Genomewide association studies (GWAS)

Chris Wallace

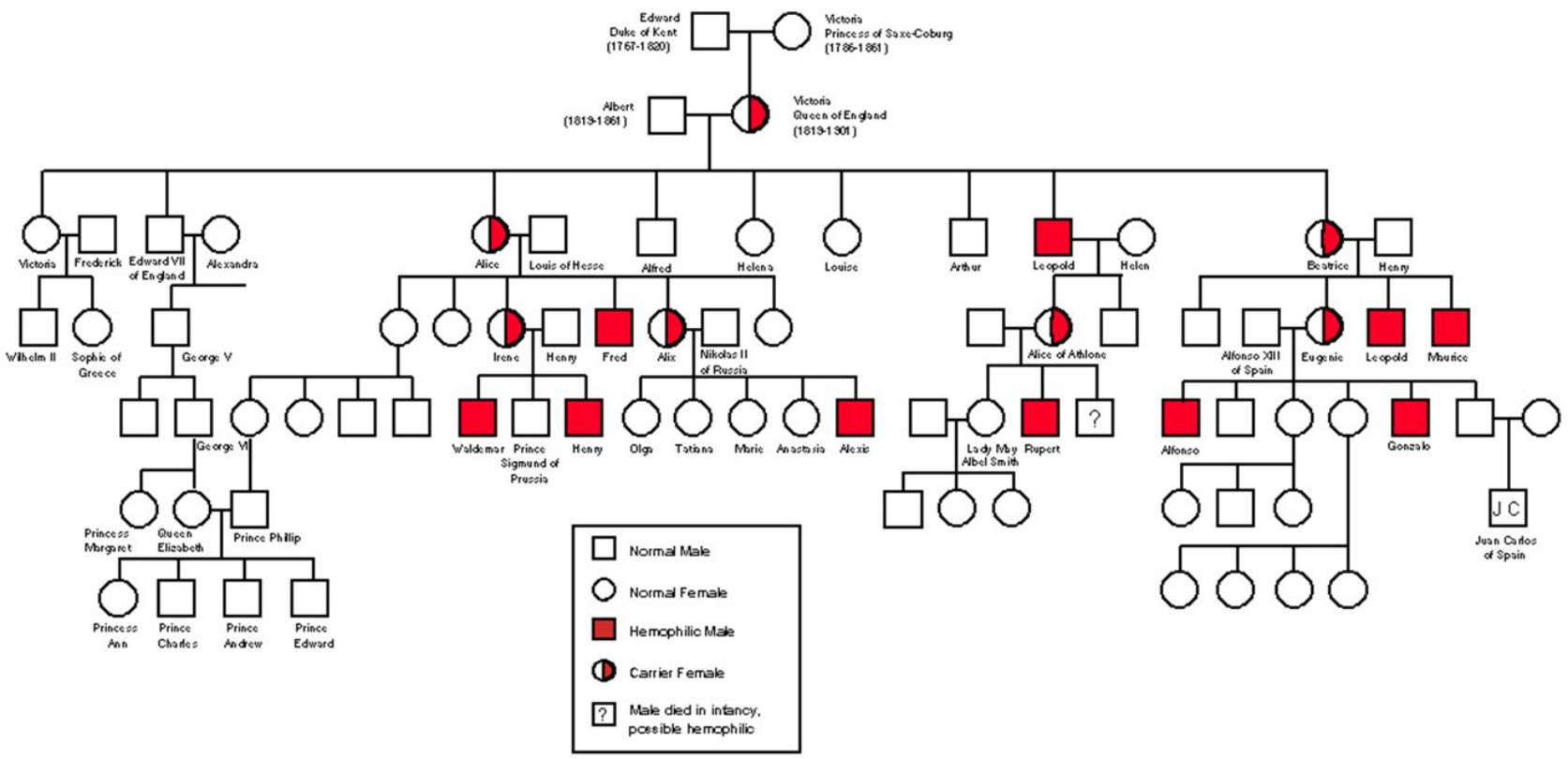
chr1swallace chr1swallace.github.io cew54@cam.ac.uk

Outline

- Inheritance of disease
- Genome-wide association studies
- Confounding in GWAS
- Imputation to extend genome coverage
- Heritability
- Inferring disease relevant cells

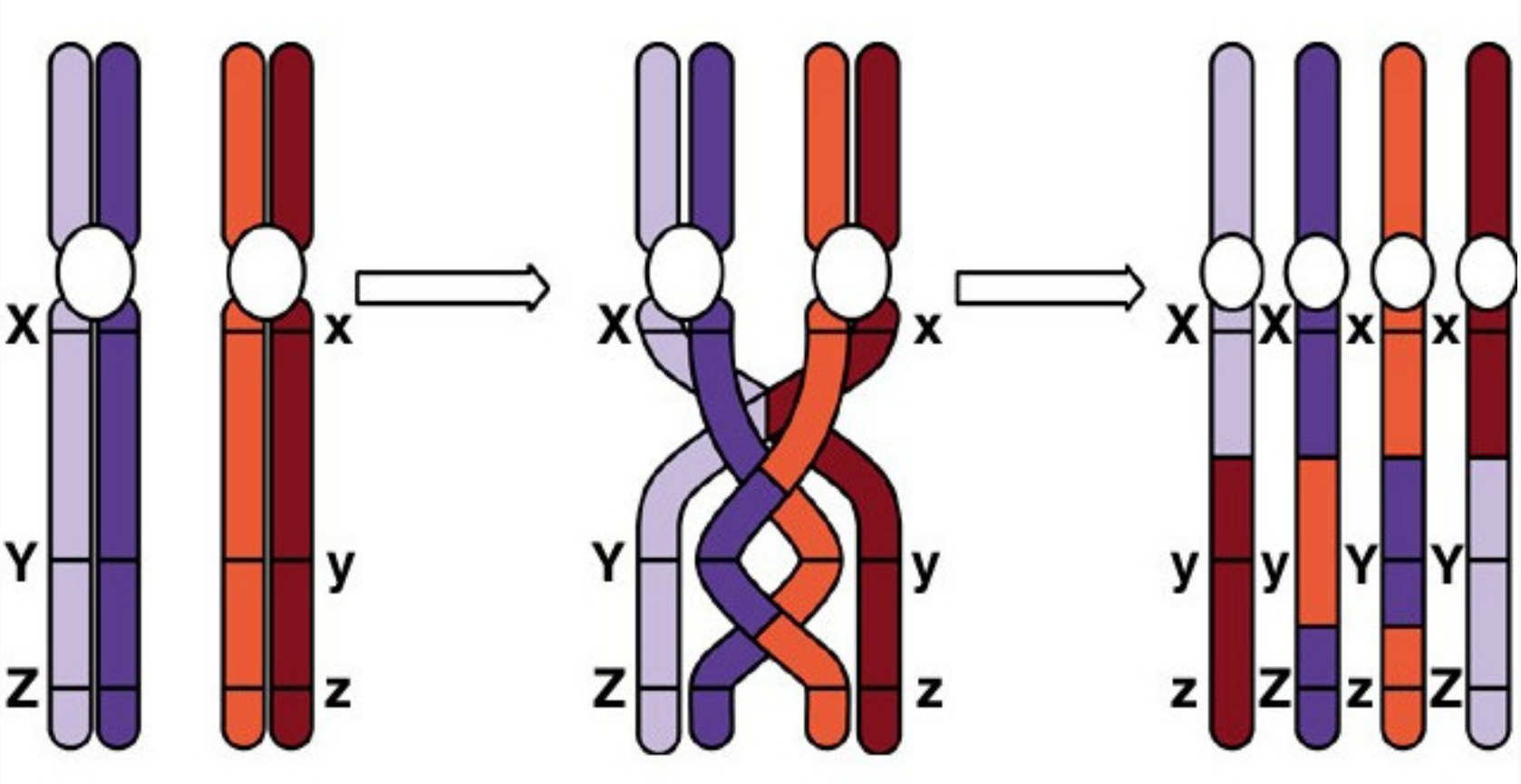
Inheritance of disease

Heritability- diseases can be passed down in families



To understand inheritance of disease, first understand inheritance of DNA

DNA replicates by meiosis during gamete formation: DNA from two (autosomal) parental chromosomes recombine to create new chromosomes.



Consequences of recombination

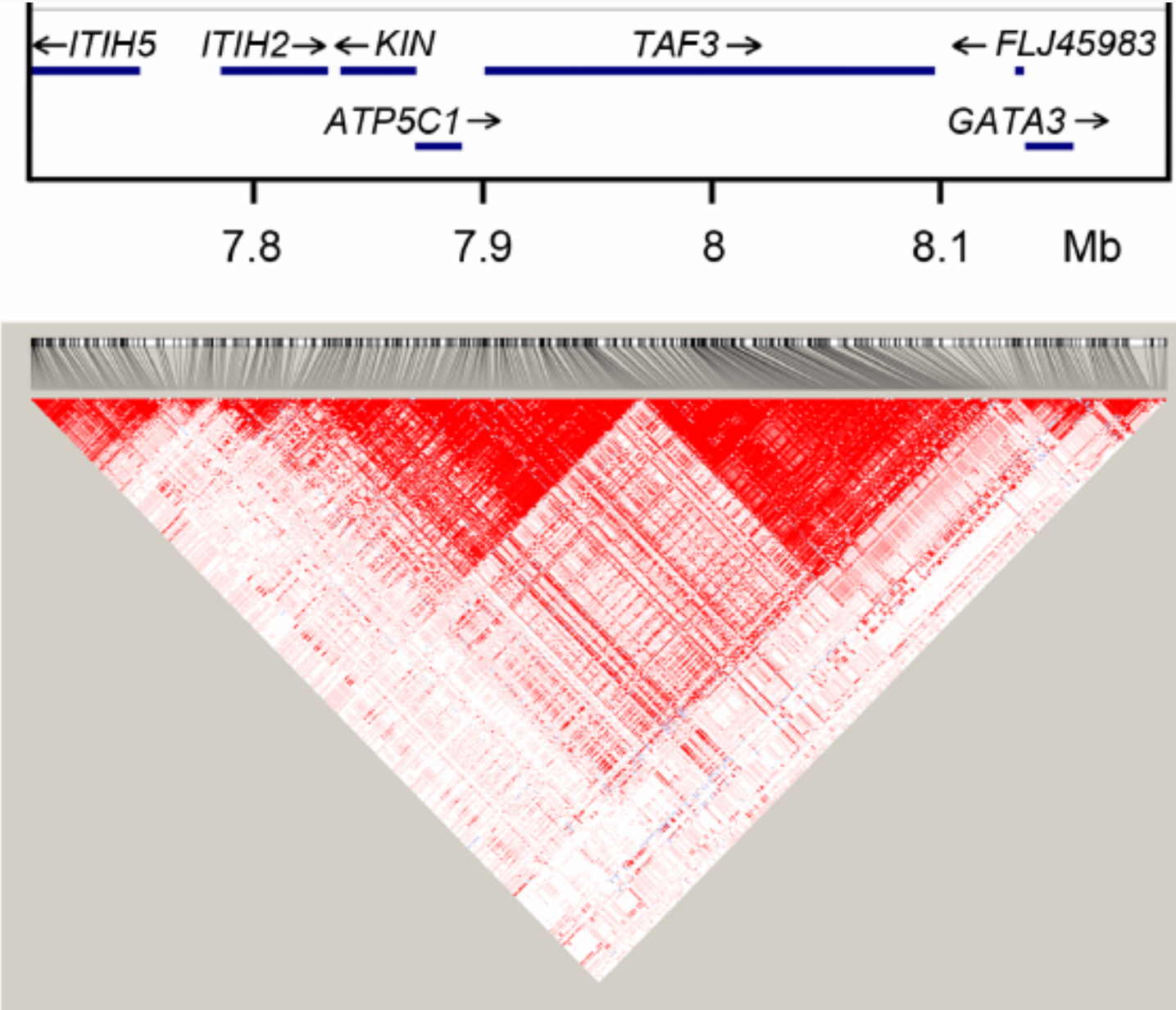
Recombination should lead to independent assortment of markers in the population (Mendel's second law)

Law doesn't hold at short genomic distances

Sequences physically close on the same chromosome tend to stay together - are linked within families and populations. Alleles are held in linkage disequilibrium (LD).

- 3 billion bp in the human genome
- $\sim$ 80 million single nucleotide polymorphisms (SNPs) in 2,504 individuals sequenced for the 1000 Genomes Project
- We can test for genetic association using only 100,000 – 1,000,000 SNPs

Consequences of recombination



Genome-wide association studies

Genetic association to common disease

Genetic association

Change in genetic code that alters **risk** of disease

Usually in a stochastic manner

GWAS

The art/science of identifying genetic positions associated with different disease risks in the population

How to conduct a GWAS

Recruit patients and controls

Genotype at 500,000 SNPs genomewide

Check data quality:

- Hardy Weinberg equilibrium
- data completeness
- sex correctly reported

Remove low quality variants and samples and document

Test H_0 : equal allele frequencies between cases and controls at each SNP

Massive multiple testing problem, use a **genome-wide significance** threshold of 5×10^{-8}.

Further check quality at every associated variant

Hardy Weinberg Equilibrium

In a population with random mating, a variant that has two alleles (a/A), with frequency

$$a : p \quad A : q = (1 - p)$$

then, the genotype frequencies are

$$aa : p^2 \quad Aa : 2pq \quad AA : q^2$$

This can be tested in a χ^2 test,

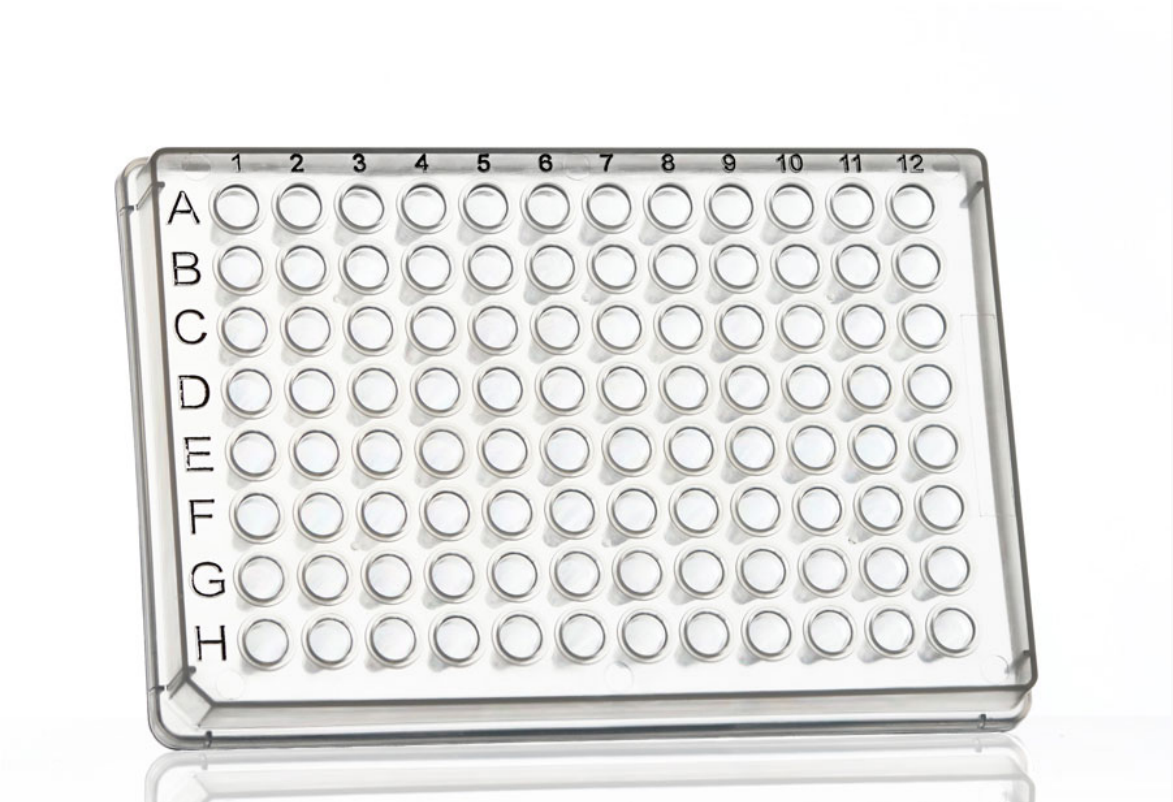
$$\frac{(\pi(aa) - p^2)^2}{p^2} + \frac{(\pi(Aa) - 2pq)^2}{2pq} + \frac{(\pi(AA) - q^2)^2}{q^2} \sim \chi^2_2$$

Quality control

- Drop SNPs which are not in HWE
- Drop SNPs or samples with a large amount of missing data (e.g. $> 1\%$)
- Drop samples with mismatch between declared sex
- Drop/rotate annotations for plates with data in control wells or many sex mismatches

Quality control

- Drop SNPs which are not in HWE
- Drop SNPs or samples with a large amount of missing data (e.g. $> 1\%$)
- Drop samples with mismatch between declared sex
- Drop/rotate annotations for plates with data in control wells or many sex mismatches



Genetic data

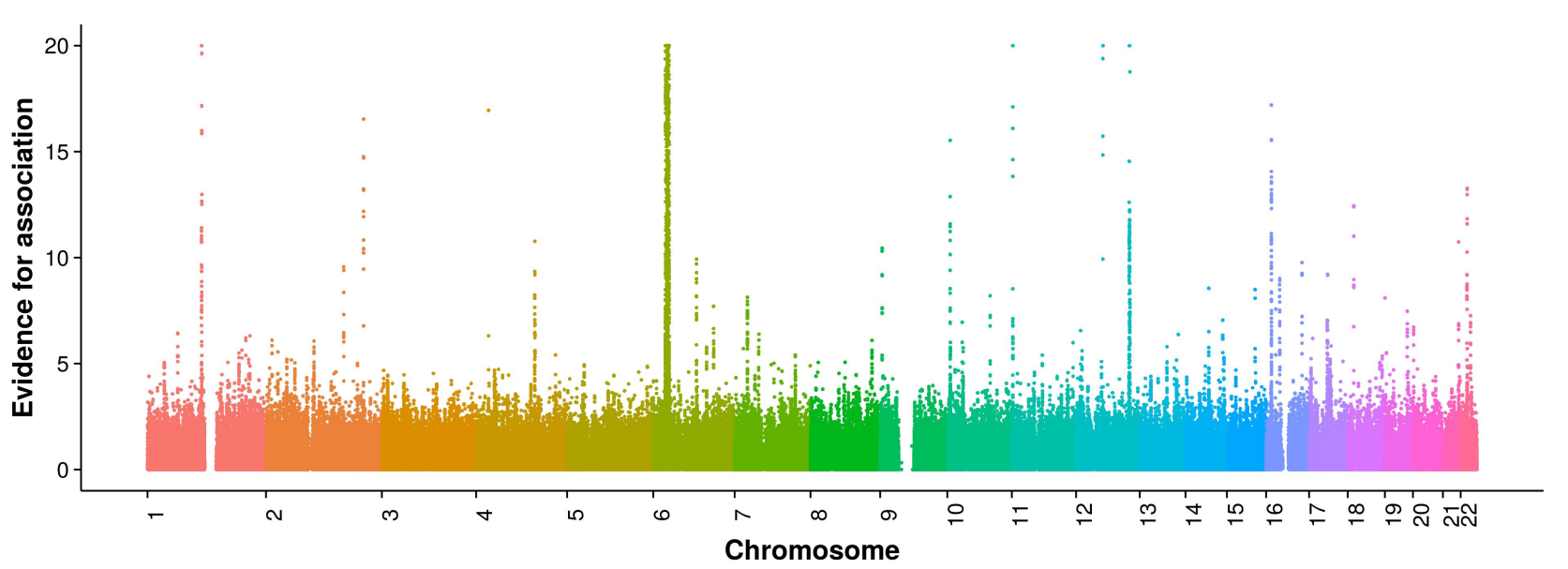
$$
\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\quad
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 2 & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\
1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & \cdots \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
2 & 2 & 0 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & \cdots \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots
\end{bmatrix}
$$

Y

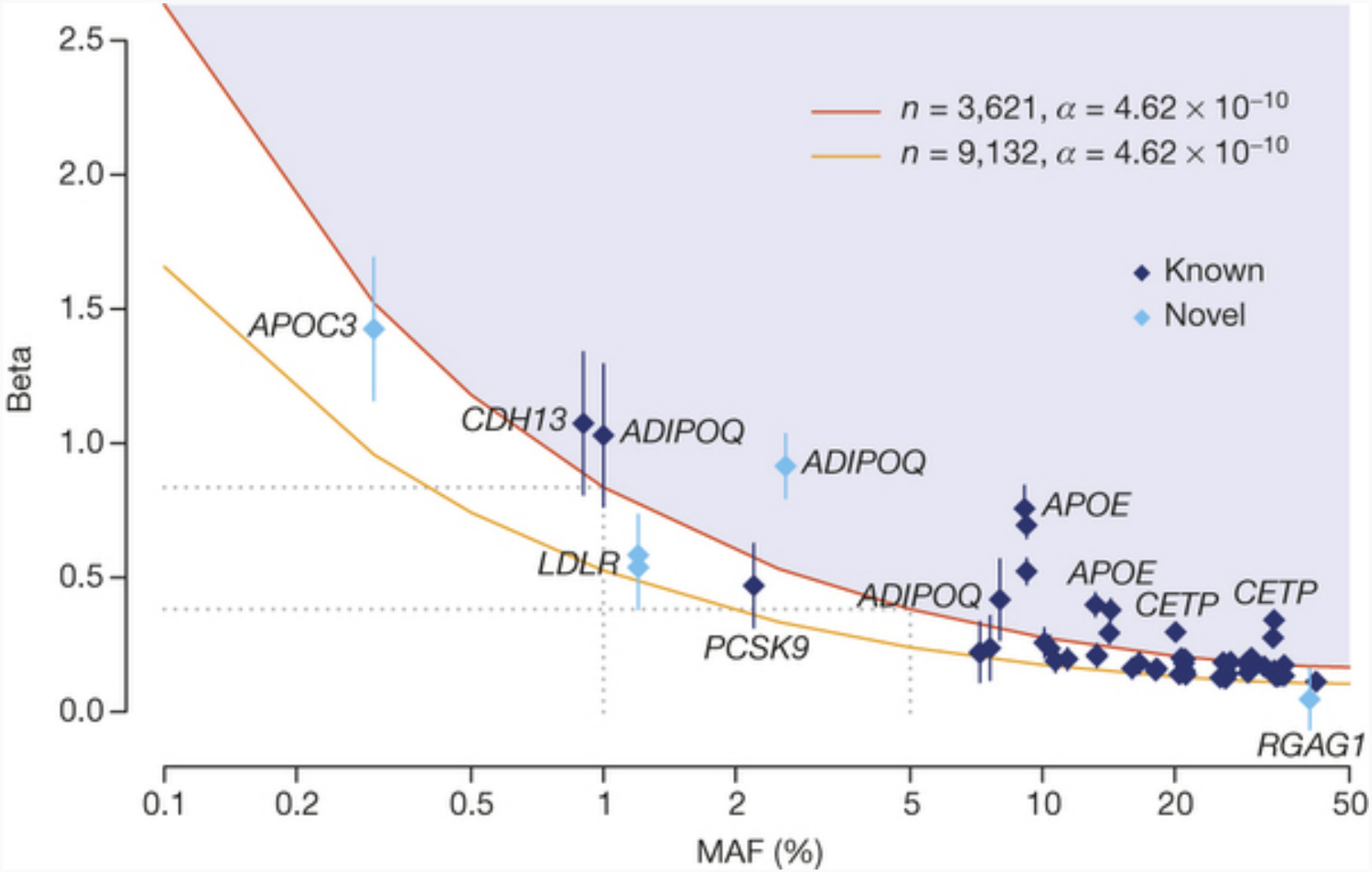
X

Genetic data

Represent results of GWAS in Manhattan plots

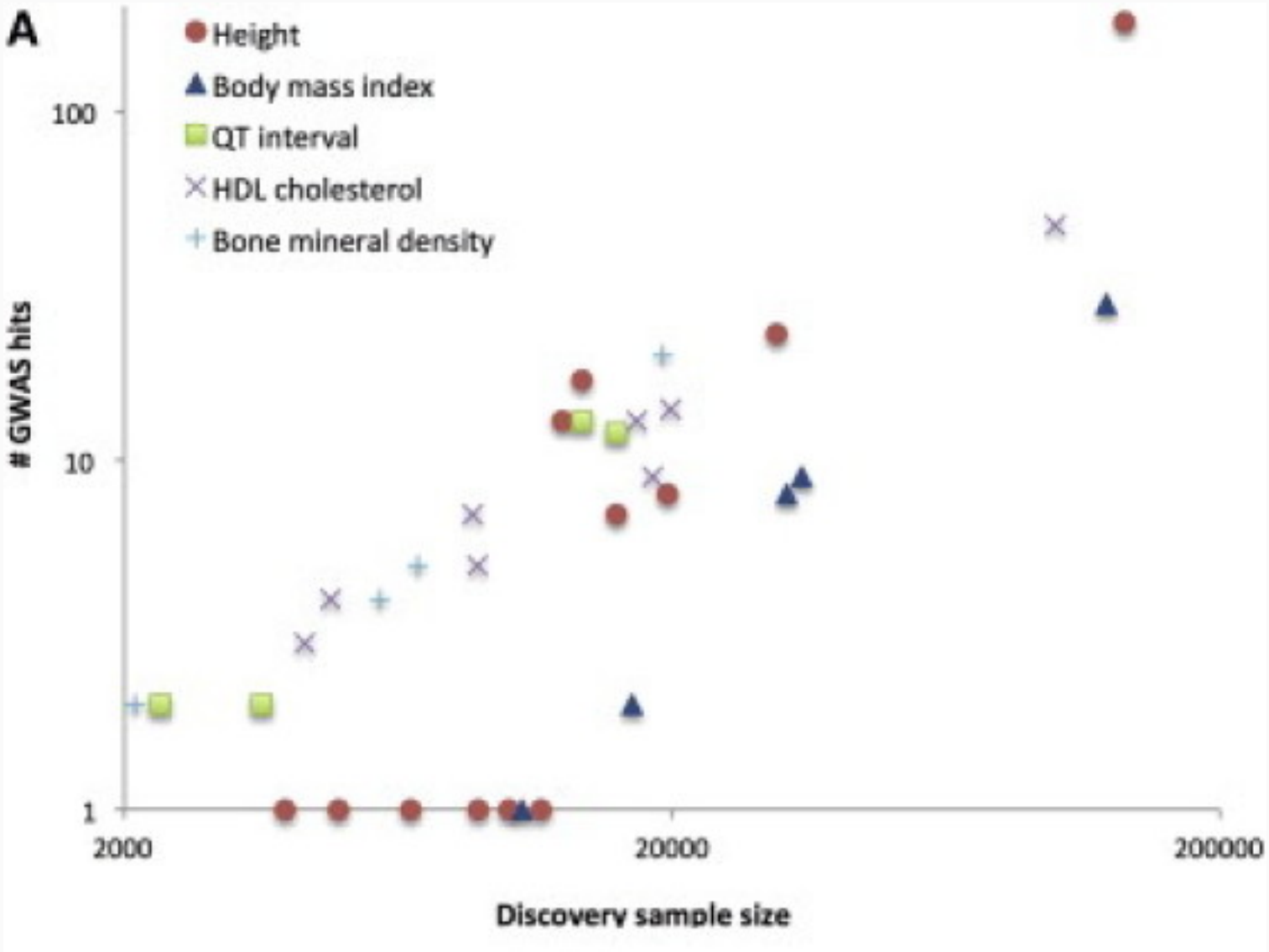


The limits of GWAS detection: effect size and population frequency



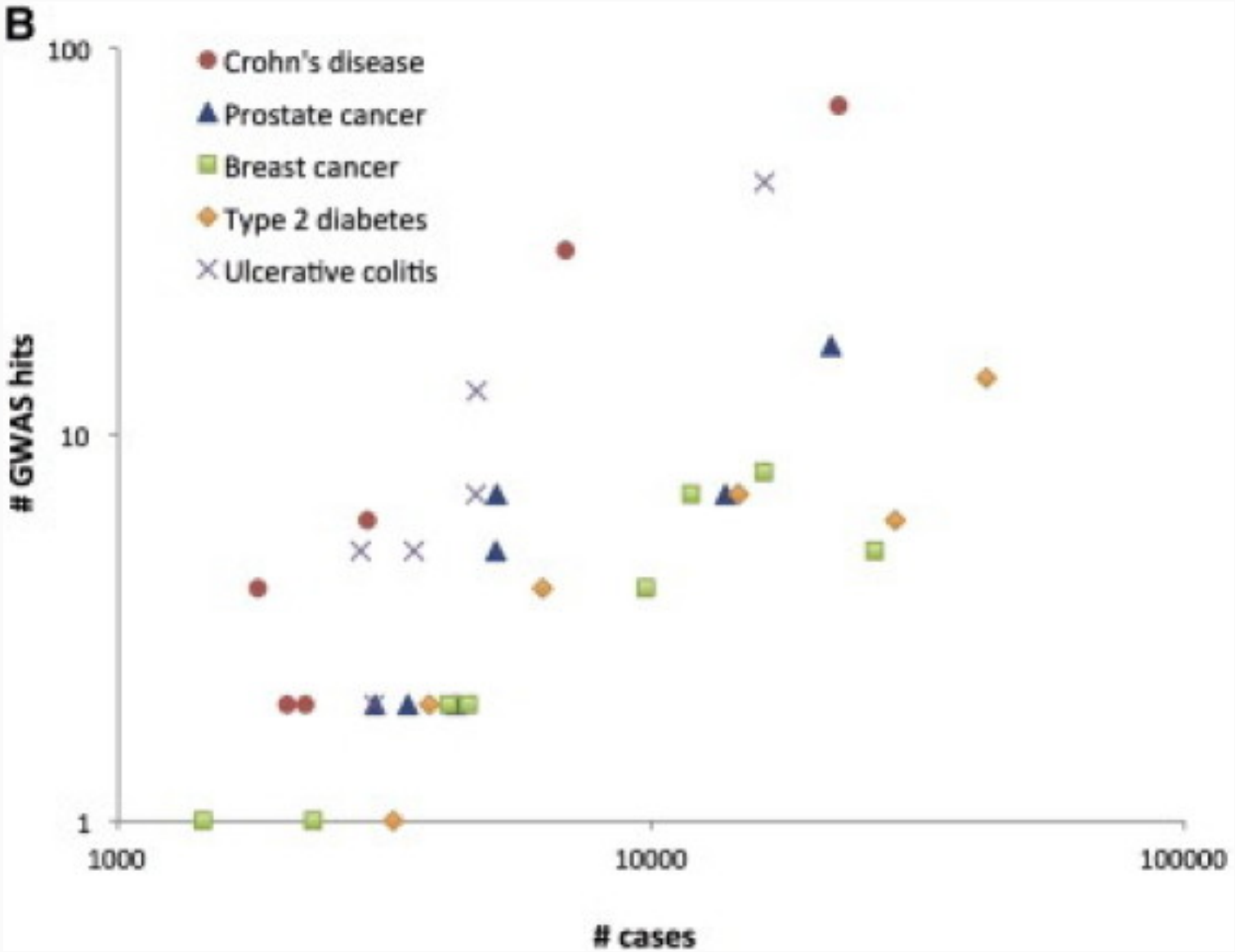
UK10K Consortium. Nature 2015

More samples = more findings (of smaller effect)



Visscher et al, Am J Hum Genet 2012

More samples = more findings (of smaller effect)



Visscher et al, Am J Hum Genet 2012

Confounding in GWAS

Confounding

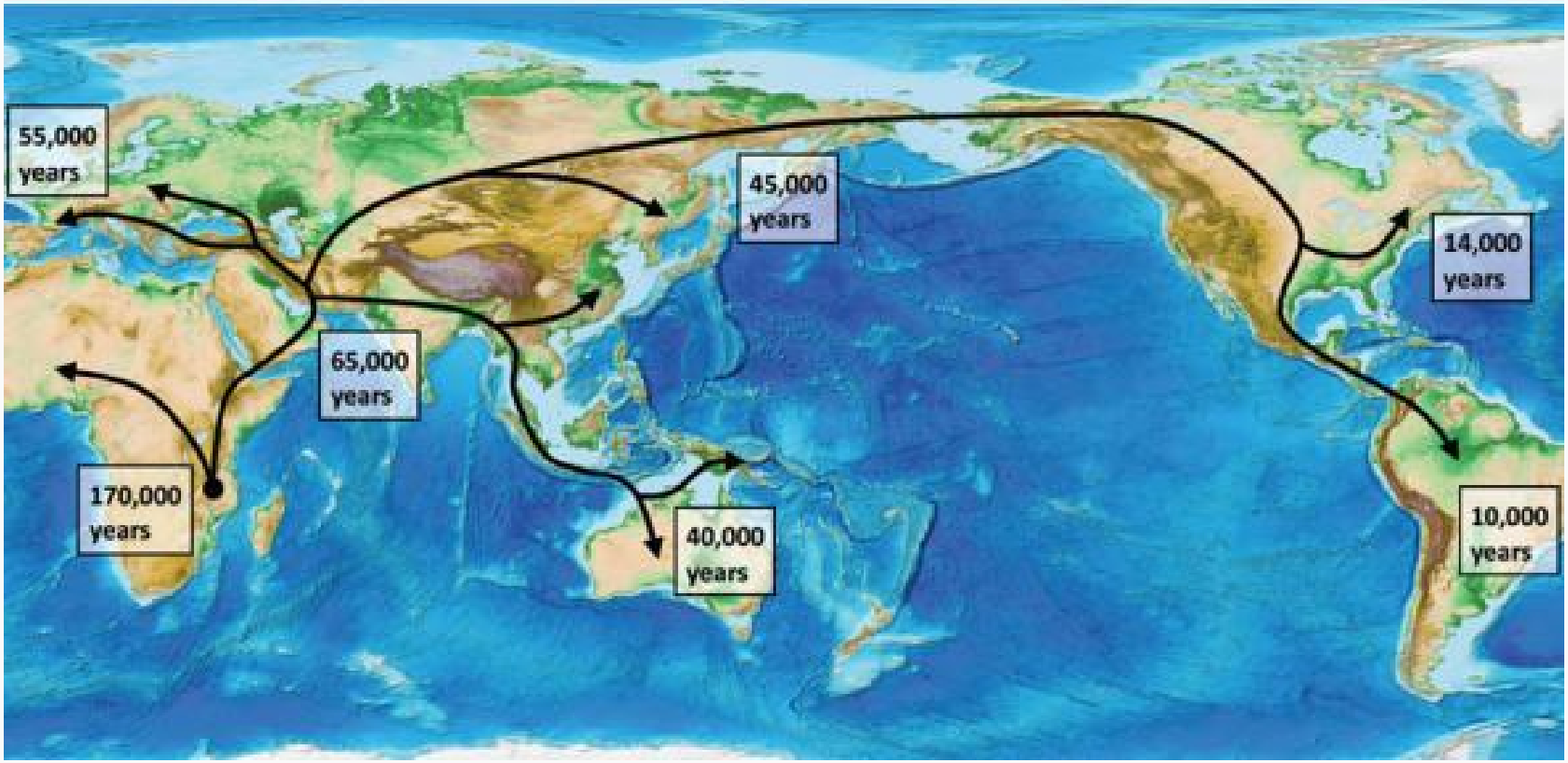
Definition

A counfounding factor is associated with both exposure and outcome

Effect

Can cause spurious association if not recognised and accounted for

Worldwide variation in allelic frequencies



Eriksson et al, PNAS, 2012

Genetic structure of each population varies across generations

Mutation: each new offspring carries ~ 38 de novo mutations

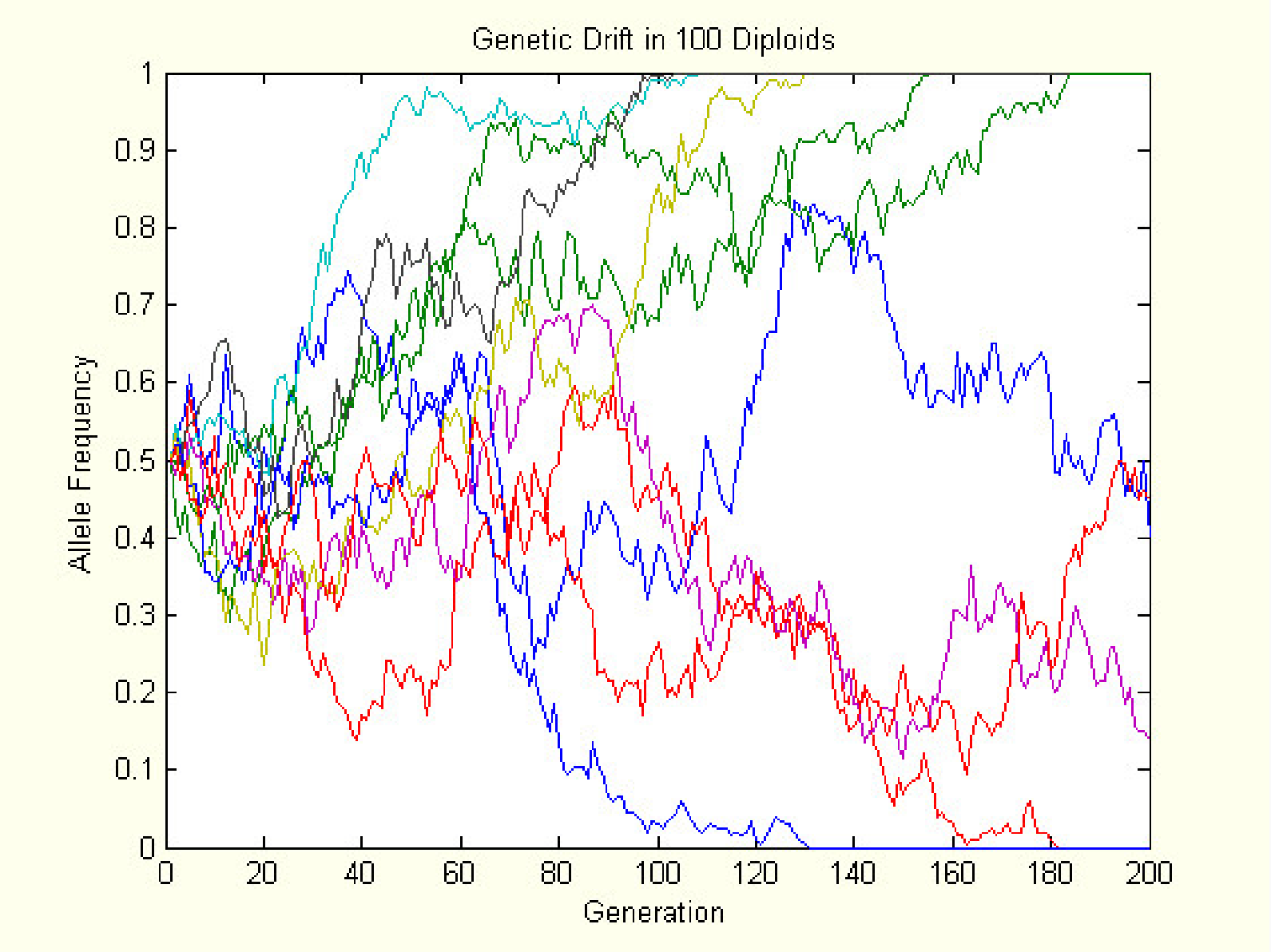
Selection: alleles which improve reproductive fitness tend to increase in frequency.
Affects $\sim 30 - 40\%$ of amino-acid changes which lead to viable offspring
Affects $\sim 8\%$ of genome overall

Drift: each generation is a random sample from the generation before (subject to selection)

Bottlenecks: when population sizes reduce, random selection of alleles has high variance

Nielsen et al, Nat Rev Genet 2007; Rands et al, PLoS Genet 2014

Drift: an under-appreciated cause of genetic difference between geographically separated populations



Ancestry is a confounding factor for GWAS

Allele frequencies **and** disease incidence differ systematically between populations

sub-population	1	2
frequency in population	0.6	0.4
frequency of disease in sub population	0.1	0.3
propn of cases from each sub population*	0.33	0.67
propn of controls from each sub population*	0.66	0.34

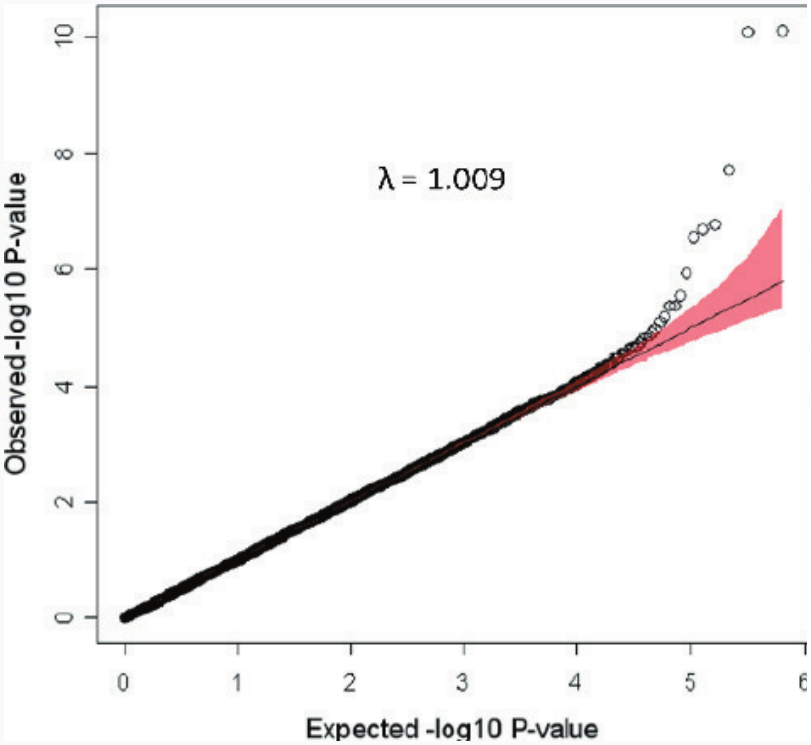
*Bayes rule:

$$P(\text{sub pop}|\text{case}) = \frac{P(\text{case}|\text{sub pop})P(\text{sub pop})}{P(\text{case})}$$

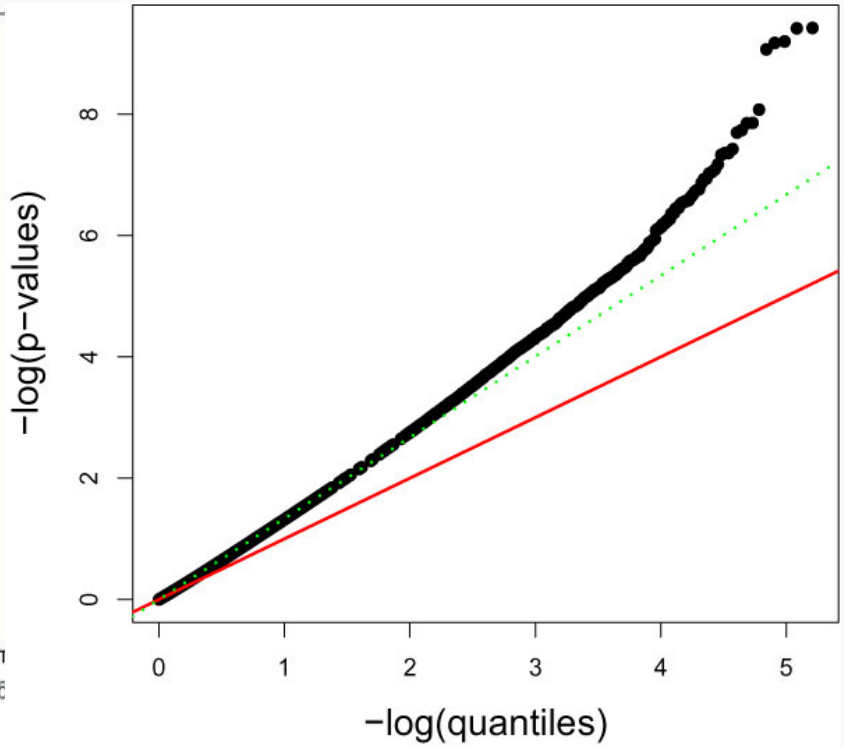
Even subtle confounding can inflate test statistics

quantile-quantile plots reveal evidence of counfounding

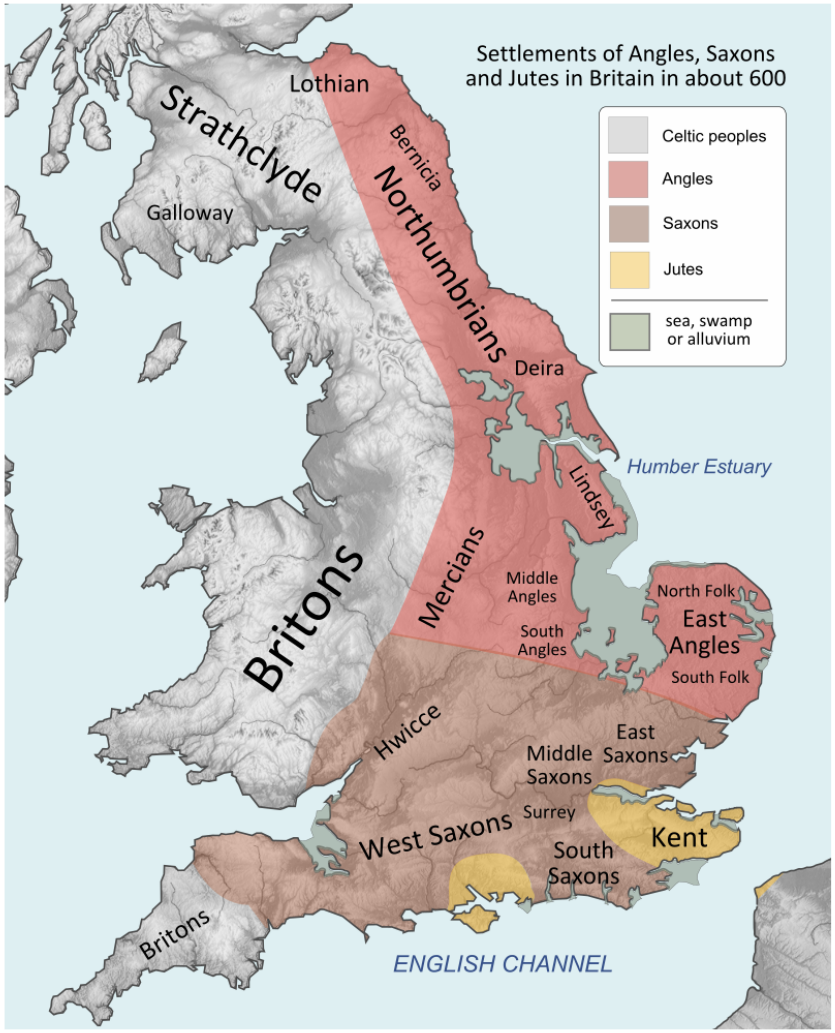
pass

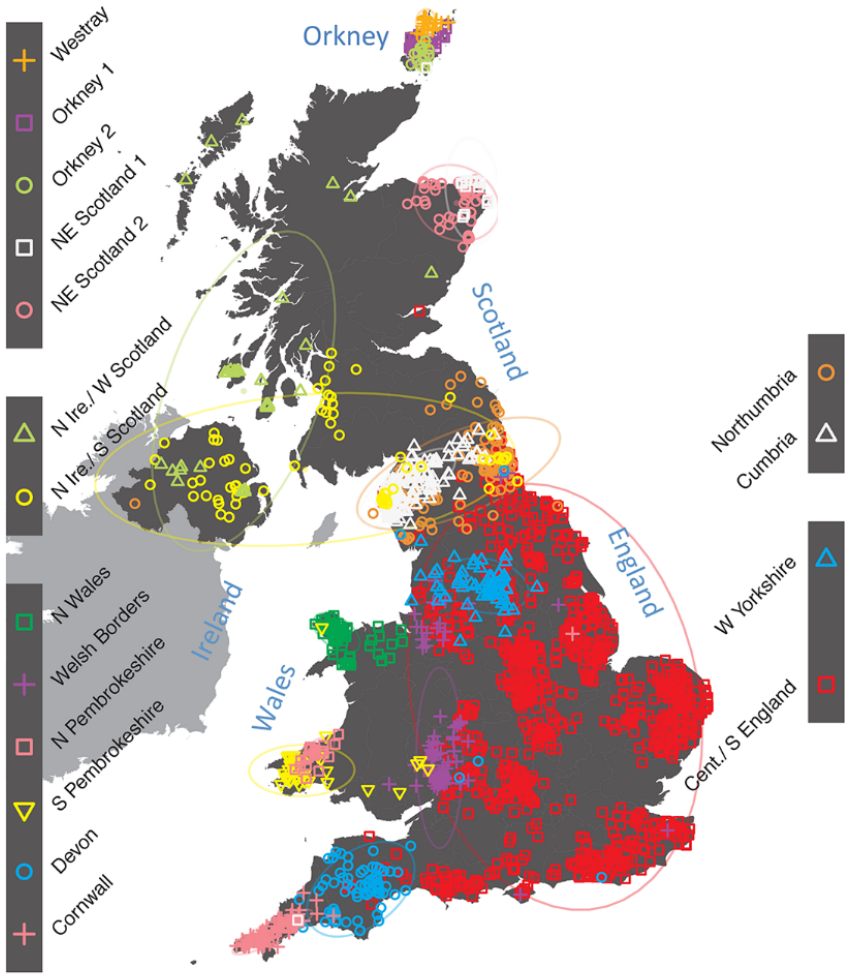


fail



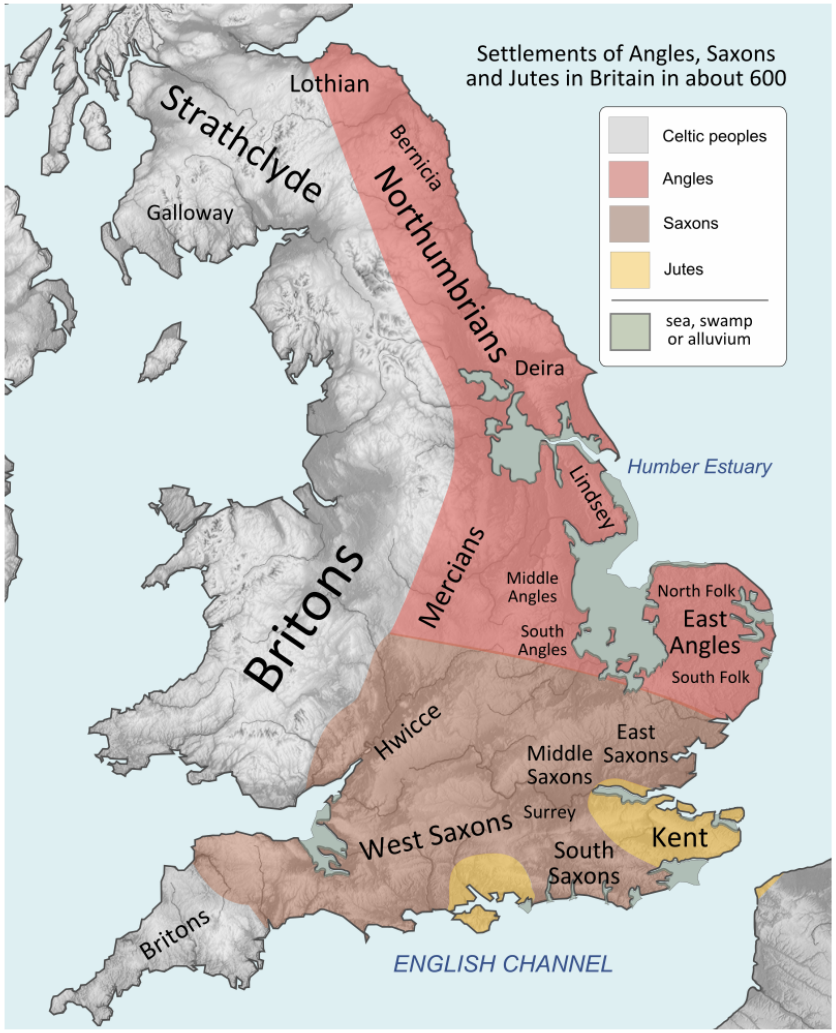
Diversion: population structure within the UK

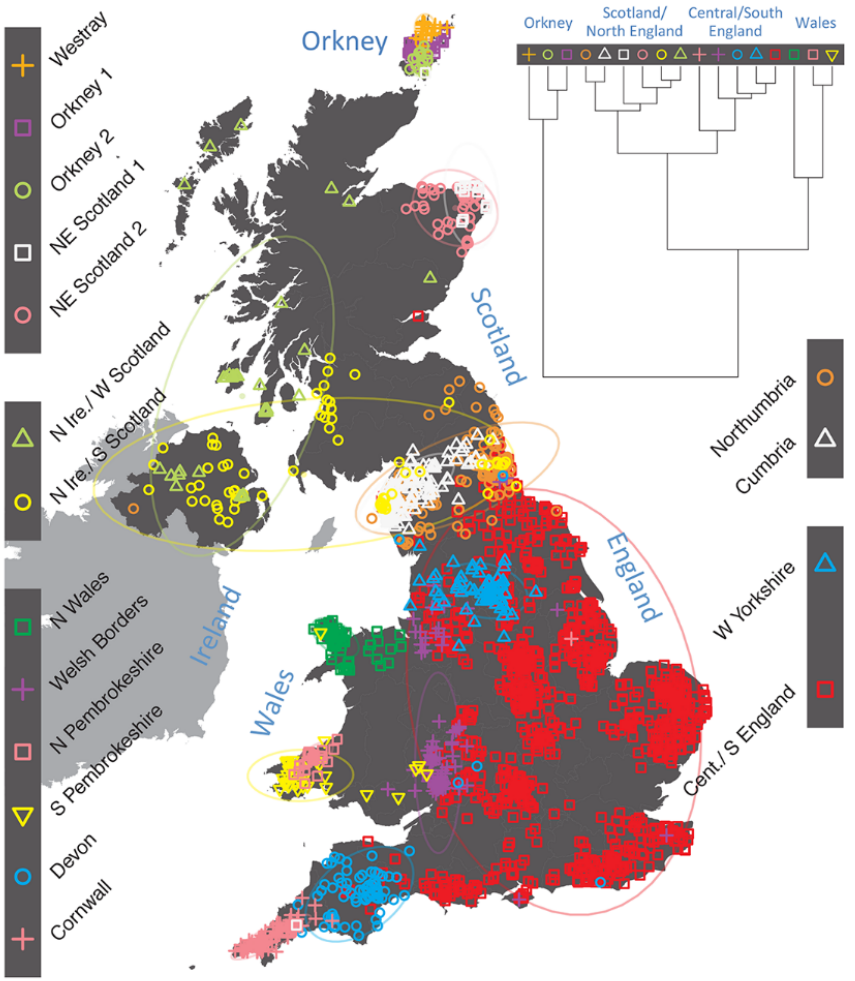




Leslie et al, Nature 2015 doi:10.1038/nature14230

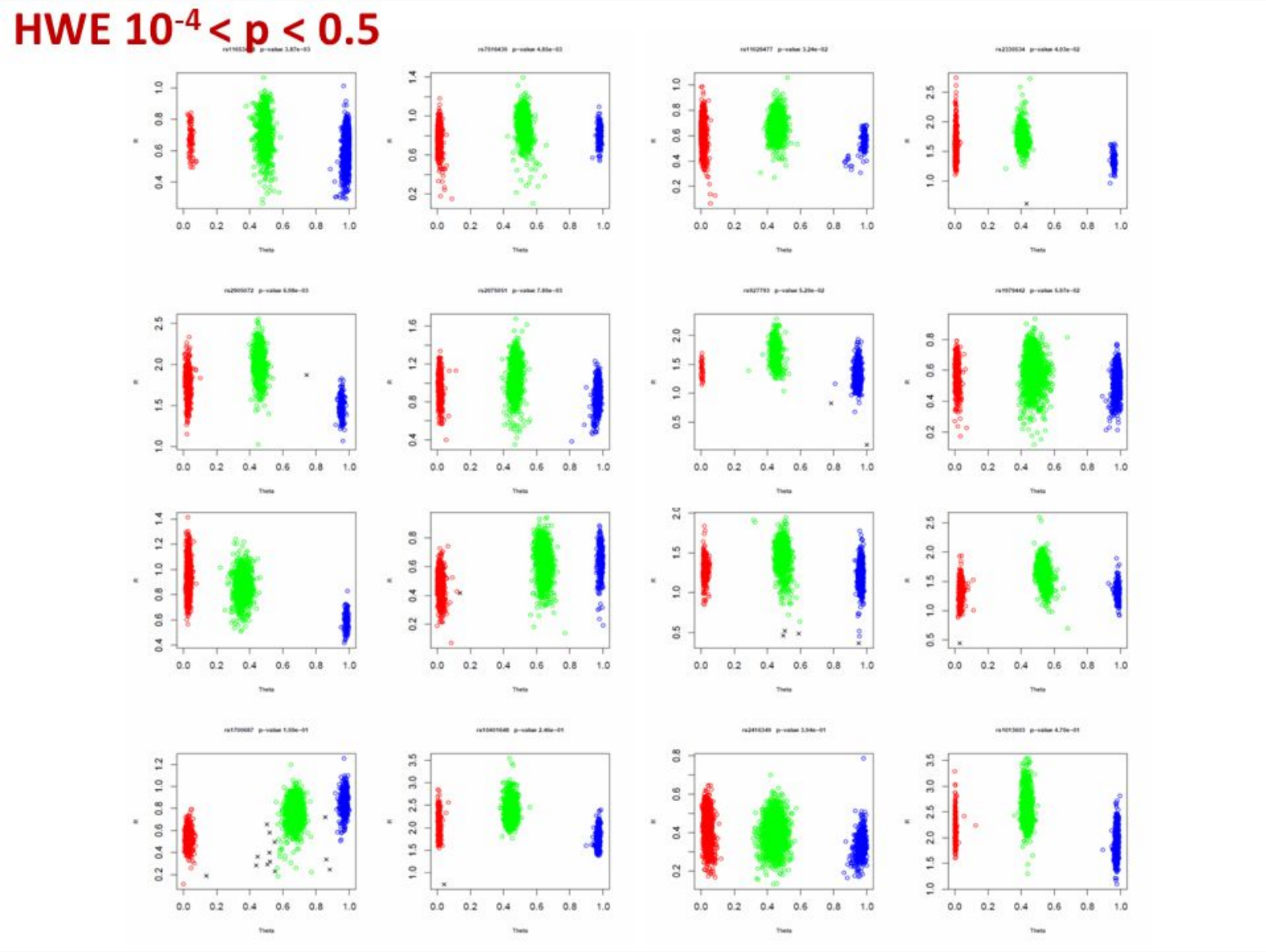
Diversion: population structure within the UK



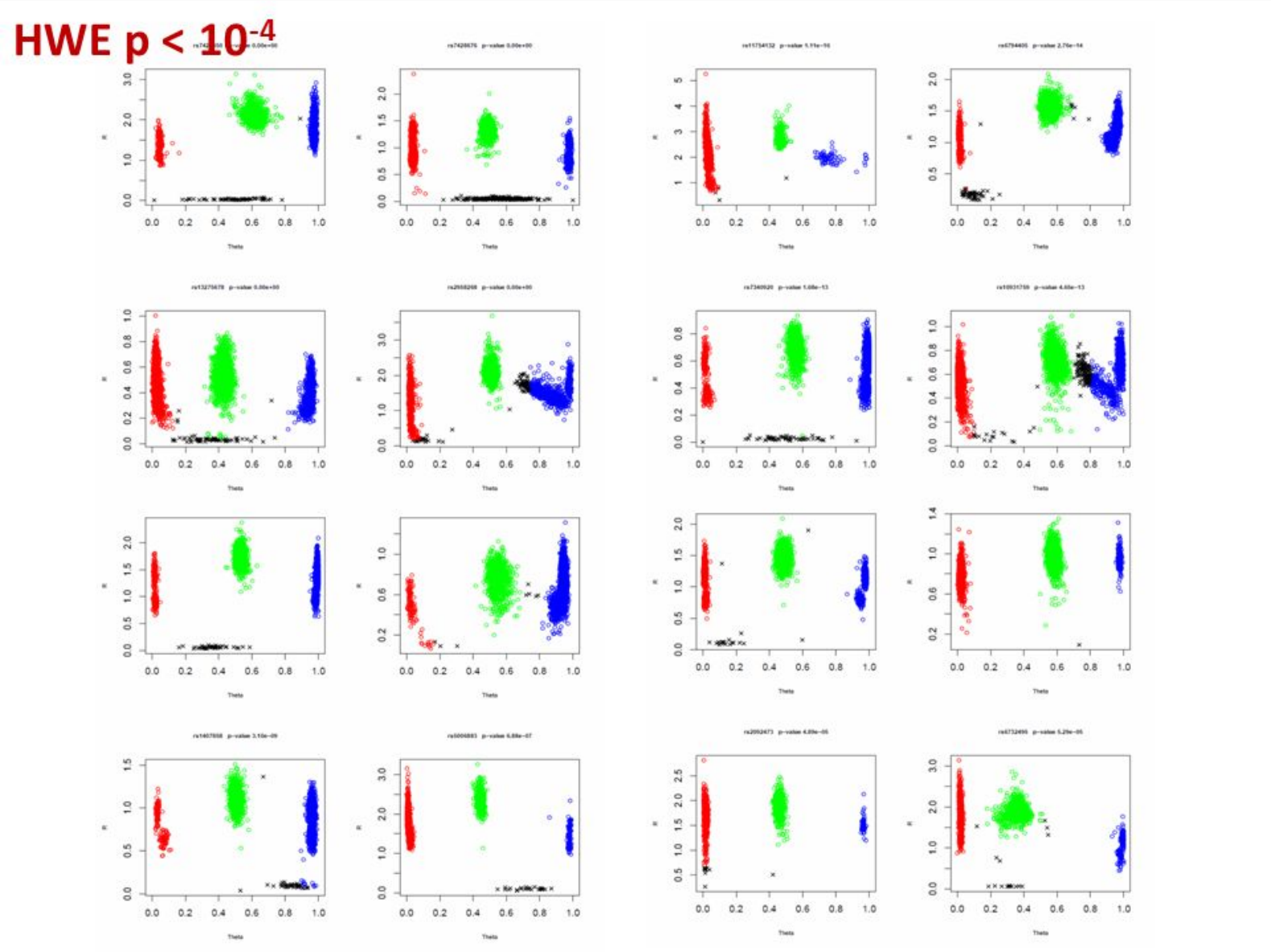


Leslie et al, Nature 2015 doi:10.1038/nature14230

Ancestry is not the only confounder



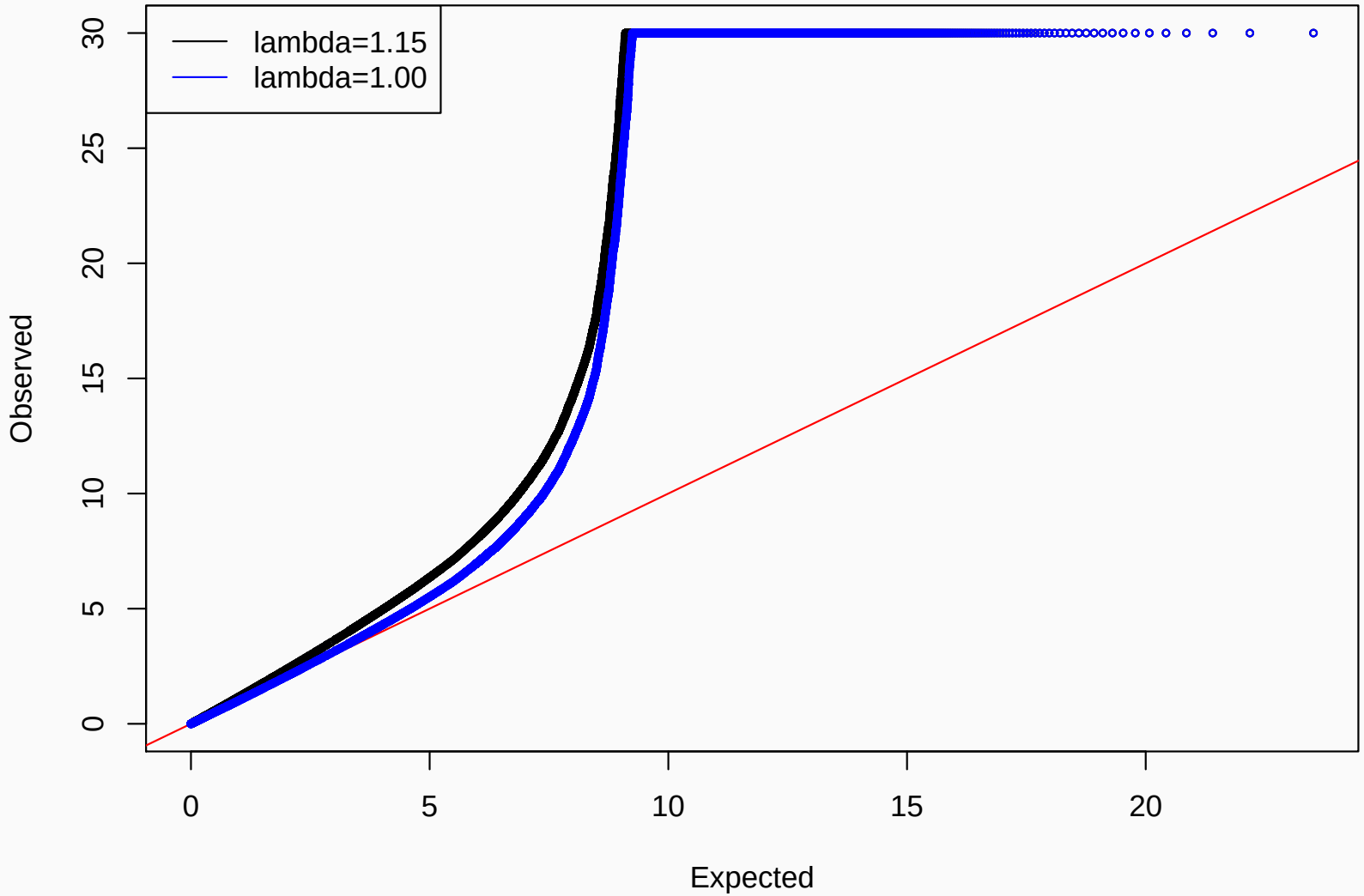
Ancestry is not the only confounder



Accounting for confounding

- by design: treat all samples equally, store in same place, mix cases and controls when running genotype arrays
- genomic control: divide all χ^2 stats by $\lambda = \frac{median(\chi^2)}{0.455}$; assumes **all** SNPs equally affected

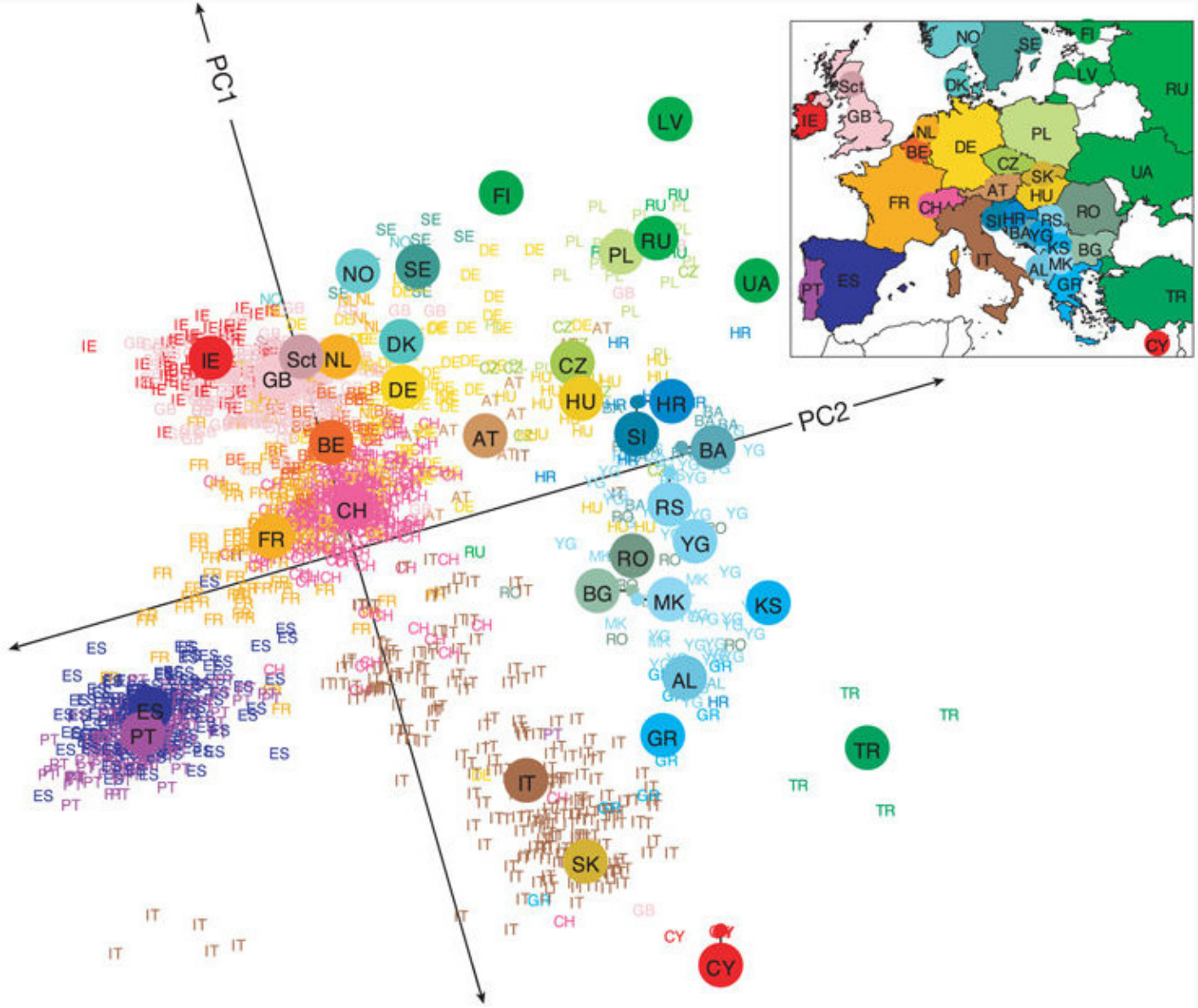
Accounting for confounding



Accounting for confounding

- by design: treat all samples equally, store in same place, mix cases and controls when running genotype arrays
- genomic control: divide all χ^2 stats by $\lambda = \frac{median(\chi^2)}{0.455}$; assumes **all** SNPs equally affected
- stratify by confounding factor (self reported ancestry/site of recruitment), eg in WTCCC
- perform PCA with individuals with known ancestry, project on gwas samples, use projected co-ordinates as covariates in testing

Accounting for confounding



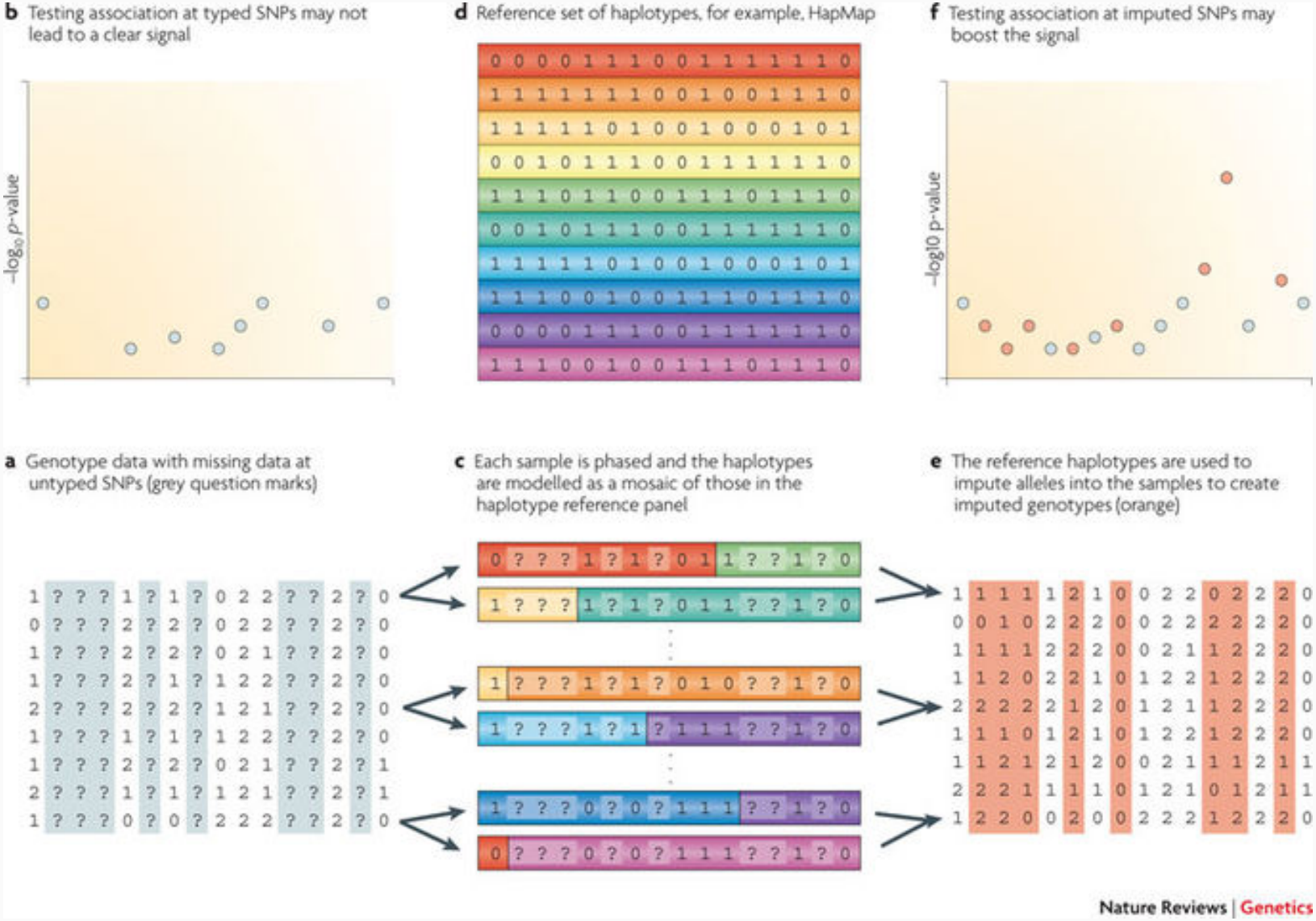
Novembre et al, Nat Genet, 2008

Accounting for confounding

- by design: treat all samples equally, store in same place, mix cases and controls when running genotype arrays
- genomic control: divide all χ^2 stats by $\lambda = \frac{median(\chi^2)}{0.455}$; assumes **all** SNPs equally affected
- stratify by confounding factor (self reported ancestry/site of recruitment), eg in WTCCC
- perform PCA with individuals with known ancestry, project on gwas samples, use projected co-ordinates as covariates in testing
- include measured non-ancestry covariates in tests, examine whether they reduce global inflation, measured by λ

Imputation to extend genome coverage

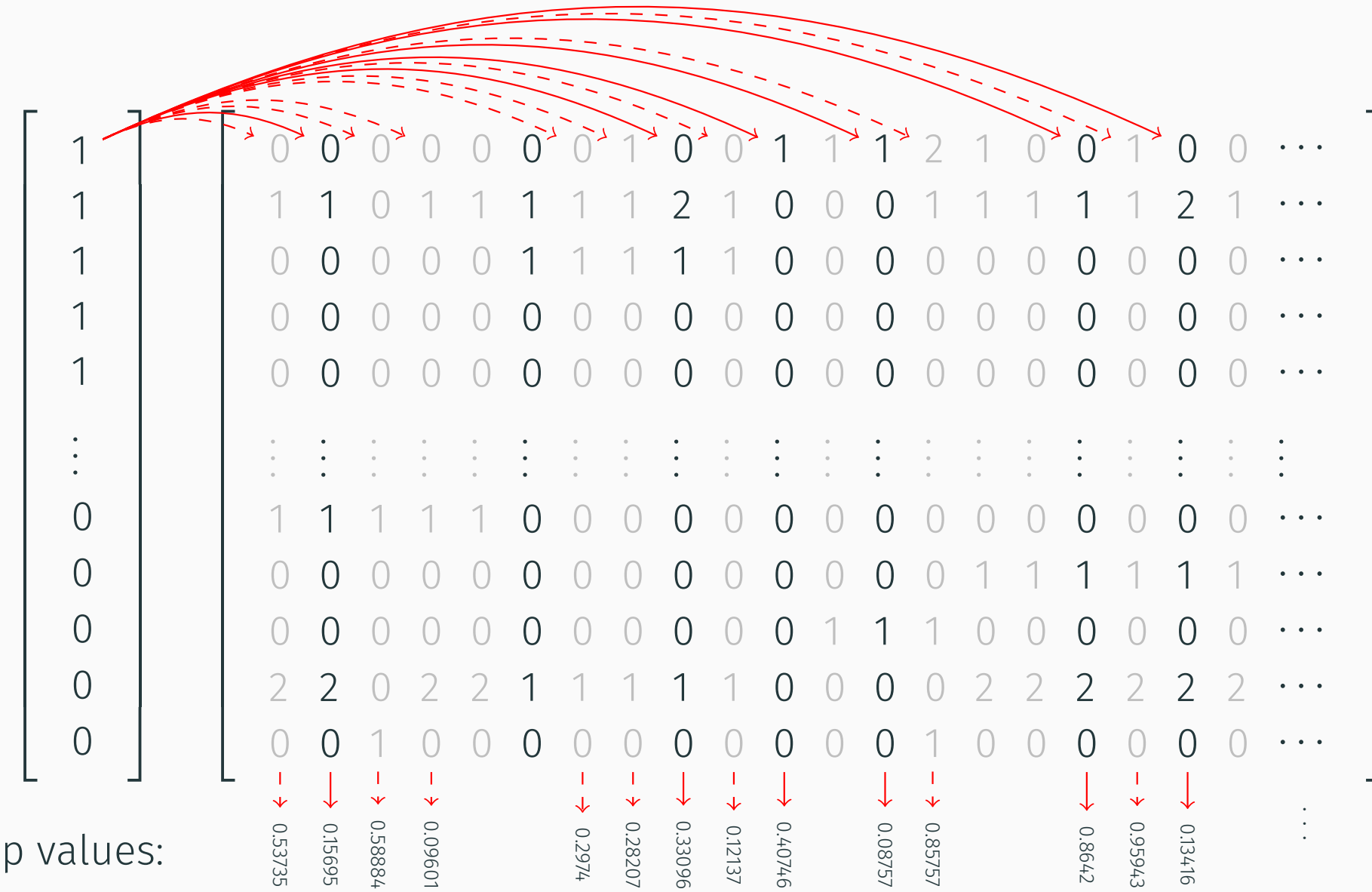
Imputation makes informed guesses about ungenotyped SNPs



Marchini and Howie, Nat Rev Genet 2010

Imputation extends genome coverage

Imputation extends genome coverage



Heritability

Most diseases are not Mendelian, yet are still partially heritable

Broad sense heritability is the proportion of population variance of a trait explained by genetics

Narrow sense heritability is the proportion of population variance of a trait explained by additive genetic effects

Most diseases are not Mendelian, yet are still partially heritable

Statistical model of a trait under polygenic control:

$$Y = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \cdots + \beta_m G_m + e$$

Variance of Y :

$$V(Y) = \underbrace{\beta_1 V(G_1) + \beta_2 V(G_2) + \cdots + \beta_m V(G_m)} + V_e = V_g + V_e$$

Covariance of Y :

$$\mathrm{Cov}(Y) = V_g R + V_e I$$

where R is relatedness matrix (can be estimated from genetic data).

This allows estimation of V_g and hence heritability

$$h^2 = V_g / V$$

Accounting for environmental bias

$$V(Y) = V_g + V_{e\sim g} + c_{g,e} + V_e$$

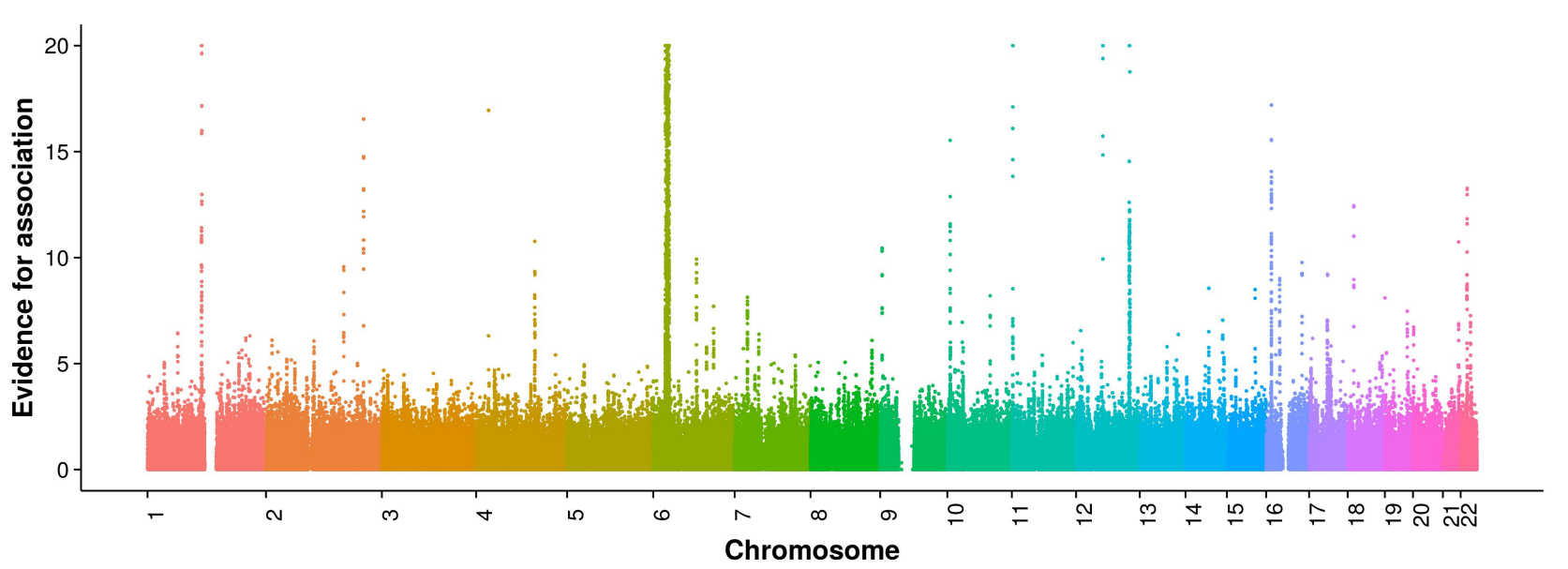
$$\mathrm{Cov}(Y) = V_g R + v_{e\sim g} R_{par} + c_{g,e} R_{o,par} + V_e I$$

Trait	n	sib-pairs	RDR	Kinship F.E.
BMI	19,589	56,461	28.9 (6.3)	46.7 (2.5)
height	21,802	64,847	55.4 (4.4)	78.0 (1.9)
AFCW	22,367	30,582	22.6 (6.0)	33.5 (2.1)
AFCM	17,117	21,729	14.9 (7.9)	16.3 (2.6)
menarche	11,242	16,621	30.9 (10.5)	41.9 (4)
education	12,035	32,542	17.0 (9.4)	52.4 (3.7)
total chol.	27,320	74,271	30.6 (5.0)	32.2 (1.8)
HDL	24,570	67,894	44.8 (5.3)	45.1 (2.1)

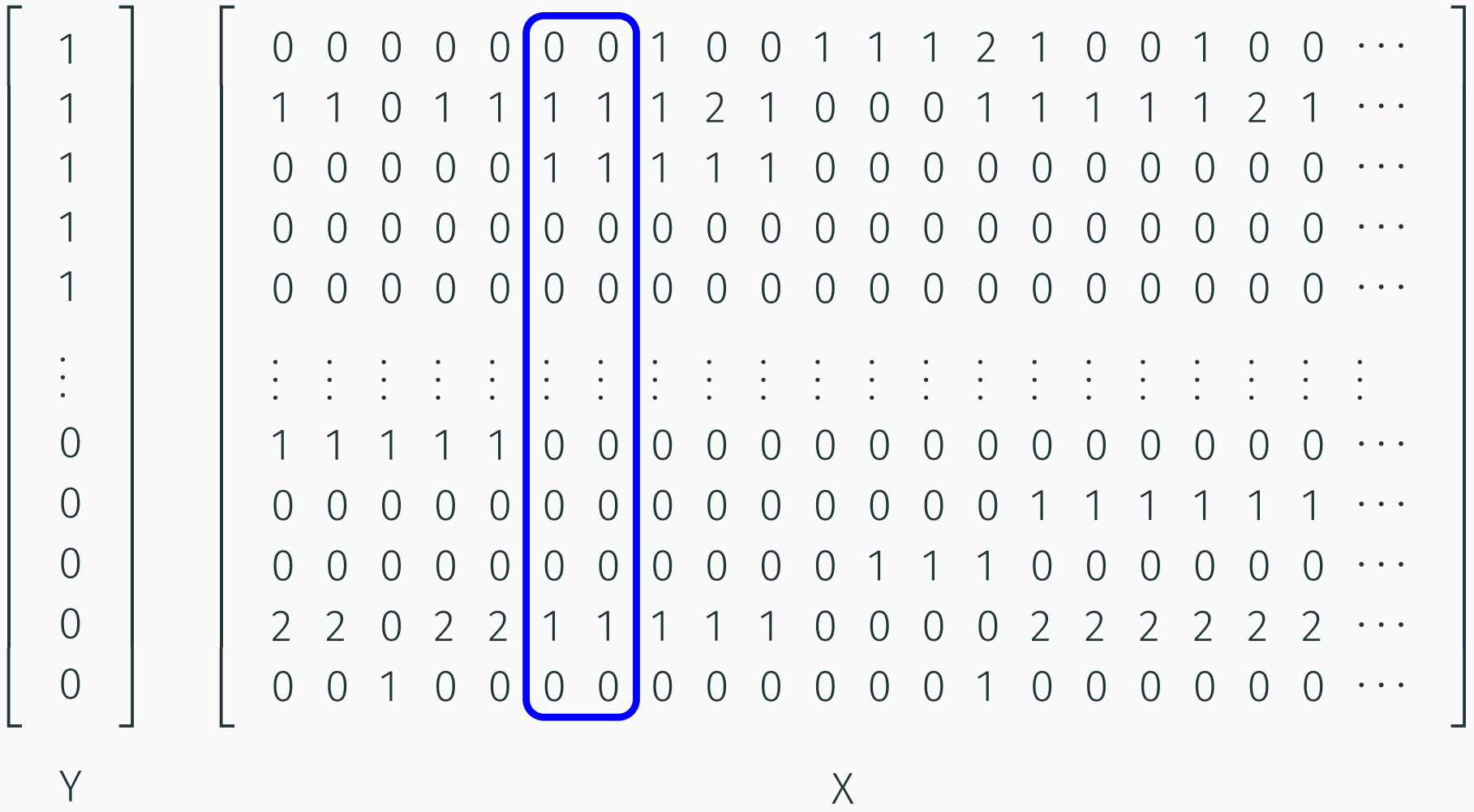
Young et al, https://www.biorxiv.org/content/early/2017/11/14/218883

Inferring disease relevant cells

Association does not identify the causal variant



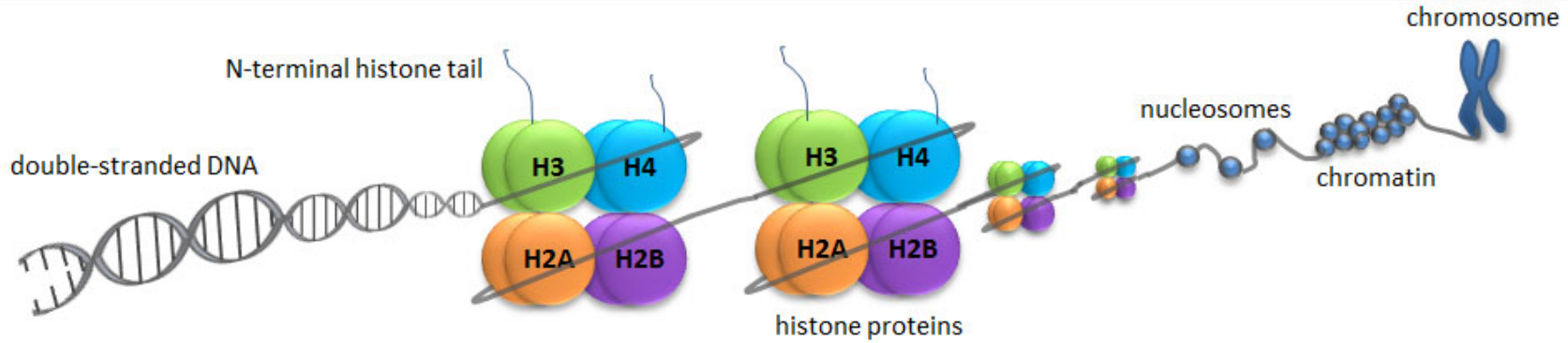
Association does not identify the causal variant



We can learn something from location of *associated* variants

DNA is packed around histone proteins into nucleosomes

Modifications of N-terminal tails of histones vary with accessibility of DNA sequence

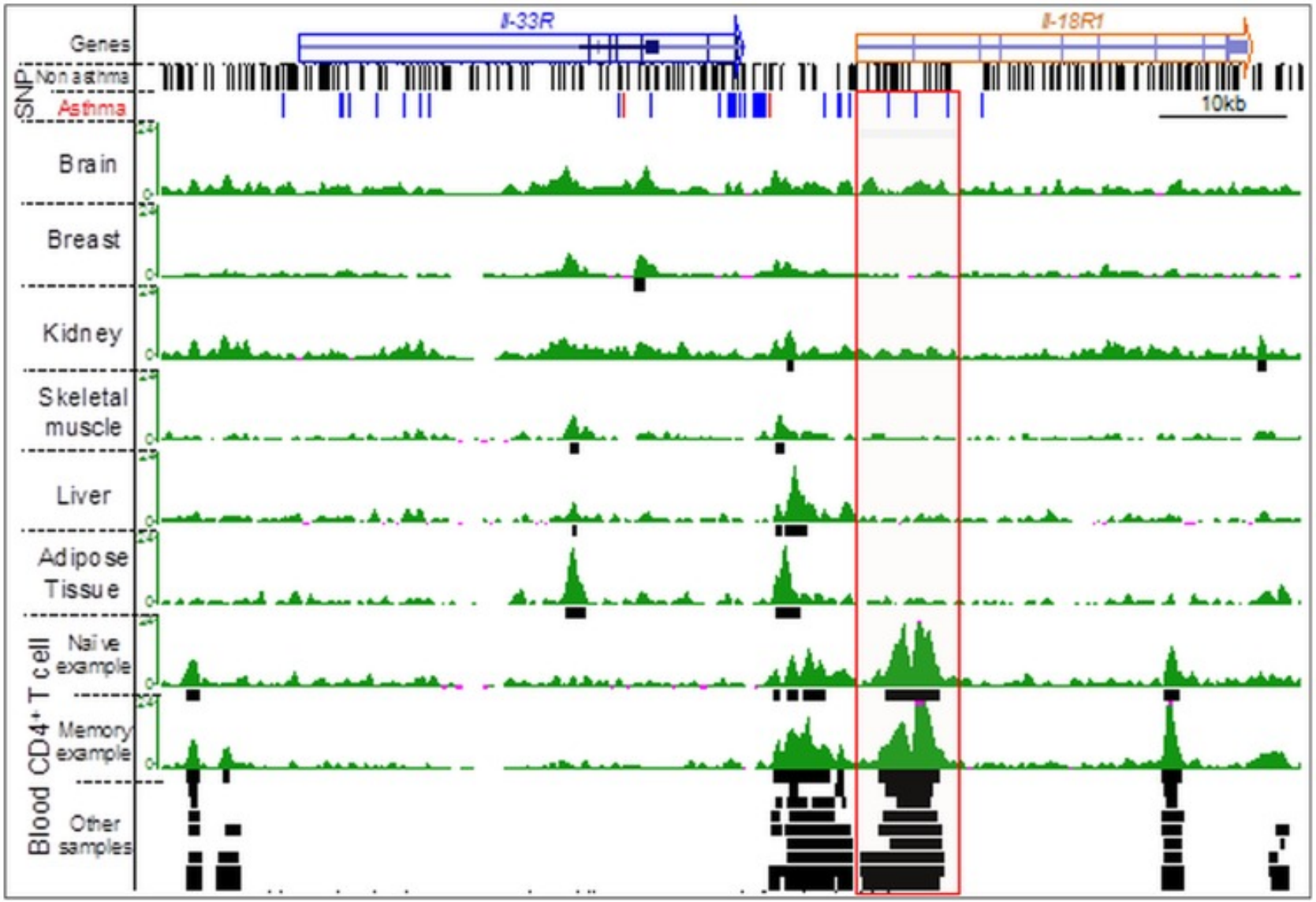


Resources:

ENCODE `https://genome.ucsc.edu/encode`

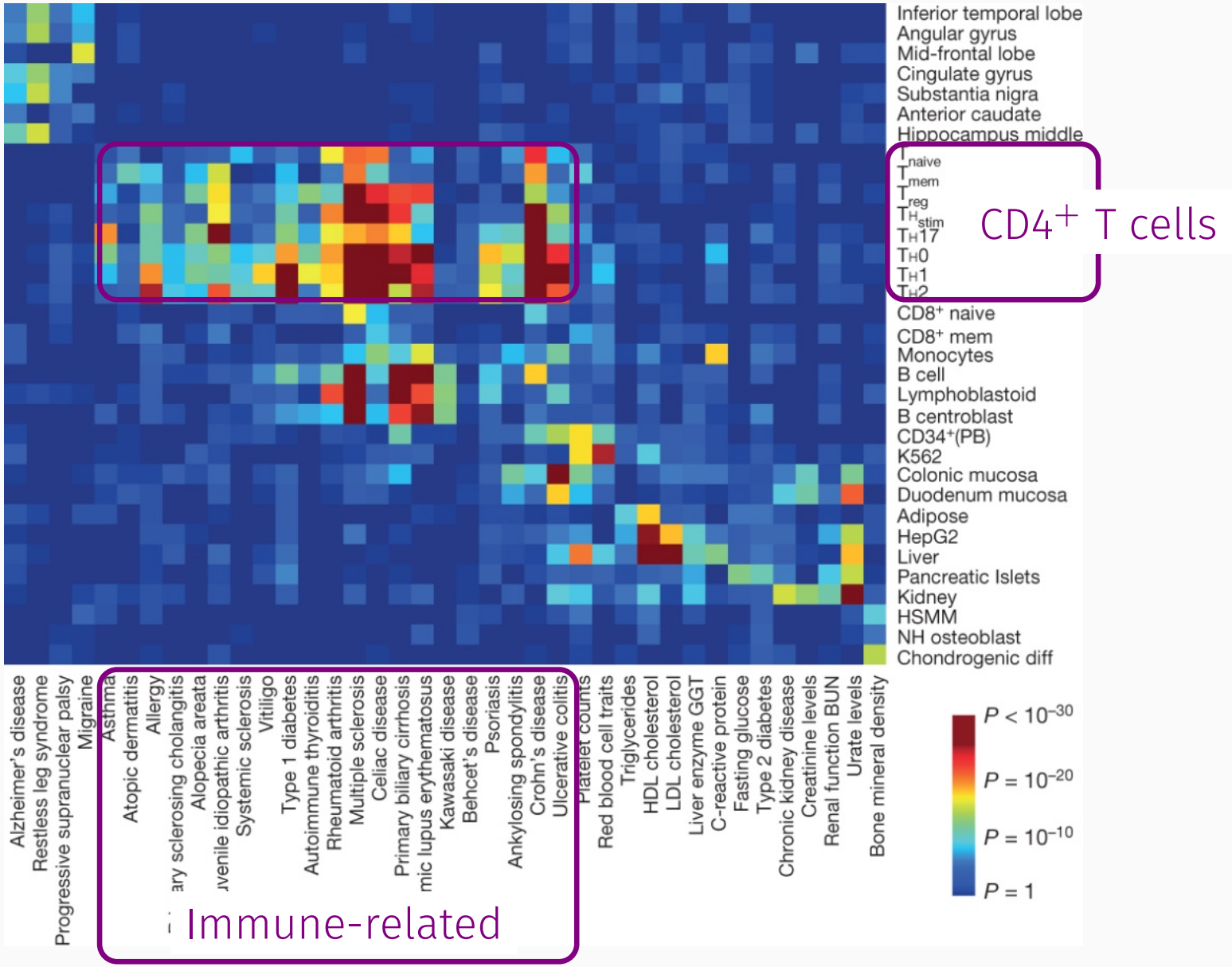
BLUEPRINT `http://www.blueprint-epigenome.eu`

We can learn something from location of *associated* variants



Gerasimova et al, PLOS One 2013

Enrichment of enhancer marks links cells to disease



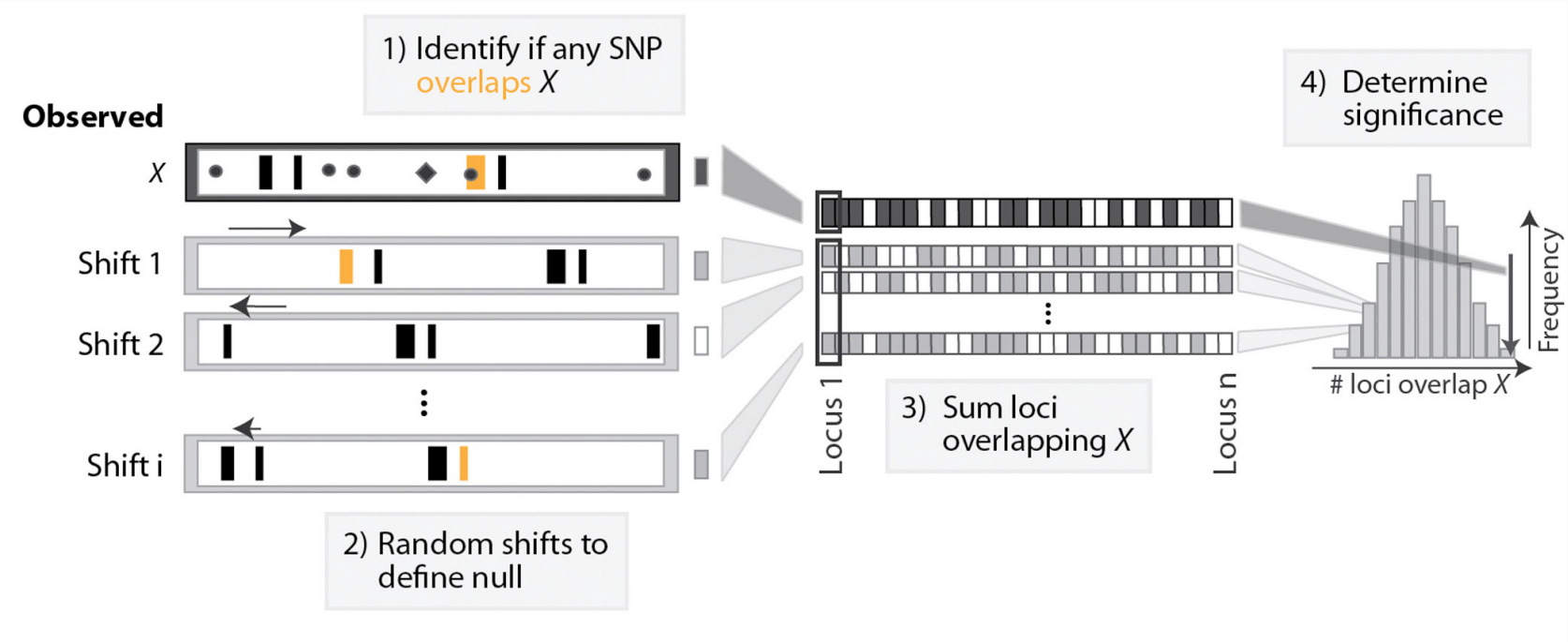
Farh et al., Nature 2014

LD (again)

SNPs are not independent (LD). Chromatin marks spatially cluster.

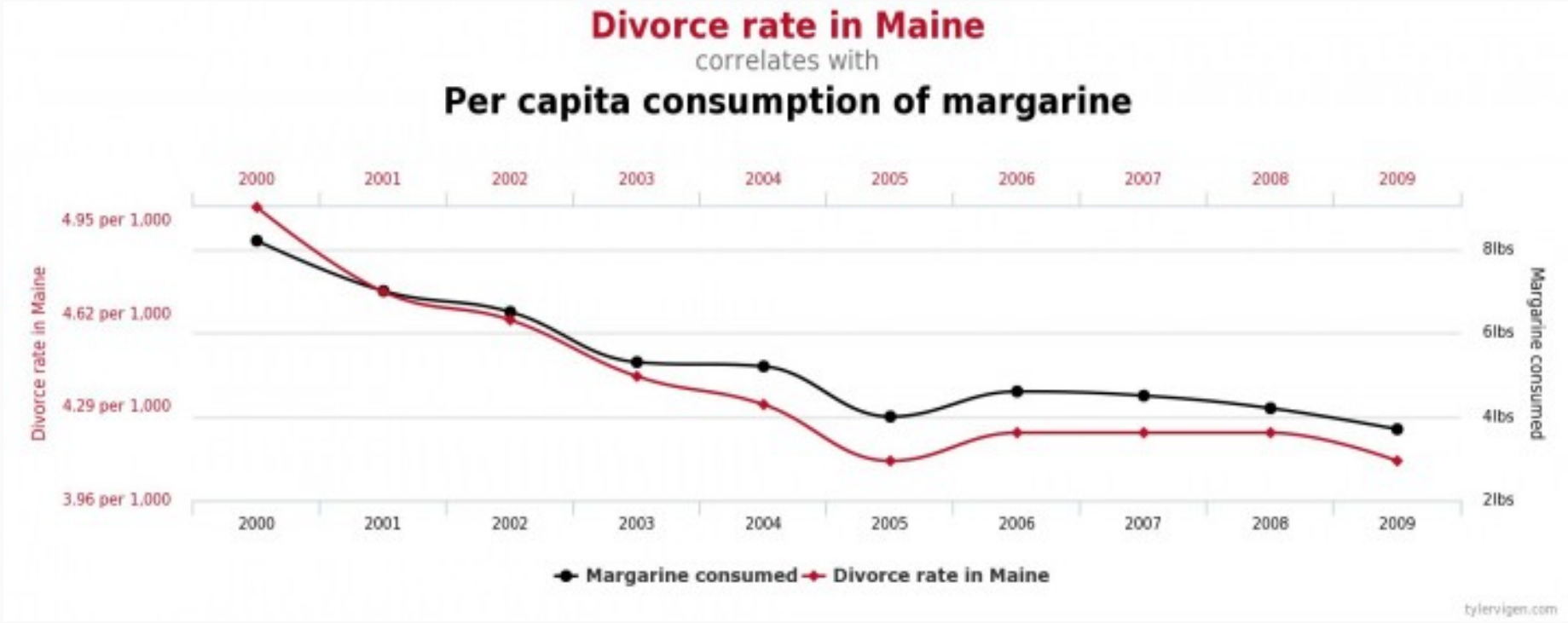
Ignoring LD makes p values more significant - but is still wrong.

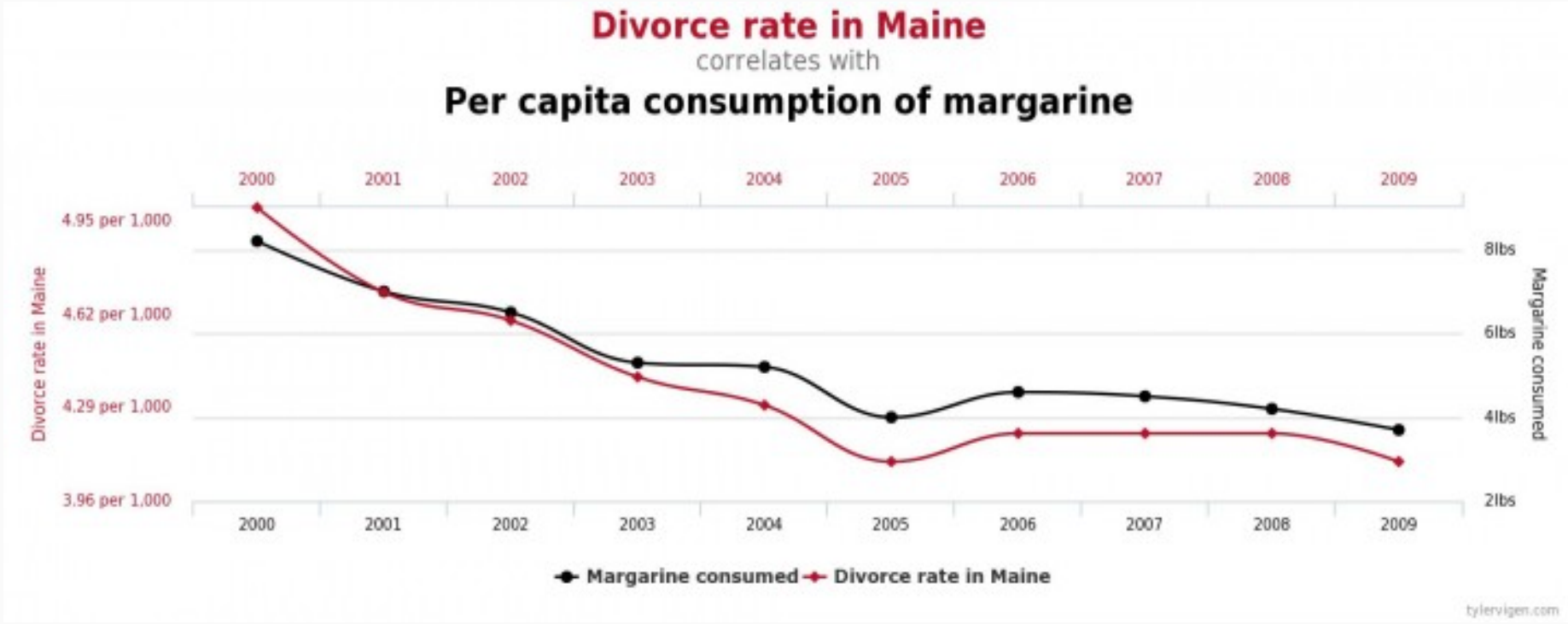
Methods to account for LD maintain the structure of SNP distances when computing null distribution of test statistics.

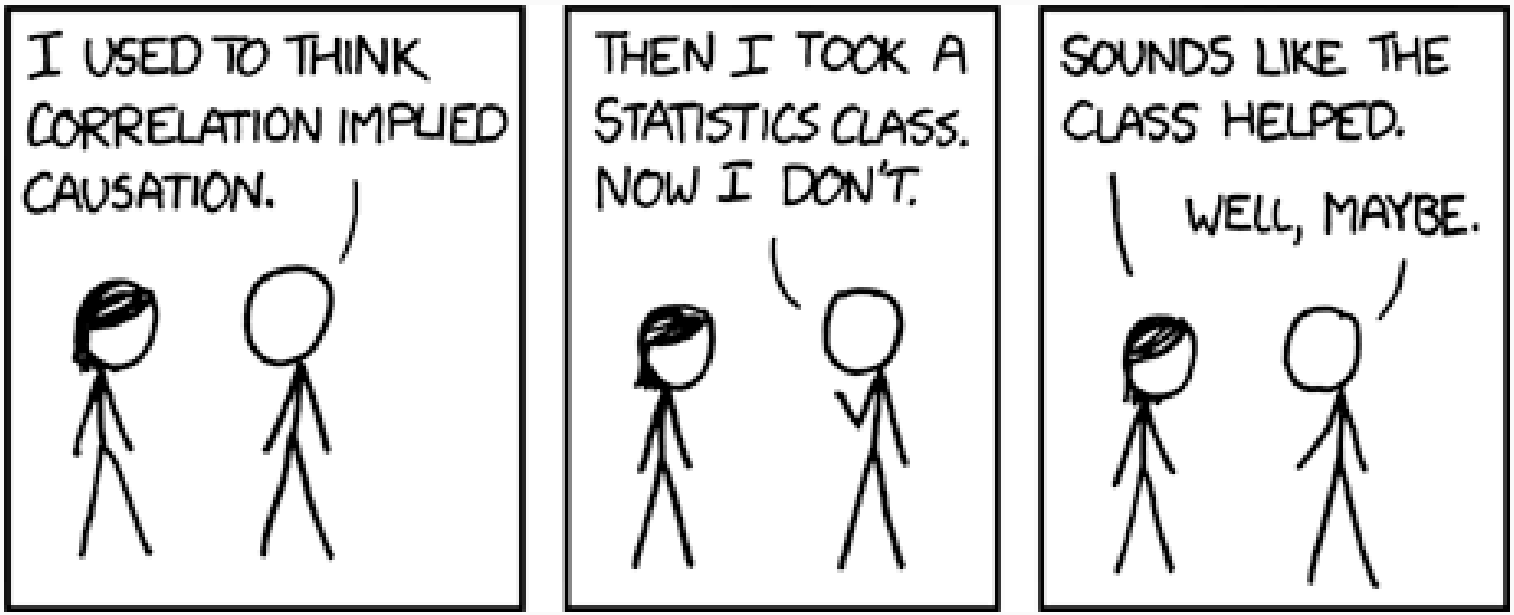


Summary

- Inheritance of disease
- Genome-wide association studies
- Confounding in GWAS
- Imputation to extend genome coverage
- Heritability
- Inferring disease relevant cells


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.