GWAS (Genotype, Genome-Wide Association)

Bioinformatic tools

Corentin Maslard

October 26, 2026

Université de Neuchâtel
Faculty of Science

SCAN ME

Quick poll: pe.app/cmaslard5

One question from Session 5 before we move on.

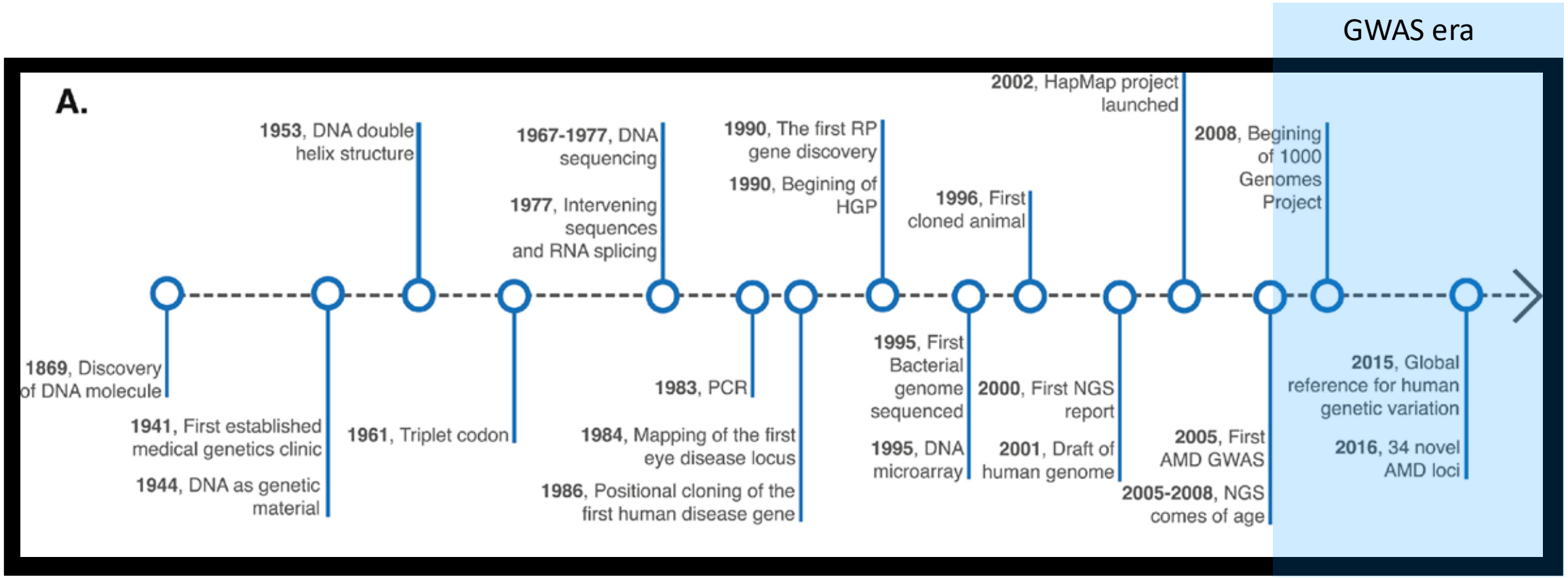
What is GWAS?

Benefits and limitations of genome-wide association studies

Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré & David Meyre

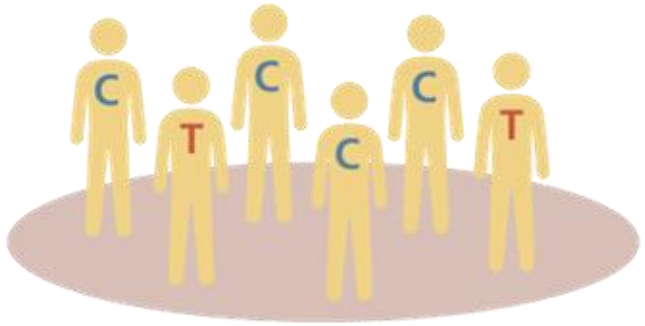
Nature Reviews Genetics **20**, 467–484 (2019) | Cite this article

GWAS: a Timeline

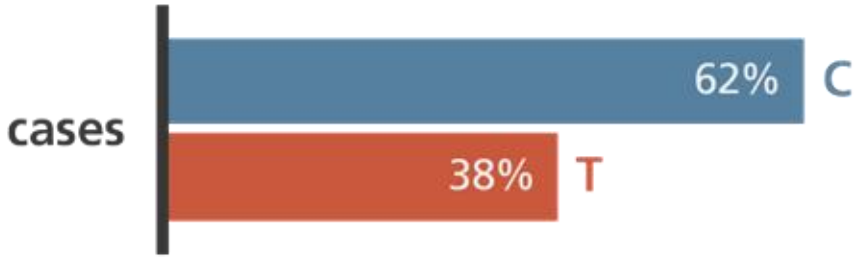


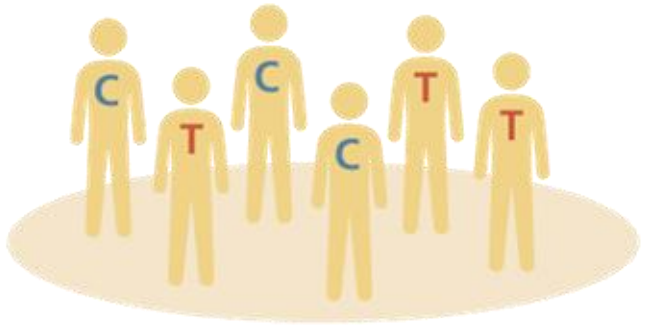
(Chaitankar et al., 2016)

The basic purpose of a GWAS



cases (n=1,000)
people with heart disease





controls (n=1,000)
people without heart disease



(yourgenome.org)

GENOME-PHENOME ASSOCIATION

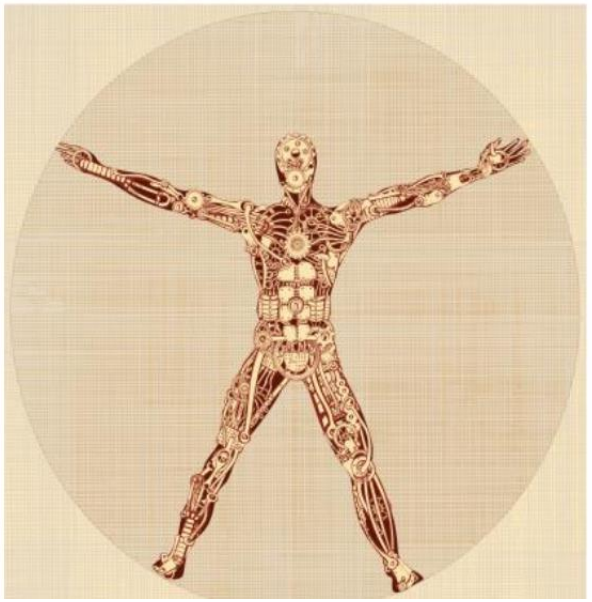
Genome

Phenome = all **phenotypes** combined

Variation in the genome between individuals stays (nearly) constant through lifetime!

”genome-wide” studies consider variation in millions of positions across the genome





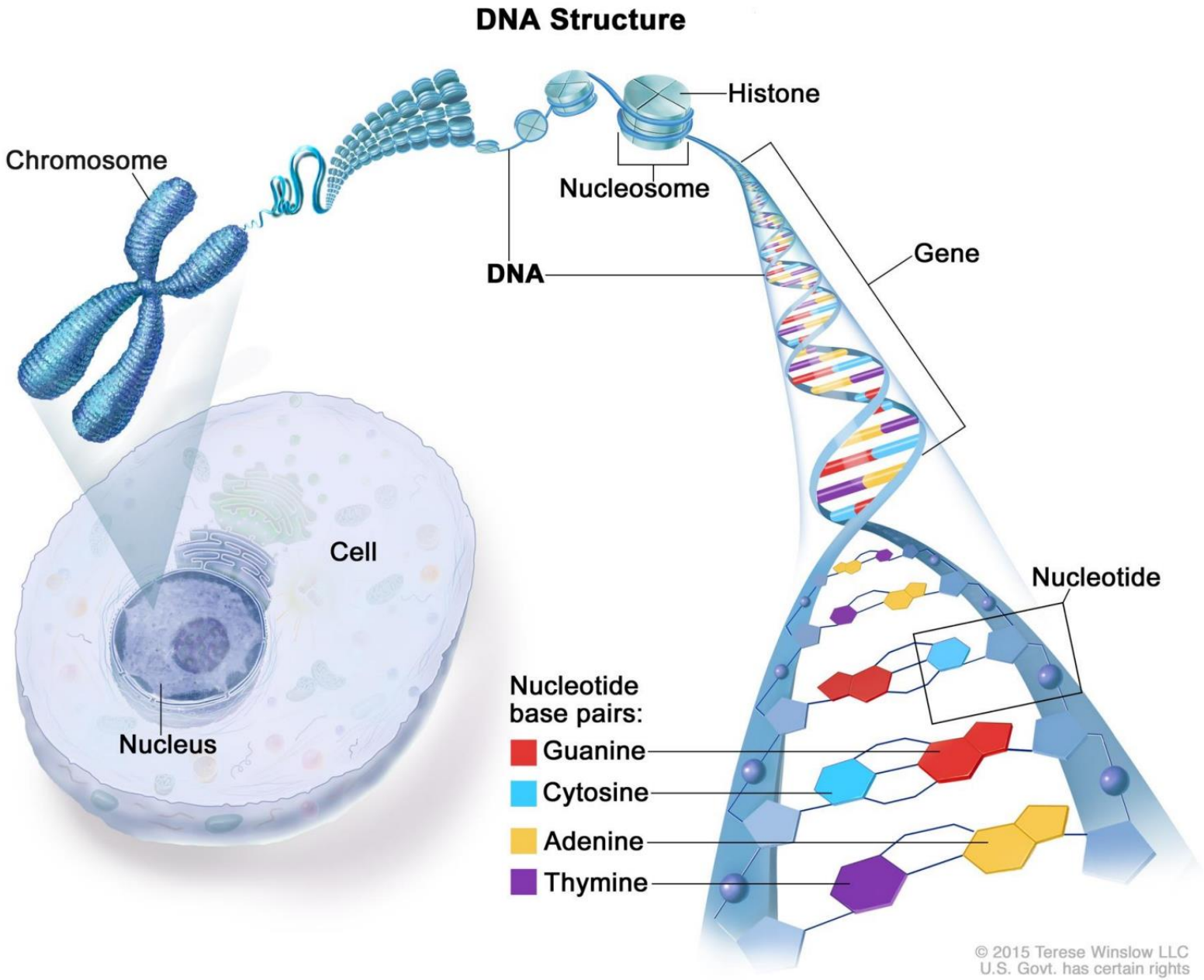
Measurable traits/ Quantitative traits (blood pressure)

Disease traits (MS-disease, diabetes)

Behavioral traits (chronotype, smoking)

Statistical association can

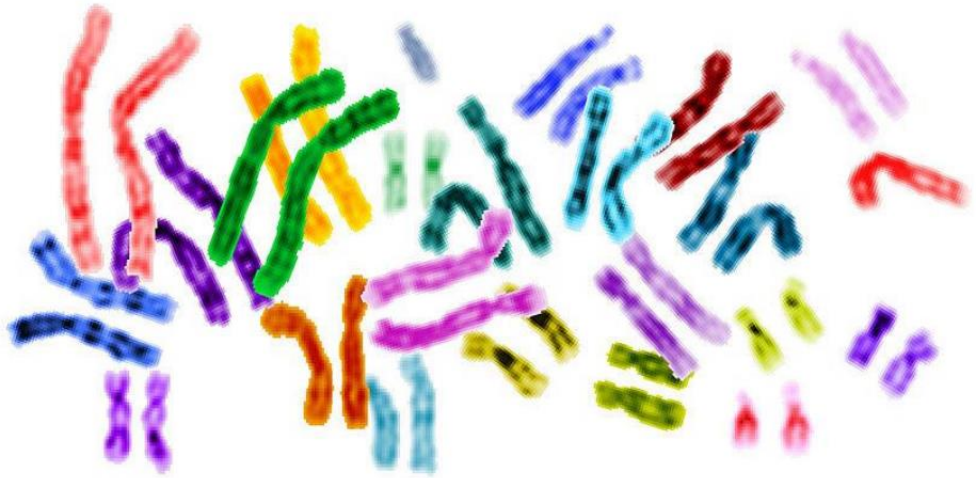
- allow predicting one from the other
- suggest causal links between the two


DNA Structure
Chromosome
Histone
Nucleosome
DNA
Gene
Cell
Nucleotide
Nucleus
Nucleotide
base pairs:
Guanine
Cytosine
Adenine
Thymine
© 2015 Terese Winslow LLC
U.S. Govt. has certain rights

Huaman genome

- Sequence of 3×10^9 letters from alphabet { A, C, G, T }

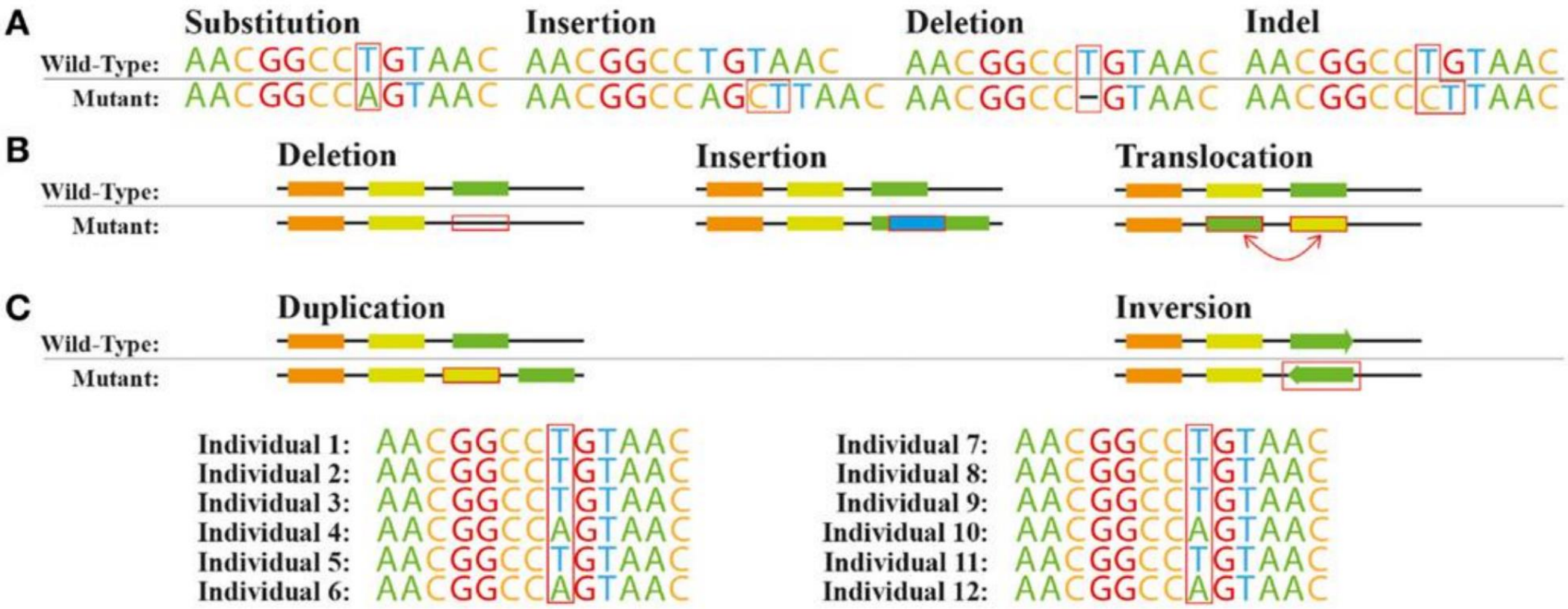
... G C G T T T A C G ...



You have two genomes: maternal and paternal.

Your genomes are physically divided into 22 pairs of autosomal chromosomes and
1 pair of sex chromosomes (males XY, females: XX)

Type of variations



Nucleotide level variation

Structural variation

SNP A/T in population

Cardoso et al. 2015
Front. Bioeng. Biotechnol., 16 February 2015 |https://doi.org/10.3389/fbioe.2015.00013

Single nucelotide polumorphisme (SNP)

On average, 1:300 positions in genome has common (MAF>1%) variation in population; these are called “SNPs”

Genomes in population

... G C G T T ... 96%

... G C T T T ... 4%

Only positive strand of genomes is shown here

This is a SNP, with alleles: G / T, minor allele frequency (MAF) = 4%

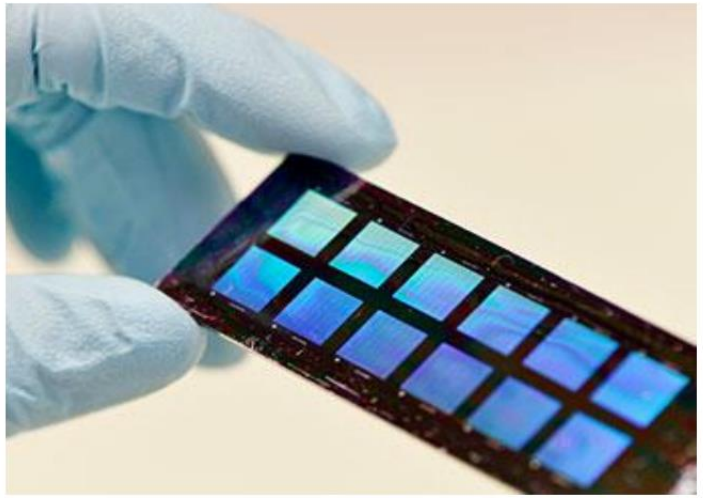
Genotypes at this SNP in population

0: GG ~ 92.1%
1: GT ~ 7.7 %
2: TT ~ 0.2 %

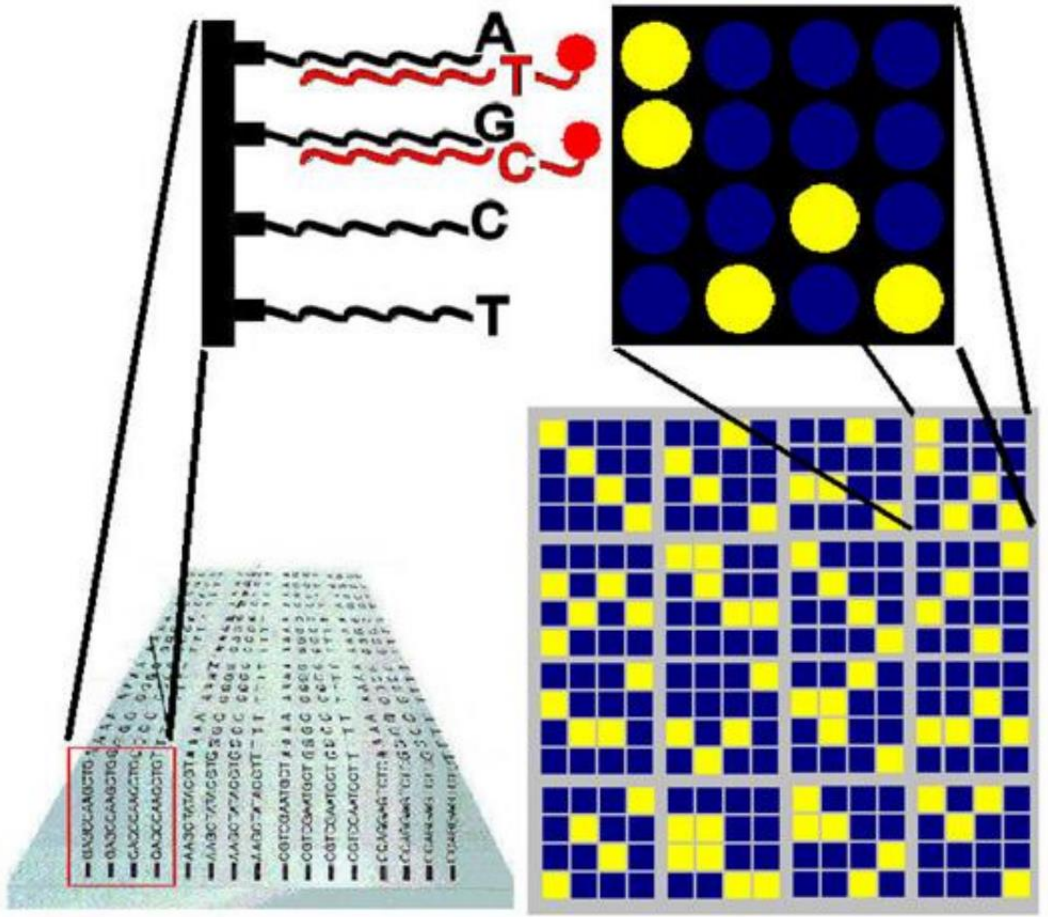
Assuming **Hardy-Weinberg equilibrium frequencies**

Reading SNP

- Human SNP array can measure 10^6 SNPs
- Cost per individual ~30 euros

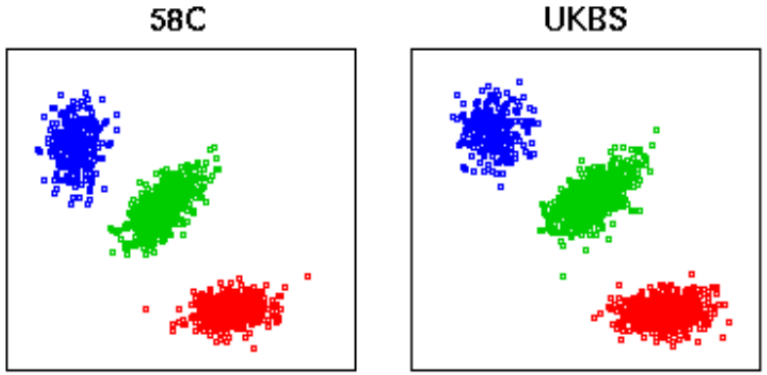


This array can genotype 12 individuals at 10^6 SNPs

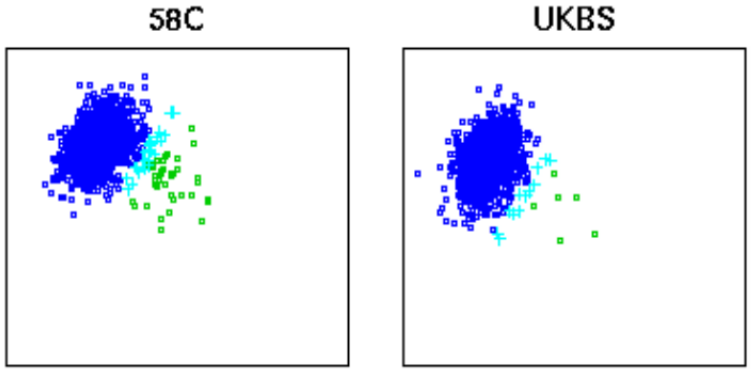


Steven M. Carr
www.mun.ca/biology/scarr/DNA_Chips.html

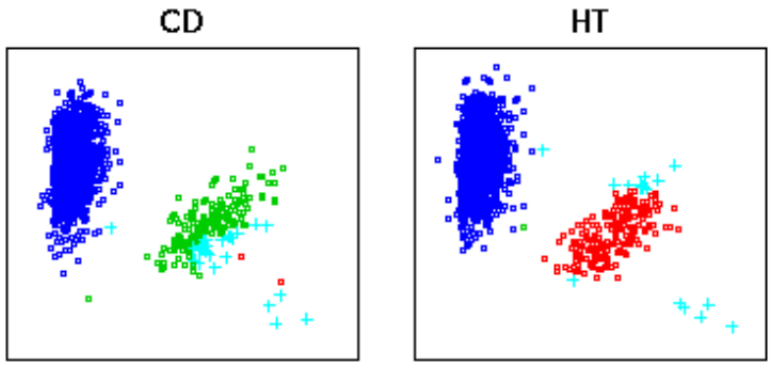
Genotype calling from snp array data



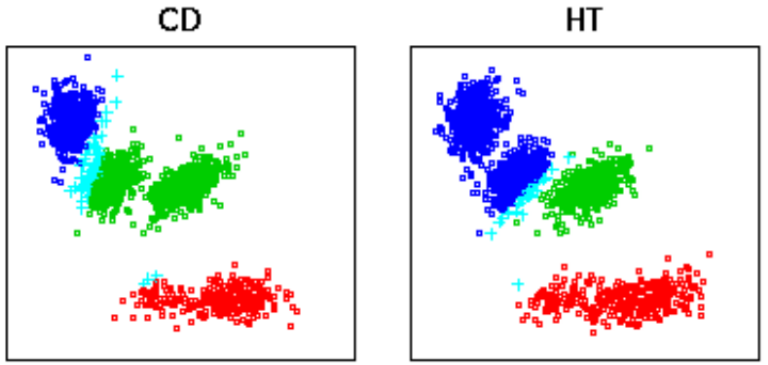
GOOD calling!



ERROR, rare variant has less than 3 clusters



ERROR, clustering algorithm performs differently in two cohorts



ERROR, structural variant has more than three genotypes

From D.Phil thesis of Damjan Vukcevic, Oxford, 2009

The calling algorithm tries to find the three genotype clusters.

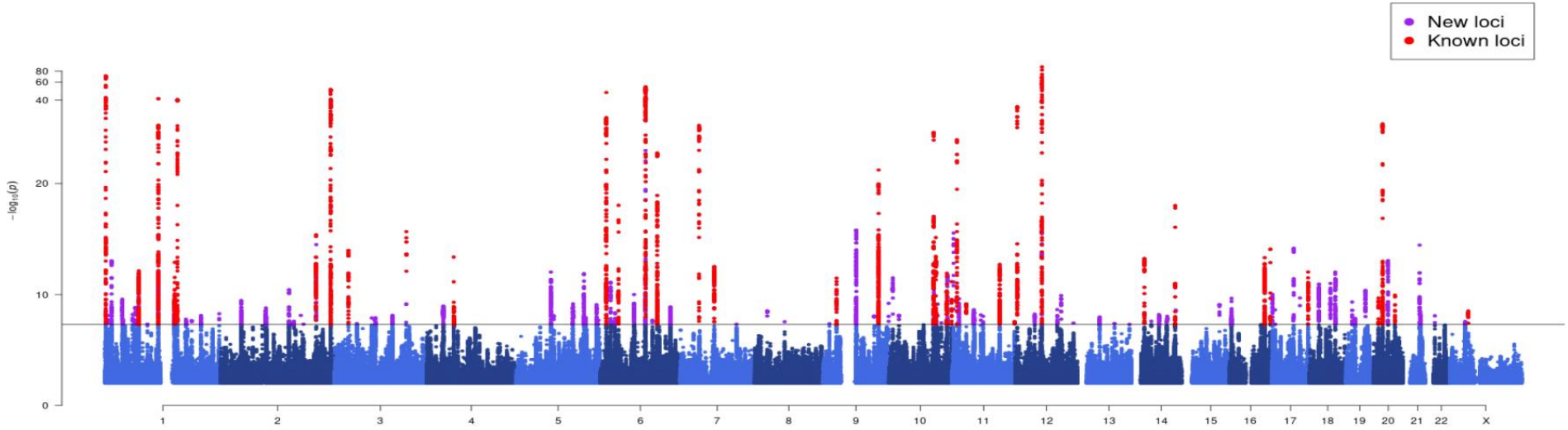
Figures shows how an algorithm has clustered individuals into three groups

Light blue means algorithm has made no call.

Bottom line errors would likely fail HWE test.

Genome-wide association study

- Statistical problem: Is genetic variation at a particular position associated with observed phenotypic variation?
  - Population cohorts with quantitative trait measurements (lipids, BMI, blood pressure)
  - Case-control studies (diseases such as schizophrenia, breast cancer or migraine)



100,000 migrain cases vs. 750,000 controls analyzed @ Helsinki / FIMM 2022

Anatomy of basic GWAS



Genotype arrays:

- Affymetrix
- Illumina

400K – 2.5M SNPs

Subjects:

- sex discordance
- missingness and heterozygosity
- duplication
- cryptic relatedness
- ancestry

Variants:

- missingness
- Hardy-Wienberg Equilibrium
- minor allele frequency

Strand alignment to reference:

- TopMed
- HRC v1.1
- 1000G phase 3

Phasing

⬇

Imputation

Imputation quality score

Variant QC as in QC #1

Dosage or best guess?

Linear or Logistic

Need subject phenotypes

Technical and biological covariates

Conditional?

p-value correction

- $p<5.0\times10^{-8}$
- sFDR

LD-based SNP clumping

Q-Q plot

Manhattan plot

LD Clumping

LocusZoom (regional) plot

(Recommend: *Anderson et al., 2010, Nature Protocols*)

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X_1$$

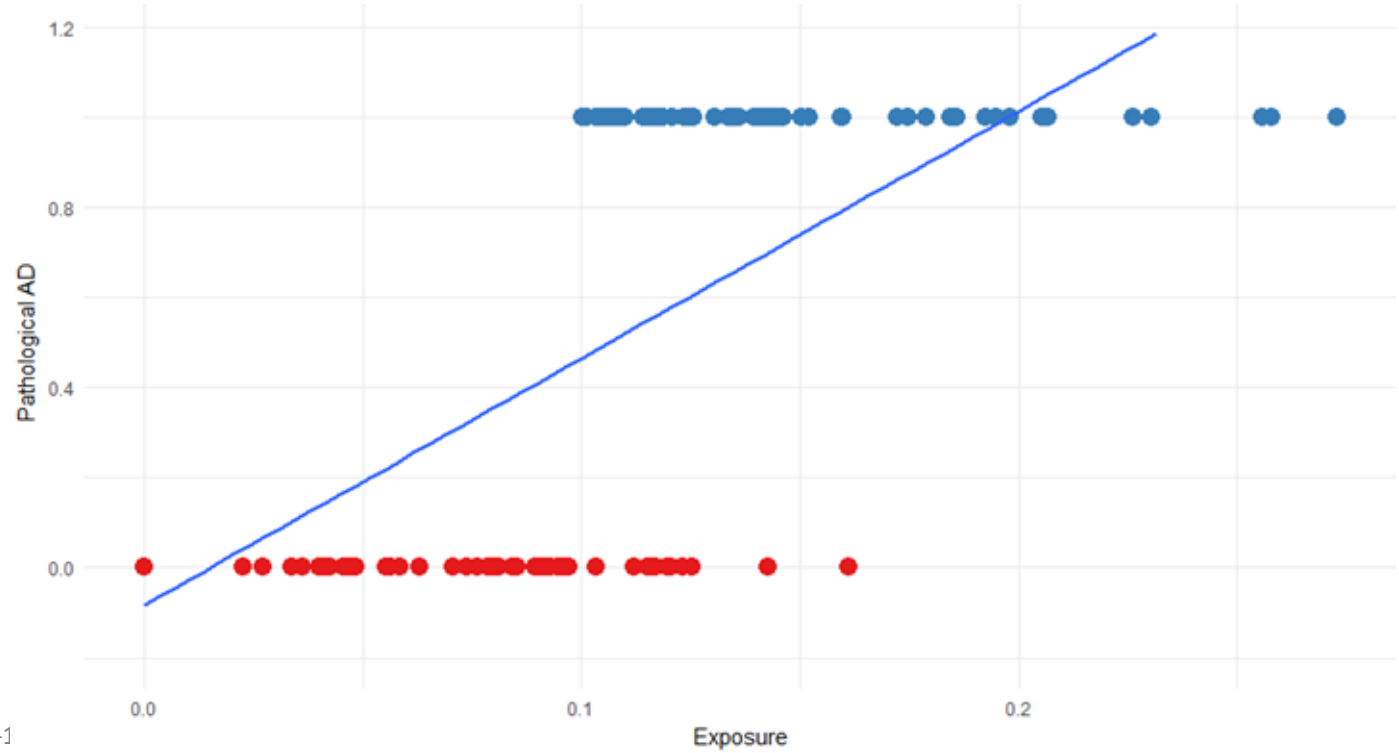
- Remember that a GWAS is just a collection of millions of individual regression tests, performed independently of each other.

- Key parameters are extracted **per SNP:**
  - Beta coefficient for effect on outcome
  - Standard error of the effect (sometimes C.I.)
  - P-value of the effect

Binary outcome - Logistic Regression

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1$$

- Q: How do you model a binary variable with a linear function?

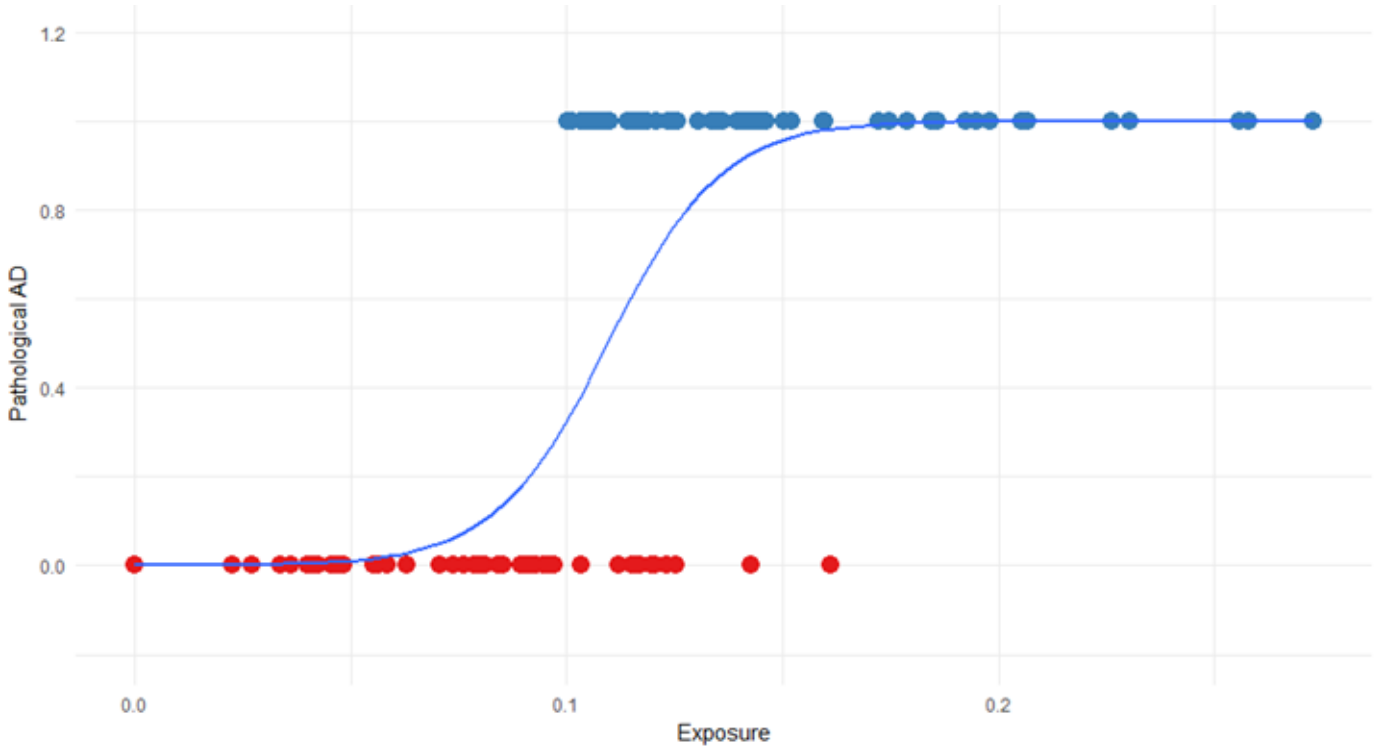


Remember:
Binary variables are considered on a scale of probability [0-1]

Binary outcome - Logistic Regression

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1$$

- A: Transformation of the Y variable – in this case with a “logit”



This transformation crucially keeps the estimated values of Y between 0 and 1.

The effect coefficients are interpreted differently since:

A) Scale is no longer linear

B) The outcome is a probability

Regression for SNPS?

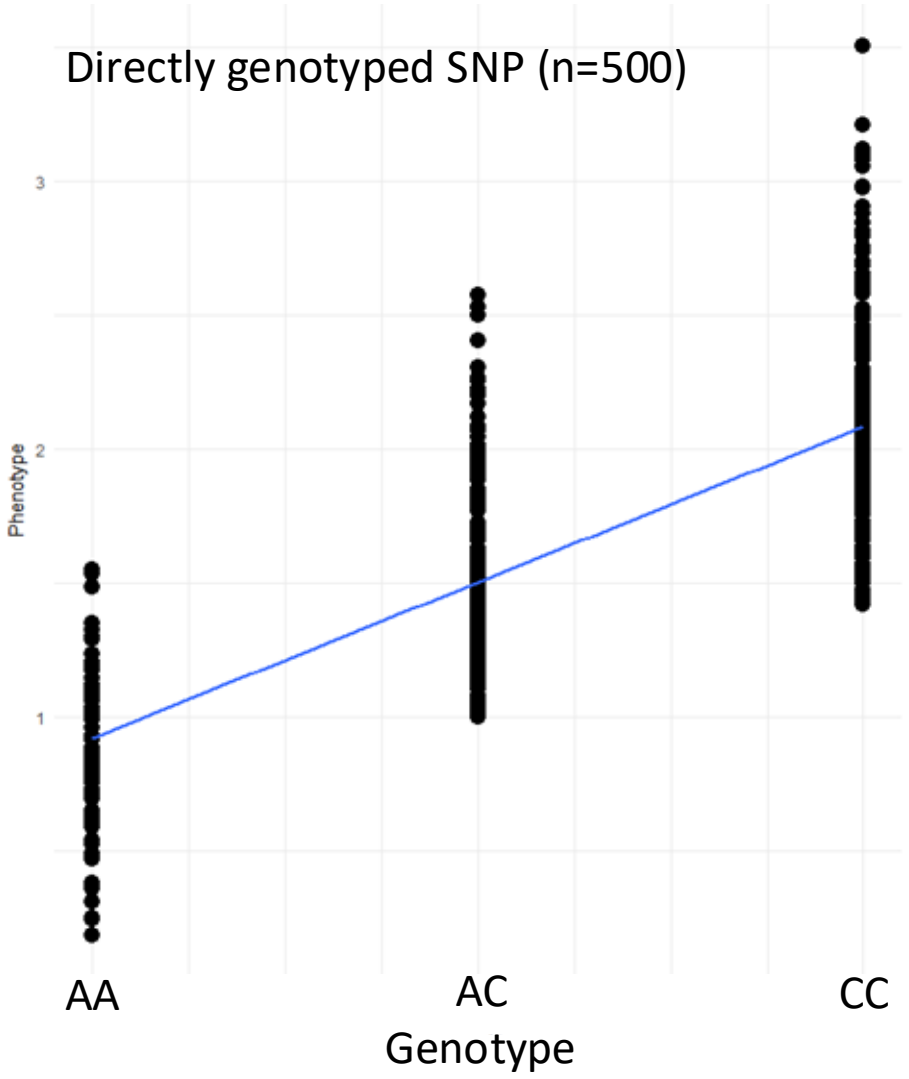
- Genotype is a discrete variable in diploid organisms (humans / some plant)
  - e.g. SNP^{AC}: A/A, A/C, C/C

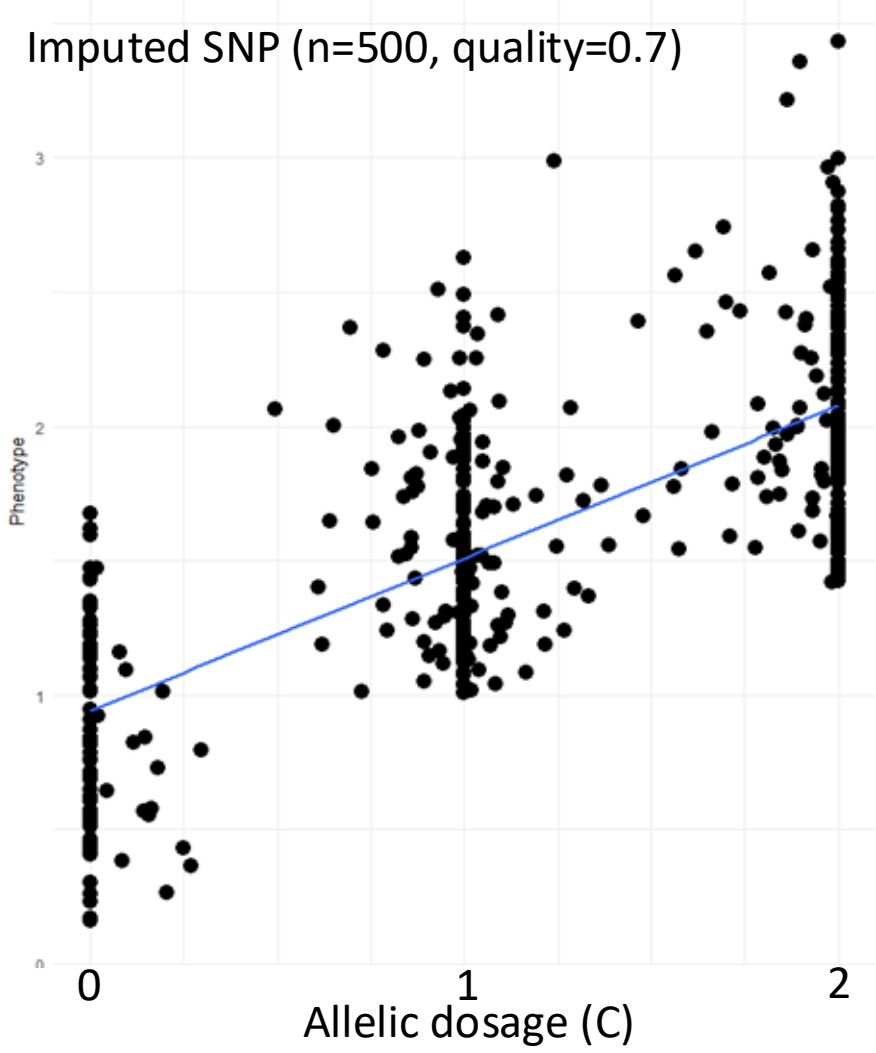
- However, the process of **imputing unobserved SNPs is imperfect**. Genotype probabilities are calculated given evidence provided in the chosen reference panel (posterior probabilities)

- Therefore we model genotype as a continuous variable (dosage), meaning the predicted ‘dosage’ of an allele at each SNP. The number ranges from 0 - 2

- If the above SNP^{AC} was **directly genotyped**, it’s C-allele dosages would be 0 (A/A), 1 (A/C), and 2 (C/C)

Regression for SNPS?





GWAS Key Considerations

- **Sample size**
  - Key influencer of statistical power in GWAS – that is, the ability of your study to detect genetic effects if they truly exist.

- **Minor allele frequency** (depends on sample size)
  - A SNP with **MAF=0.4** means that you would expect 160/1000 individuals to have the rare homozygous genotype
  - A SNP with **MAF=0.01** means that you would expect 0.1/1000 individuals to have the rare homozygous, and only ~20/1000 to have the heterozygous genotype.

- **Phenotype distribution**
  - For linear regression, normality of residuals is assumed
  - For logistic regression, you must have enough cases *and* controls (i.e. 1/0)

Challenges & Developments

Stats: ordinary regression methods assume i.i.d.

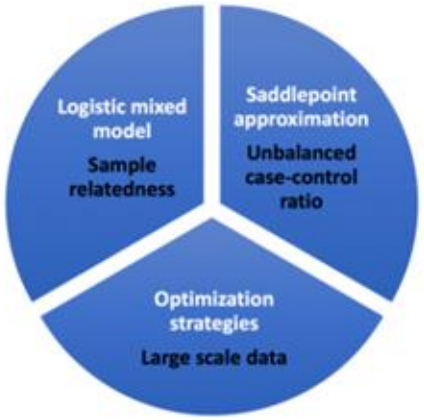
independent and **i**dentically **d**istributed

Problems with large, population-based samples

- **Shared ancestry**
  - geography, recent migration, historical social/religious groups

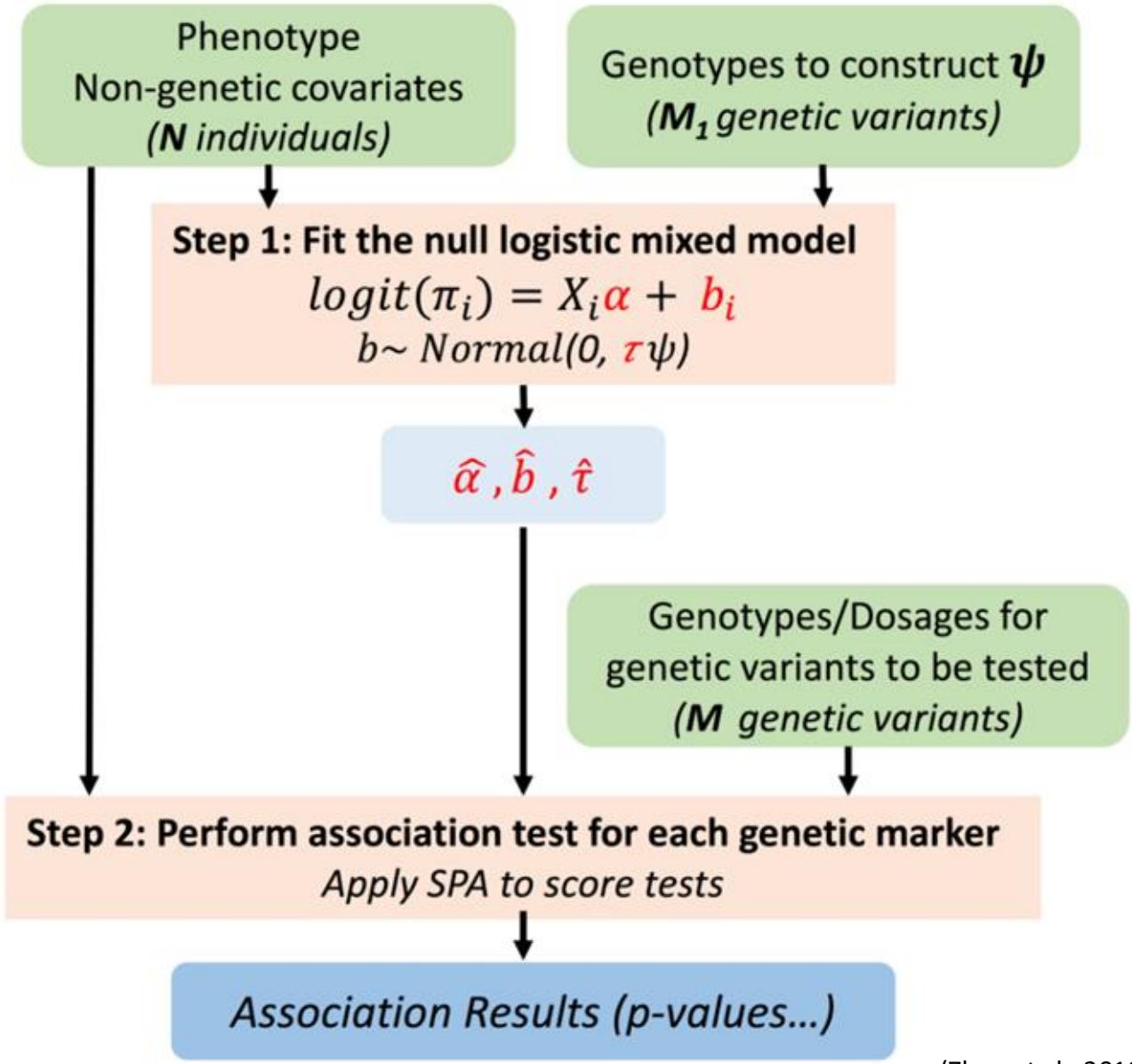
- **Cryptic Relatedness**
  - missed close relations between individuals in your sample

- **Major imbalances in case-control sample sizes**



SAIGE

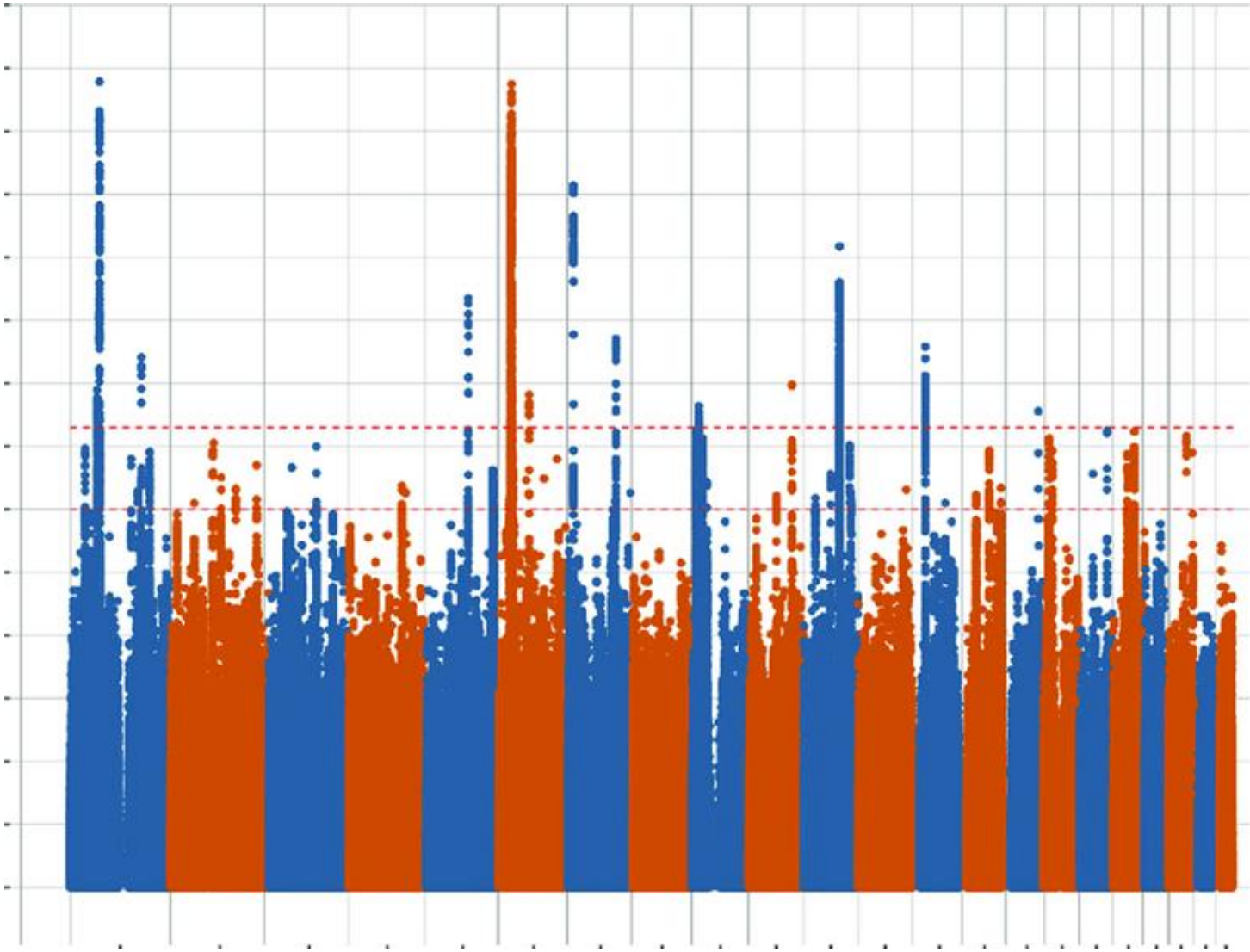
Scalable and **A**ccurate **I**mplementation of **GE**neralized mixed model



(Zhou et al., 2018, Nature Genetics)

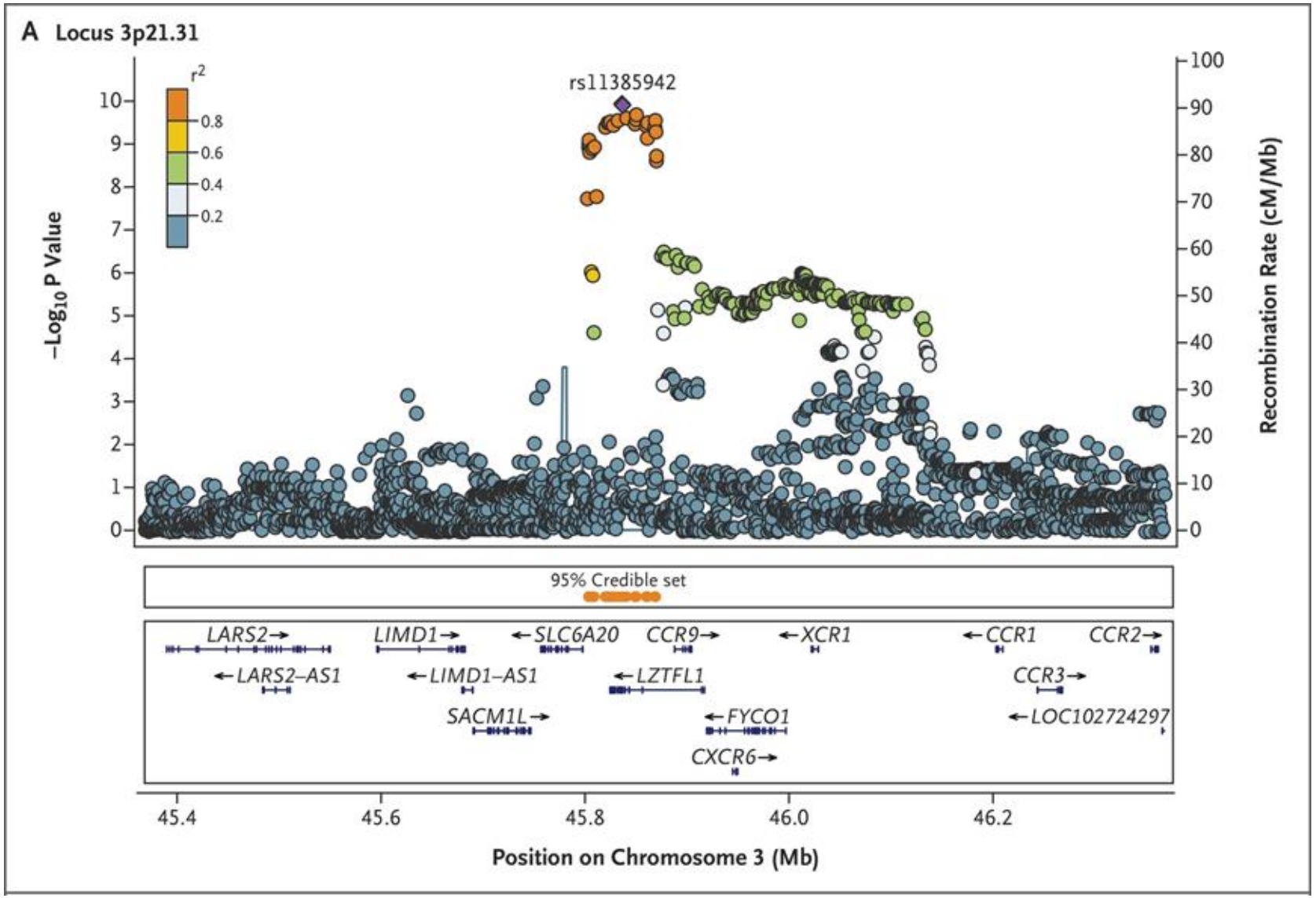
Linkage Disequilibrium

Notice peaks and valleys



Linkage Disequilibrium

Zoom in...

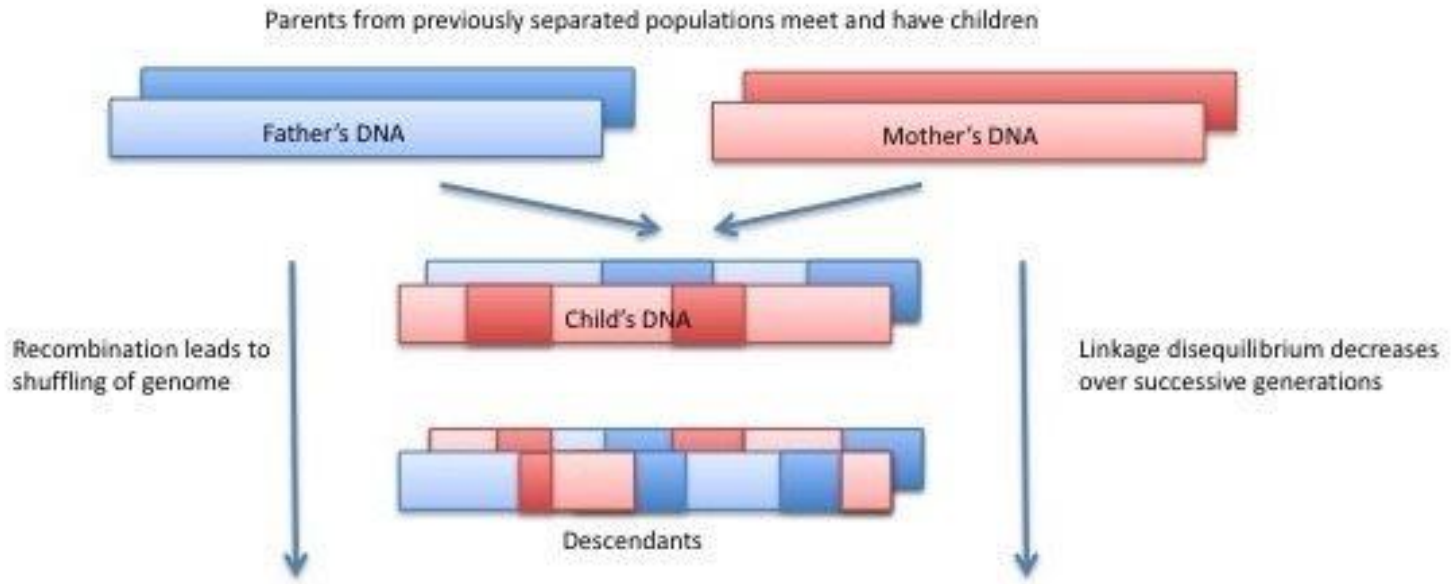


Linkage Disequilibrium

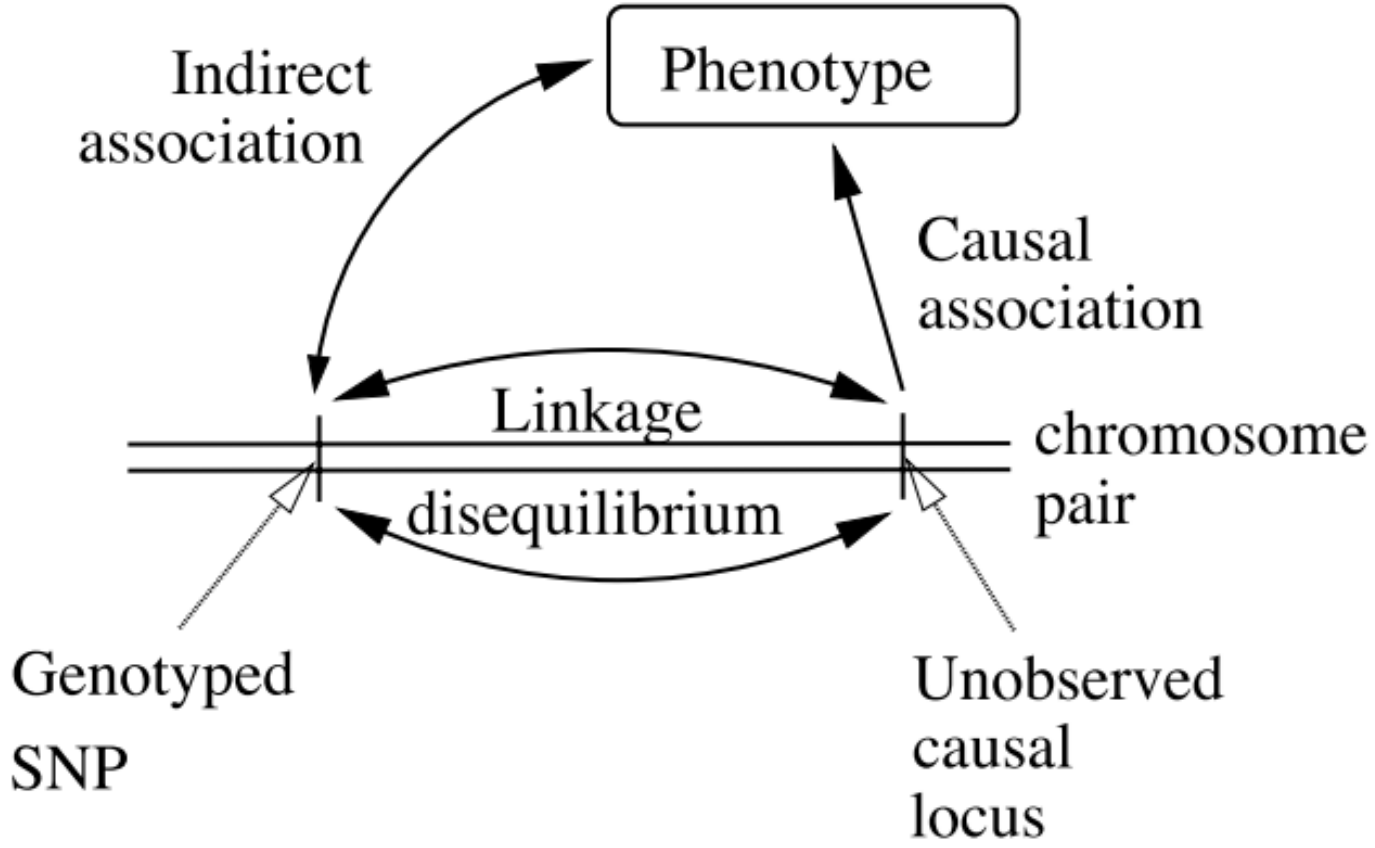
SNPs are passed on non-randomly in blocks known as haplotype blocks.

We can estimate the amount of disequilibrium (the non-randomness) by evaluating correlations of alleles in large populations.

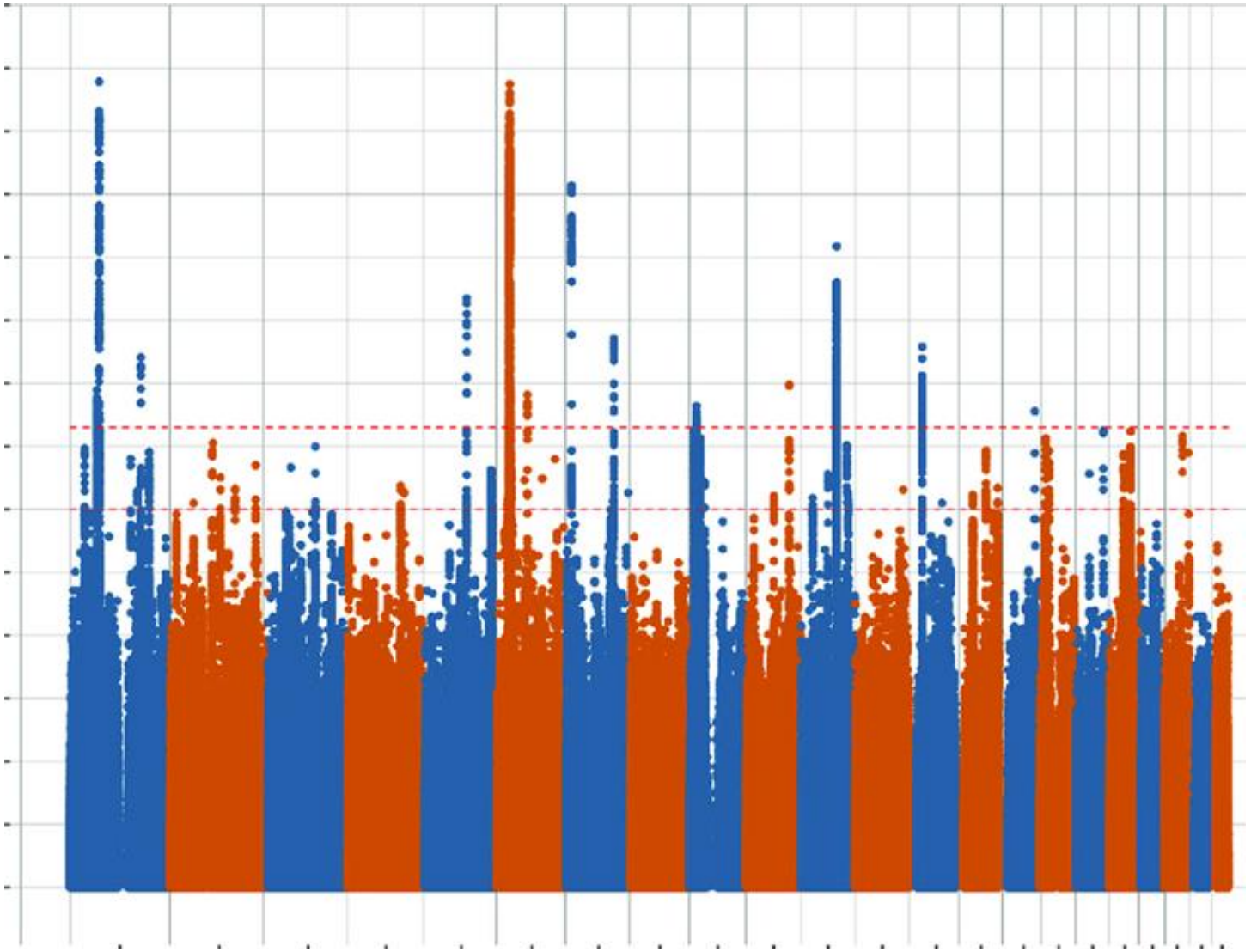
Causal vs. in linkage?



Linkage Disequilibrium

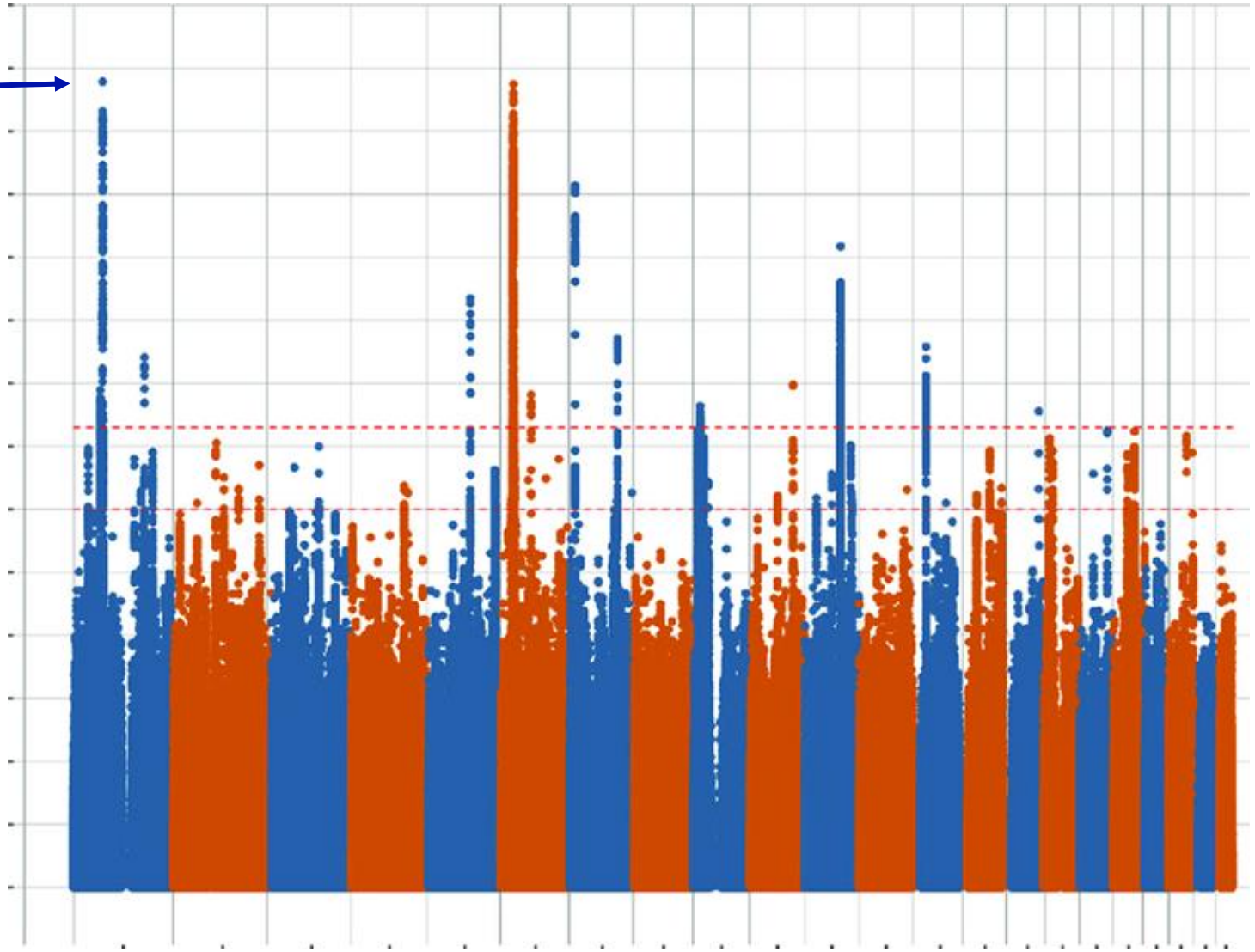


Linkage Disequilibrium

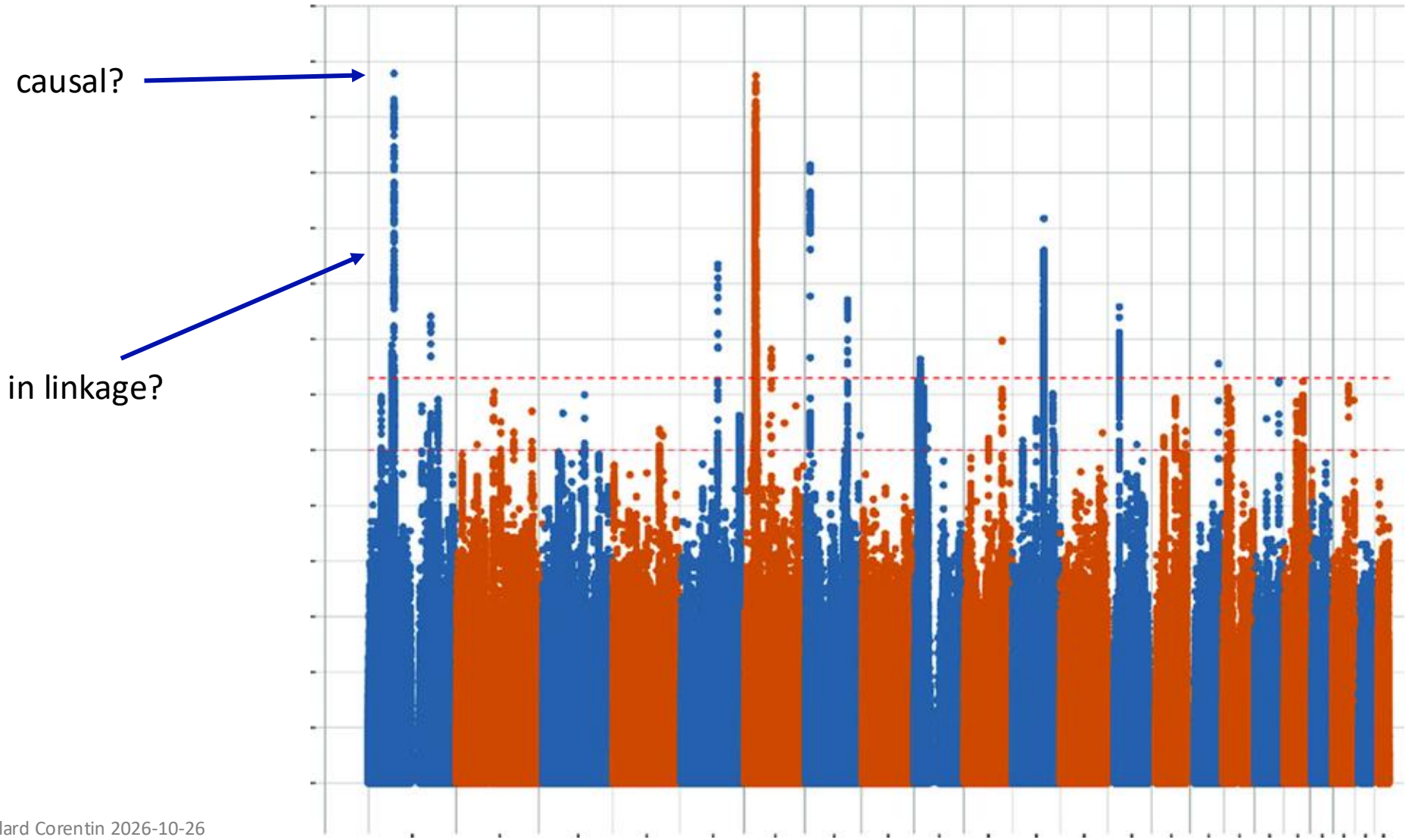


Linkage Disequilibrium

causal?



Linkage Disequilibrium



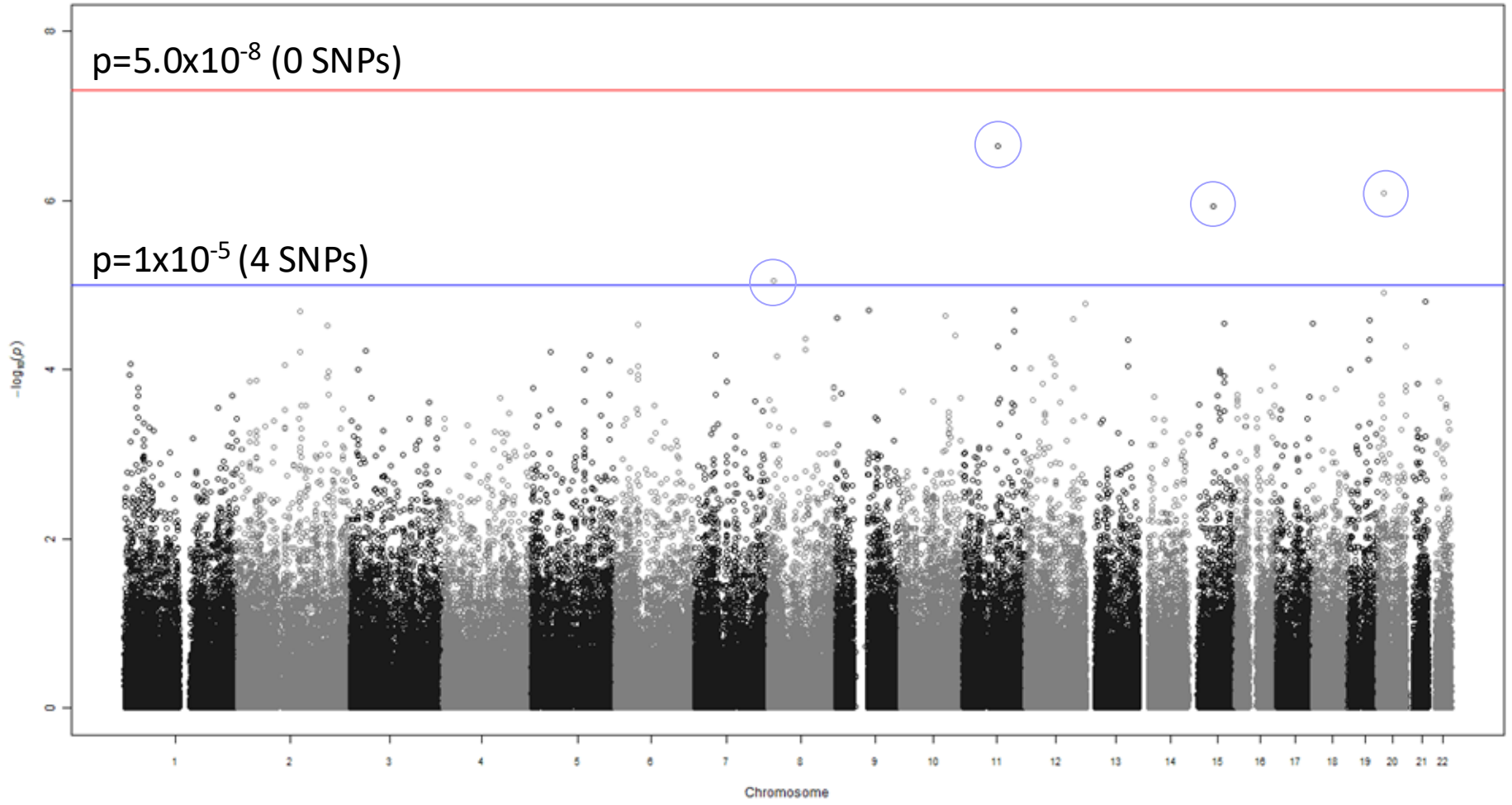
Polygenic Risk

What are we trying to do?

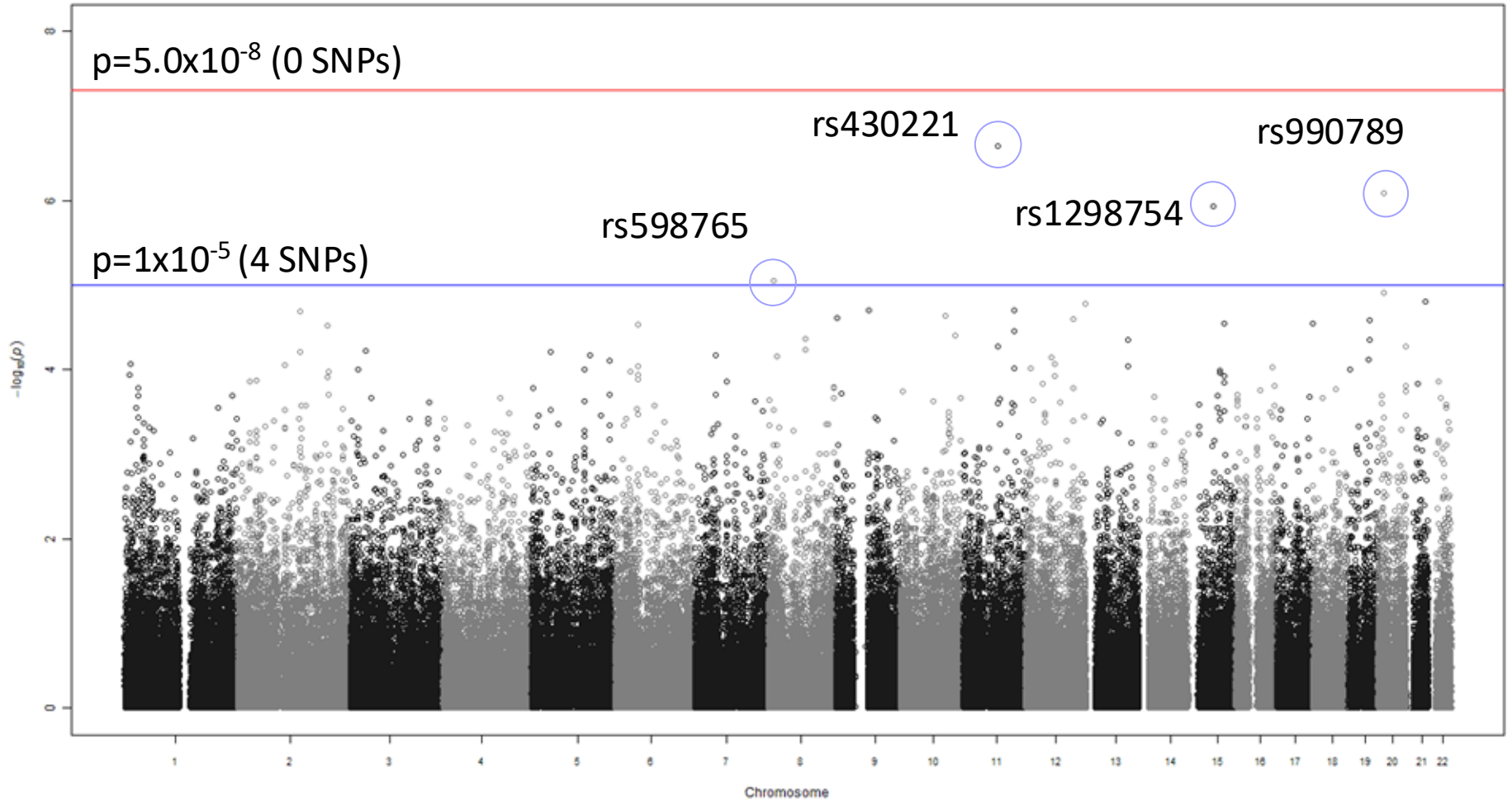
- A polygenic score is just a way of *distilling* or *simplifying* data for many genetic variants in one person to a single number.
- In essence it is a method of *dimensionality reduction*, i.e. creating a single number from many.

Example: we have 100 SNPs associated with height in a GWAS. Can we use all 100 SNPs to generate a single score per subject that summarizes their genetically-determined ‘risk’ for taller or shorter height?

How might we do this simply?

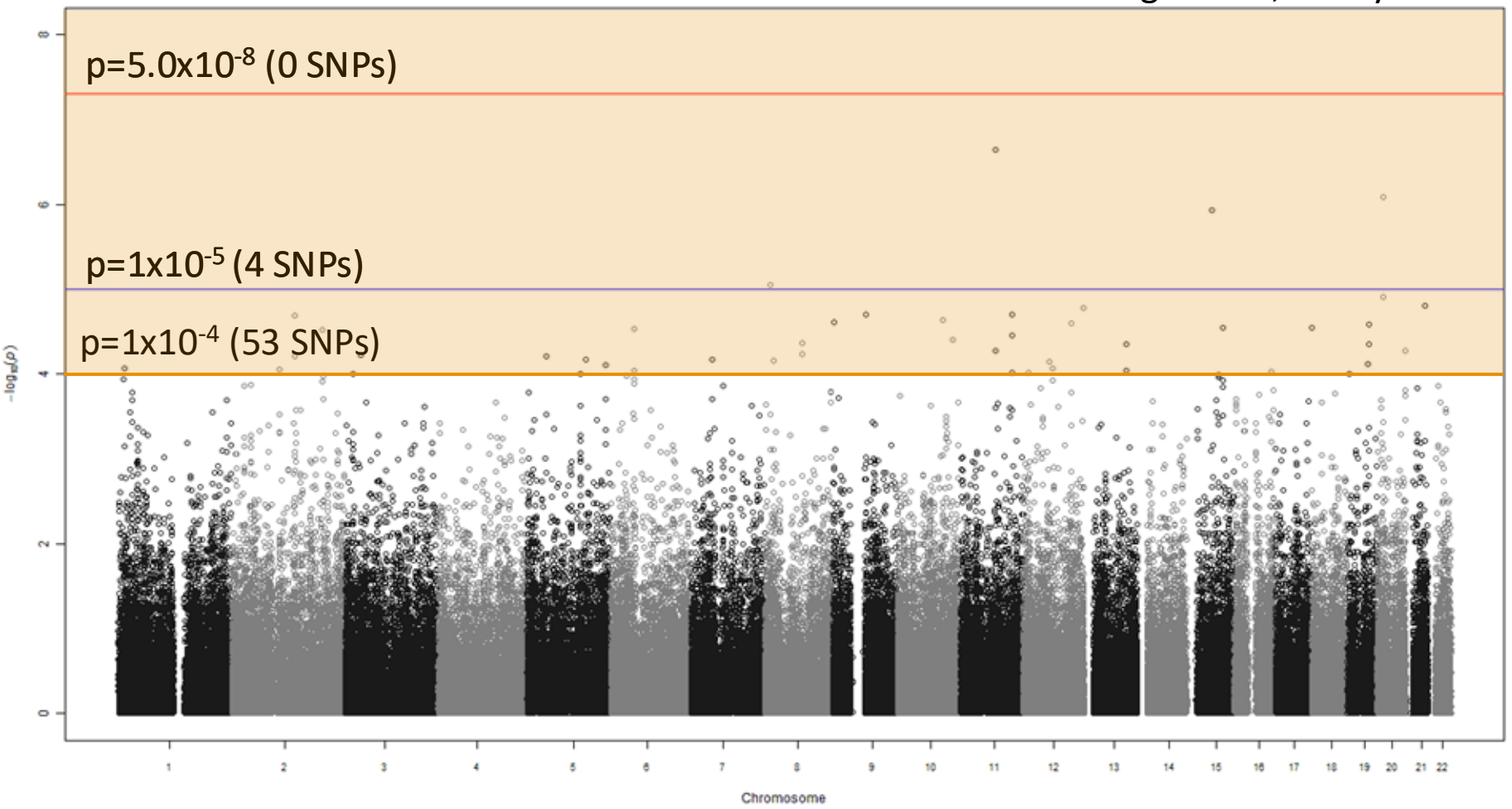


How might we do this simply?



Including more SNPs

A GWAS tests the association between each SNP and a trait or disease. To declare an association statistically significant across the genome, a very strict threshold is often used: $\mathbf{p < 5 \times 10^{-8}}$.



Increasing the threshold therefore allows more SNPs to be included. This may improve prediction by capturing more small effects, but it may also add noise. We need to evaluate which threshold works best using independent data.

Weighting by effect

$$\text{PRS} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{n_{SNPs}} X_{n_{SNPs}}$$

We input the % of effect alleles for $X_1 \ldots X_{nSNPS}$ and use beta coefficients from a GWAS as $\beta_1 \ldots \beta_{nSNPs}$

Weight each SNP by its effect: the β coefficient
Simply counting risk-associated alleles gives each allele the same weight. The slide proposes accounting for their estimated effects:

$$\text{PRS} = \sum_i \beta_i \times n_i$$

Here, n_i is the **number of copies of the effect allele** carried by the person: 0, 1, or 2. This is the additive form of the score

For example, for **rs598765**, assuming that A1 = A is the effect allele and β = 1.2:

Person's genotype	Copies of A	Contribution to the score
GG	0	0
AG	1	1.2
AA	2	2.4

We perform this calculation for each selected SNP, then add the contributions together.
The columns in the table mean:

- **A1 / A2:** the two alleles; we must identify which allele β refers to.
- **Beta:** the estimated effect of each copy of the effect allele.
- **SE:** the standard error, which measures the uncertainty in the estimated effect.
- **P-value:** the strength of the statistical evidence against the absence of an association.

$$\mathrm{PRS} = \sum_i \beta_i \times n_i$$

The p-value is used here to select SNPs; β determines their weight in the score. A PRS calculated this way is not directly a percentage risk.
One caveat about this slide: the **β, SE, and p-value values appear inconsistent with one another** under a standard statistical test. For example, \(1.2/0.037 \approx 32\), which would produce a much smaller p-value than the one shown. The table therefore appears to be mainly illustrative.

A simple example using type 2 diabetes

"The genetic risk of diabetes depends on many variants. Each can contribute a little to the risk. A PRS adds these contributions together, taking into account the weight of each variant."
To illustrate this, imagine only **three variants**, with **entirely fictional weights**:

Risk-associated allele	Weight per copy	Copies in Alice	Copies in Paul
A	1 point	2	0
B	2 points	0	1
C	3 points	0	1

The calculation gives:

- **Alice:** $2 \times 1 = 2$ points.
- **Paul:** $1 \times 2 + 1 \times 3 = 5$ points.

Both carry two copies of risk-associated alleles, but Paul has a higher score because his alleles have larger effects.

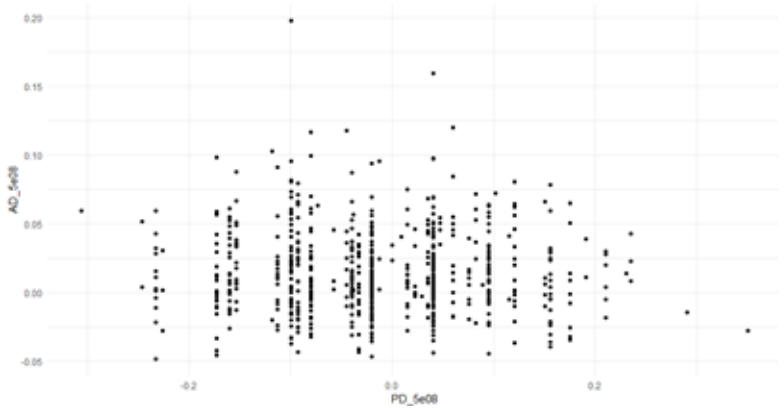
"Paul therefore has a higher genetic predisposition according to this score. This does not mean that he will necessarily develop diabetes: age, diet, physical activity, and other factors also play a role. And 5 points does not mean a 5% risk!"

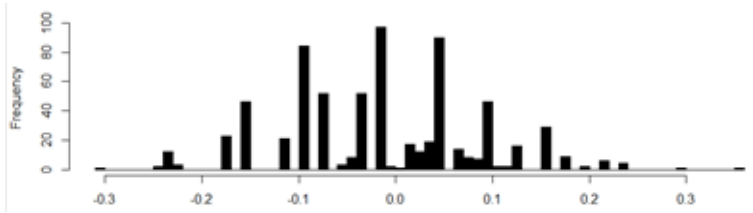
A real PRS combines many more variants, and their weights are estimated through a GWAS. The score is then generally interpreted **relative to a reference population**, for example as being among the highest 10% of scores.

Resolution (PD example)

$p<5x10^{-8}$

$N_{snps}=4$

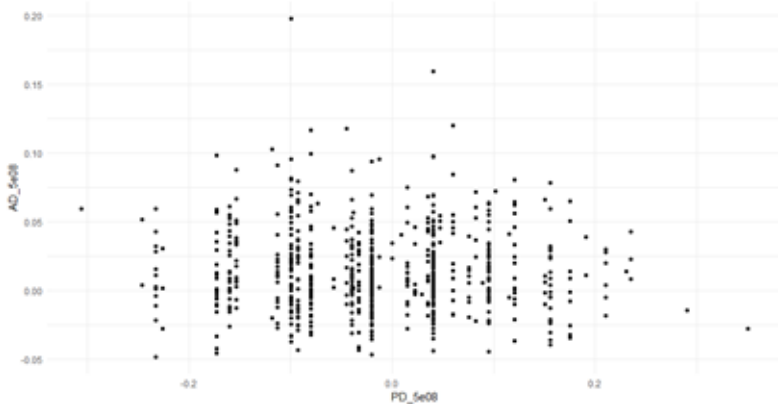




Resolution (PD example)

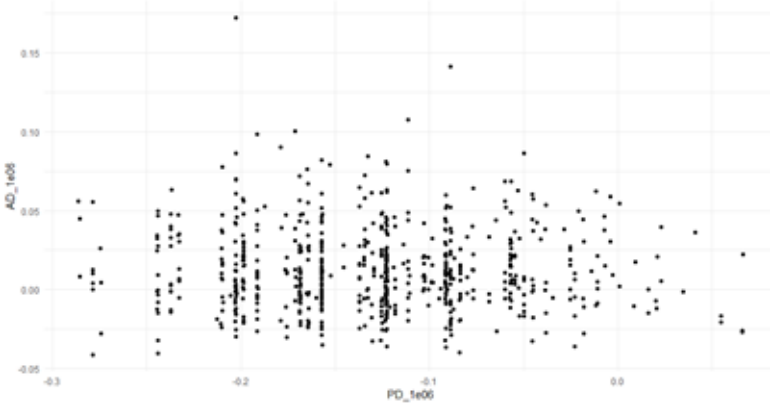
$p<5x10^{-8}$

$N_{snps}=4$



$p<1x10^{-6}$

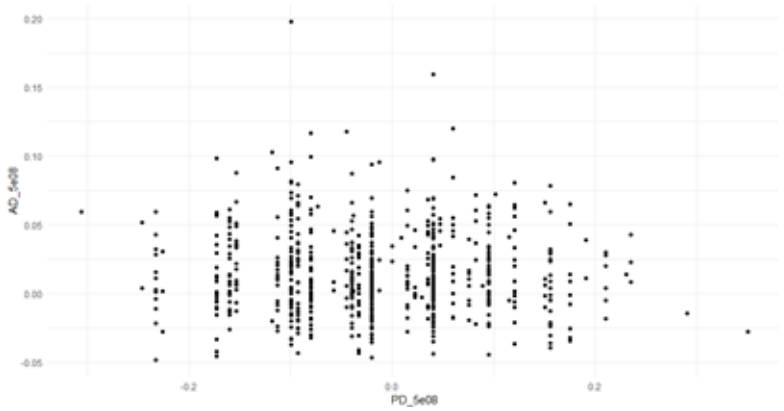
$N_{snps}=7$



Resolution (PD example)

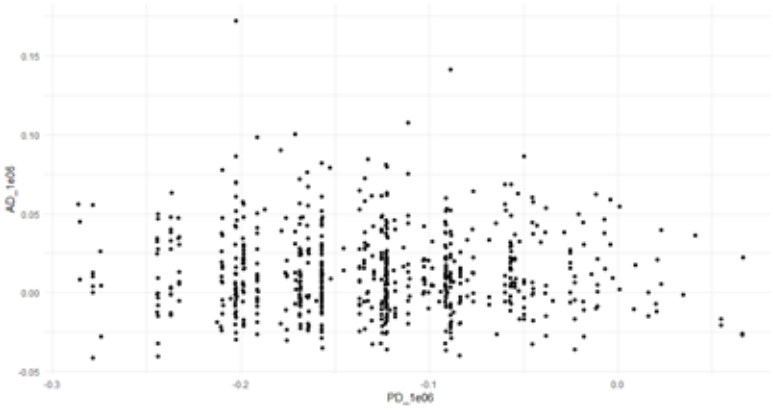
$p<5x10^{-8}$

$N_{snps}=4$



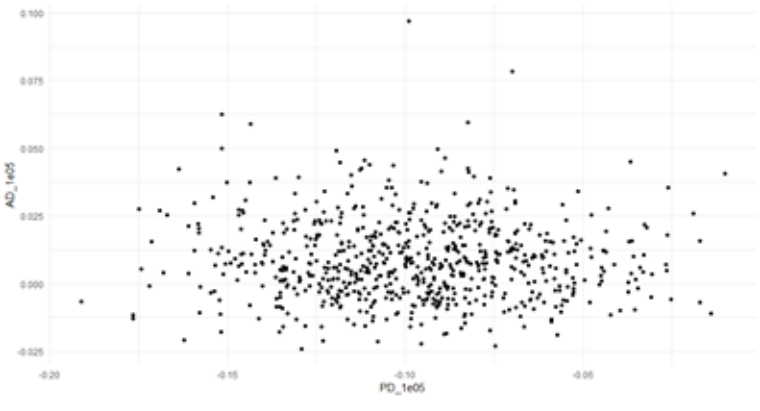
$p<1x10^{-6}$

$N_{snps}=7$



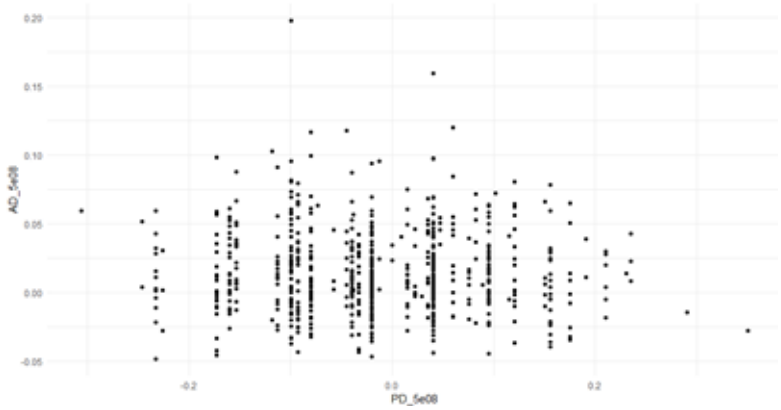
$p<1x10^{-5}$

$N_{snps}=22$

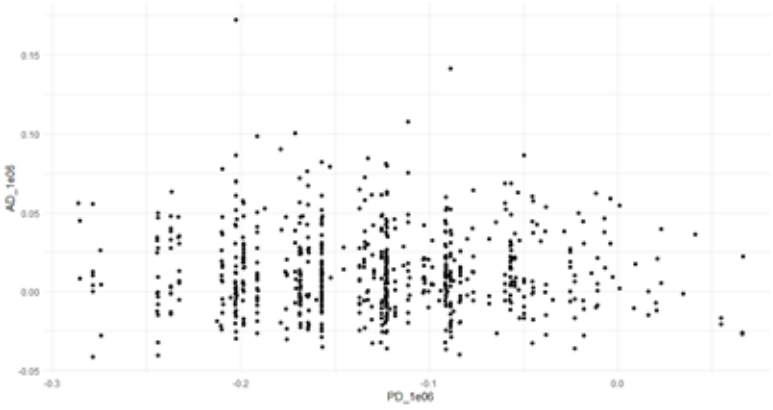


Resolution (PD example)

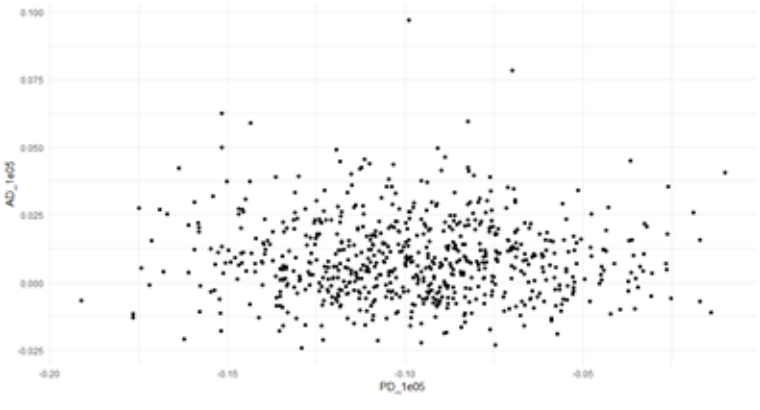
$p<5x10^{-8}$
$N_{snps}=4$



$p<1x10^{-6}$
$N_{snps}=7$



$p<1x10^{-5}$
$N_{snps}=22$

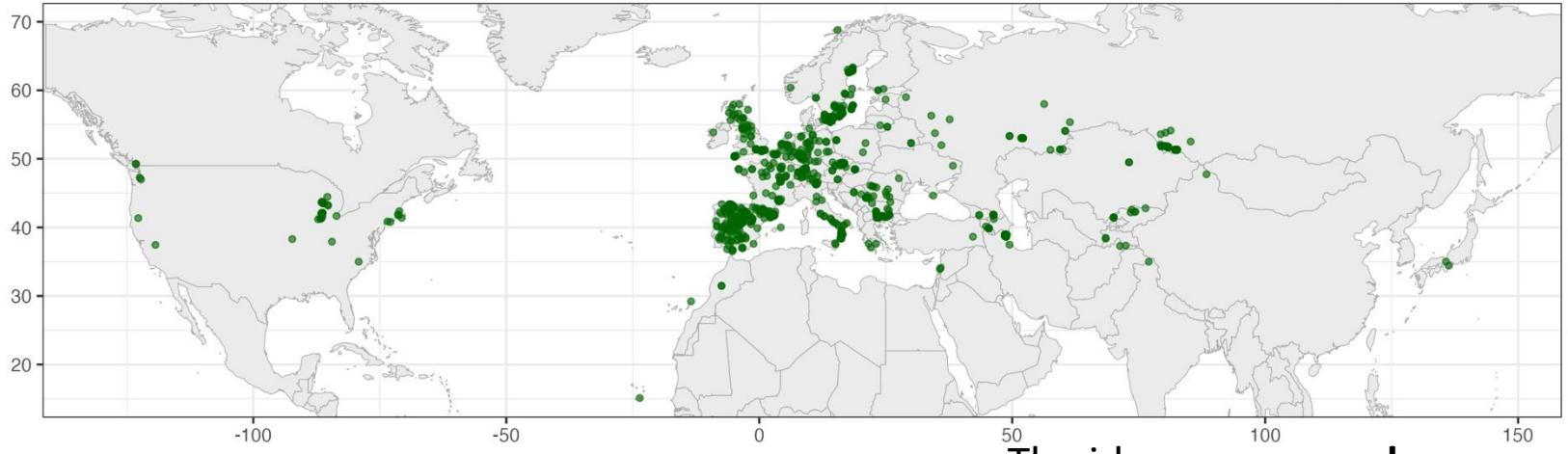


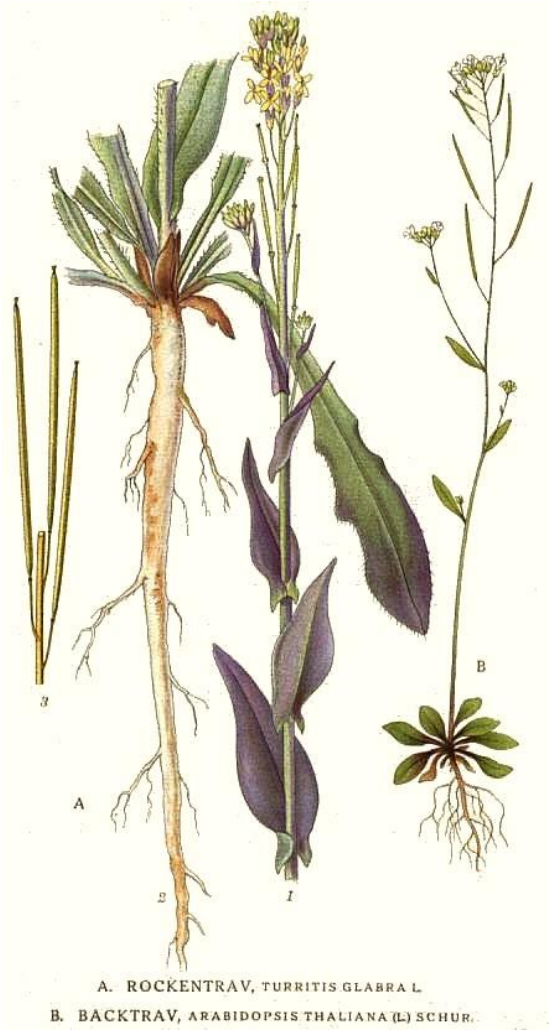
Including more SNPs increases the resolution of the PRS, but does not necessarily improve its predictive accuracy

Where today's data come from

Where today's data come from

The plants: wild *Arabidopsis* accessions collected across the world, each one an inbred line. The 1001 Genomes Project sequenced them once; any lab can grow the same lines again (1001 Genomes Consortium, *Cell*, 2016).
The DNA: the SNP table of the 1001 Genomes, reduced to common SNPs spread along the genome, so that it runs in class.
The traits (AraPheno database): flowering time at 10 °C and 16 °C (1001 Genomes Consortium, 2016); leaf sodium (Baxter et al., *PLoS Genetics*, 2010).
The genes: Ensembl Plants (TAIR10 genome, Araport11 annotation).
The idea: **sequenced once, grown many times**. Every new trait measured on these lines can be mapped with the same genomes.



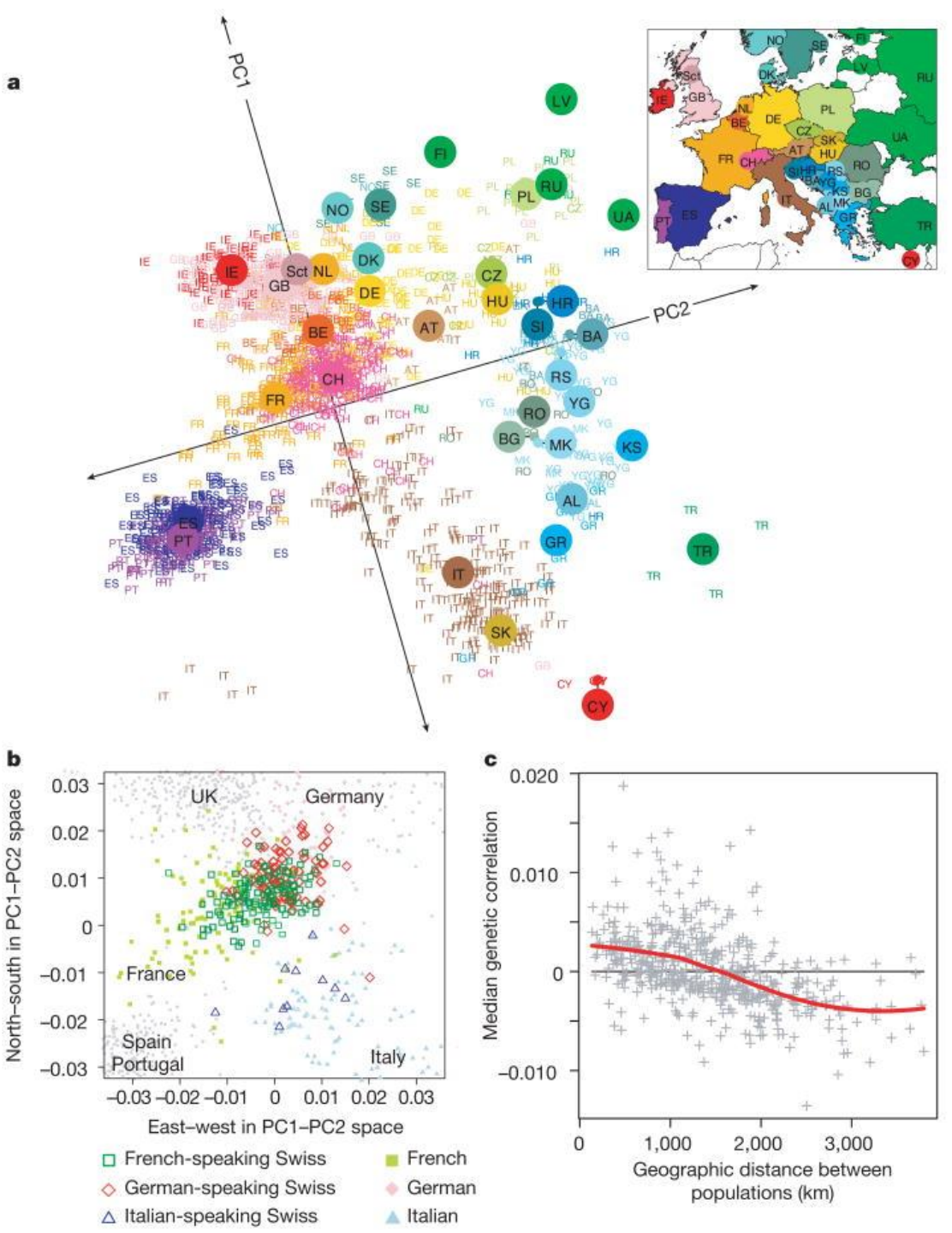


The idea: **sequenced once, grown many times**. Every new trait measured on these lines can be mapped with the same genomes.

Population structure: genetic variation reflects geography

- **Panel a:** Principal component analysis (PCA) reveals genetic patterns that closely resemble the geographic map of Europe.
- **Panel b:** Fine-scale population structure is visible even within Switzerland.
- **Panel c:** Genetic similarity tends to decrease as geographic distance increases.

Population structure can confound genetic associations. Including ancestry principal components as covariates helps reduce this bias



The question

A single grow room, where some plants are already flowering and others are still in the rosette stage. This is an example of “same place, different flowering.”



