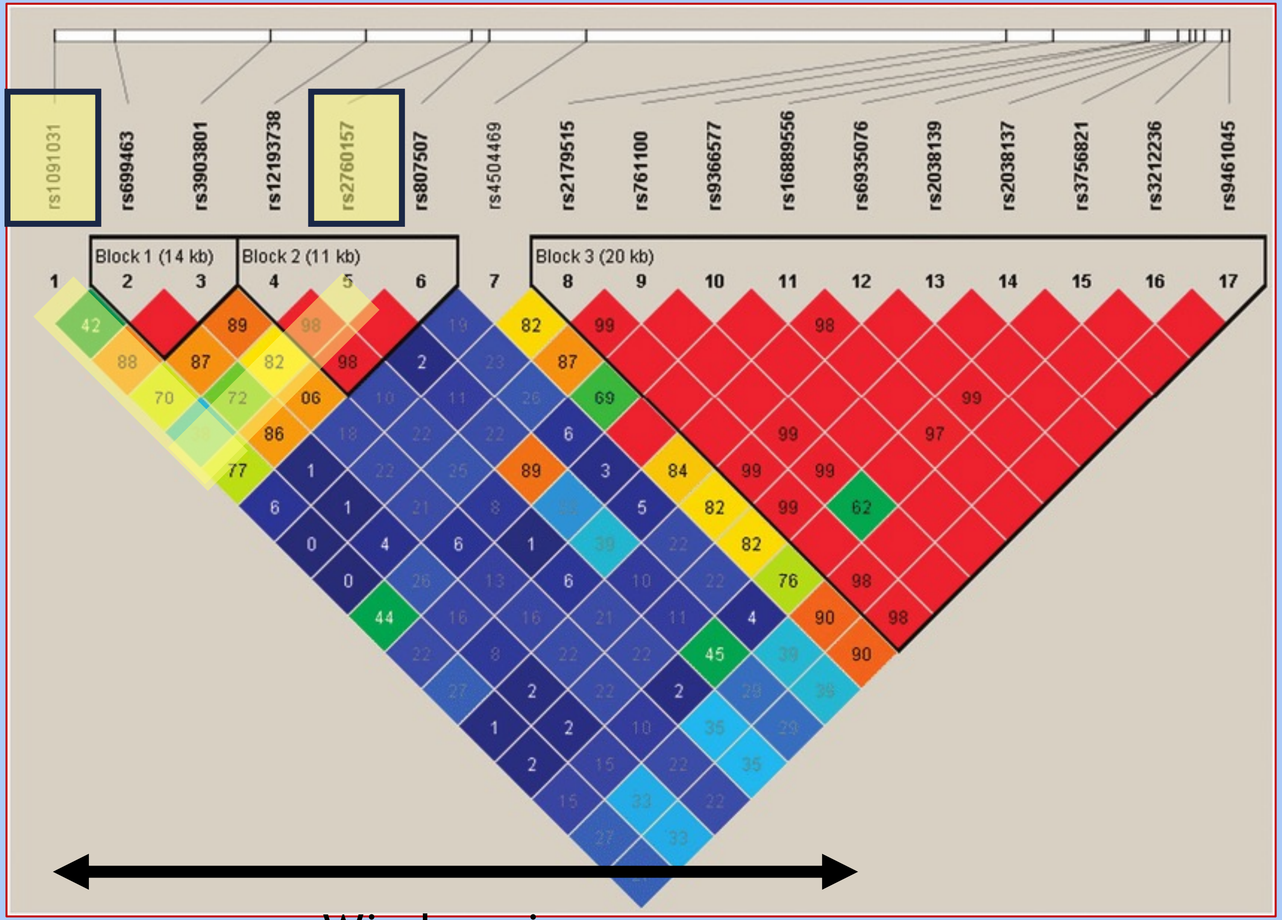
- Compute the LD with every other SNP within a sliding window of predetermined size

An algorithm for computing an independent subset of alleles



- If LD > threshold, remove the SNP

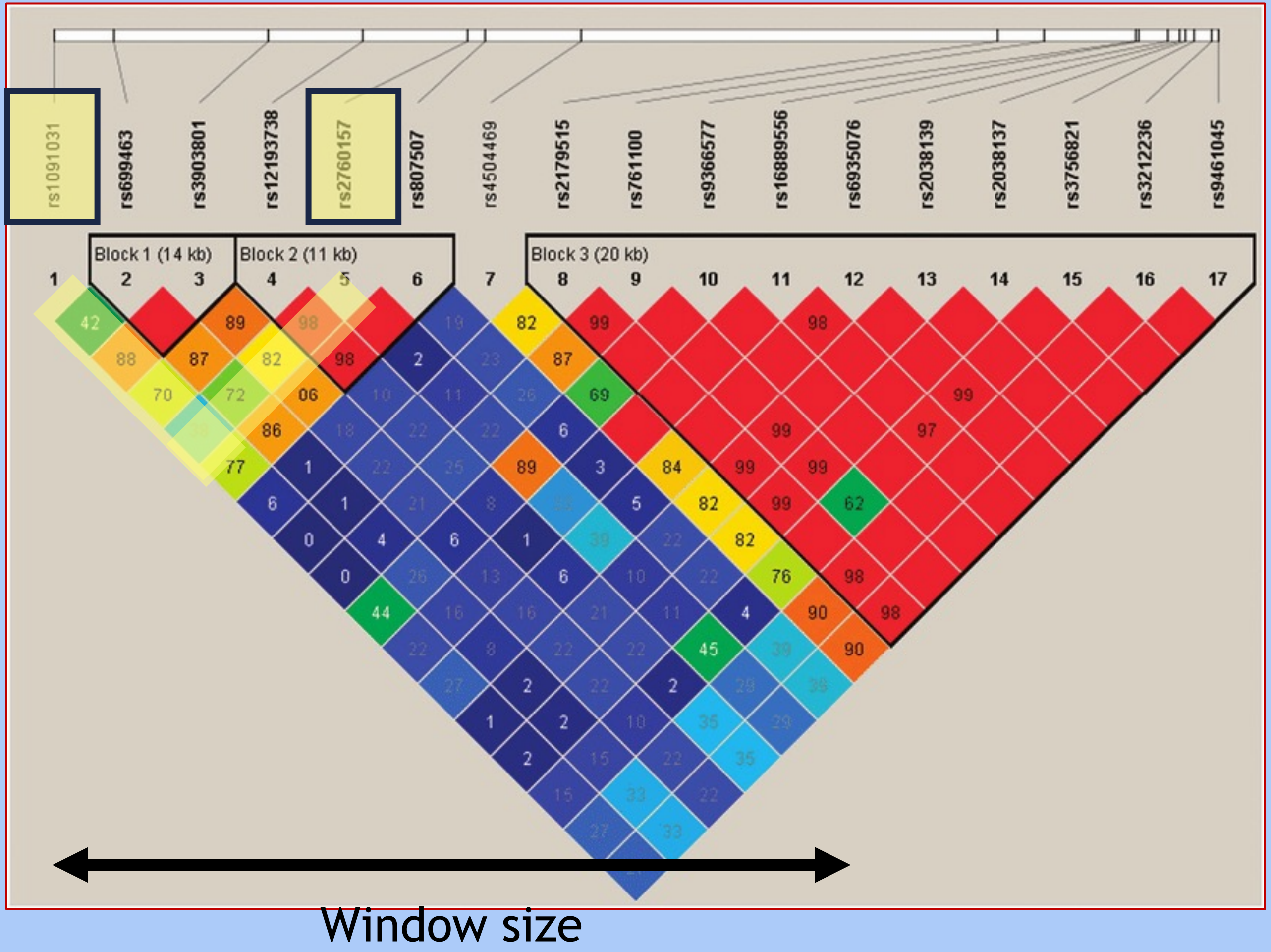
An algorithm for computing an independent subset of alleles



Window size

- Else it becomes a new independent SNP

An algorithm for computing an independent subset of alleles

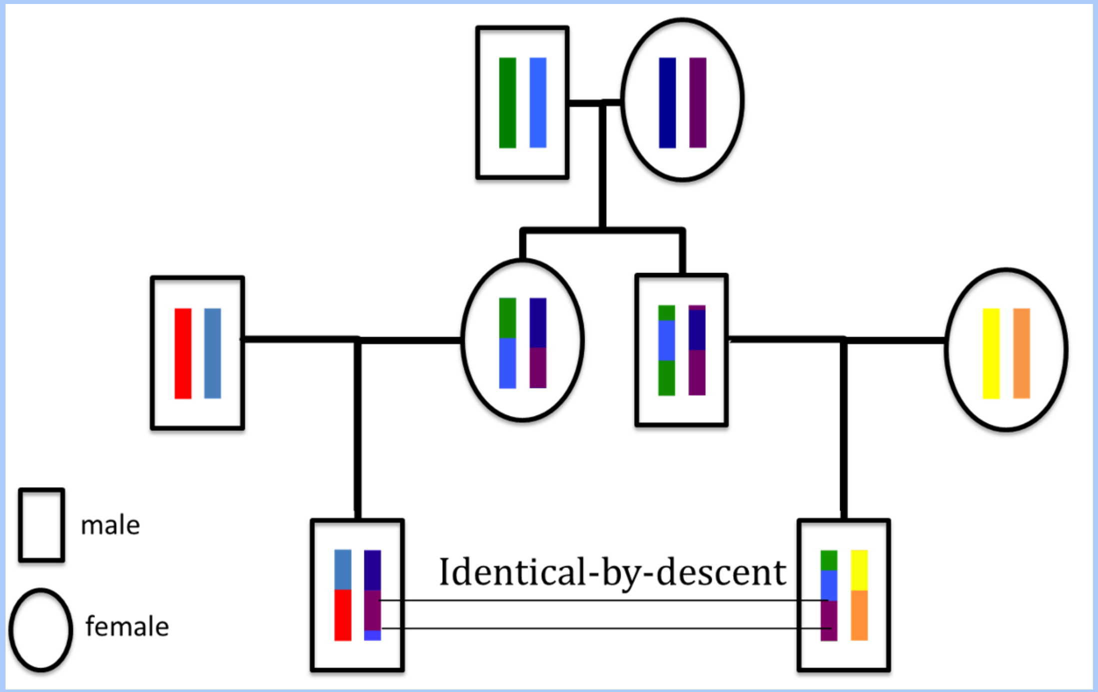


- The algorithm is random, and should be initiated from a fixed seed

Kinship analysis

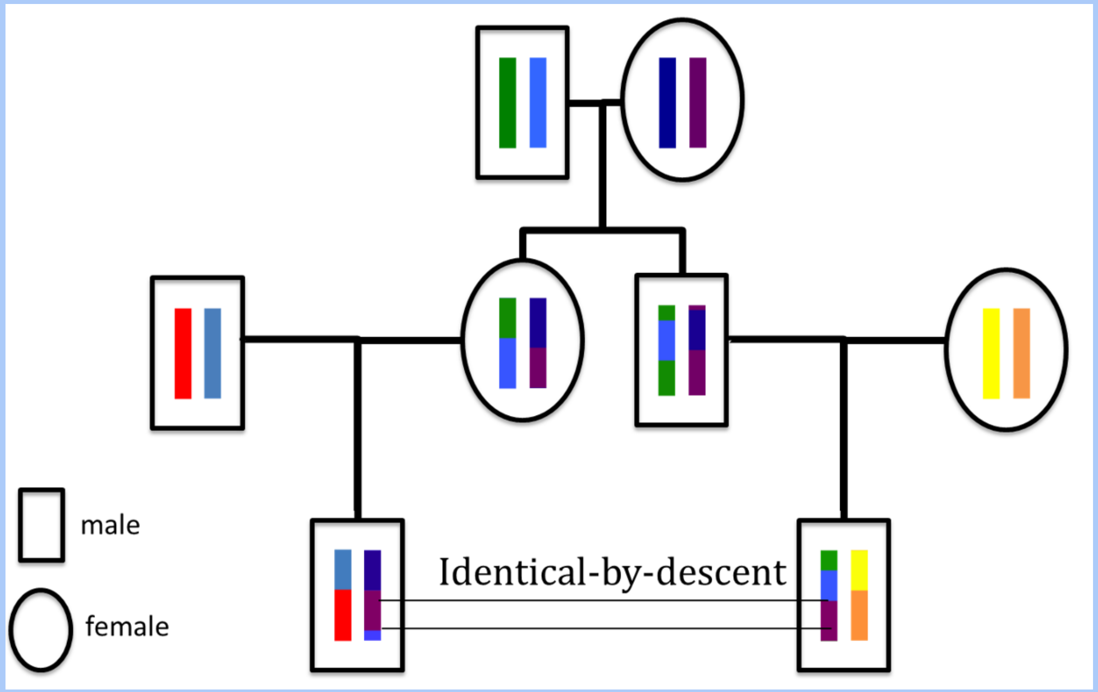
The concept of genetic relatedness

Relatives share haplotypes IBD



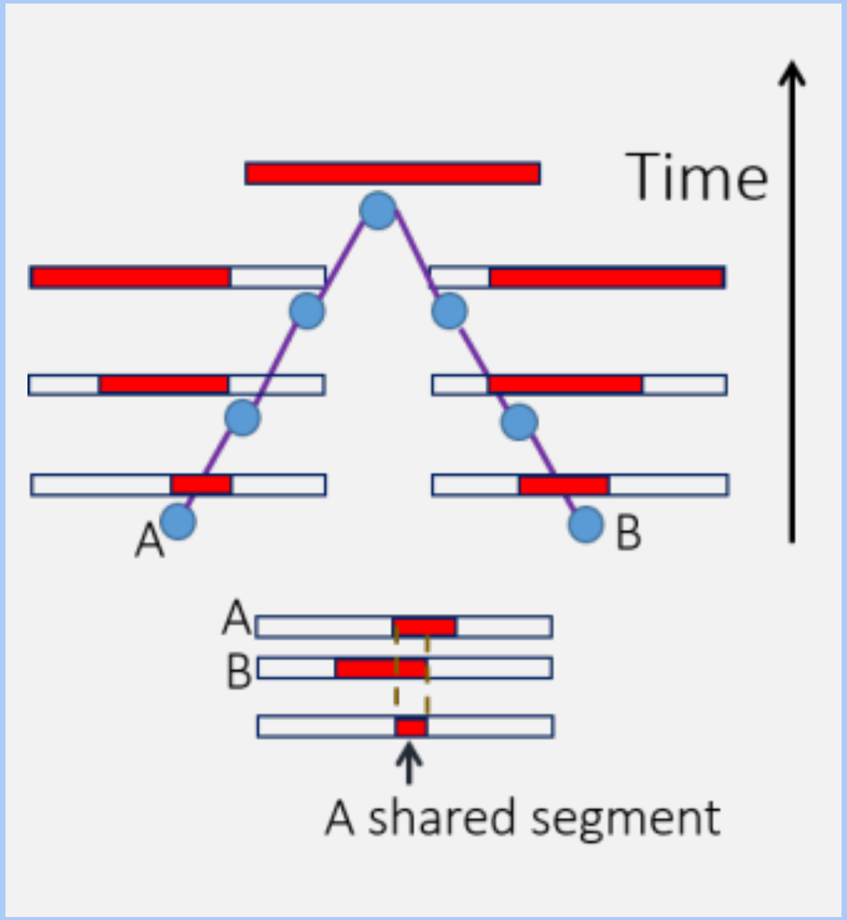
- Segments of DNA inherited from a common ancestor are said to be **identical by descent (IBD)**

Relatives share haplotypes IBD



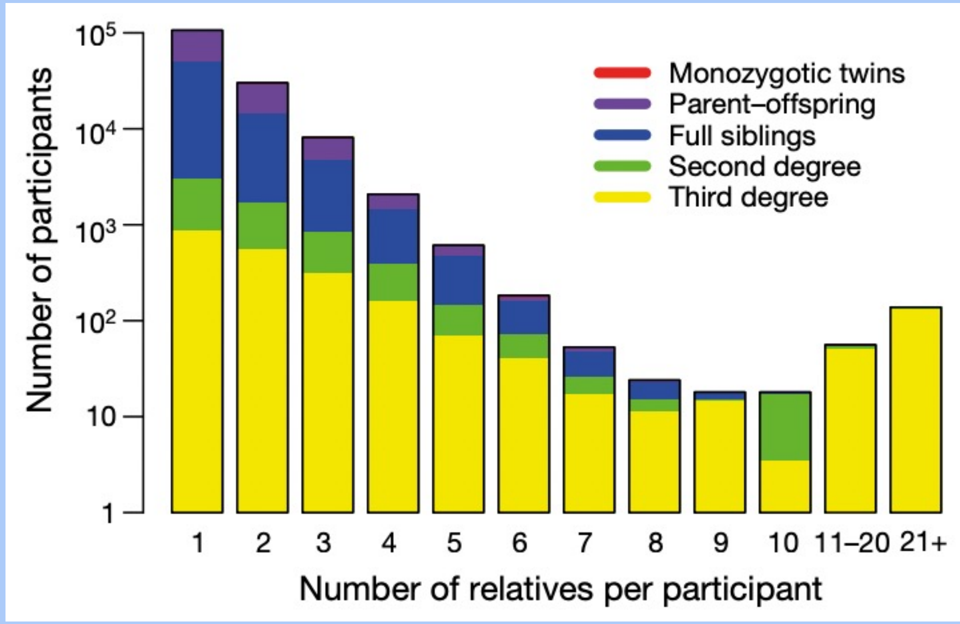
- DNA that just happens to be the same is **identical by state (IBS)**

Haplotype sharing decays over time



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

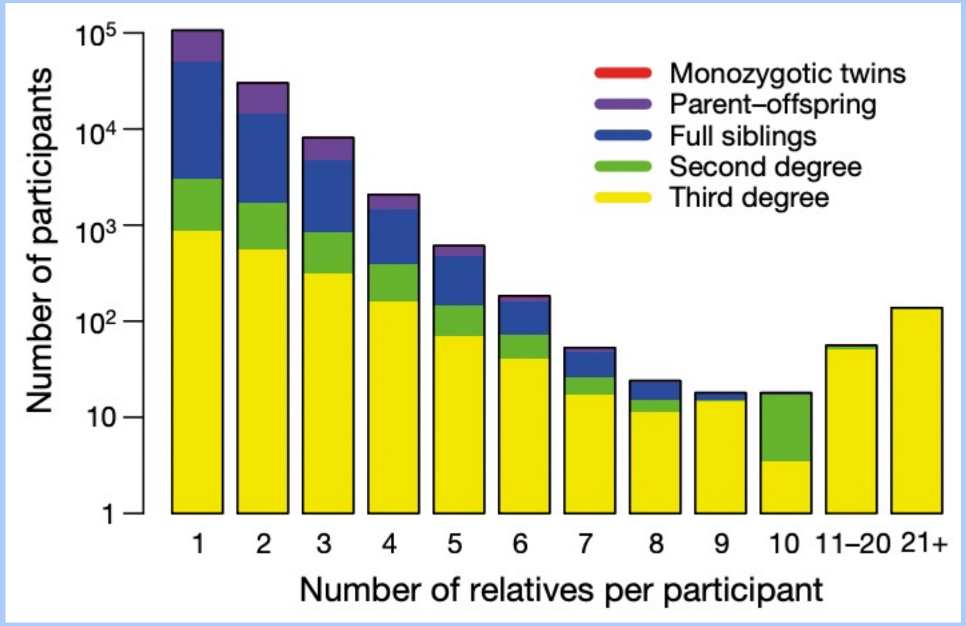
Kinship in genetic association studies



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

- Large genomic datasets, such as the UK Biobank, contain related individuals

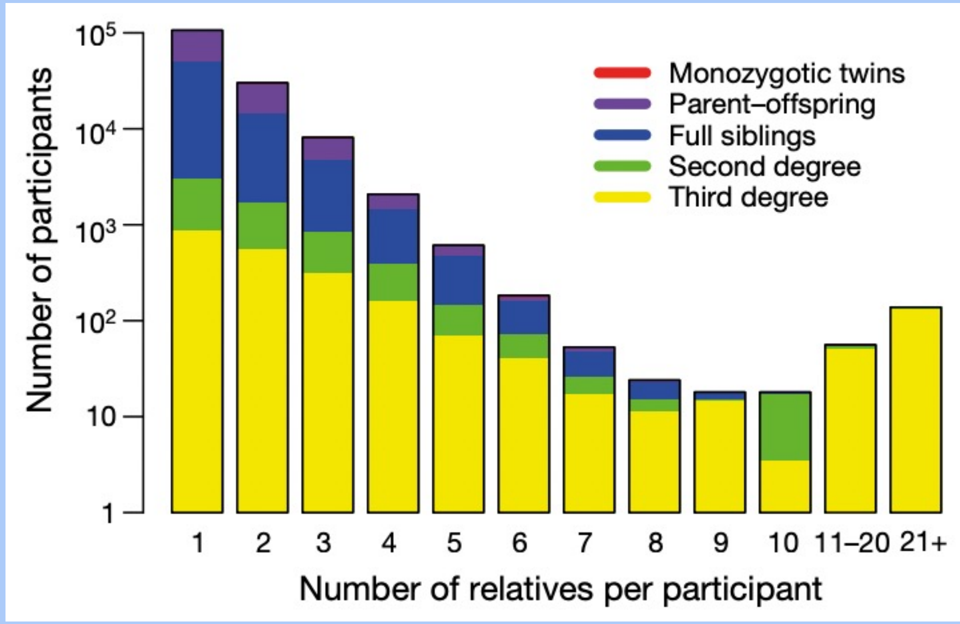
Kinship in genetic association studies



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

- Sometimes there is even “cryptic” relatedness

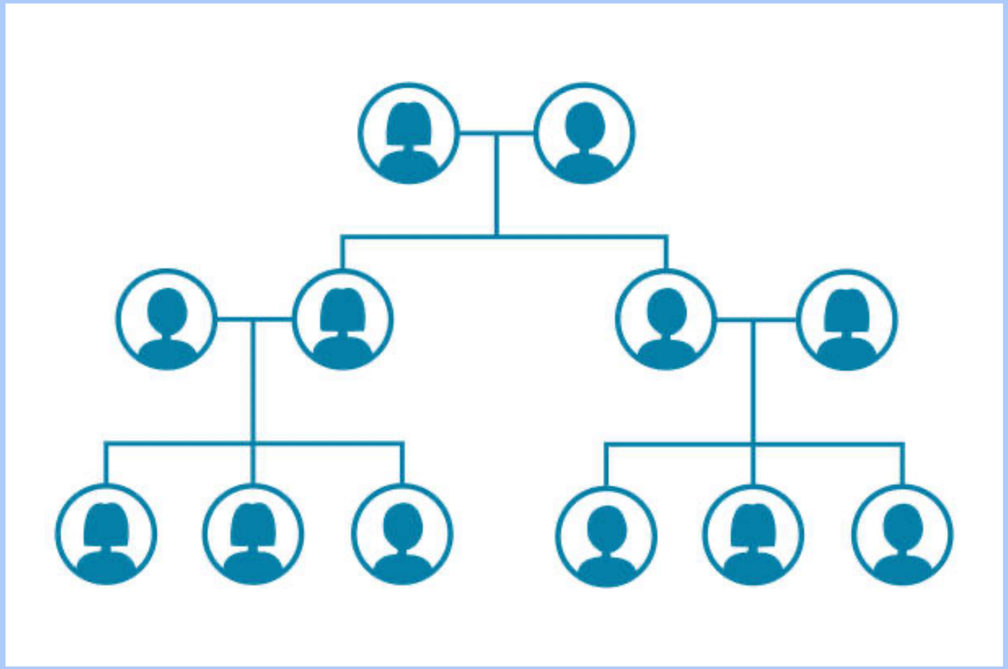
Kinship in genetic association studies



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

- Because of IBD sharing, not all the observations are independent, and genotype-phenotype associations may be **confounded**

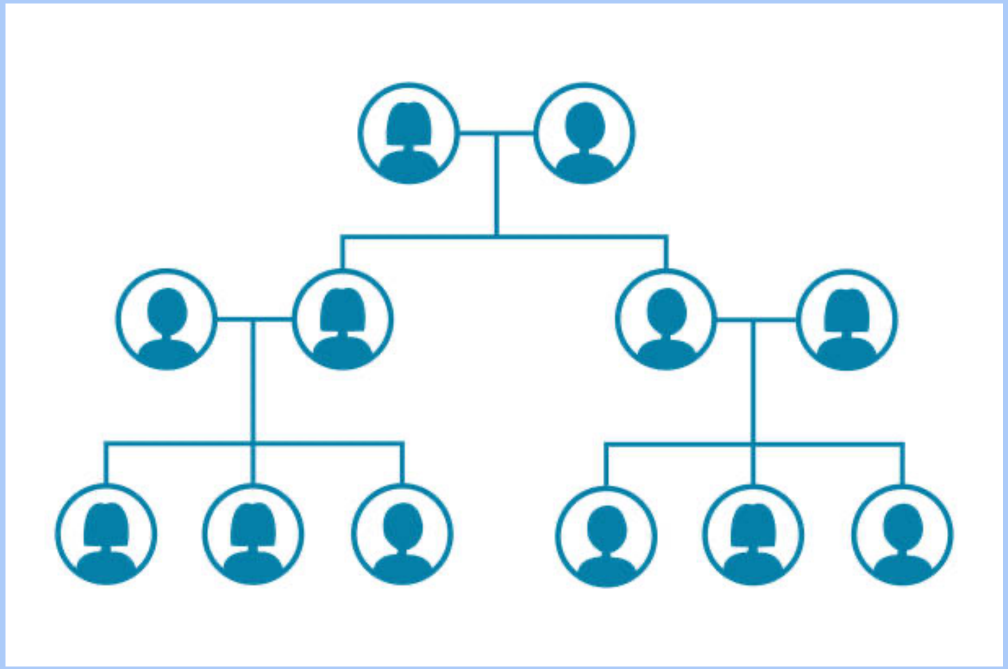
Degree of relatedness



- R is the **effective number of meioses** separating two individuals through their two parents 1 and 2

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

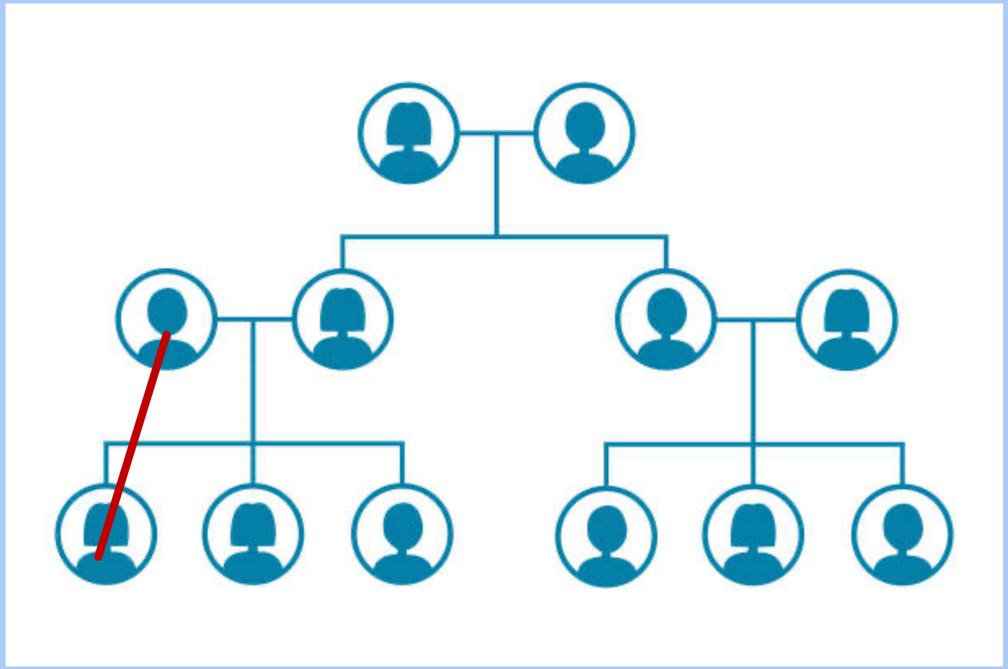
Degree of relatedness



- R → ∞ for unrelated individuals

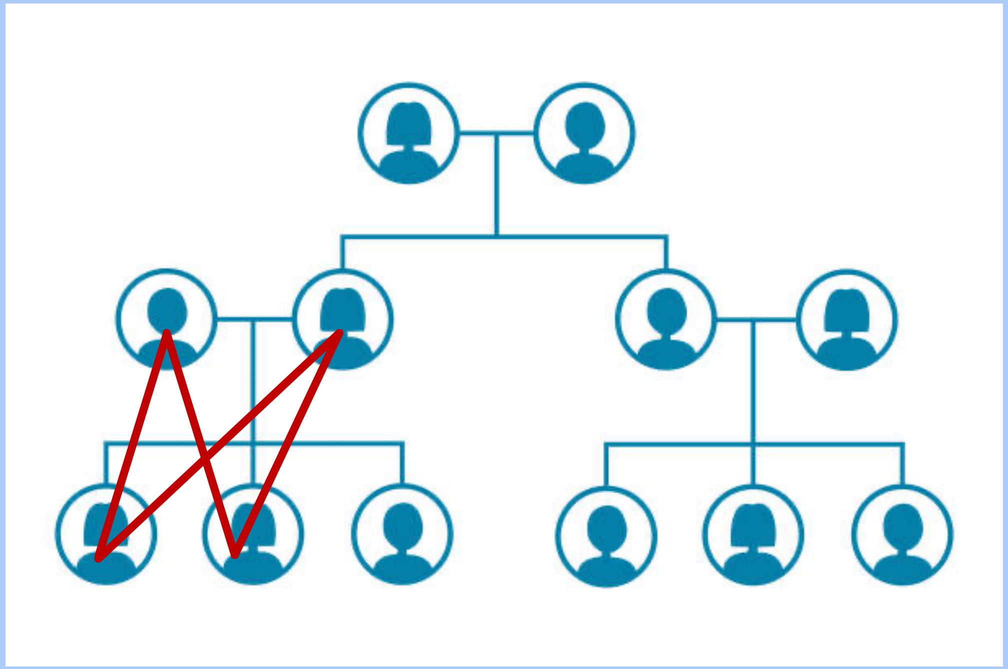
$$\frac{1}{2^R} = \frac{1}{2^{R_1}} + \frac{1}{2^{R_2}}$$

Degree of relatedness



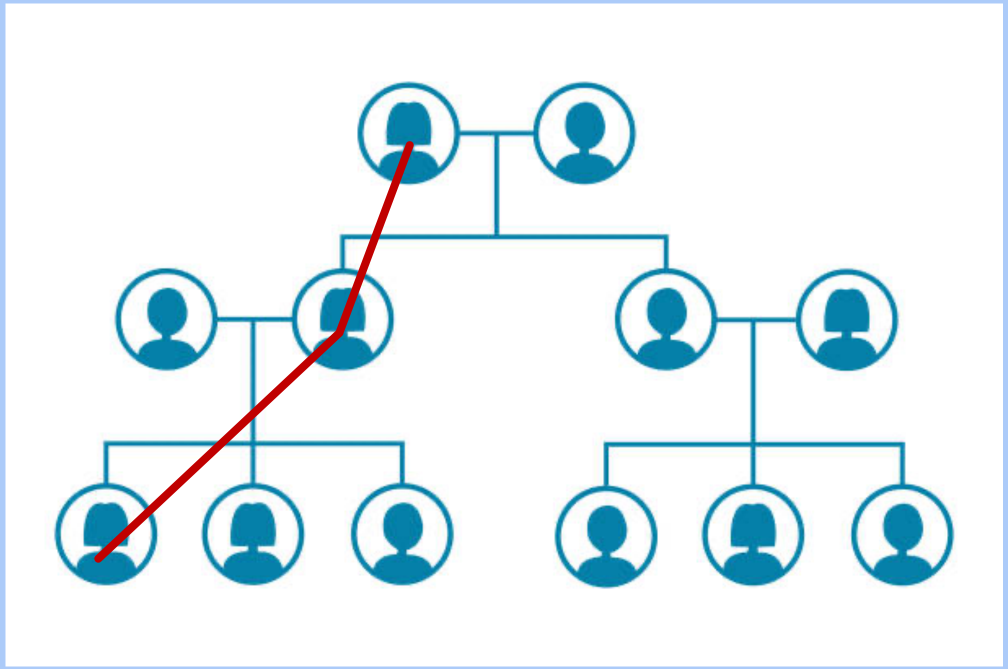
- Parent-child:
  - R = 1 meiosis

Degree of relatedness



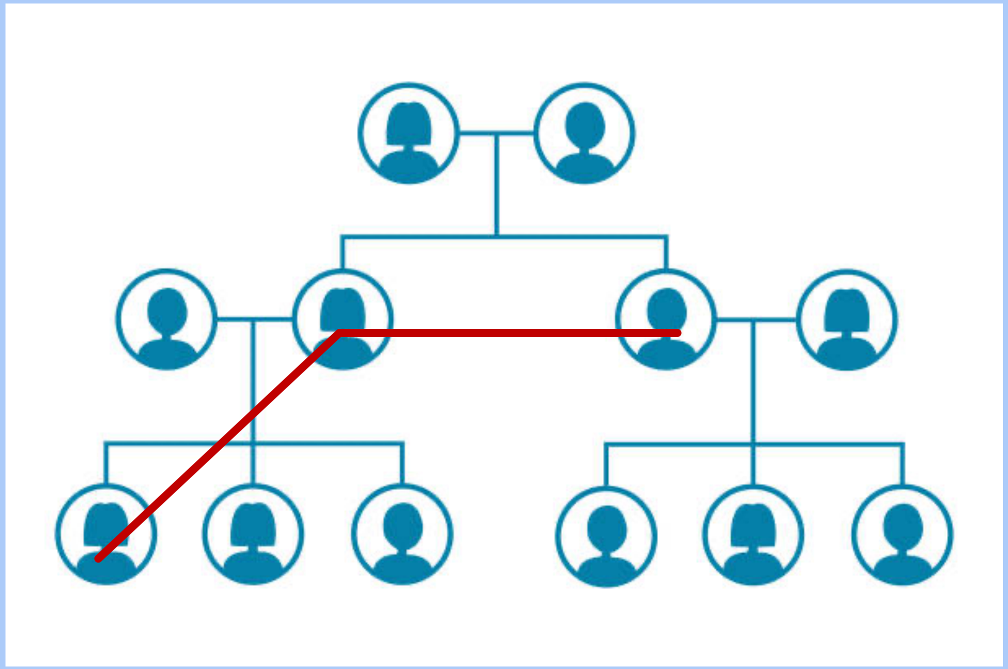
- Siblings: R = 1 “effective” meiosis:
  - $1 / 2^1 = 1 / 2^2 + 1 / 2^2$

Degree of relatedness



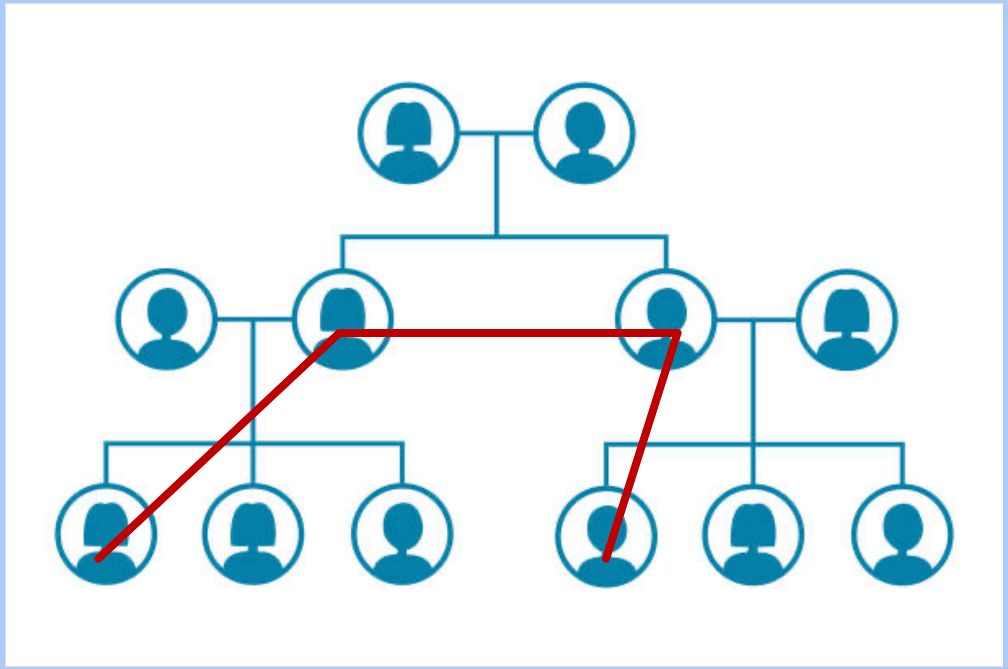
- Grandparent-grandchild:
  - R = 2 meioses

Degree of relatedness



- Avuncular:
  - R = 2 meioses

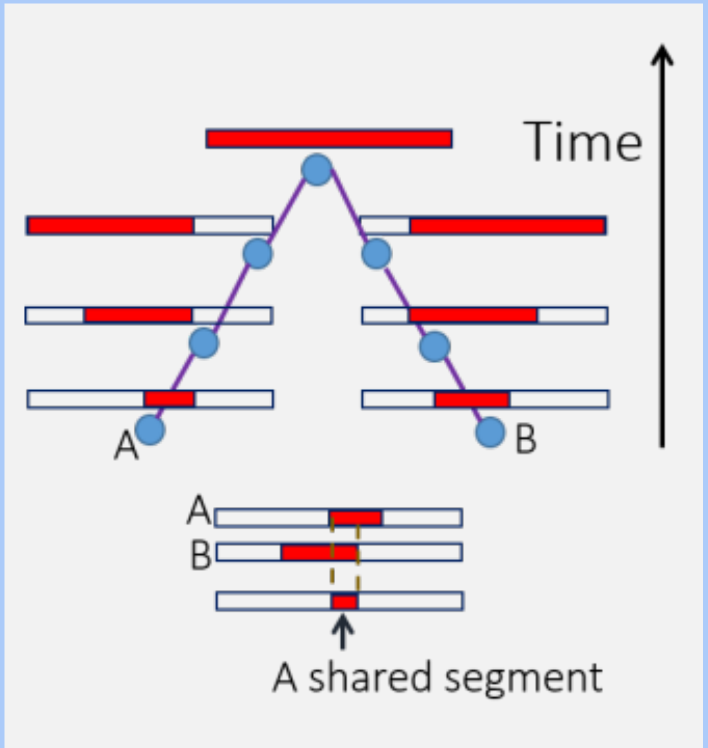
Degree of relatedness



- Cousins:
  - R = 3 meioses

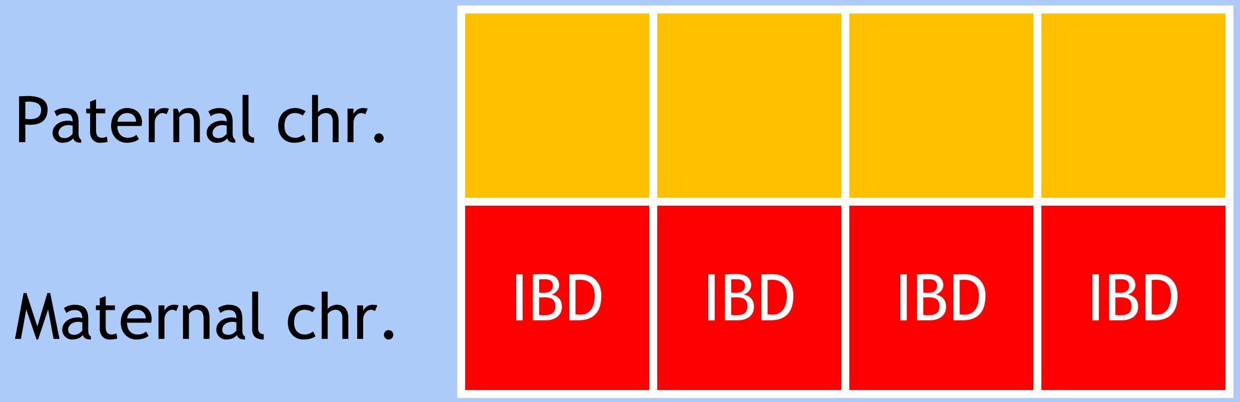
Degree of relatedness and the fraction of the genome shared IBD

- $r = 1 / 2^R$ 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



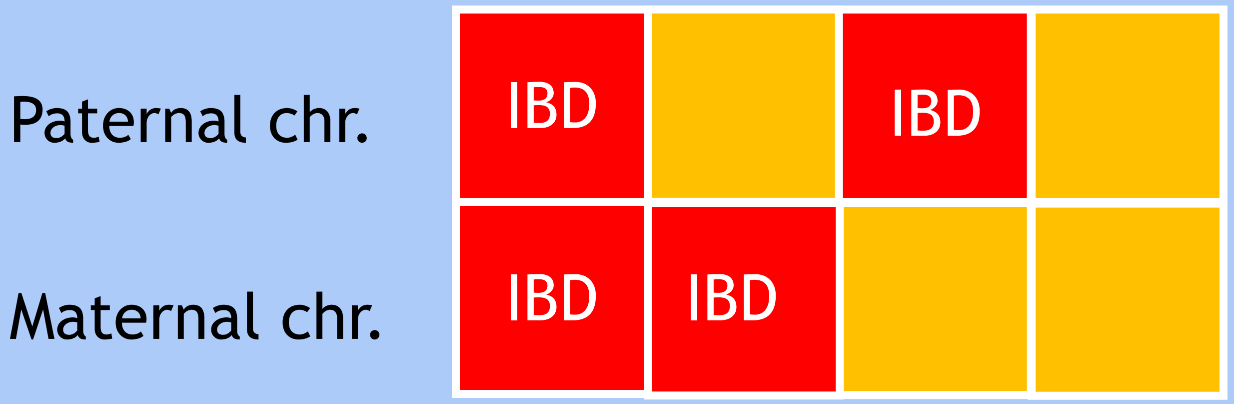
Degree of relatedness and the fraction of the genome shared IBD

- A child shares **half** of its DNA with its parent



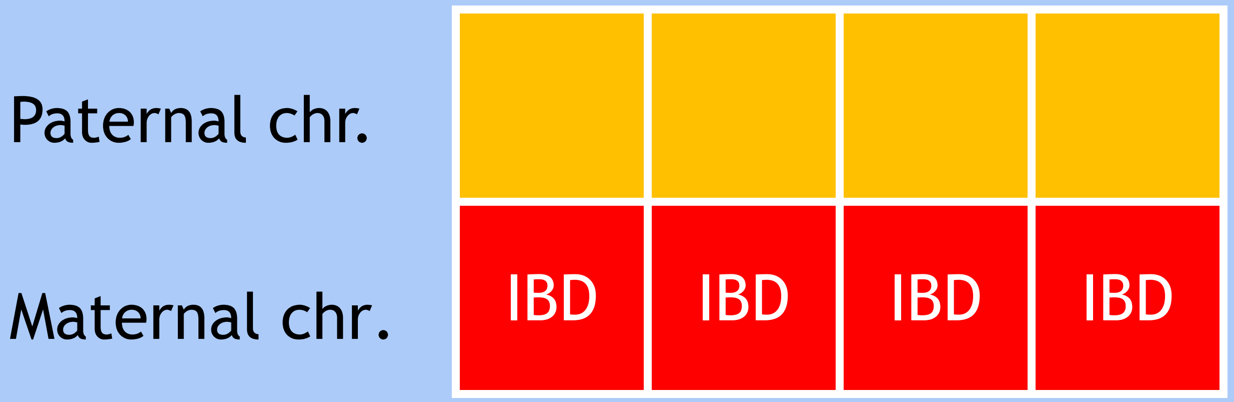
Degree of relatedness and the fraction of the genome shared IBD

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



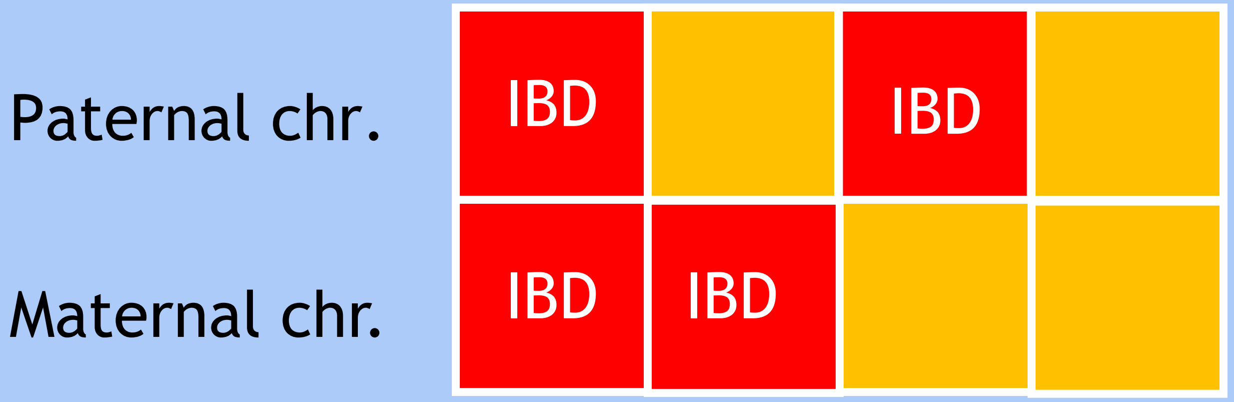
Degree of relatedness and the fraction of the genome shared IBD

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



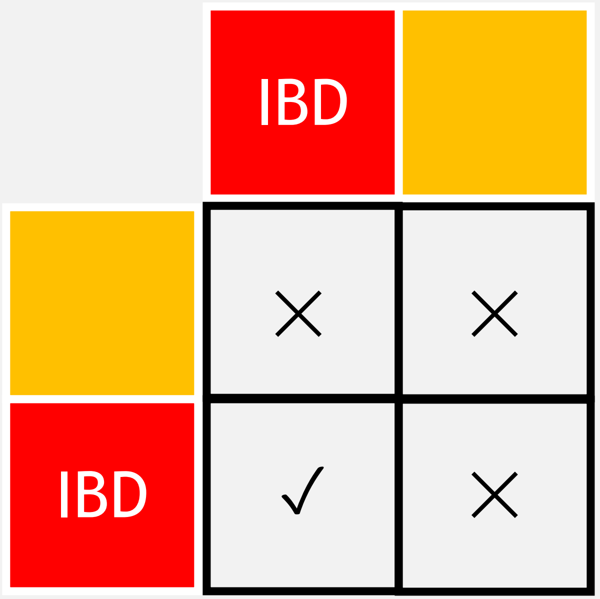
Degree of relatedness and the fraction of the genome shared IBD

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



The coefficient of relatedness φ

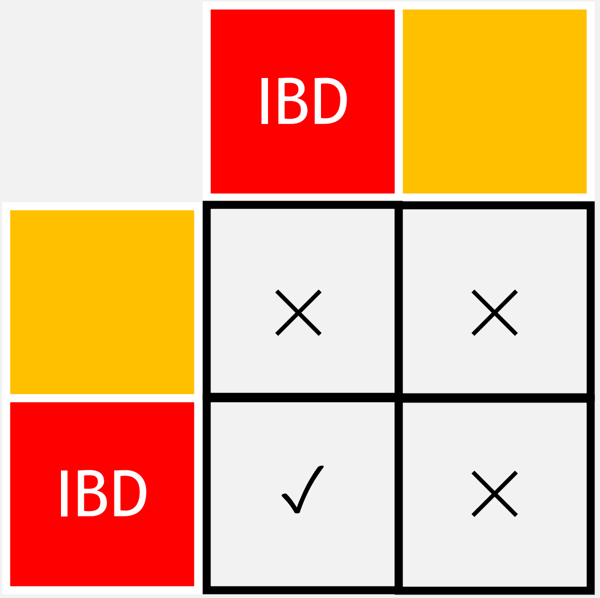
- φ is the probability that any two alleles at a single locus chosen from two individuals are shared IBD



$R = 1$, $r = \frac{1}{2}$, $\varphi = \frac{1}{4}$

The coefficient of relatedness φ

- φ is equal to half of $r = 1 / 2^R$



$R = 1$, $r = ½$, $φ = ¼$

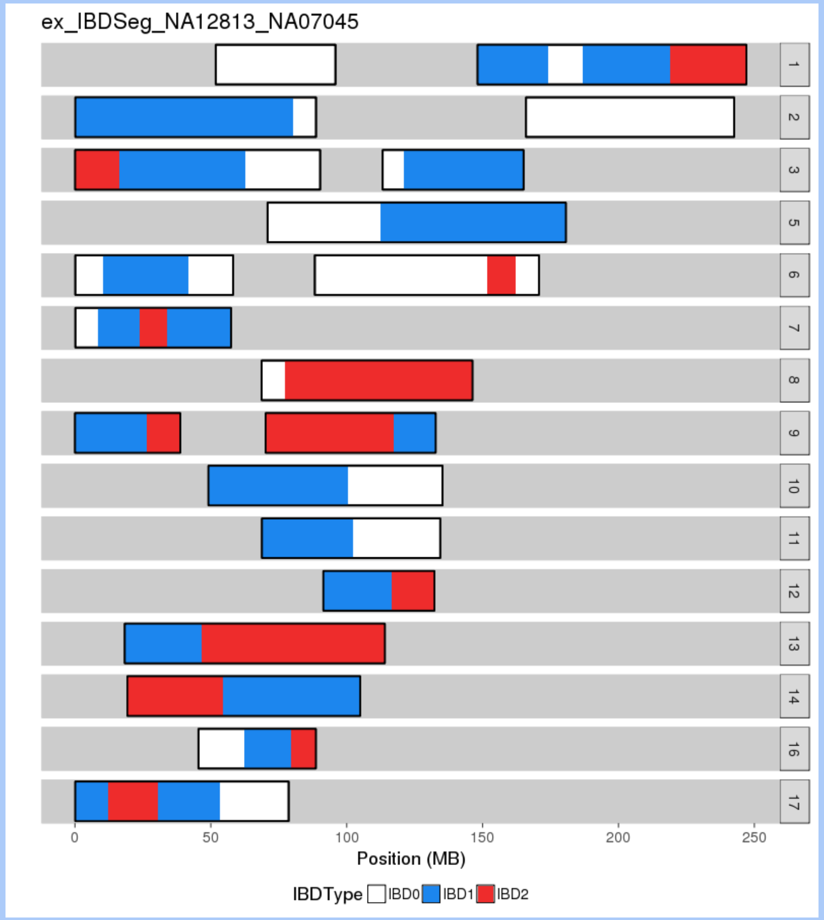
Coefficient of relatedness and IBD = 0

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

Relationship	R	φ	IBD = 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

Kinship-based Inference for GWAS (KING)

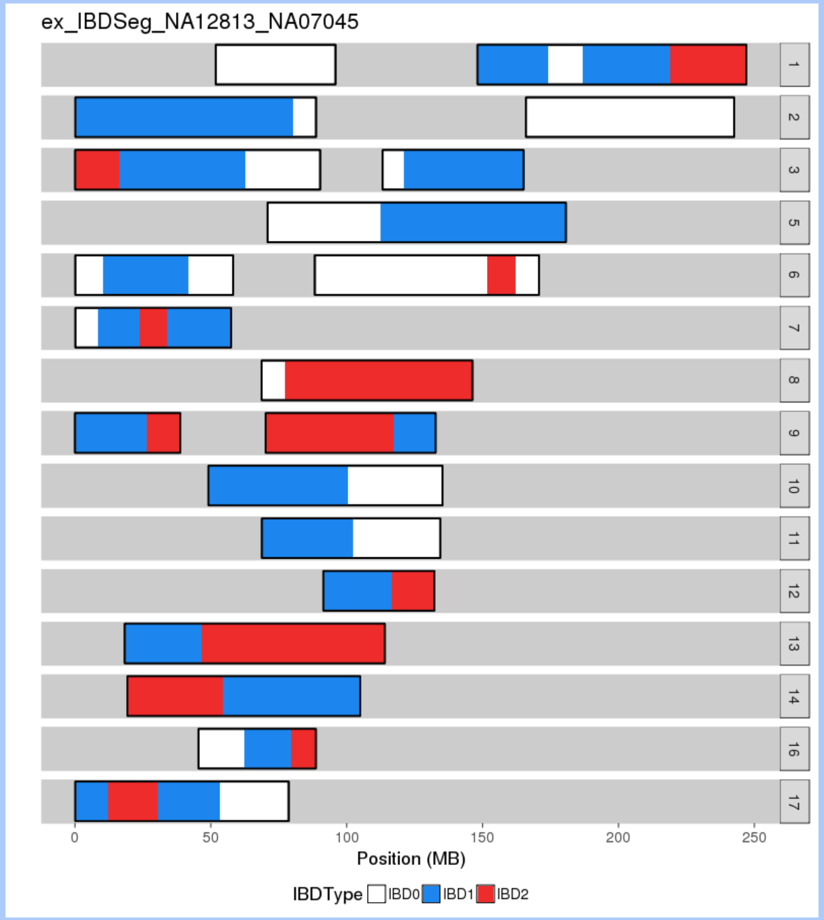
- Estimate φ and IBD sharing from the number of sites at which two individuals are both heterozygotes (Aa,Aa) or opposite homozygotes (AA,aa)



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

Kinship-based Inference for GWAS (KING)

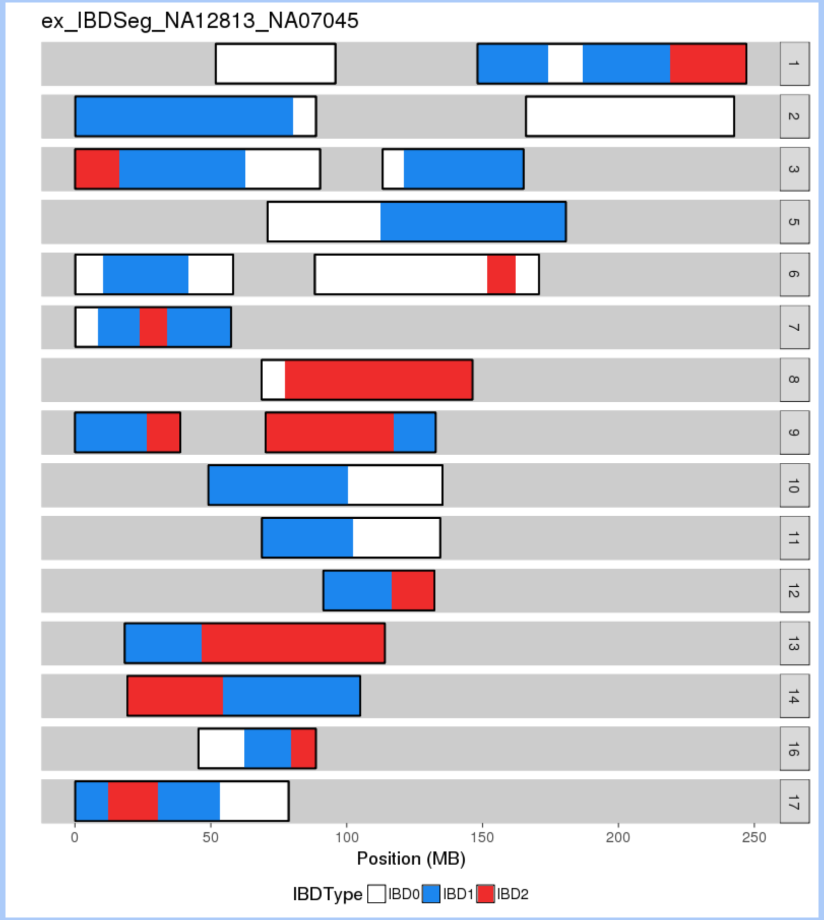
- A **robust** method that avoids estimating population allele fractions, just focuses on pairs



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

Kinship-based Inference for GWAS (KING)

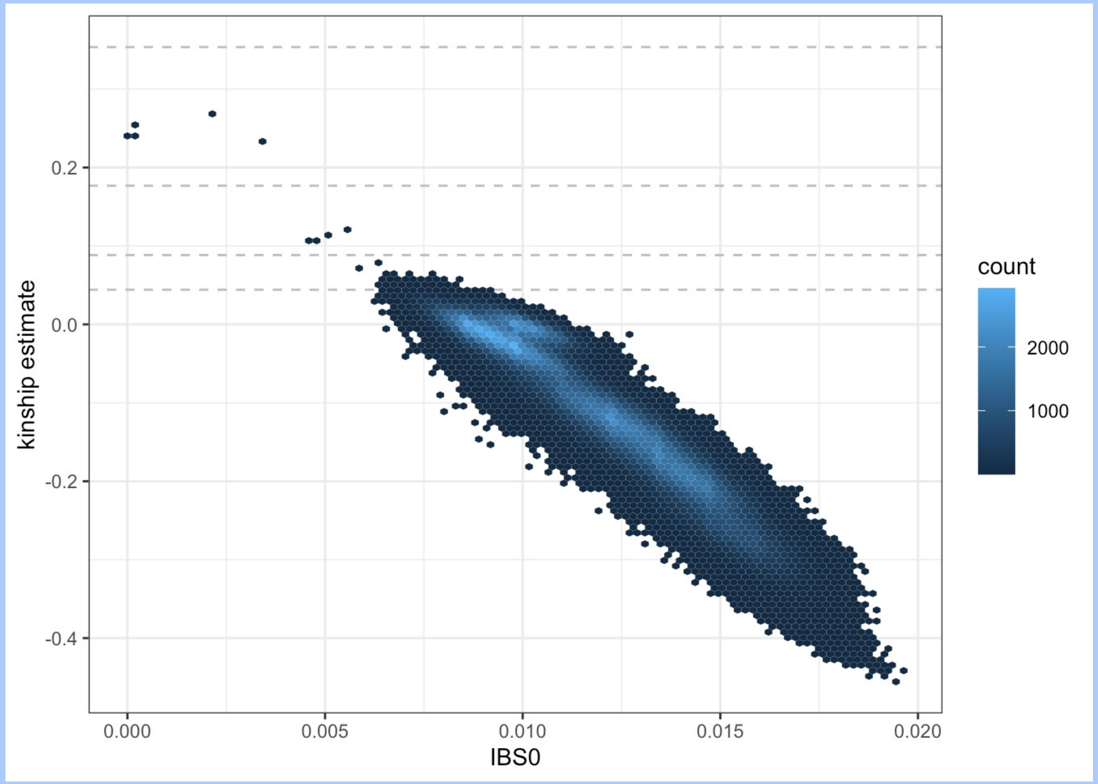
- Can generate negative estimates of φ, indicating individuals are from distinct populations



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

Kinship-based Inference for GWAS (KING)

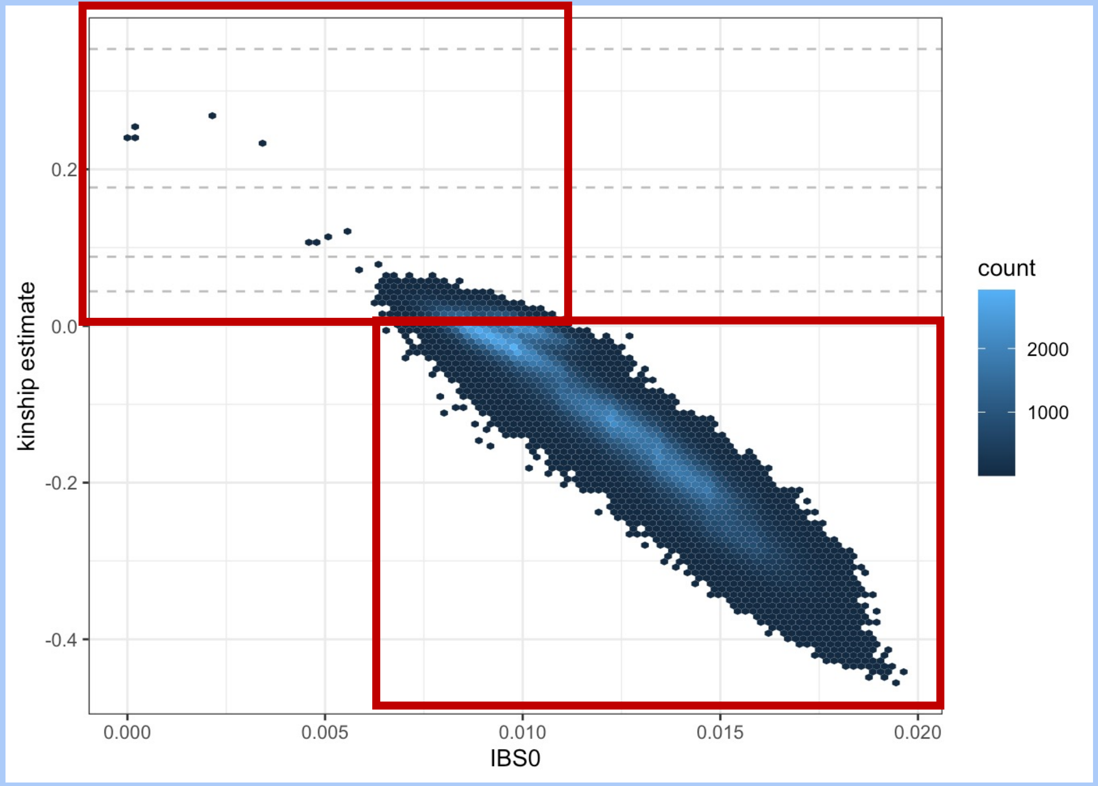
- φ is plotted vs. the fraction of IBS = 0 sites (AA,aa)



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

Kinship-based Inference for GWAS (KING)

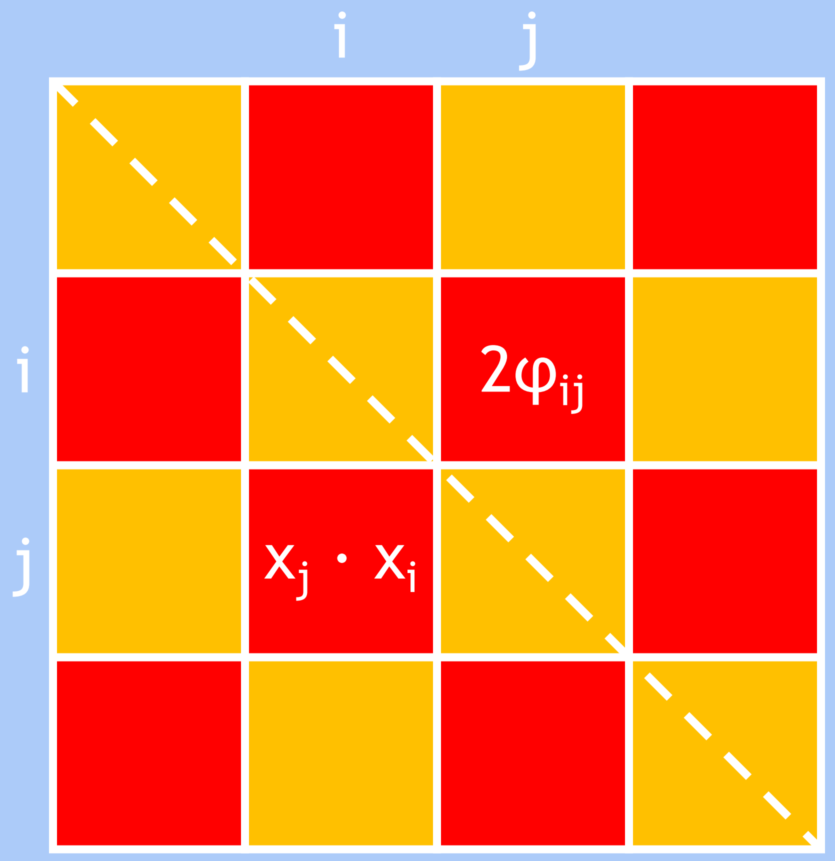
- Negative estimates indicate unrelated individuals from different populations



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

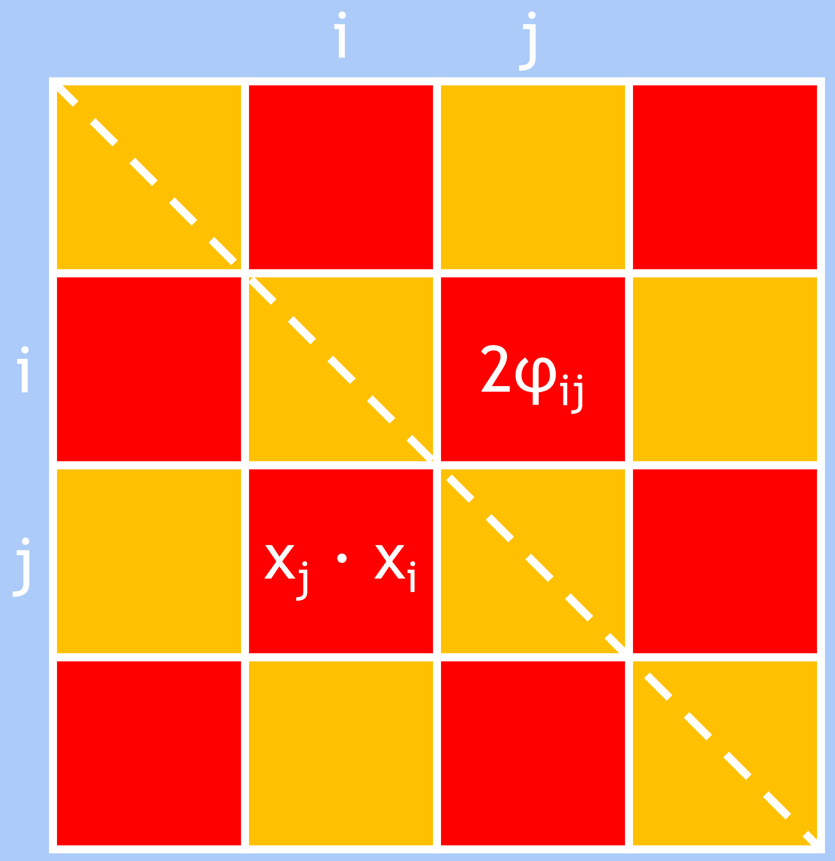
Updating the GRM

- The KING kinship coefficients 2φ are approximately equal to the GRM



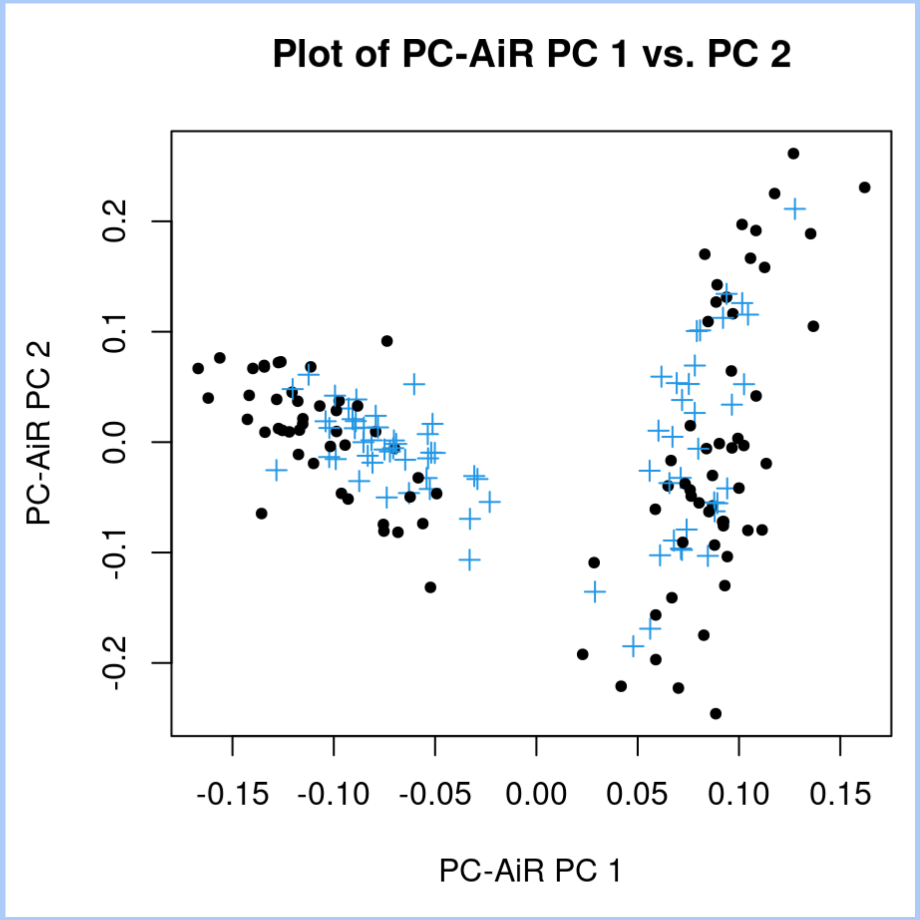
Updating the GRM

- 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)

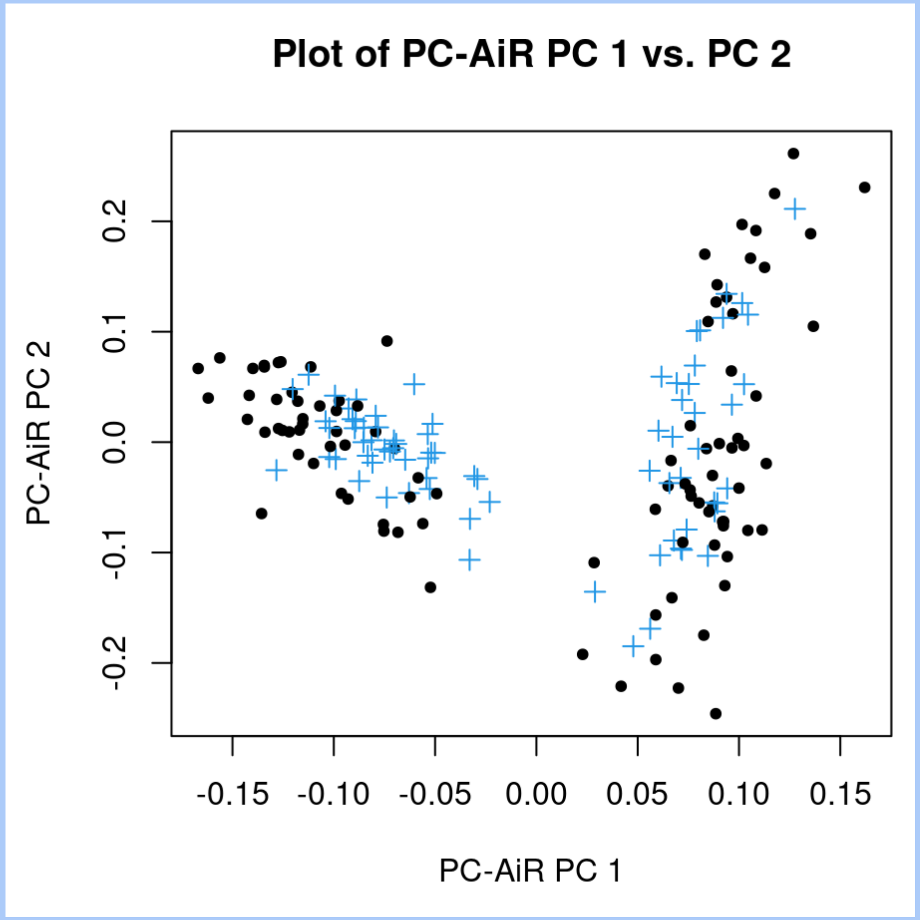


https://bioconductor.org/packages/devel/bioc/vignettes/GENESIS/inst/doc/pcair.html

PC-AiR: PCA in Related Samples

- Based on the KING estimates, PC-AiR computes PCs **U** for a set of unrelated individuals (black) with genotype matrix **X**

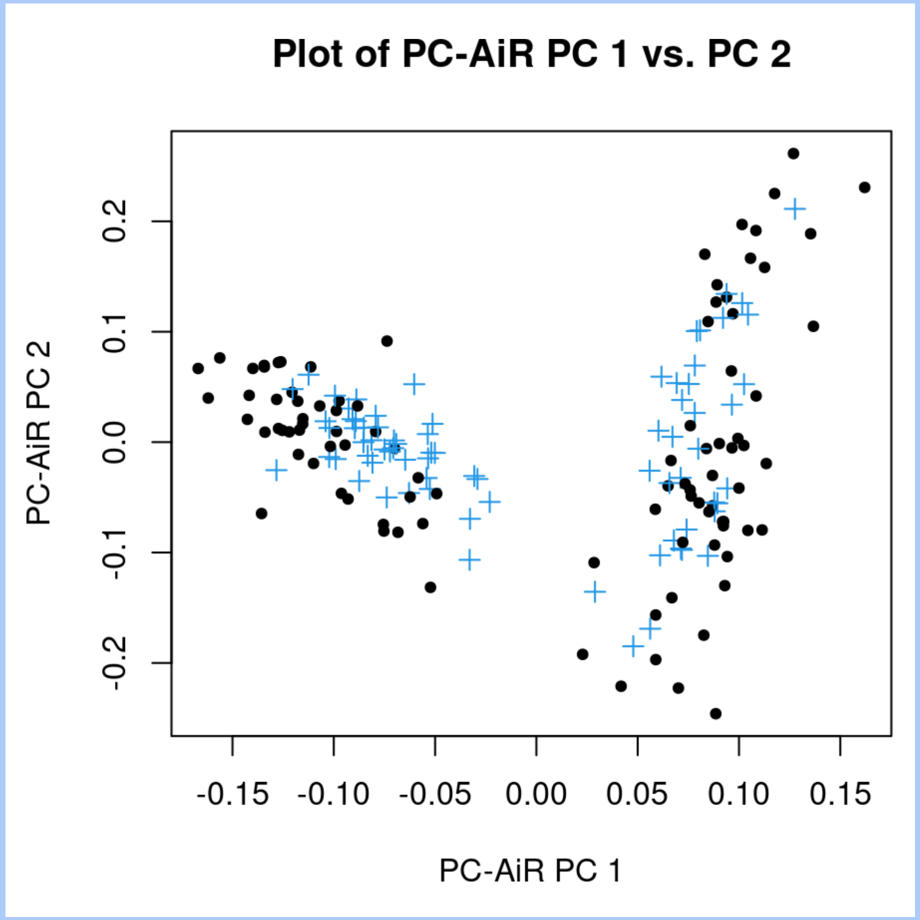
$$\mathbf{X}^T\mathbf{U} = \mathbf{V}\boldsymbol{\Sigma}$$



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

PC-AiR: PCA in Related Samples

- PCs for the remaining samples (blue) are imputed into the remaining subset (blue)

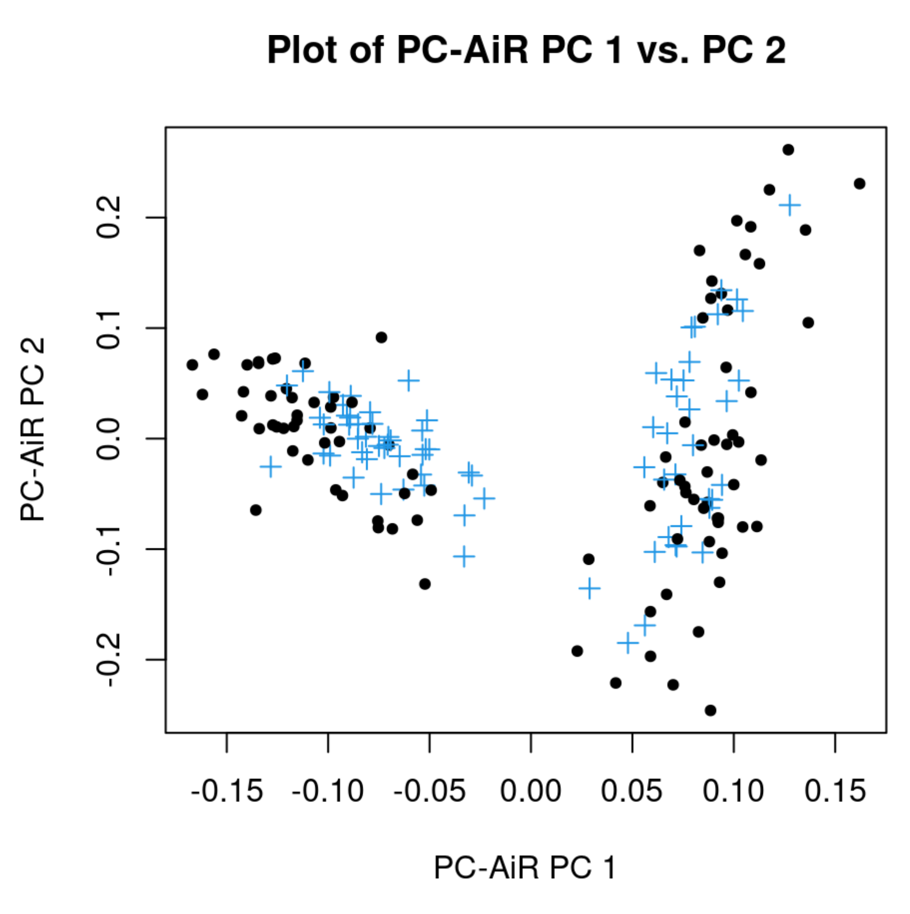


https://bioconductor.org/packages/devel/bioc/vignettes/GENESIS/inst/doc/pcair.html

PC-AiR: PCA in Related Samples

- PCs **U'** for the remaining samples (blue) with genotype matrix **X'** are imputed into the remaining subset (blue)

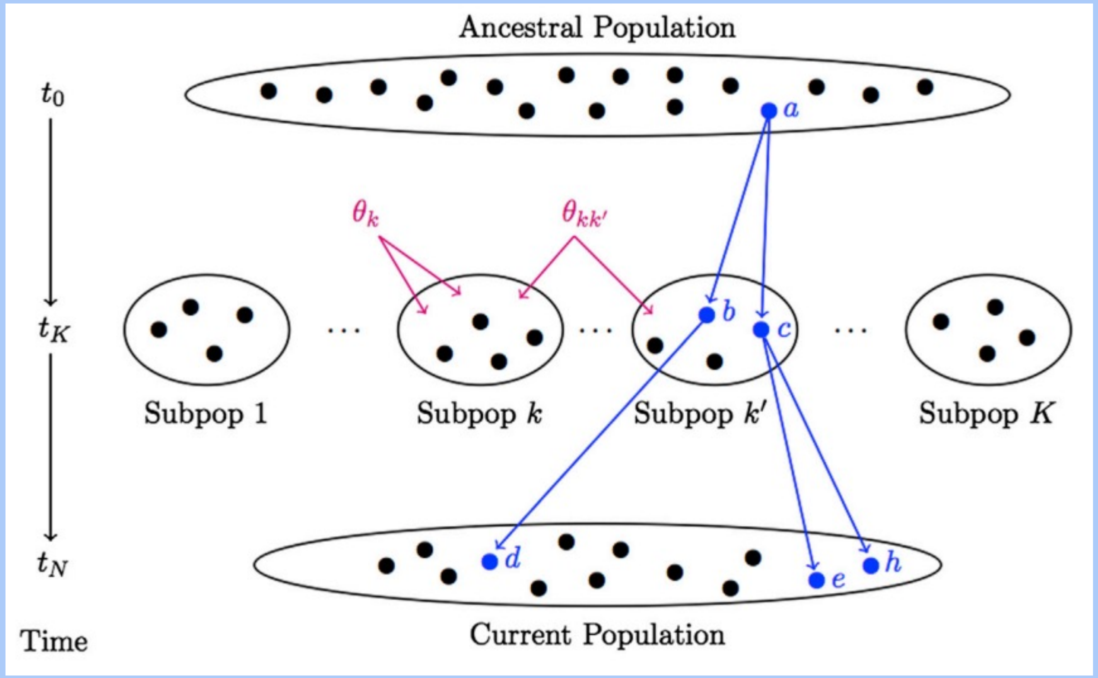
$$\mathbf{X}'\left(\mathbf{X}^T\mathbf{U}\boldsymbol{\Sigma}^{-1}\right) = \mathbf{X}'\mathbf{V}$$



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

PC-Relate

- PC-Relate uses the updated PCs to distinguish shared genetic ancestry from recent common ancestors



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

PC-Relate

- Each individual's "best-fit" genotype is predicted from its PCs

$$\mathbb{E}\left(g_{ik} \mid u_{ij}\right) = 2p_k + 2p_k\left(1 - p_k\right) u_{ij} \lambda_{jj} \cdot v_{kj}$$

PC-Relate

- Each individual's "best-fit" genotype is predicted from its PCs

$$\mathbb{E}\left(g_{ik} \mid u_{ij}\right) = 2p_k + 2p_k\left(1 - p_k\right) u_{ij}\lambda_{jj} \cdot v_{kj}$$

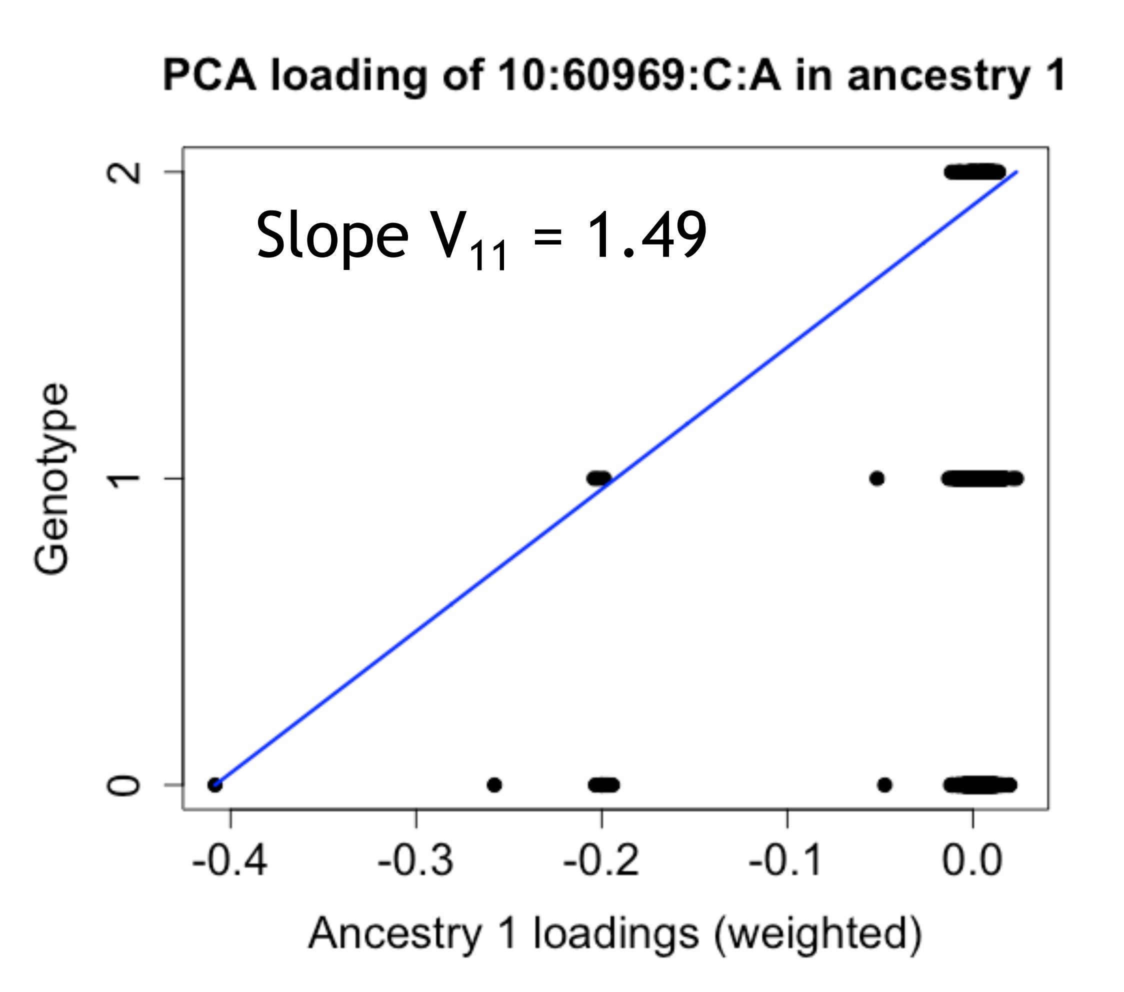
Population allele frequency

Amount of ancestry j

SNP k genotype in ancestry j

PC-Relate

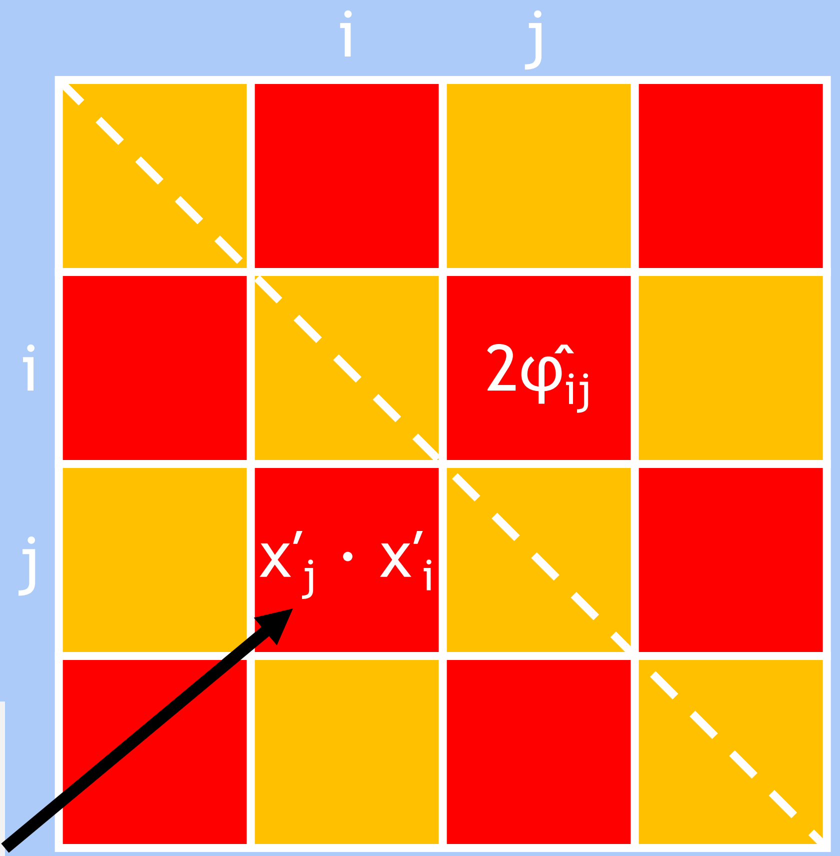
- The slope of the best fit line of genotype vs. (weighted) PC1 is equal to the expected SNP genotype in ancestry 1



PC-Relate

- An updated GRM that reflects only recent common ancestry can be constructed using the “best-fit” genotypes 2pik for each individual I at SNP k

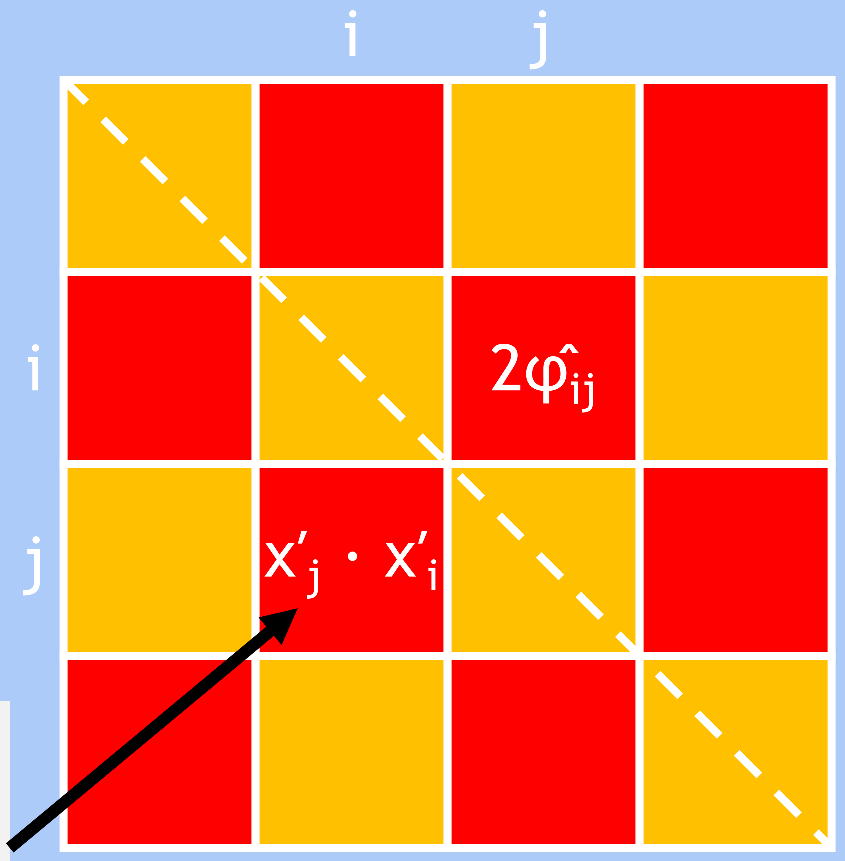
$$2\hat{\varphi}_{ij} = \frac{\sum_k \left(g_{ik} - 2p_{ik}\right)\left(g_{jk} - 2p_{jk}\right)}{2\sqrt{p_{ik}\left(1 - p_{ik}\right)p_{jk}\left(1 - p_{jk}\right)}}$$



PC-Relate

- The updated GRM will be used for fitting a generalized linear model during association testing

$$2\hat{\varphi}_{ij} = \frac{\sum_k \left(g_{ik} - 2p_{ik}\right)\left(g_{jk} - 2p_{jk}\right)}{2\sqrt{p_{ik}\left(1 - p_{ik}\right)p_{jk}\left(1 - p_{jk}\right)}}$$

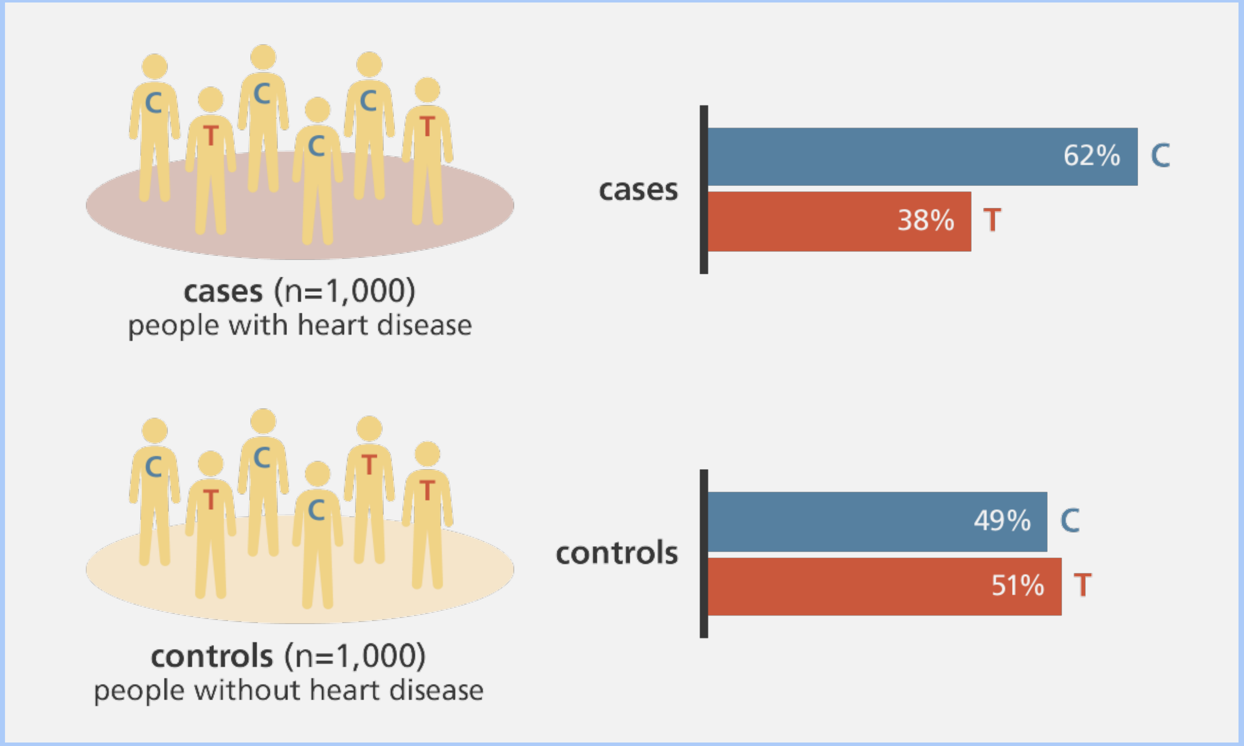


Association testing

Logistic regression and linear mixed models

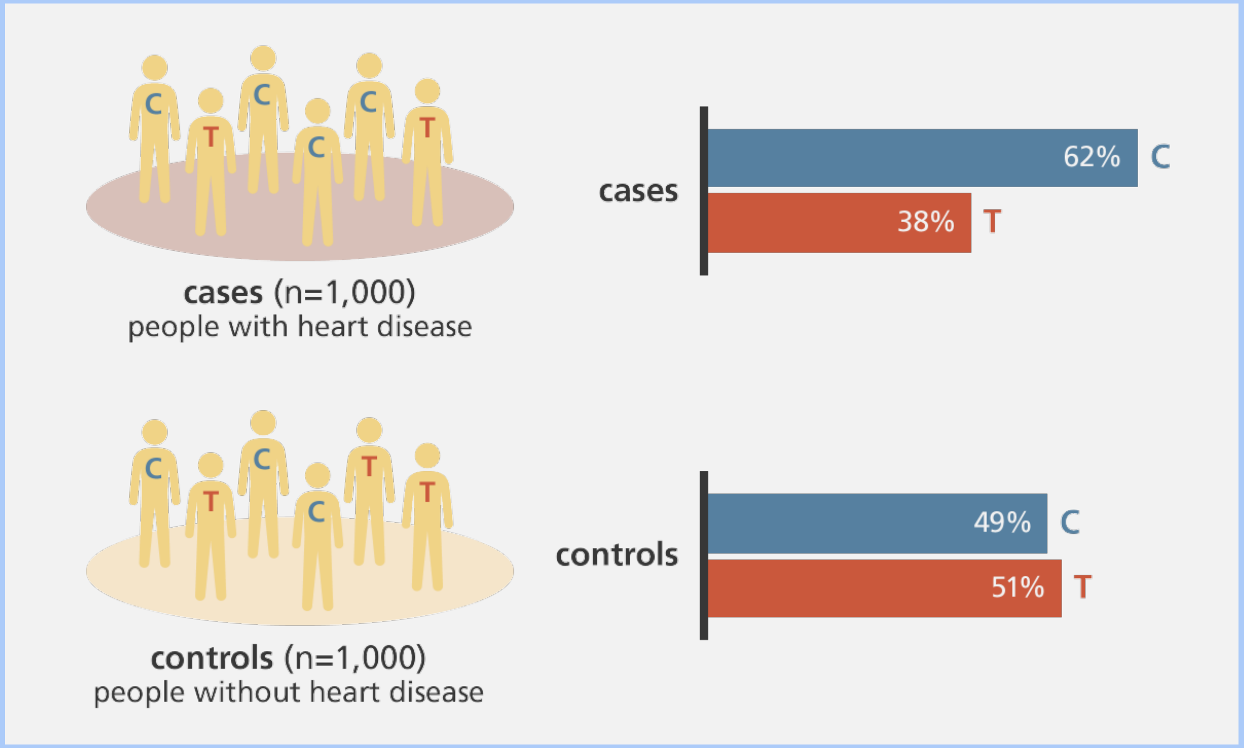
Case-control studies

- Is a genetic variant associated with disease?



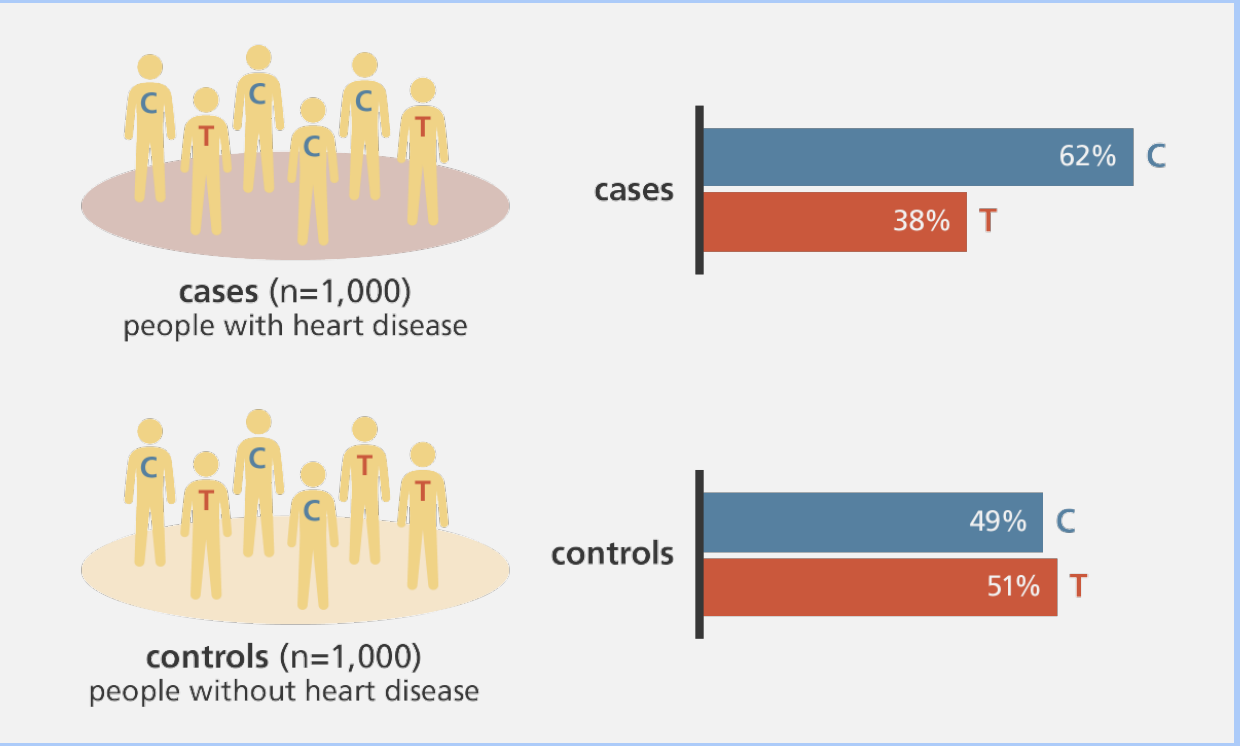
Case-control studies

- Is a genetic variant enriched in people with disease compared to people without?



Case-control studies

- 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 (a / c) to the odds that Controls have the risk allele (b / d)

	Cases	Controls
C	a	b
T	c	d

$$OR = (a / c) / (b / d) = (ad) / (bc)$$

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)

	Cases	Controls
C	620	490
T	380	510

OR = (620 × 510) / (490 × 380) = 1.70

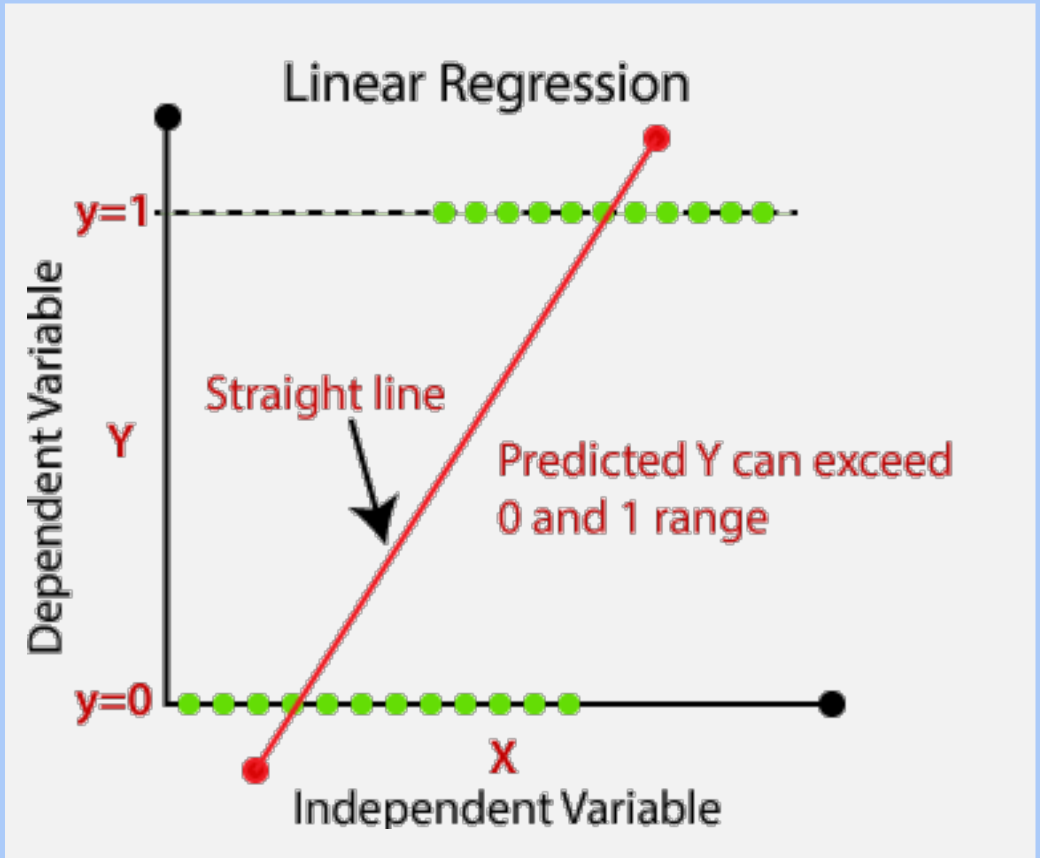
The odds ratio

- 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

- Best estimate of the slope of Y vs. X

$$\hat{\beta} = \frac{\sum_i \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sum_i \left(X_i - \overline{X}\right)^2}$$

Linear vs. logistic regression

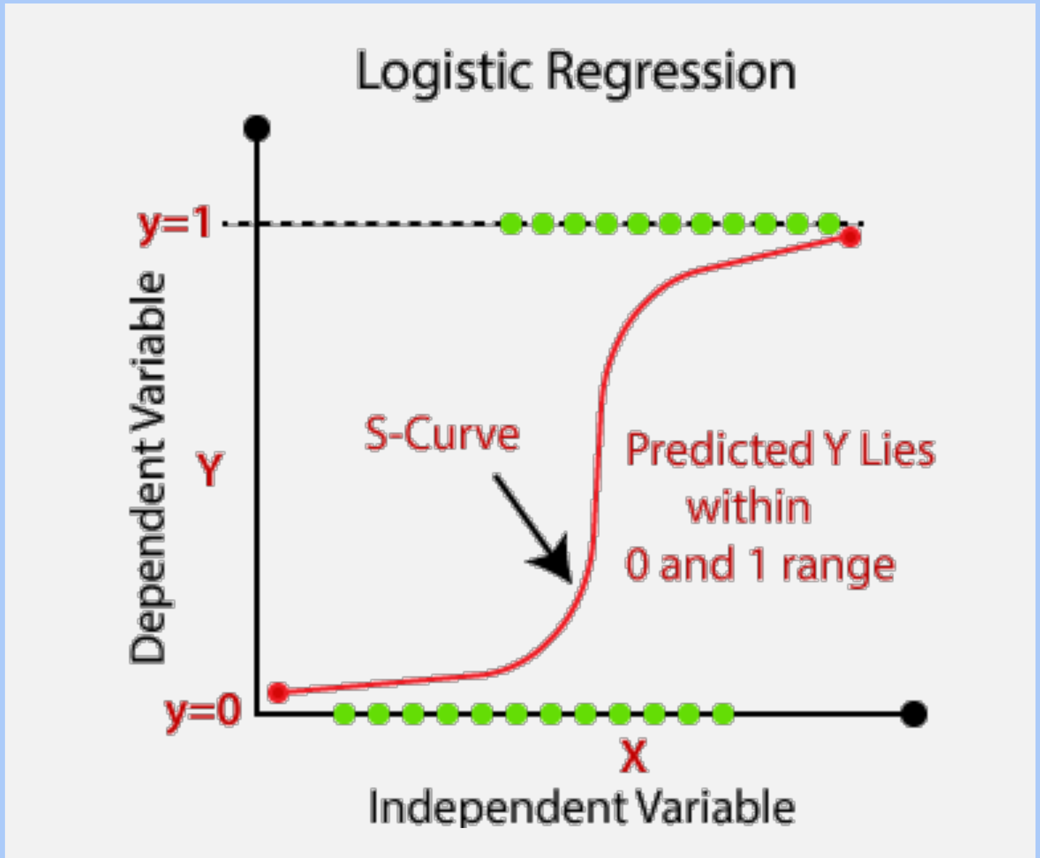
- Standard error of the estimate

$$s_{\hat{\beta}} = \sqrt{\frac{\sum_i \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2 / (N-2)}{\sum_i \left(X_i - \overline{X}\right)^2}}$$

Linear vs. logistic regression

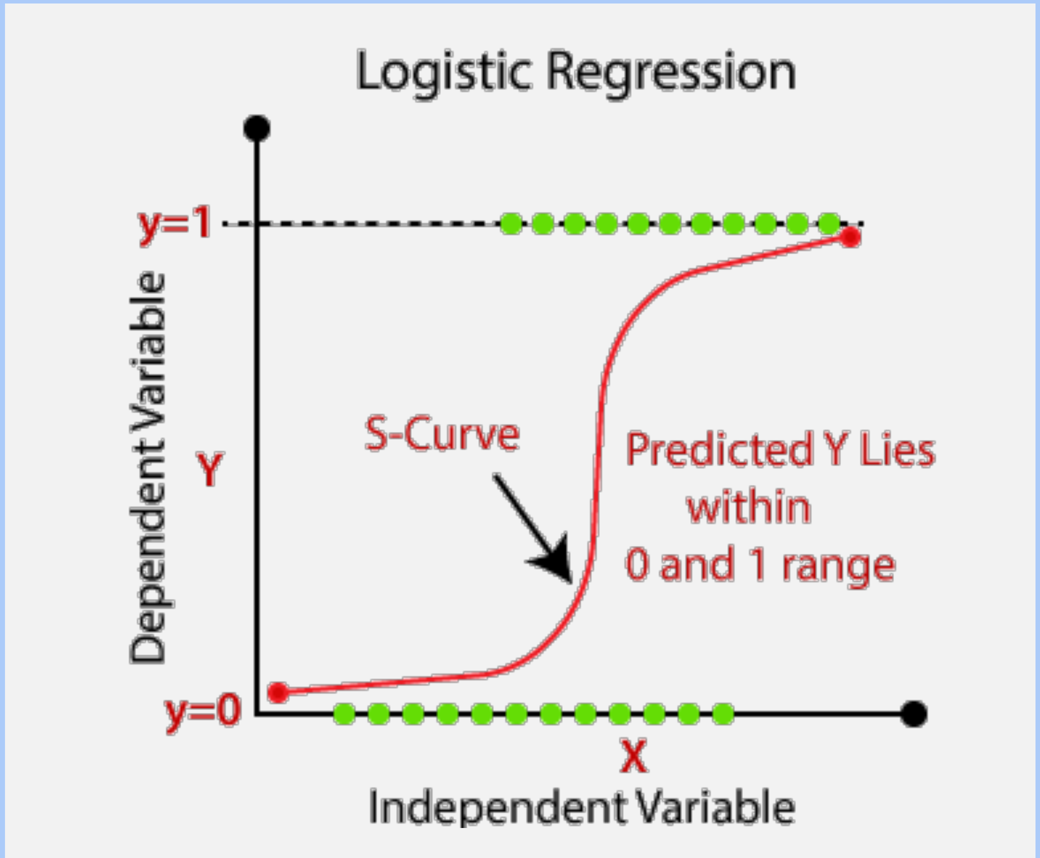
- Using the t-test, we can find out if $\hat{\beta} / s$ is statistically significantly different from 0

Linear vs. logistic regression



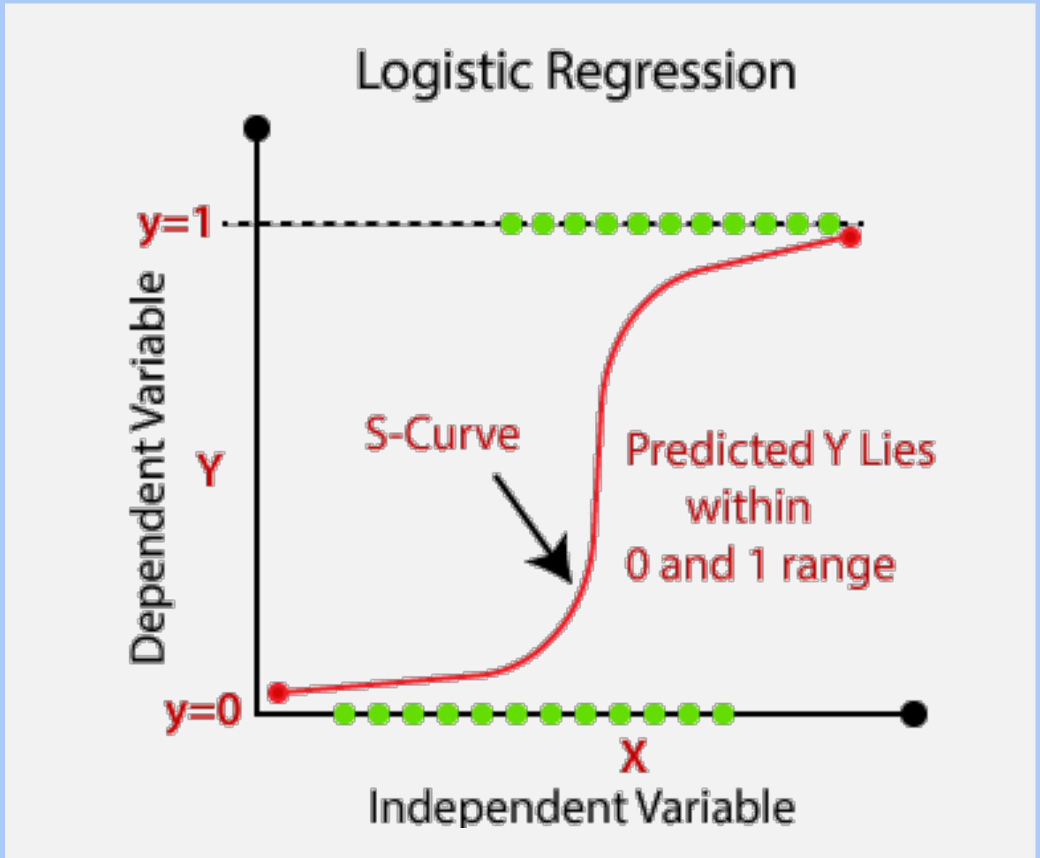
- In logistic regression, we can find the association of a **binary variate** Y with a predictor X_1 and other covariates X_2, X_3, etc.

Linear vs. logistic regression



- The sigmoid curve is an individual's probability of developing disease

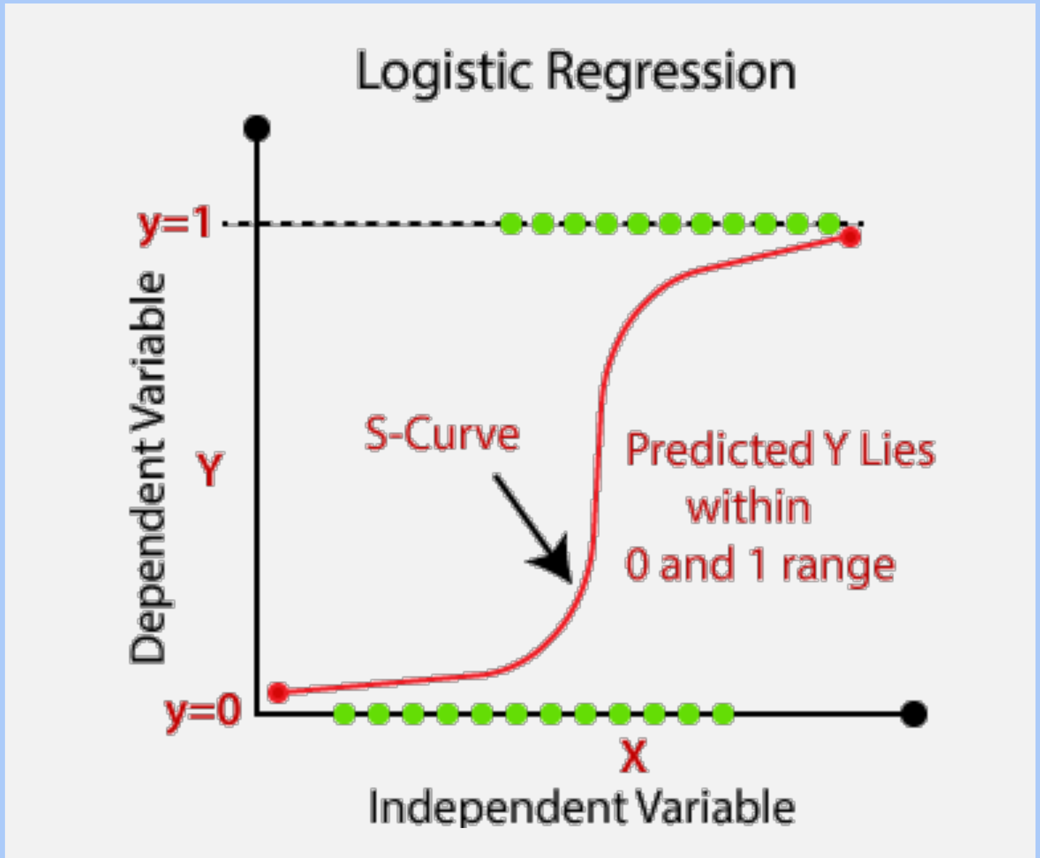
Linear vs. logistic regression



- The logistic model describes an individual's **unobserved** disease risk

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

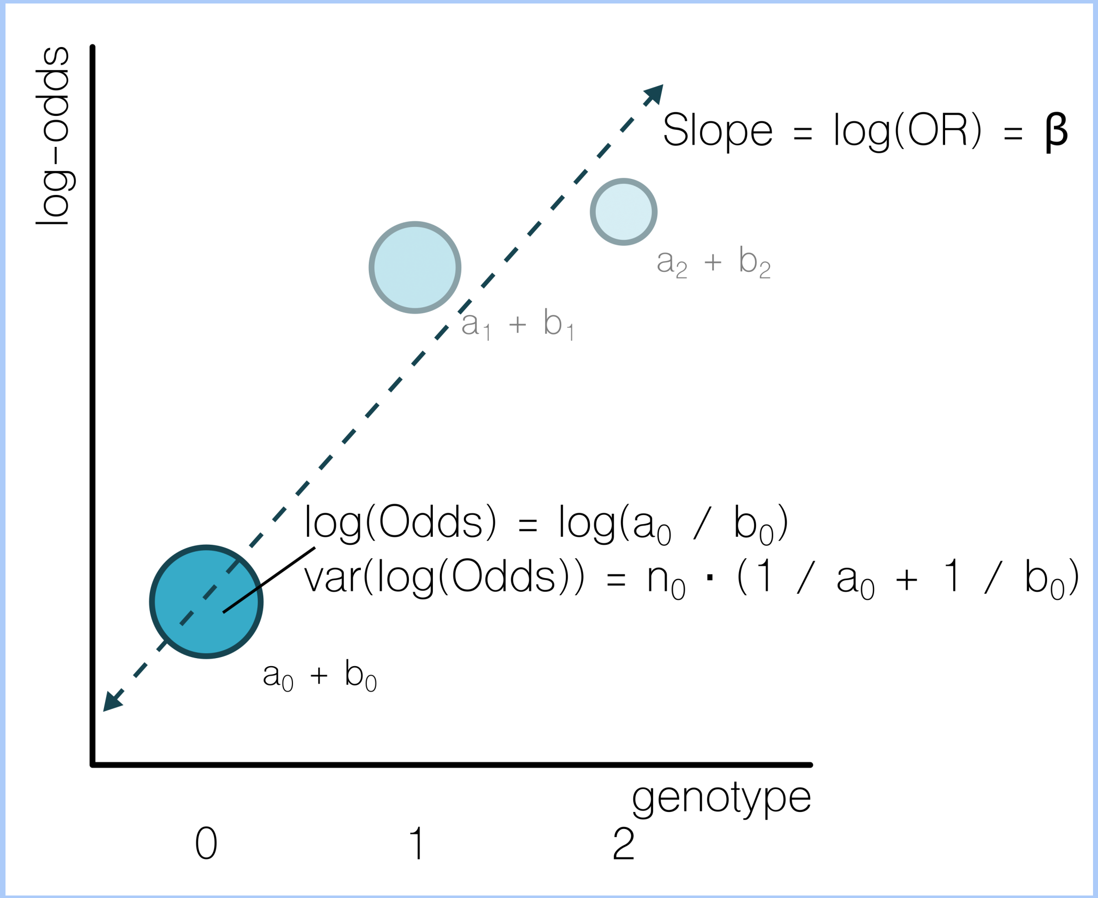
Linear vs. logistic regression



- The **logit** function turns this problem into one of linear regression

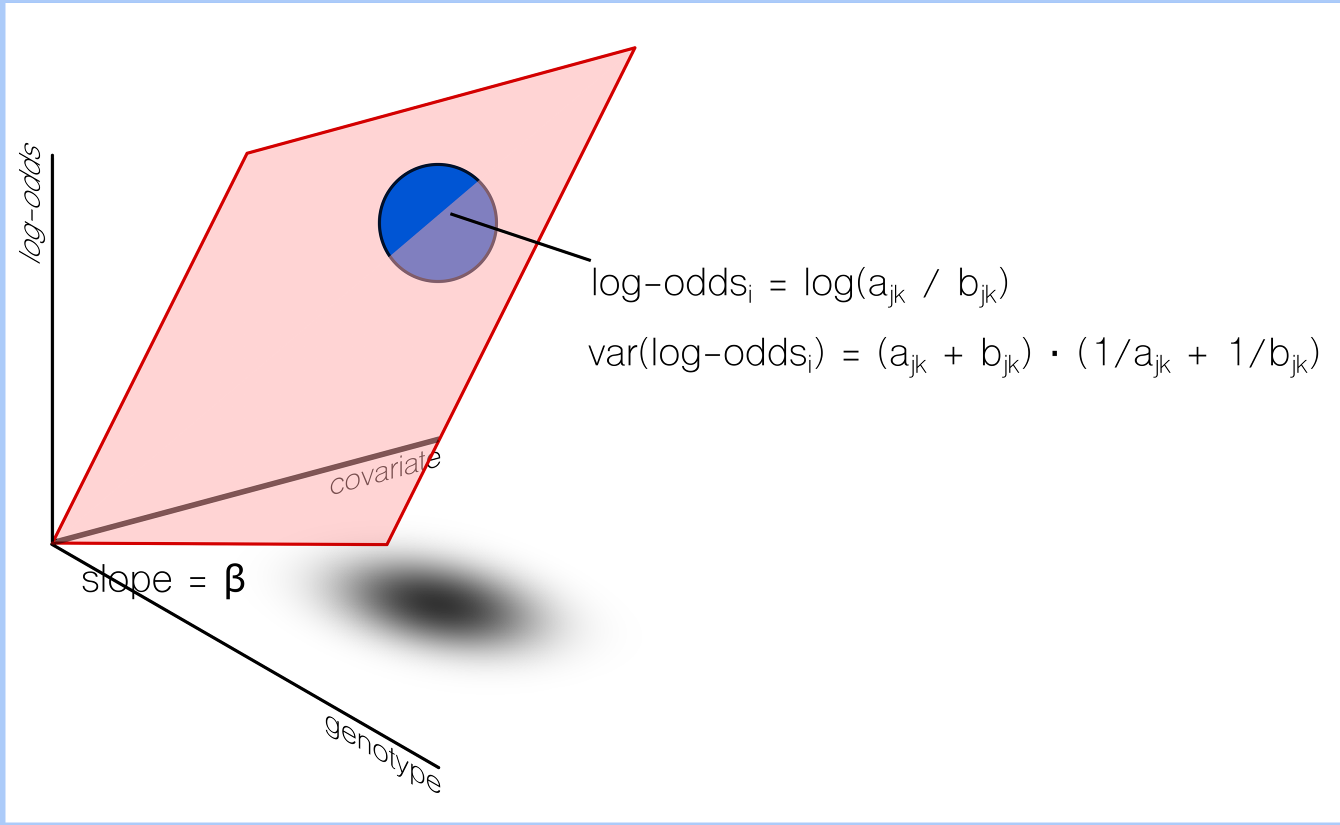
$$\log \frac{p}{1-p} = \beta_0 + X_1\beta_1$$

Linear vs. logistic regression



- Logistic regression can be thought of as 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

Linear vs. logistic regression

- The model is actually fit using the principle of **maximum-likelihood**

Linear vs. logistic regression

- Y_i is a binary indicator of disease for individual i, and p_i is the unobserved (conditional) probability of disease

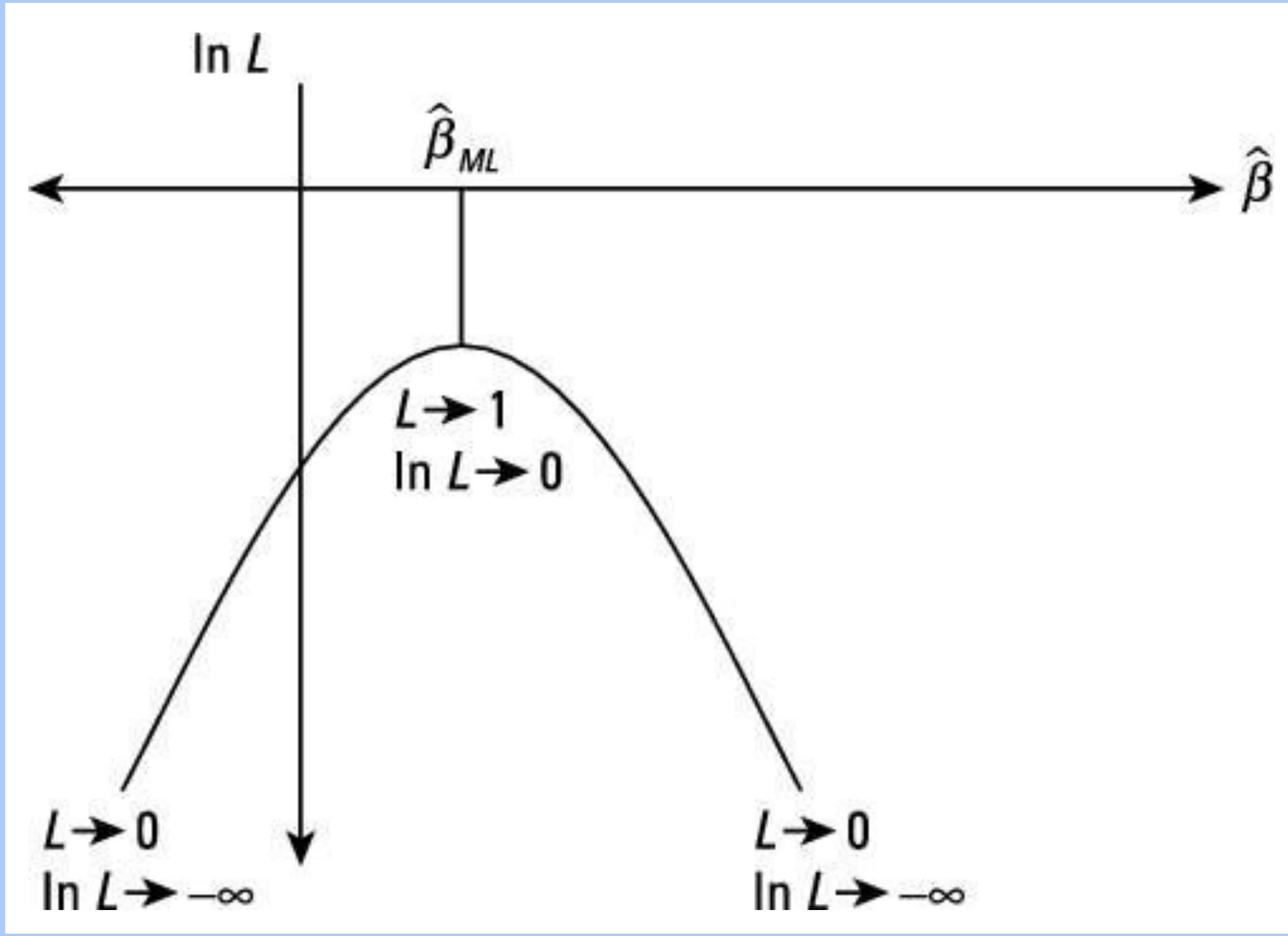
$$\mathcal{L} = \prod_i p_i^{y_i} \left(1 - p_i\right)^{1-y_i}$$

Linear vs. logistic regression

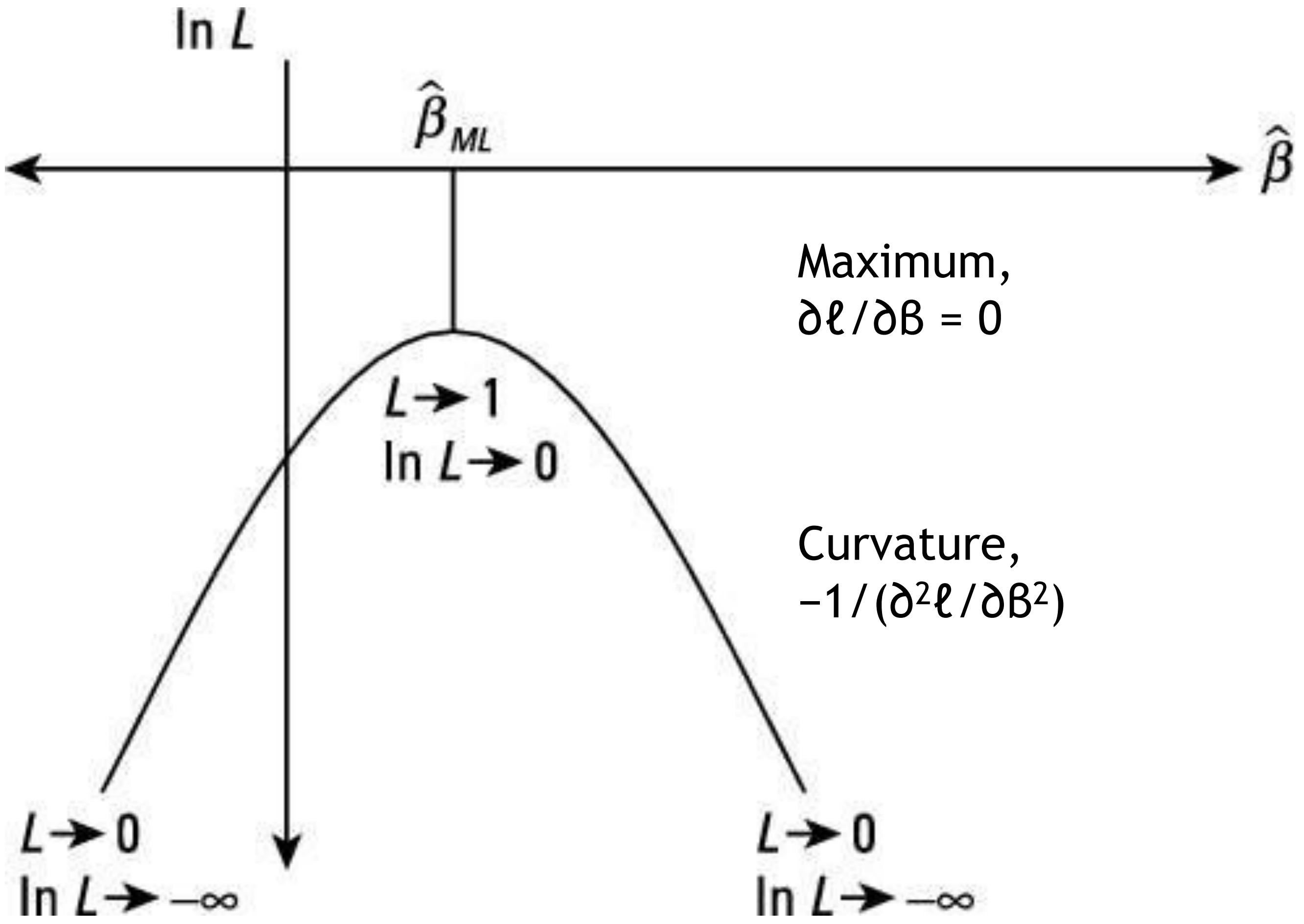
- The **log-likelihood** is a convex function of the parameters β which we can maximize

$$\ell = \sum_i y_i \left(\beta_0 + X_1\beta_1\right) + \log\left(1 + e^{\beta_0 + X_1\beta_1}\right)$$

Linear vs. logistic regression



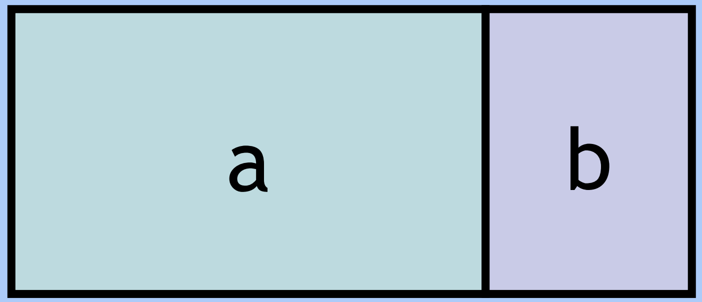
Linear vs. logistic regression



Maximum,
$\partial \ell / \partial \beta = 0$

Curvature,
$-1/(\partial^2 \ell / \partial \beta^2)$

Simulating a binary phenotype

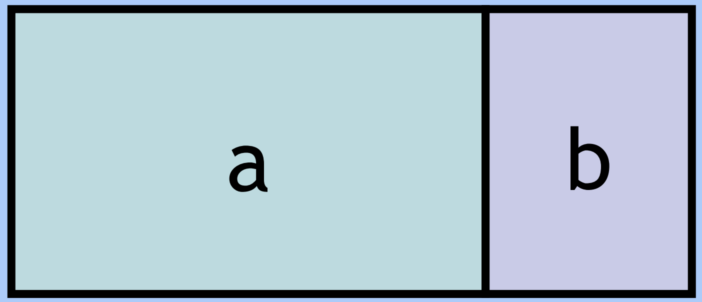


- If odds = a / b, then prob = a / (a + b) = odds / (1 + odds)

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

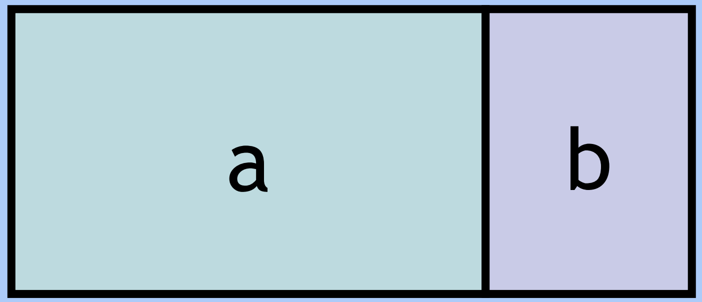


- If odds = a / b, then prob = a / (a + b) = odds / (1 + odds)

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)

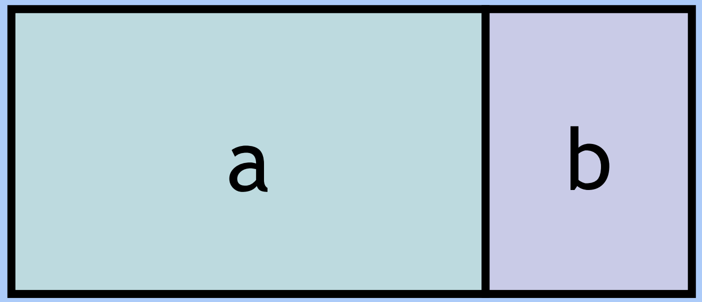


- If odds = a / b, then prob = a / (a + b) = odds / (1 + odds)

Simulating a binary phenotype

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

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

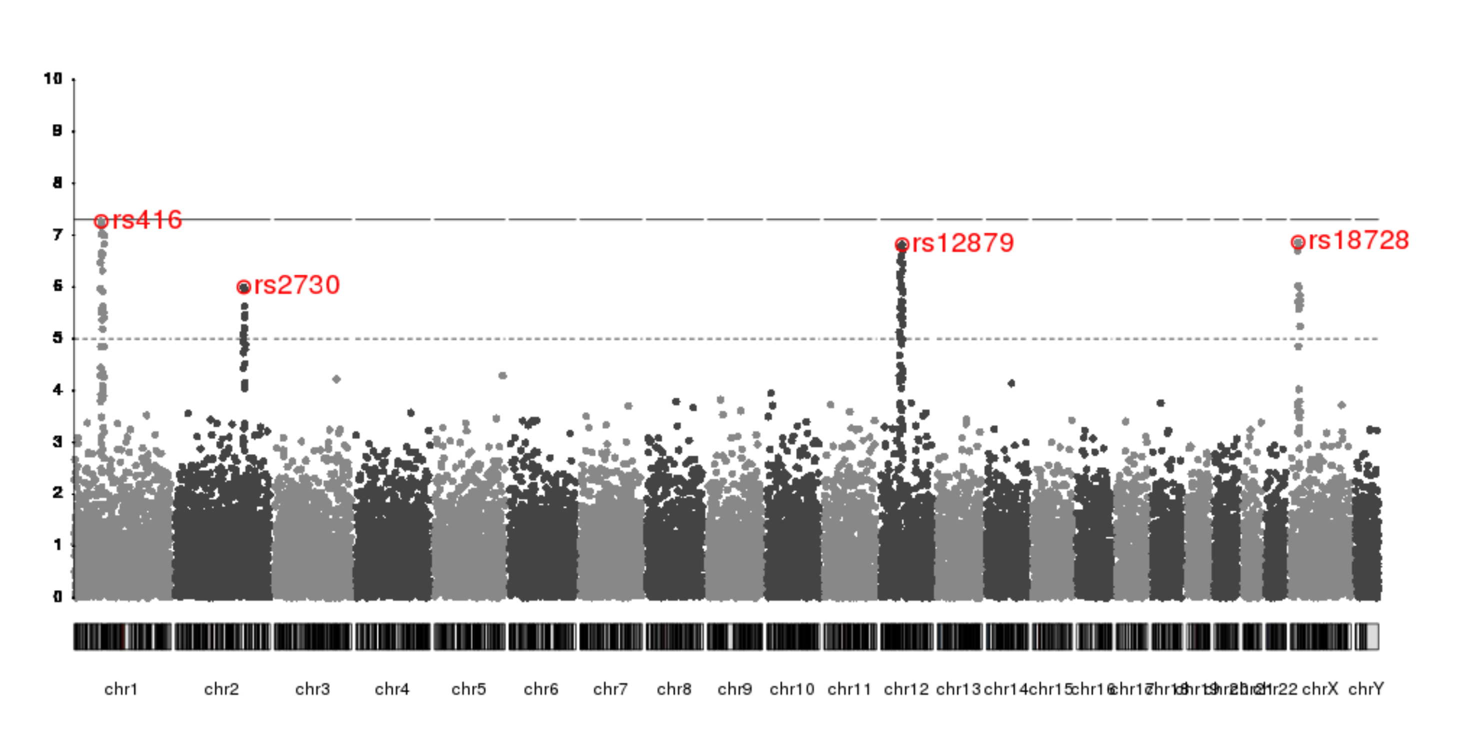


- If odds = a / b, then prob = a / (a + b) = odds / (1 + odds)

Estimating the SNP effect

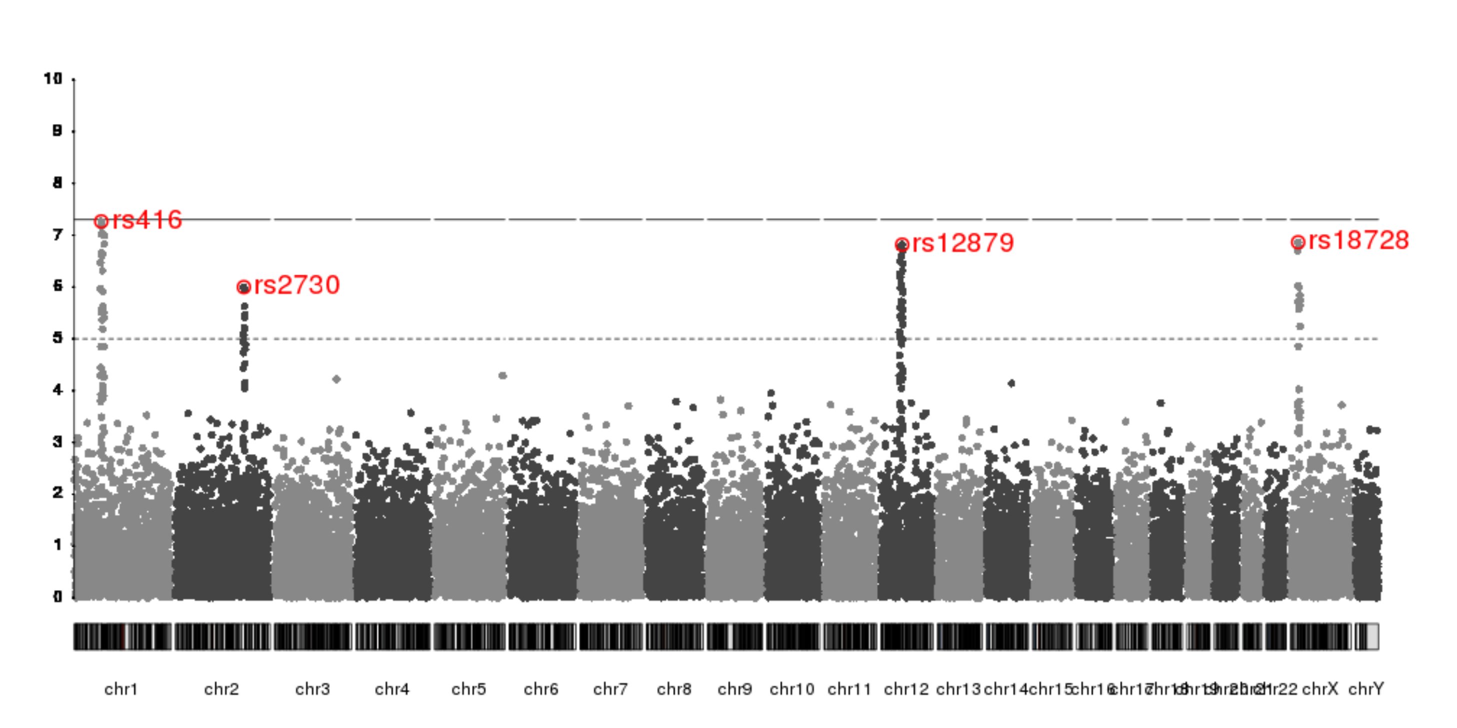
- We want to be able to **detect** the association of one SNP with disease by fitting the model $Y = \beta_0 + \beta_1 X_1 + \cdots$ and finding a slope β_1 significantly different from 0

Estimating the SNP effect



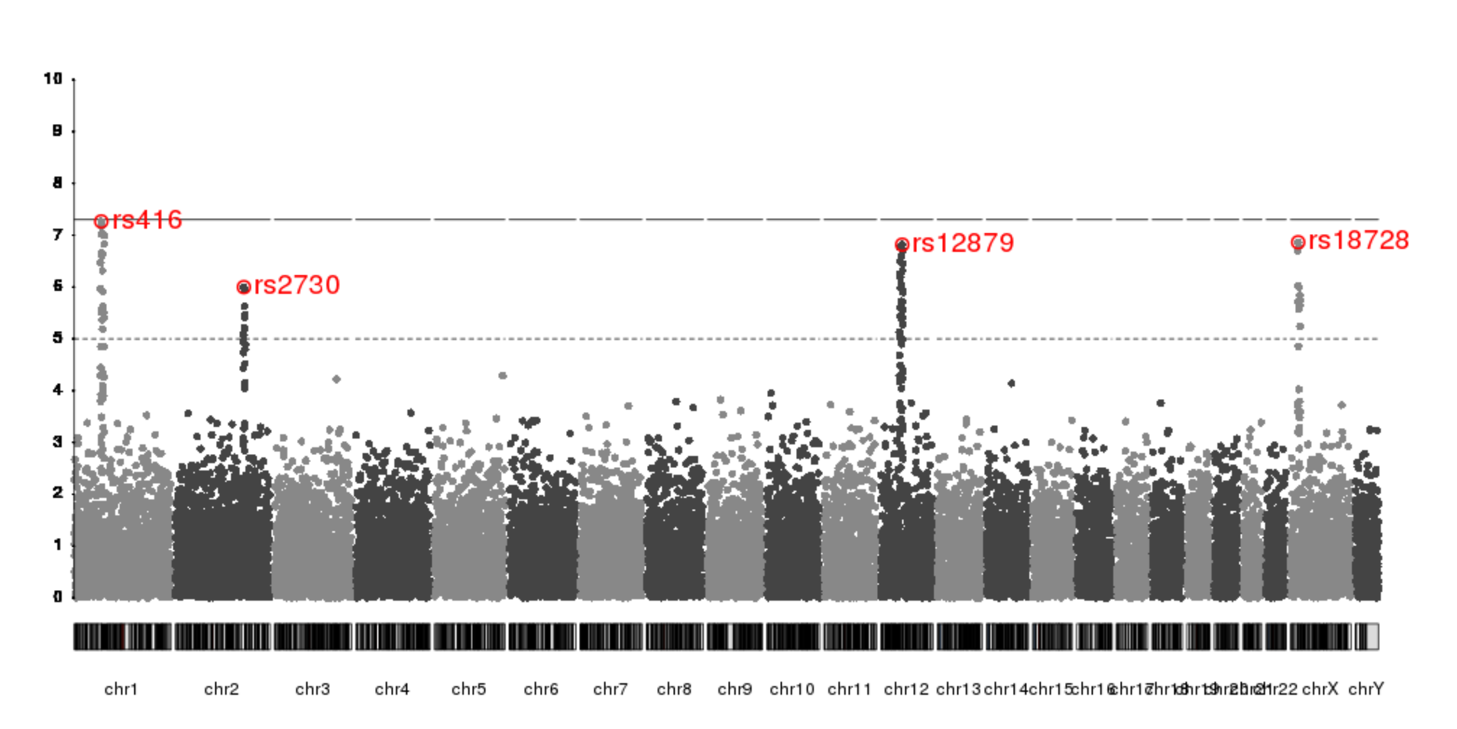
- A **Manhattan plot** gives the p-value of the log-OR estimate for each SNP

Estimating the SNP effect



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

Estimating the SNP effect

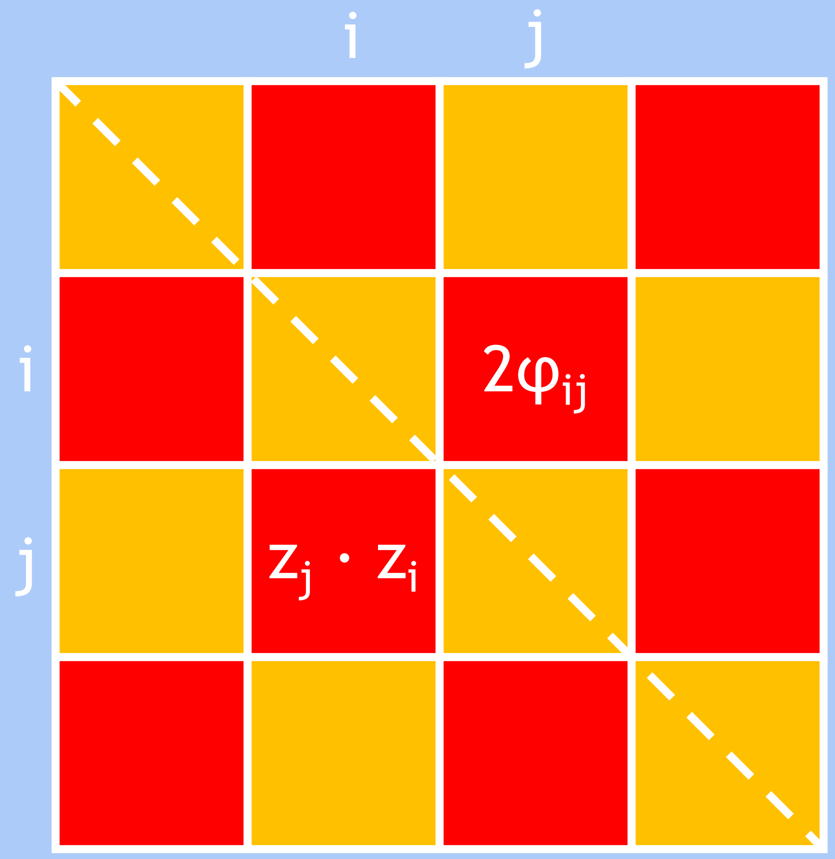


- But we can fit one SNP plus the “average” effect of all the remaining SNPs

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

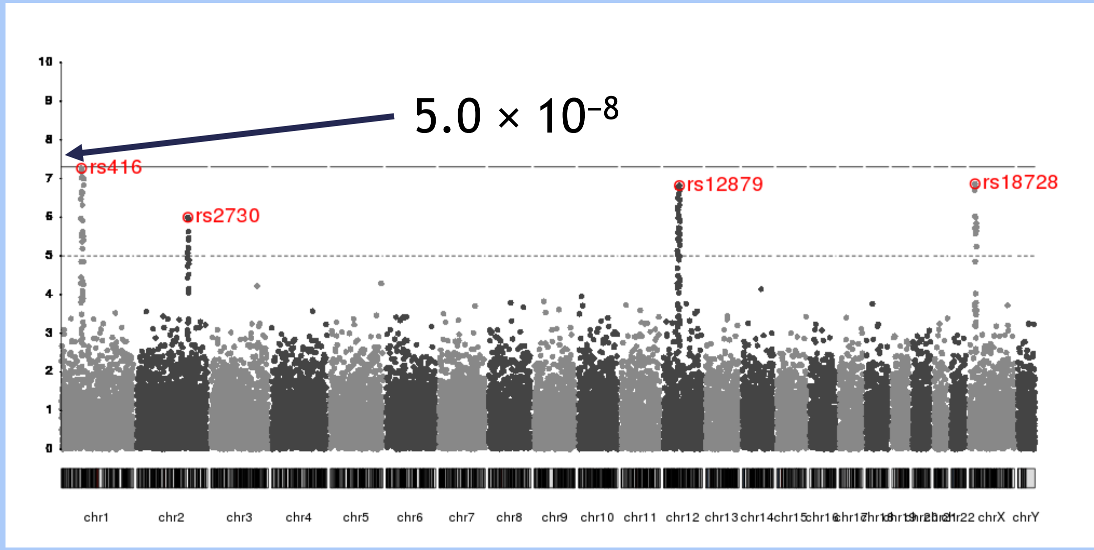
- Other covariates X commonly included in the model are age, sex, and the **first few genotype principal components** (from PC-AiR)

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

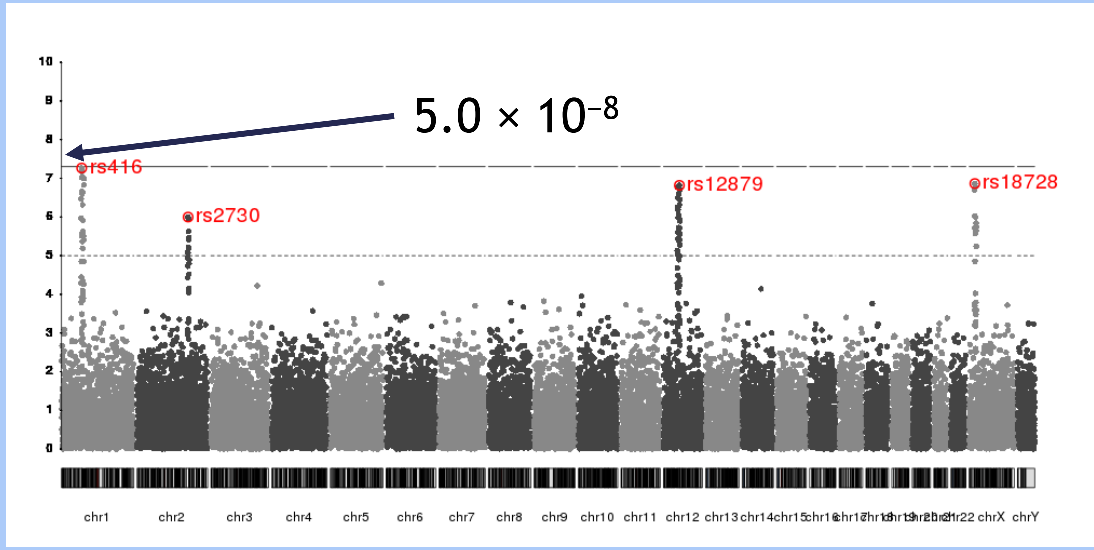
Linear mixed models

- But because of **multiple-testing**, our p-value threshold is $0.05 / 10^6$ (i.e., you perform the same test 10^6 times)



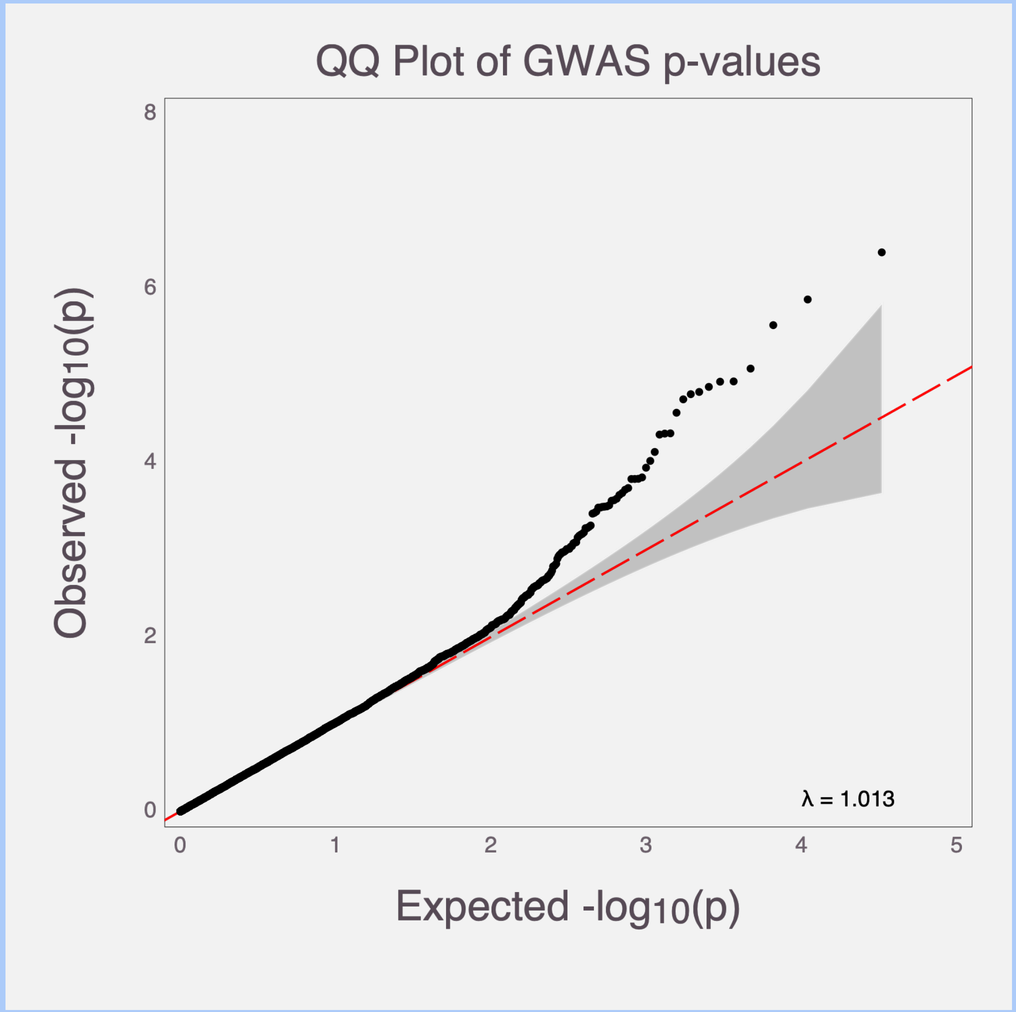
Linear mixed models

- SNPs with $p < 5.0 \times 10^{-8}$ are said to achieve **genome-wide significance**



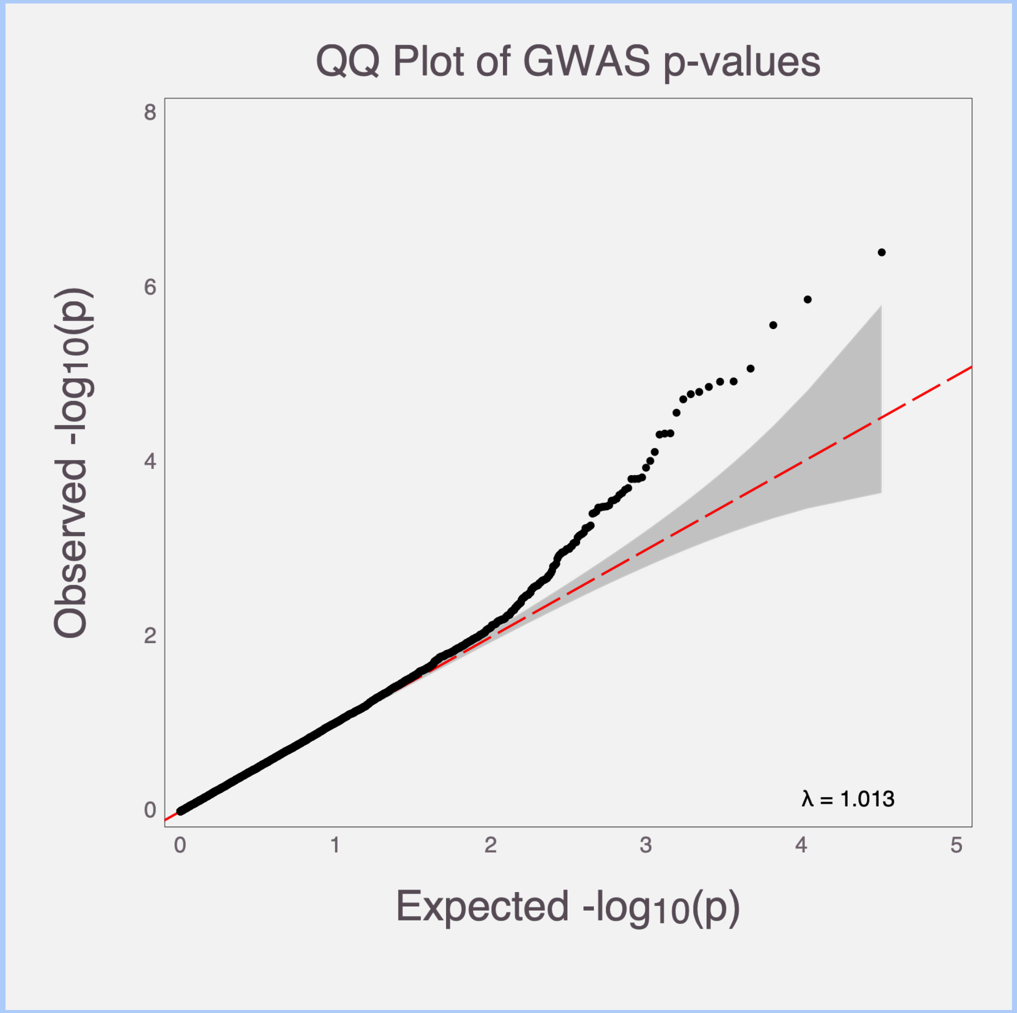
QQ plots

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



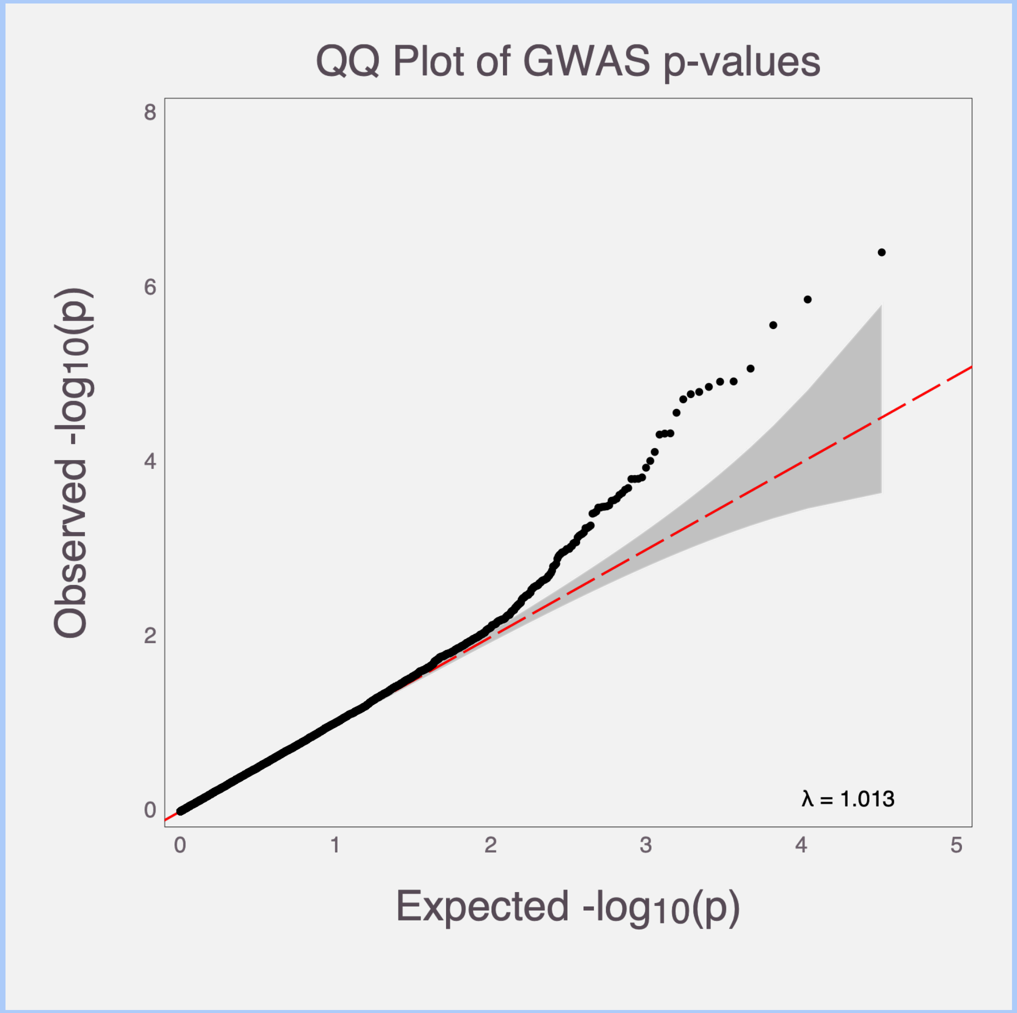
QQ plots

- The **expected** p-values for the quantiles of m SNPs are 1/m, 2/m,…, 1



QQ plots

- 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 (β's) with small p-values