GWAS Experimental Design: phenotypes

Outline of lecture

- Types of phenotypes
  - Quantitative vs. binomial traits
- Genetic architecture
- Population structure
  - Dealing with population structure & confounders
- QC of phenotypes

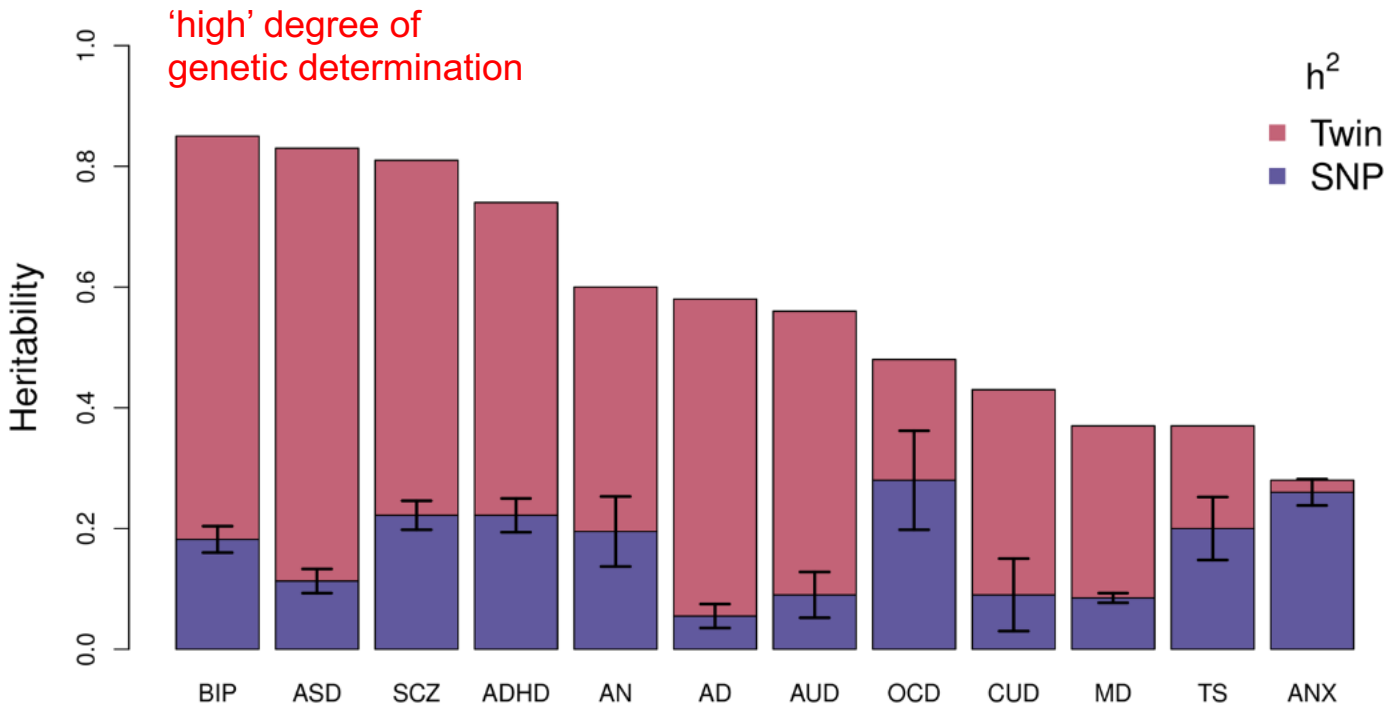
What is a phenotype?

- A **phenotype** is an observable trait
  - it is influenced by both genetic and environmental (non-genetic) factors

- Traits are typically either:
  - A **quantitative trait** is a trait that shows (measured) continuous variation, e.g. height, weight
  - A **binary trait** is a trait where individuals can be classified into two groups, e.g. disease status

Genetic influence on a trait

- The degree of genetic influence it is quantified by the **heritability** of a trait
- **heritability** is defined as the proportion of phenotypic variance explained by genetic variance
  - Ranges from 0 to 1
  - varies between traits
  - varies between estimation approaches



- *psychiatric* (**BIP**, bipolar disorder; **SCZ**, schizophrenia; **ADHD**, attention-deficit/hyperactivity disorder; **MD**, major depression; **ANX**, generalized anxiety disorder),
- *behavioural* (**AN**, anorexia nervosa; **AUD**, alcohol use disorder; **CUD**, cannabis use disorder), or
- *neurological* (**ASD**, autism spectrum disorder; **AD**, Alzheimer's disorder; **OCD**, obsessive-compulsive disorder; **TS**, Tourette's syndrome).

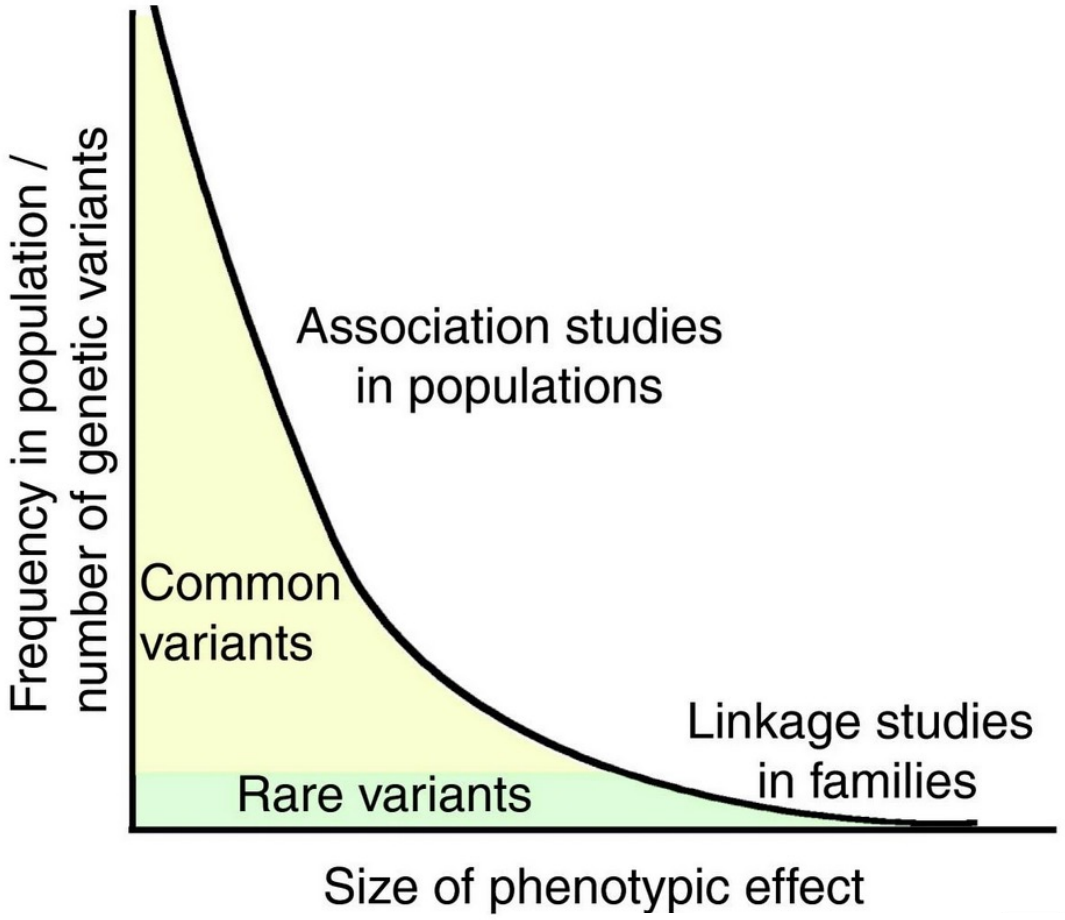
O'Connell & Coombes (2021) *Psychol Med.*

Genetic architecture

- **Genetic architecture** refers to the joint distribution of allele effect size and allele frequency, i.e. the number of loci, their effect size and frequency

- In GWAS we have best (statistical) power to detect *common variants*, e.g. alleles with frequency > 1%

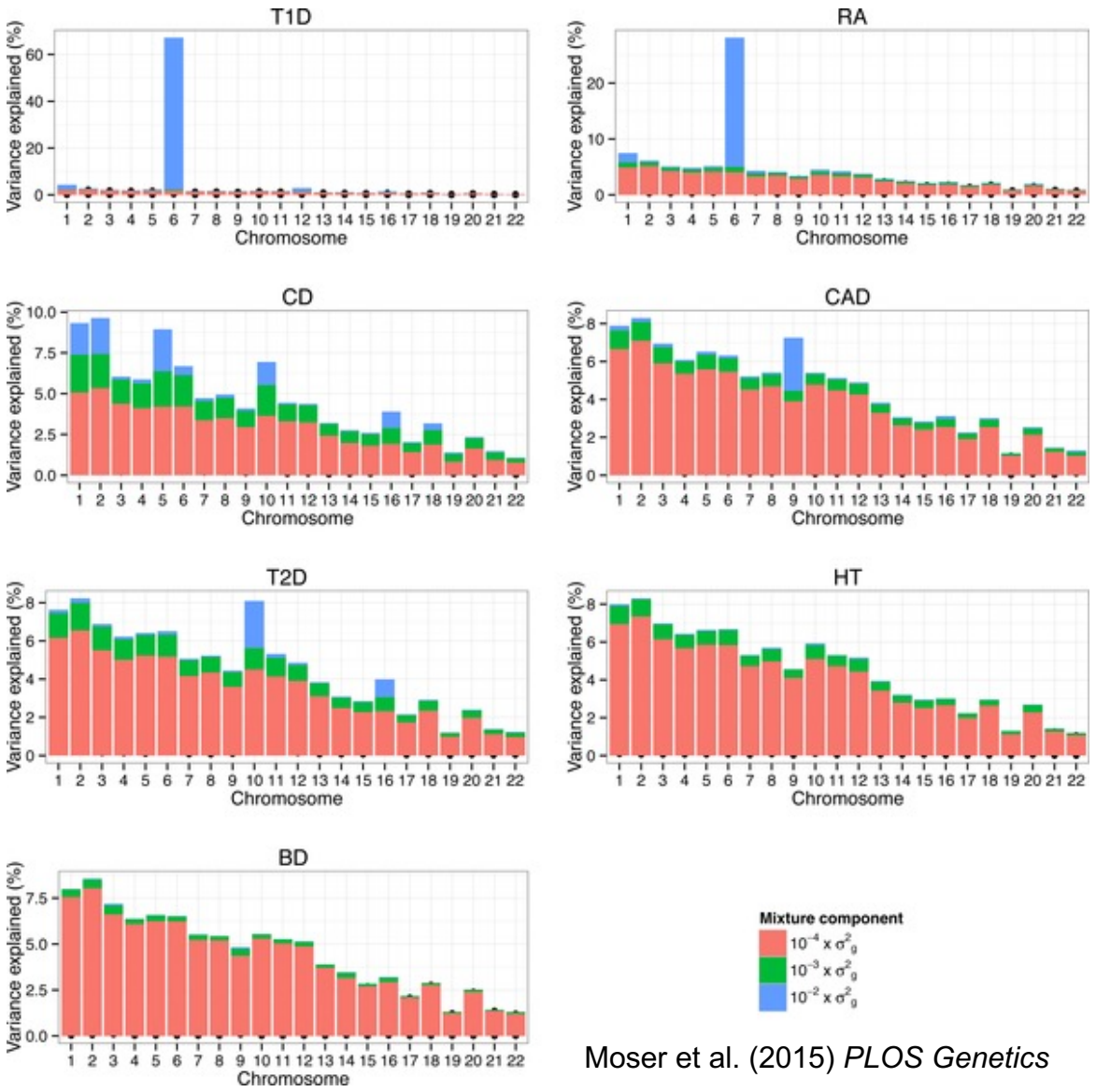
- *Common variants* tend to have smaller effect sizes



Rahim et al. (2008) *Genome Biology*

Genetic architecture

- **Genetic architecture** differs between traits, even when heritability is similar
- Some traits (e.g. T1D or RA) have loci with big effects + many loci with small effects
- Other traits (e.g. HT = height) have small effects spread evenly throughout the genome
  - i.e. variance of HT explained per chromosome is proportional to chromosome length

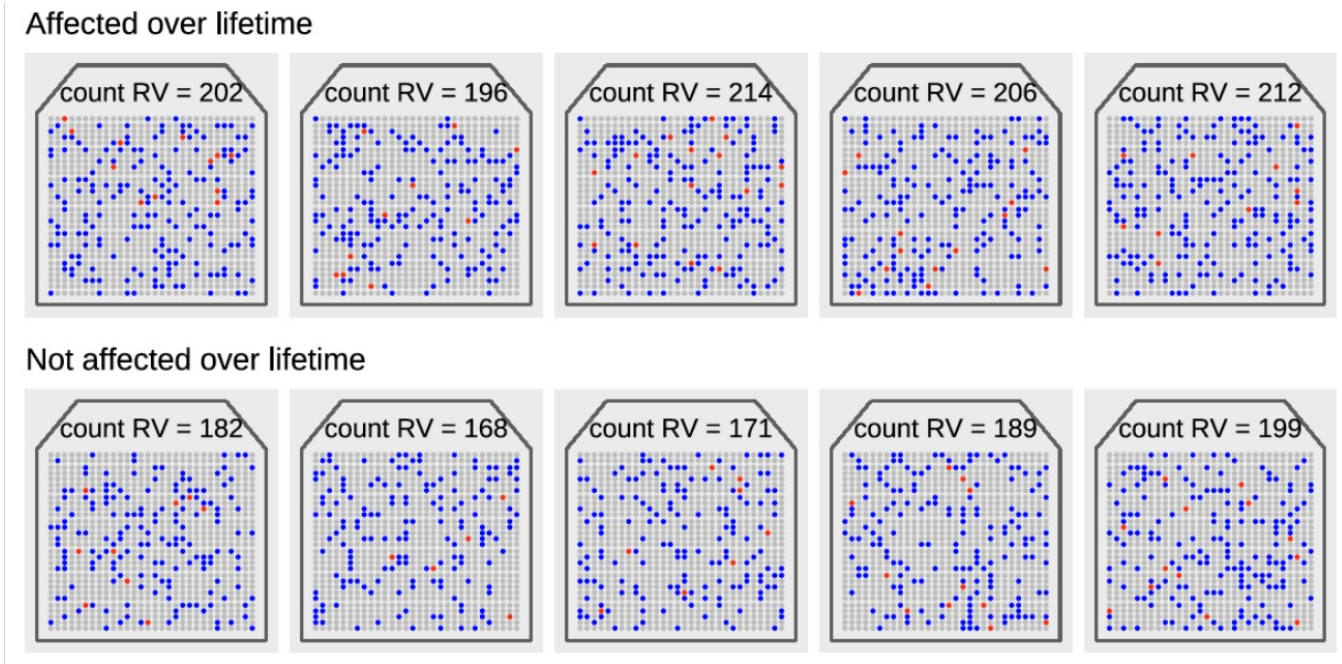


Moser et al. (2015) *PLOS Genetics*

Most traits are affected by many loci

What does this mean for disease traits?

- Everybody carries risk variants
- On average, affected individuals have higher burden of risk alleles
- Non-genetic (environmental) factors contribute to risk as well
- Each individual carries a unique risk profile



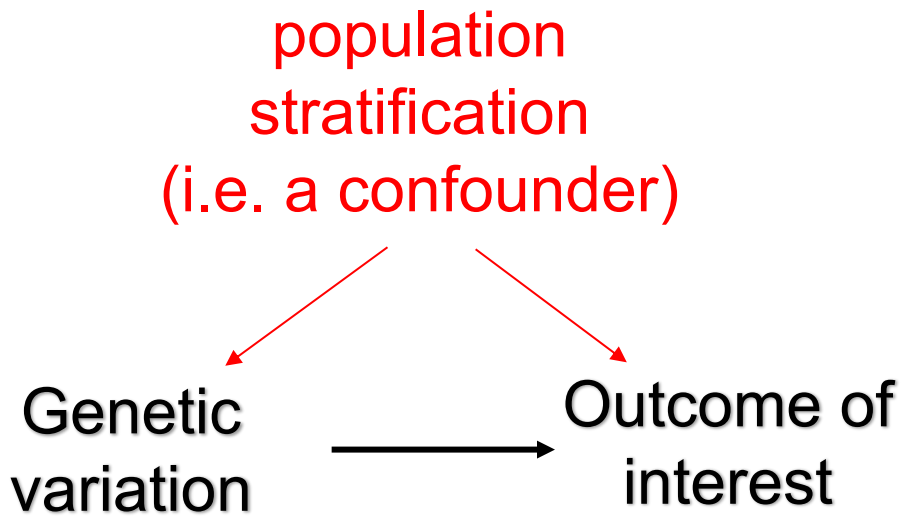
RV = risk variant

Slide adapted from Prof Naomi Wray

Population stratification

- **Population stratification** is a major source of bias in GWAS
- it creates spurious genotype-phenotype associations

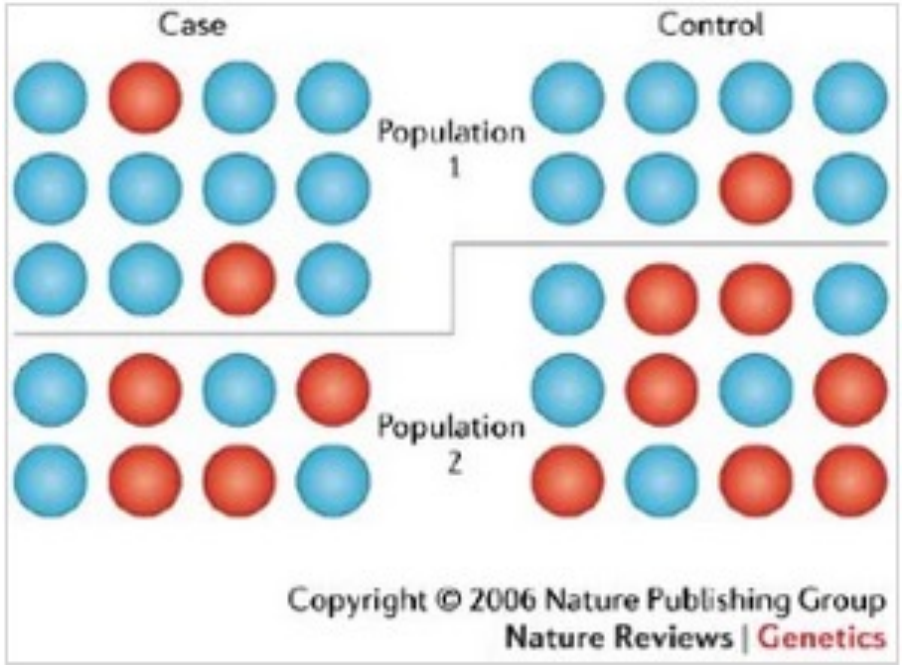
- Occurs when there are (unknown) subpopulations within the study sample which have *systematic differences in both ancestry and phenotypes*



Population stratification

- **Population stratification** is a major source of bias in GWAS
- it creates spurious genotype-phenotype associations

- Occurs when there are (unknown) subpopulations within the study sample which have *systematic differences in both ancestry and phenotypes*
- e.g. when one subpopulation contributes more cases to a case-control GWAS



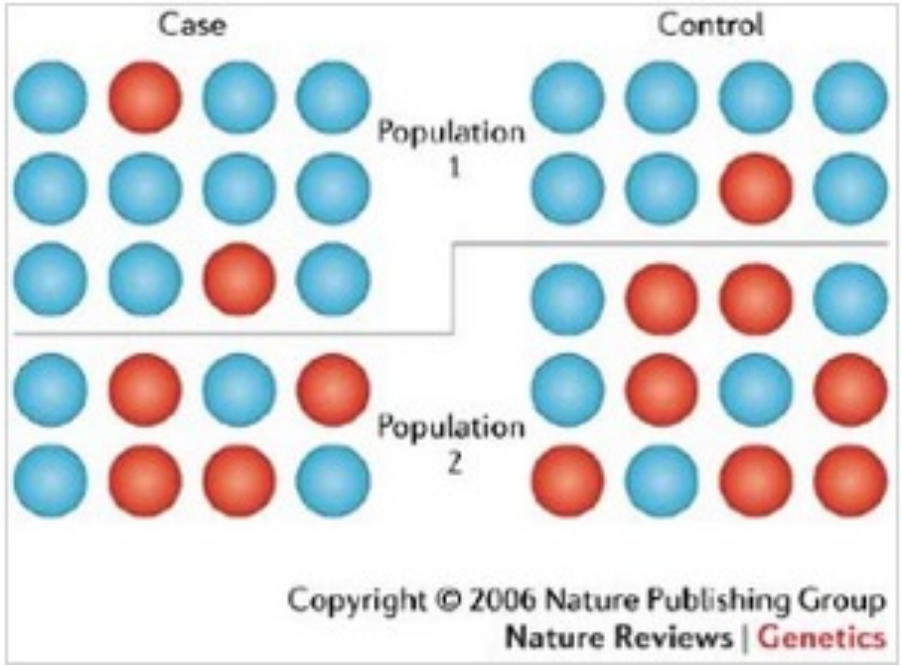
	Case	Control
ALL	14/20 = 0.7	12/20 = 0.6

Balding (2006) *Nature Rev. Genetics*

Population stratification

- **Population stratification** is a major source of bias in GWAS
- it creates spurious genotype-phenotype associations

- Occurs when there are (unknown) subpopulations within the study sample which have *systematic differences in both ancestry and phenotypes*
- e.g. when one subpopulation contributes more cases to a case-control GWAS



	Case	Control
Pop 1	10/12 = 0.83	7/8 = 0.87
Pop 2	4/8 = 0.5	5/12 = 0.41
ALL	14/20 = 0.7	12/20 = 0.6

Balding (2006) *Nature Rev. Genetics*

Population stratification

- Can also occurs for quantitative/continuous traits when systematic differences in means between subpopulations

- e.g. Campbell et al. performed a GWAS on two groups of individuals of European descent that were discordant for height and identified an association with the LCT (lactase) locus

	Height (Adult men)	Lactose Tolerance
Northern (Sweden)	5 ft 11 1/2 in	98%
Southern (Italy)	5 ft 9 1/2 in	$\sim$ 50%

Campbell et al. (2005) *Nature Genetics*

Close relatives can also cause bias

- Close relatives tend to share genetic variants AND environmental effects.
- This can bias the GWAS results → just like population stratification
- Close relatives tend to have similar genotypes & phenotypes, they are not independent

- e.g. if we have two related cases in a case-control analysis, their genotypes being on average more similar to each other than the rest of the cohort will provide a slight bias to the estimate of the allele frequency in cases and its associated standard error

- Even this small bias is important when considering the number of statistical tests being performed.

Dealing with population structure

1. Study design, match case-control samples for ancestry or other confounders

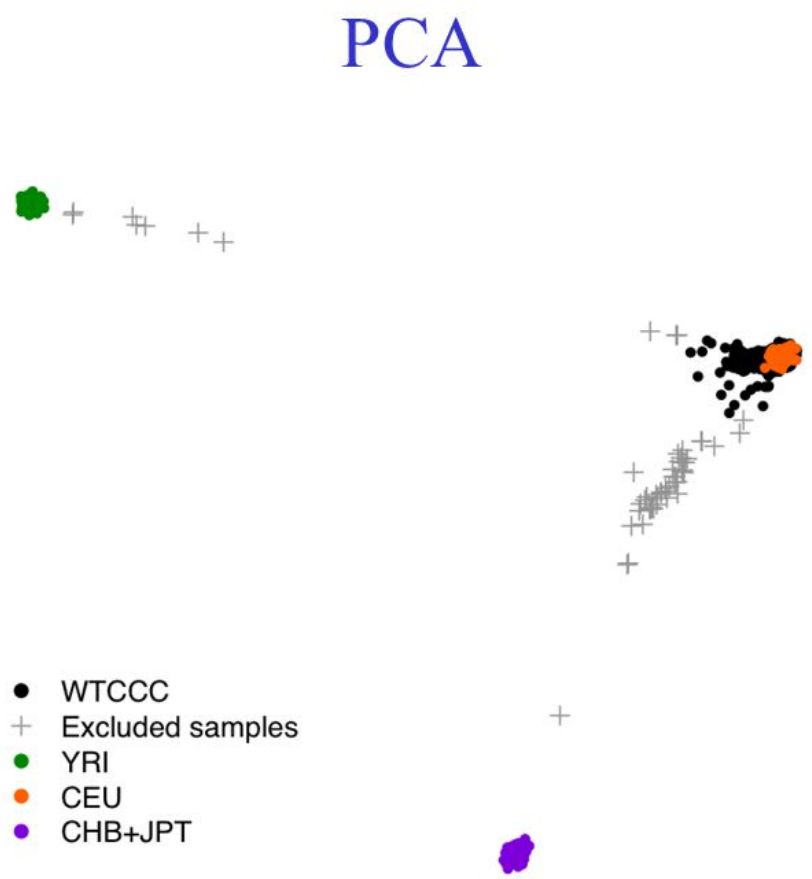
2. Remove individuals, e.g. ancestral outliers or one member of close relative pair

3. Attempt to account for the structure during statistical tests, e.g. fitting PCs as covariates to account for ancestry differences, or use a mixed model (with a genomic relationship covariance matrix) to account for close relatives

Dealing with population structure (1)

remove ancestry outliers

1. Perform PCA on the genotypes of a diverse set of individuals with known ancestry, e.g. 1000 Genomes
2. Project your samples onto PCs
3. Exclude outliers from further analysis



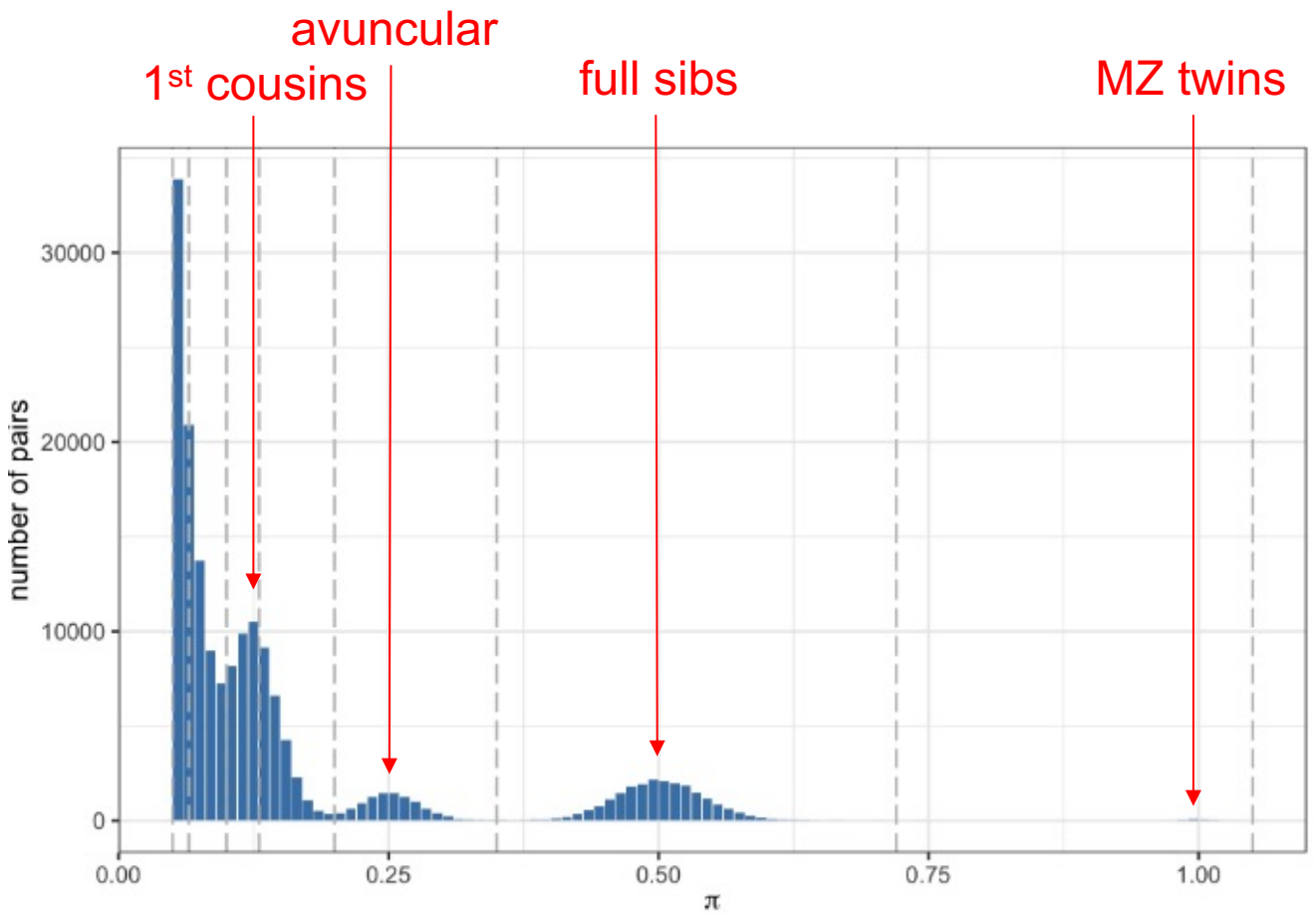
Dealing with population structure (2)

remove 1 member of a close relative pair

- We can detect related individuals by calculating their 'genomic relationship' (π)

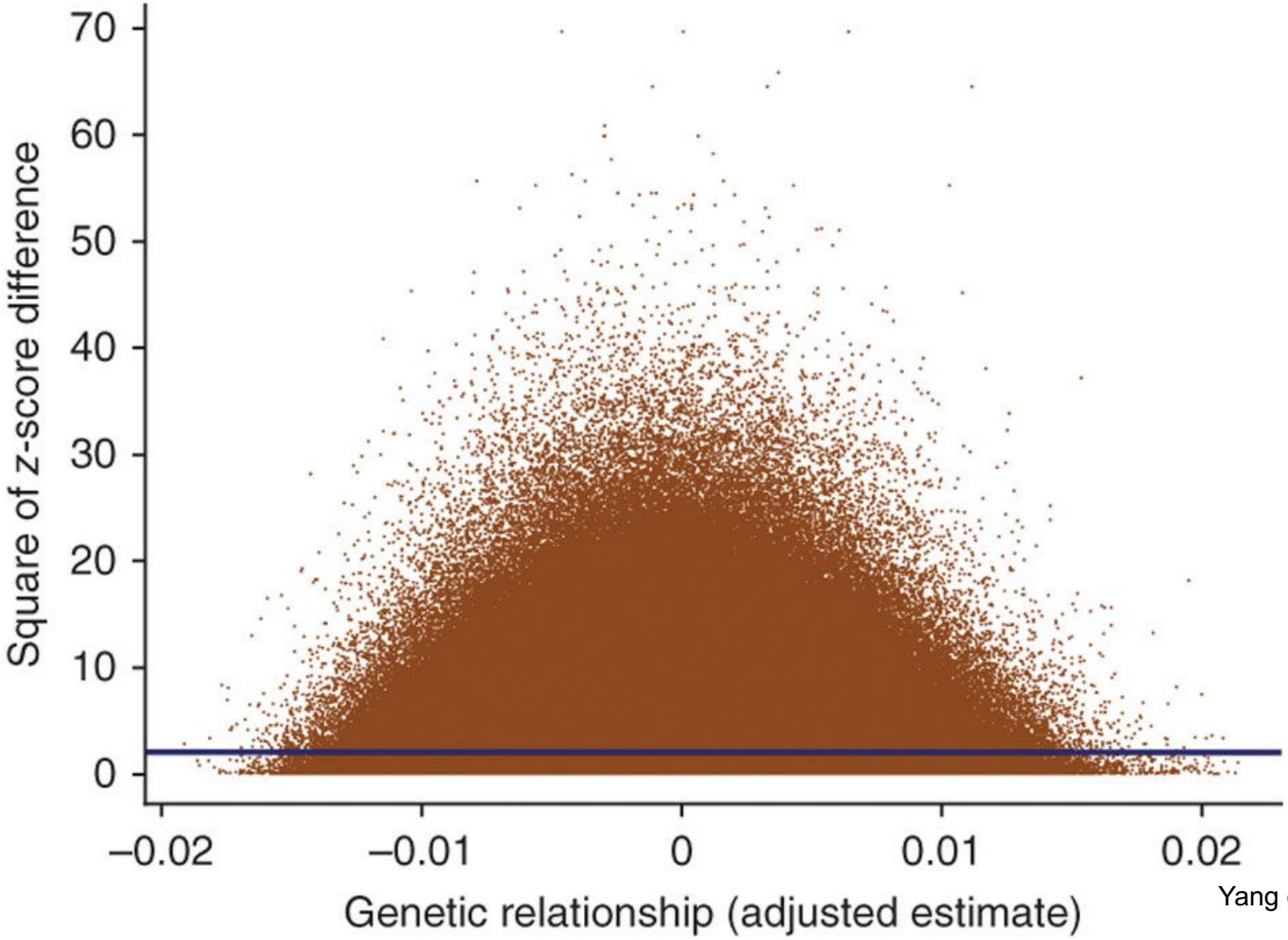
-- can think of π as average allele sharing between individuals

- For any pair with $\pi > 0.05$, remove the one with the lowest genotyping rate



Kemper et al. (2021) *Nature Communications*

Everyone is related to some extent

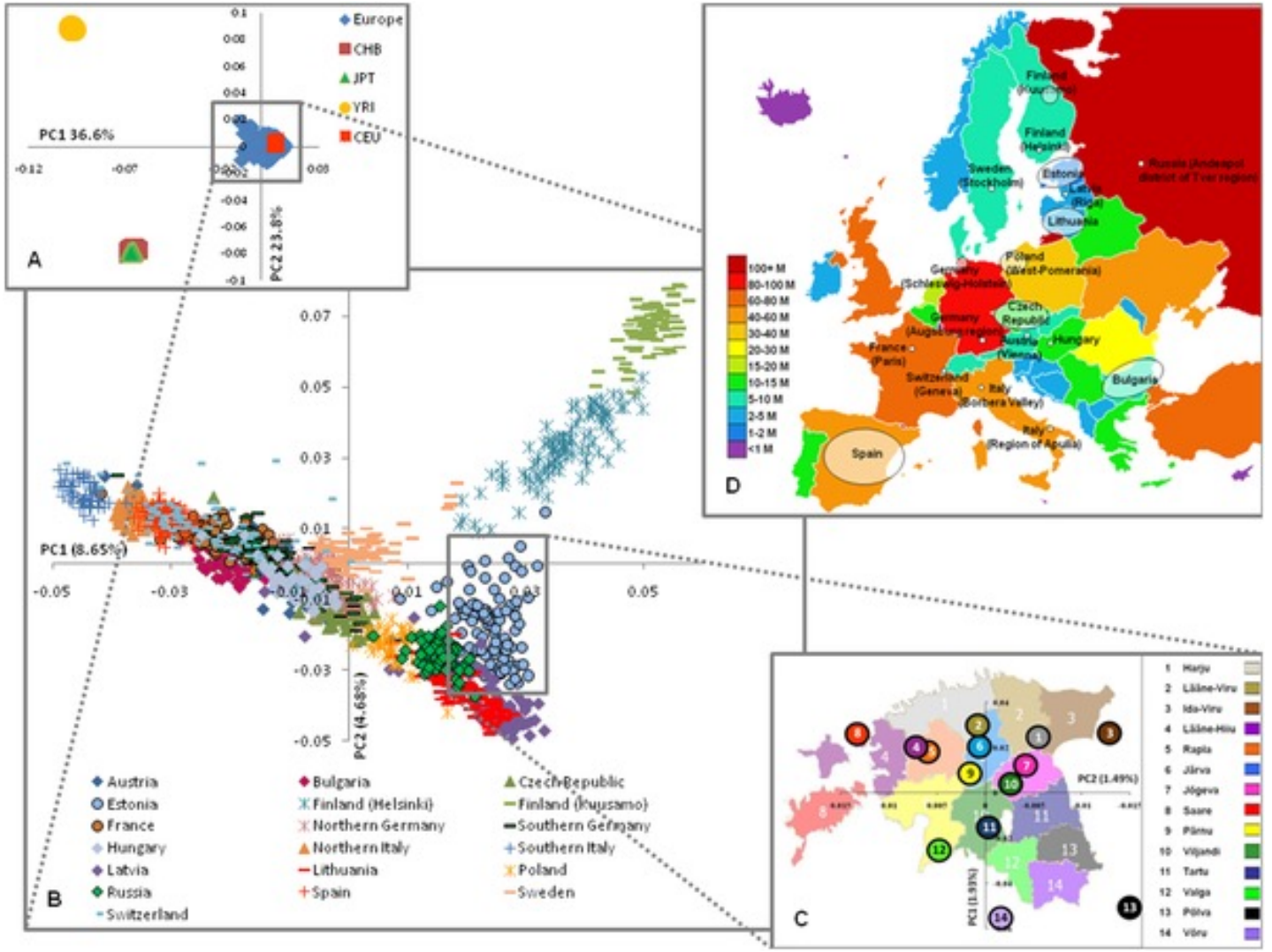


Yang et al. (2010) *Nature Genetics*

Dealing with population structure (3)

fit PCs as covariates

1. Perform PCA on your samples
2. Fit first (say) 10 PCs as covariates in your model



Nelis et al. (2009) PLOS ONE

Dealing with population structure (4)

use mixed model

Use a genomic relationship matrix (GRM) to model the covariance between closely related individuals

$$y = \alpha + x\beta + g + e$$

phenotypes

intercept

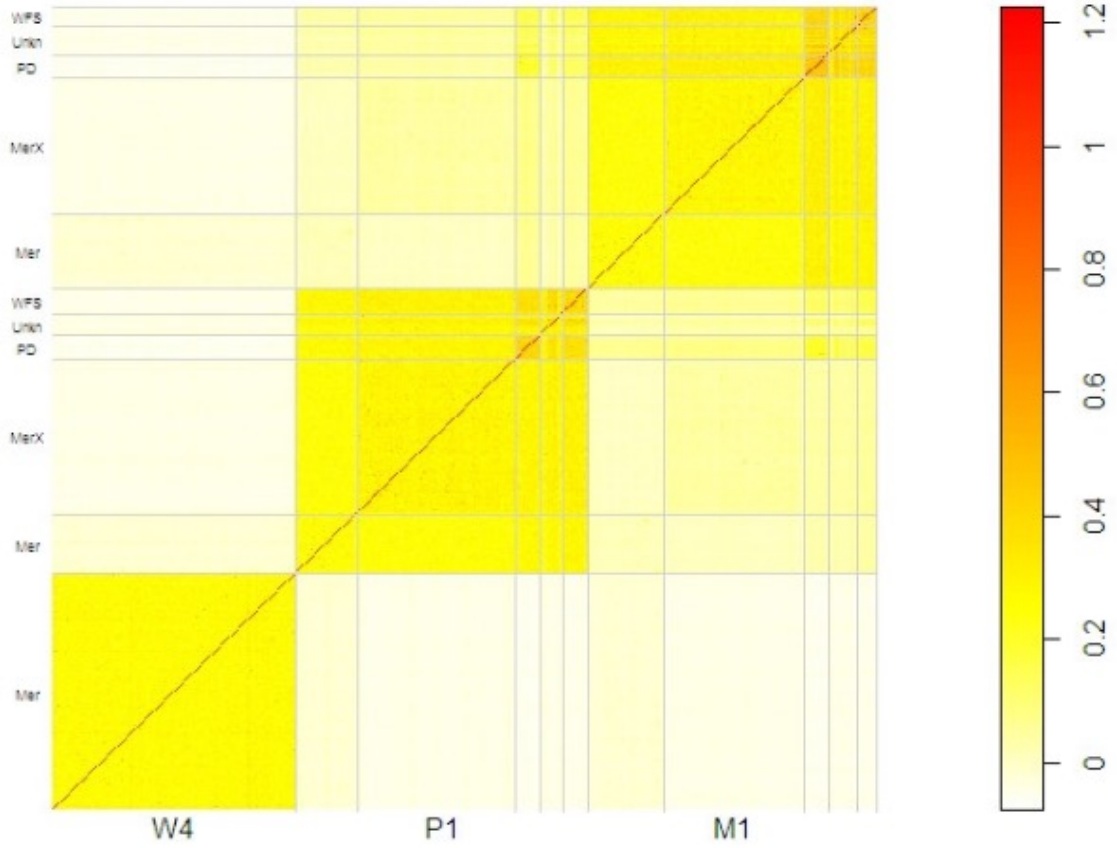
genotypes

SNP effect

(random) additive genetic effect

error

Example GRM from 3 sheep ½ sib families

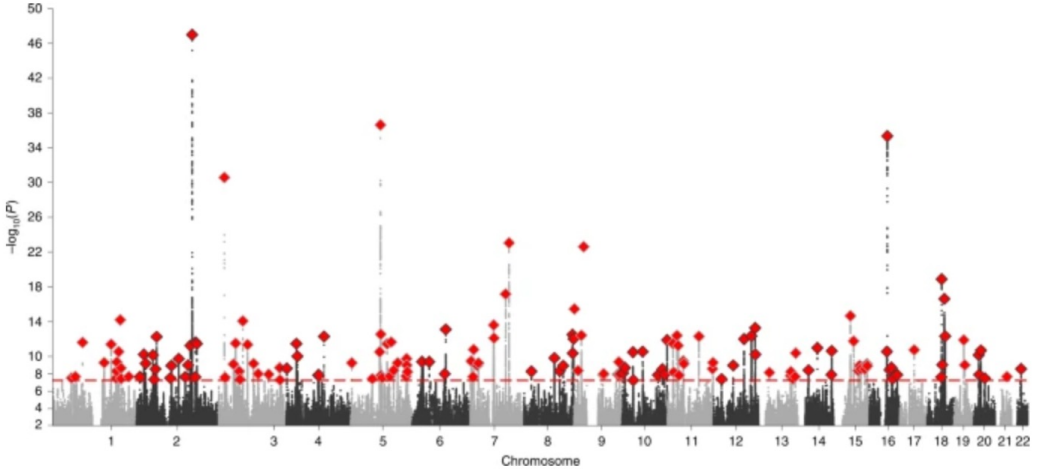


Kemper et al. (2011) *Genetics Research*

Population structure – participation bias

- Although we have some tools to deal with population structure, as sample sizes increase more subtle substructure becomes apparent

Fig. 1: Manhattan plot for a GWAS of sex in 2,462,132 participants from 23andMe.



The plot reports all identified loci, including those filtered by the extremely stringent quality control applied to directly genotyped SNPs.

Genetic analyses identify widespread sex-differential participation bias

Nicola Pirastu, Mattia Cordioli, Priyanka Nandakumar, Gianmarco Mignogna, Abdel Abdellaoui, Benjamin Hollis, Masahiro Kanai, Veera M. Rajagopal, Pietro Della Briotta Parolo, Nikolas Baya, Caitlin E. Carey, Juha Karjalainen, Thomas D. Als, Matthijs D. Van der Zee, Felix R. Day, Ken K. Ong, FinnGen Study, 23andMe Research Team, iPSYCH Consortium, Takayuki Morisaki, Eco de Geus, Rino Bellocco, Yukinori Okada, Anders D. Børglum, … Andrea Ganna ✉ + Show authors

Nature Genetics **53**, 663–671 (2021) | Cite this article

Abstract

Genetic association results are often interpreted with the assumption that study participation does not affect downstream analyses. Understanding the genetic basis of participation bias is challenging since it requires the genotypes of unseen individuals. Here we demonstrate that it is possible to estimate comparative biases by performing a genome-wide association study contrasting one subgroup versus another. For example, we showed that sex exhibits artifactual autosomal heritability in the presence of sex-differential participation bias. By performing a genome-wide association study of sex in approximately 3.3 million males and females, we identified over 158 autosomal loci spuriously associated with sex and highlighted complex traits underpinning differences in study participation between the sexes. For example, the body mass index–increasing allele at *FTO* was observed at higher frequency in males compared to females (odds ratio = 1.02, $P = 4.4 \times 10^{-36}$). Finally, we demonstrated how these biases can potentially lead to incorrect inferences in downstream analyses and propose a conceptual framework for addressing such biases. Our findings highlight a new challenge that genetic studies may face as sample sizes continue to grow.

QC quantitative phenotypes

- Most of the time phenotypes are ‘pre-corrected’ for fixed effects (such as age and sex) and standardised to N(0,1) within sex
- A transformation to normalise residuals may be necessary

  e.g. log-transformation for right skewed traits, log(y)

  e.g. RINT (rank-inverse normal) transformation

- Some loss in power, but greatly reduces analysis time

Summary

- **Genetic architecture** is the number, effect size and frequency of loci affecting a trait

- Varies between traits

- Population structure is a **major** source bias in GWAS

- Best addressed at recruitment phase

- Statistical tools can help but very difficult to remove/correct for everything