Acknowledgement of Country

- The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.
- We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.
- We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

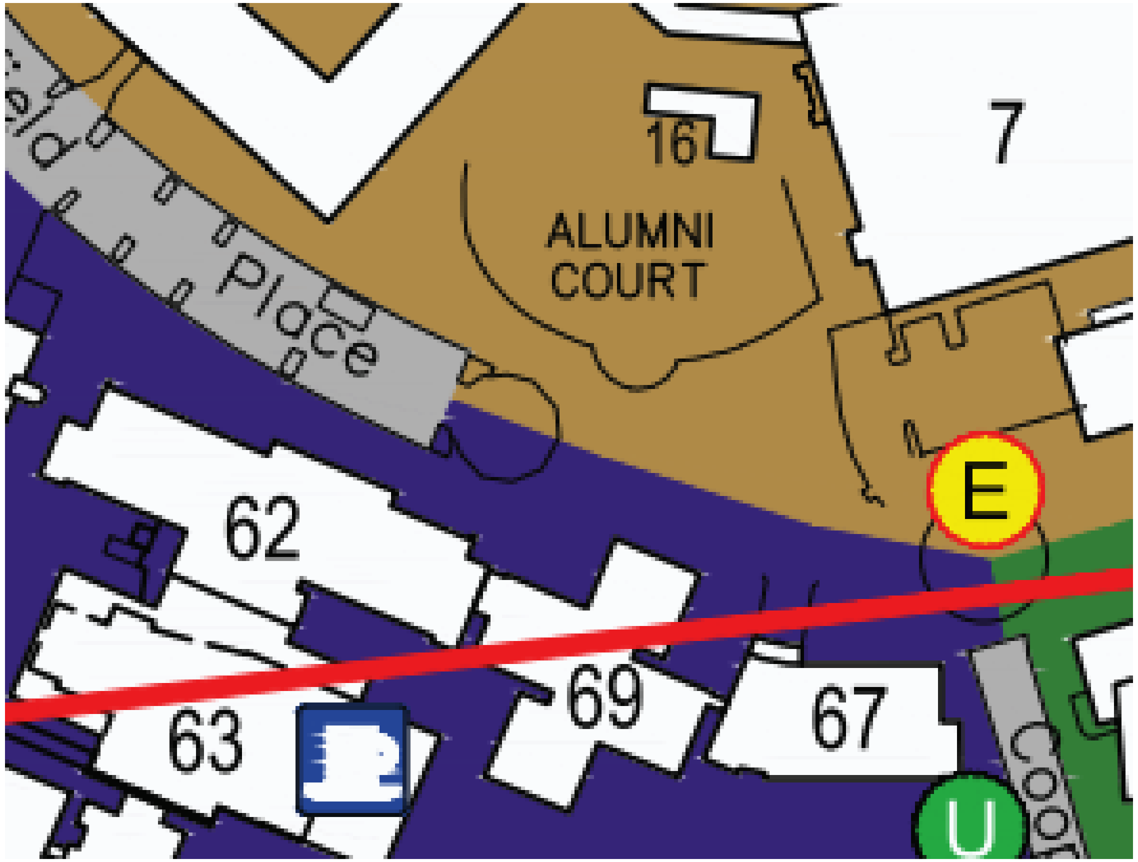
General Information:

- We are currently located in Building 69

E Emergency evacuation point

- Food court and bathrooms are in Building 63

- If you are experiencing cold/flu symptoms or have had COVID in the last 7 days *please* wear a mask for the duration of the module



Data Agreement

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

To maximize your learning experience, we will be working with genuine human genetic data, during this module.

Access to this data requires agreement to the following in to comply with human genetic data ethics regulations

If you haven't done so, please email ctr-pdg-admin@imb.uq.edu.au with your name and the below statement to confirm that you agree with the following:

"I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts."

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Plan for the Module: Genetic Mapping

DAY 1 – GWAS, theory & practice	
1pm	1. Intro + Phenotypes
2pm	2. Genotypes
3pm	3. Statistical tests
4pm	**End of day 1**
DAY 2 – GWAS & Post-GWAS analysis	
9am	4. Data Cleaning
10:30am	5. GWAS
11:30am	6. Summary Statistics
12pm	LUNCH
1pm	7. Meta-analysis & 8. Finding independent loci
2:30pm	9. Fine mapping 10. Interpreting GWAS results
4pm	**End of module**

Kath
(lecturer)

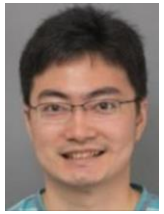
Alesha
(lecturer)

Fleur
(lecturer)

Monalisa
(teaching assistant)

Tian
(teaching assistant)

Ruolan
(teaching assistant)


JZ
(WS coordinator)

Solal
(cluster admin)

Acknowledgements

I wish to acknowledge the following people for allowing us to use/adapt their material (slides and practical) in this module:

Allan McRae
Ben Hayes
Joanna Revez
Jian Zeng
Alesha Hatton
Naomi Wray
Fleur Garton

Various internet resources & GWAS protocol papers

Introduction to Genome Wide Association Studies (GWAS)

What is a GWAS?

- A **G**enome **W**ide **A**ssociation **S**tudy is a method for identifying associations between locations in the genome and a trait of interest

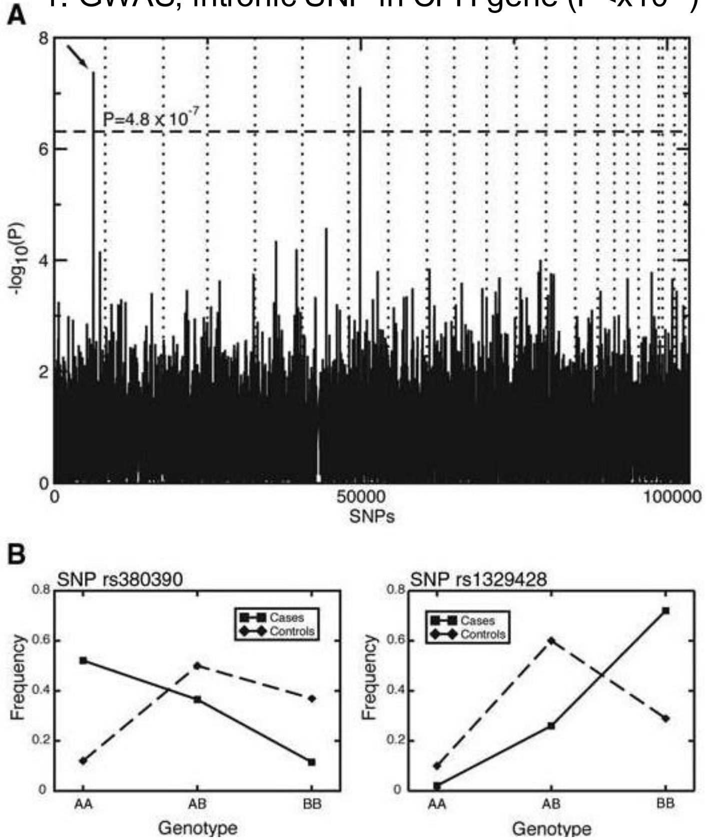
- Three key parts to a GWAS:
    - ○ A trait of interest or phenotype
    - ○ Genetic markers measured across the genome
    - ○ Statistical test of association between markers & phenotype

The first GWAS

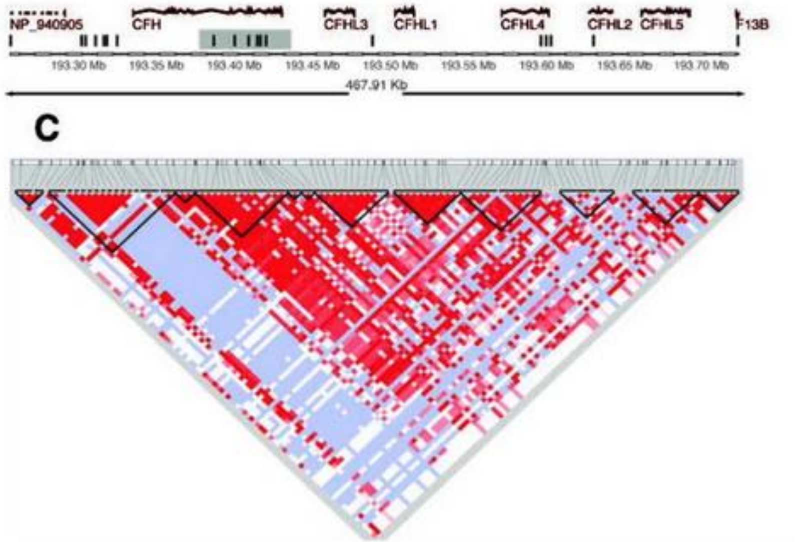
2005: one of the first GWAS, age-related macular degeneration

- Klein et al. 2005 *Science;* 96 cases and 50 controls; 116,204 SNPs

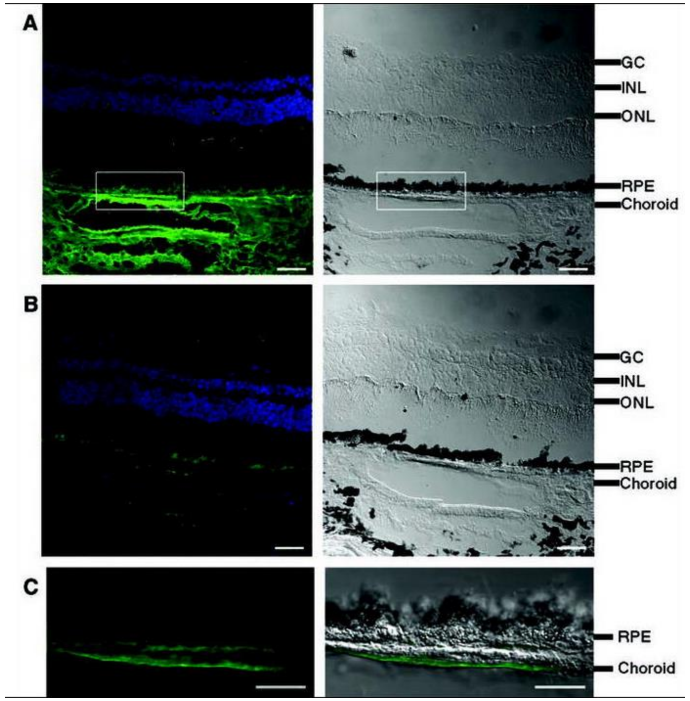
1. GWAS, intronic SNP in CFH gene ($P<x10^{-7}$)



2. re-sequencing region to identify missense variant



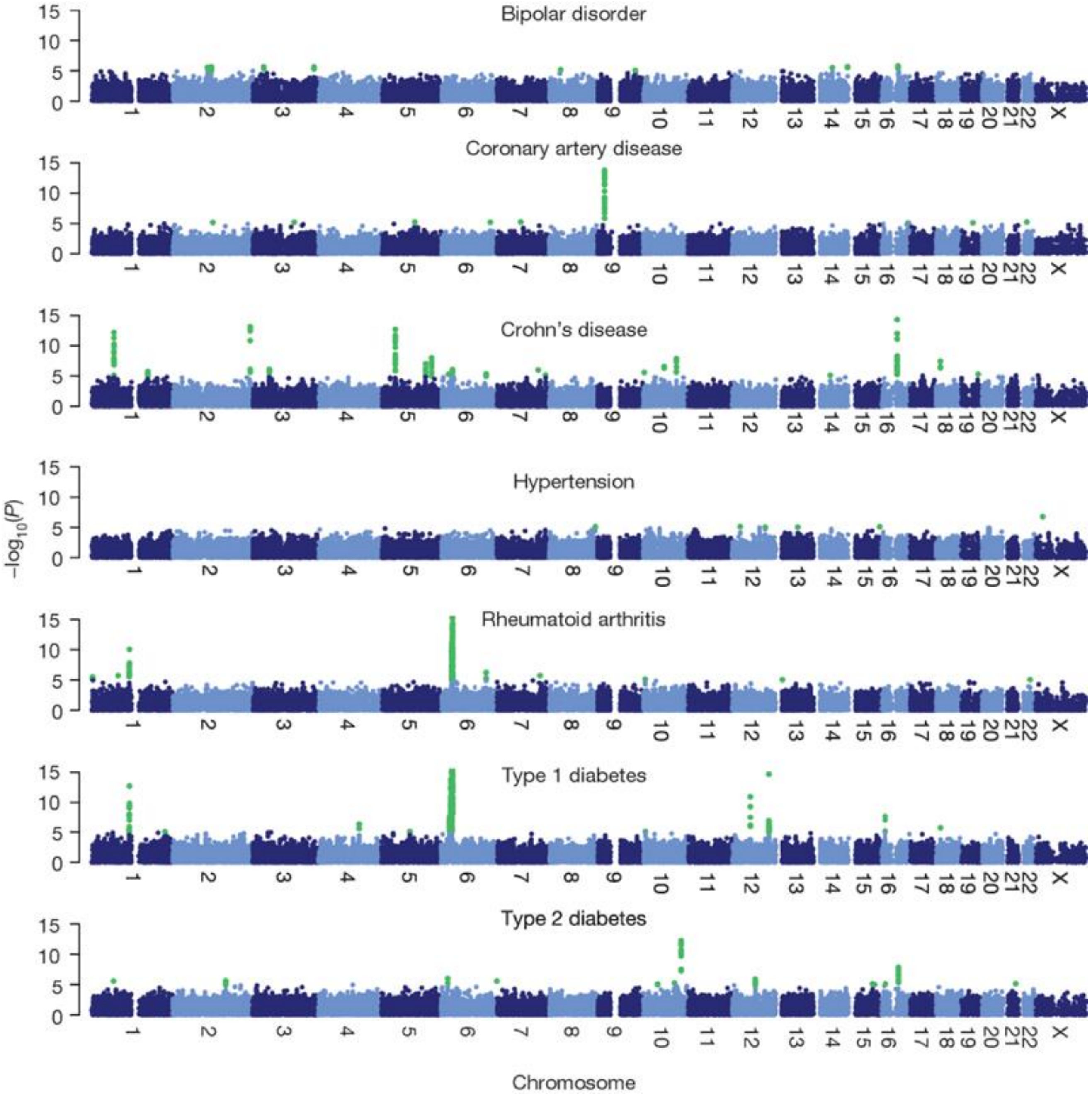
3. functional follow-up of CFH gene in retina



WTCCC

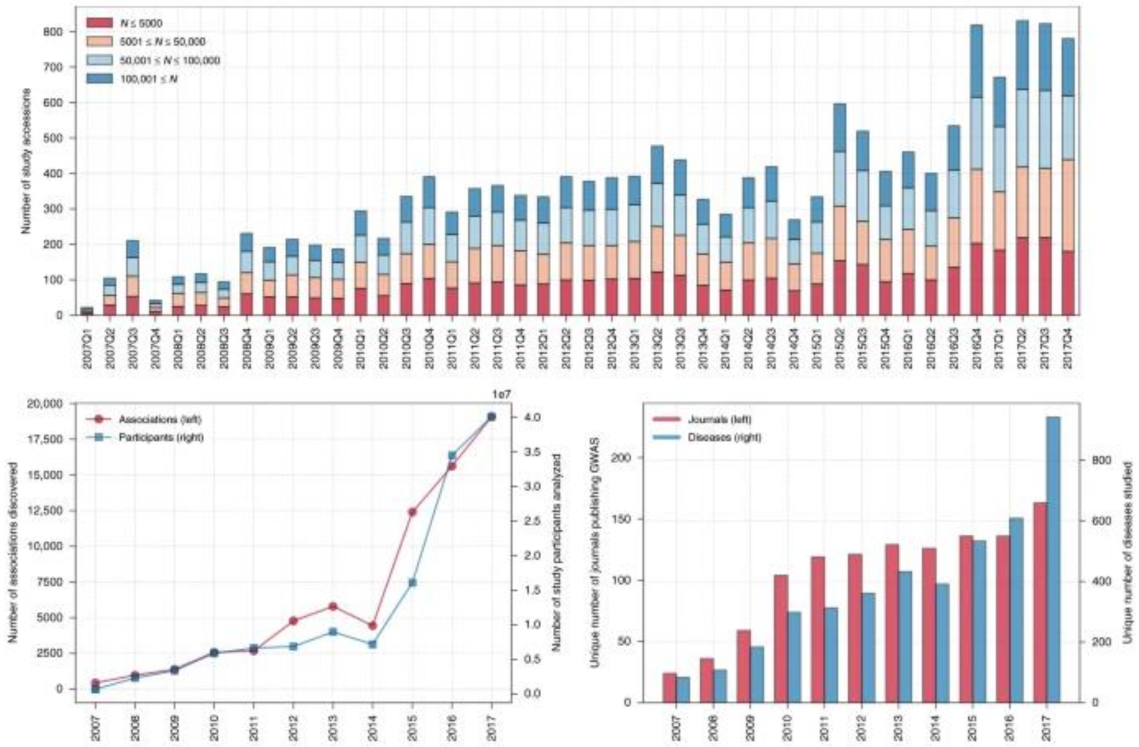
Wellcome Trust Case-Control Consortium

- First large scale GWAS (2007)
- 14,000 cases over 7 diseases
- 3,000 shared controls
- 500K Affymetrix GeneChip



Trends in GWAS

Fig. 1

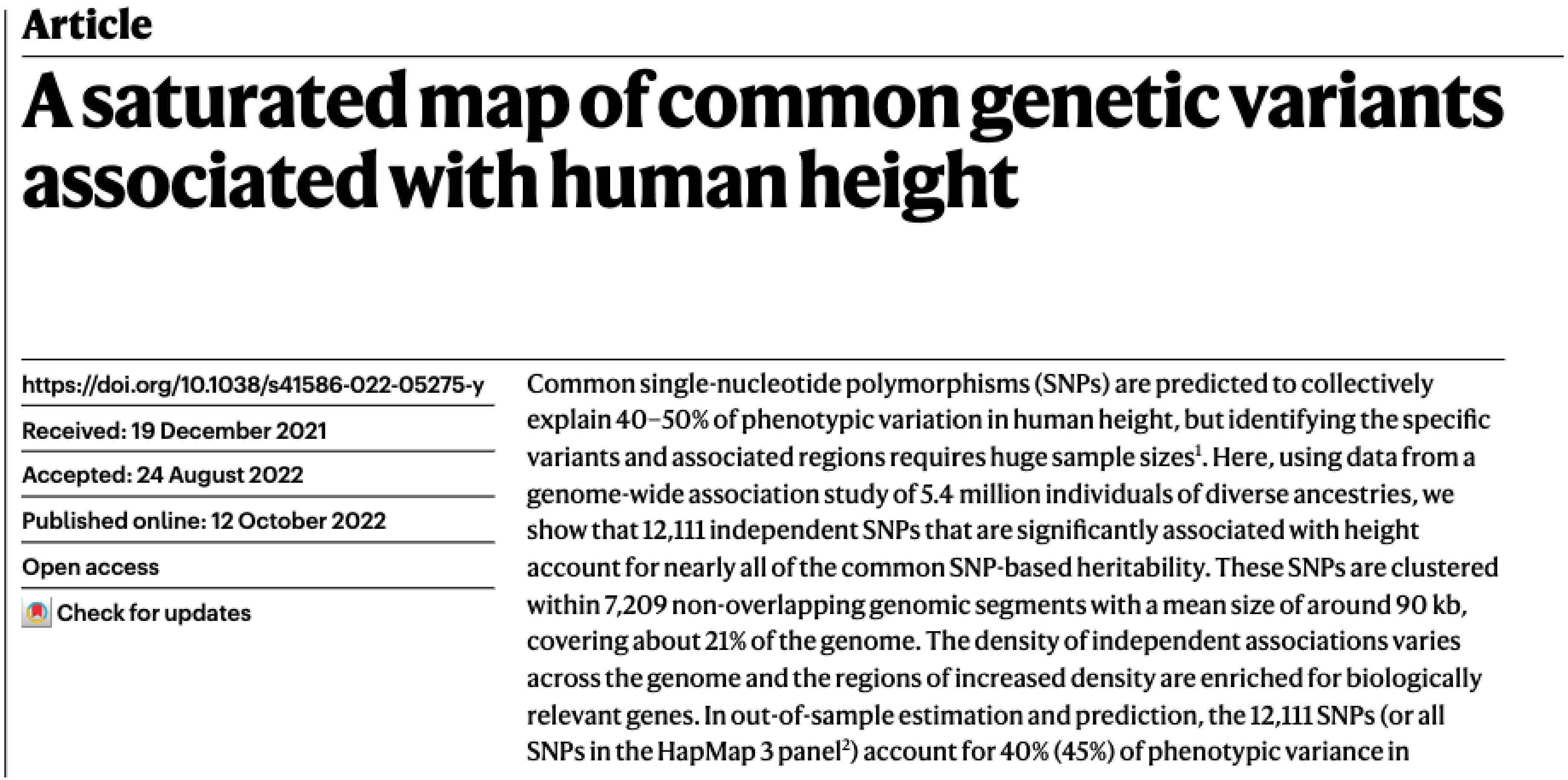


The growth of GWAS, 2007–2017. The upper panel shows the number of study accessions published per quarter over time colored according to sample size to show the growth of larger (100,001≤N) GWAS. The lower left panel shows the strong positive correlation between the number of associations found and the number of participants used in GWAS over time. The lower right panel shows the growth in the number of unique traits examined as well as the number of unique journals publishing GWAS over time. 2007–2017 is selected since only 10 entries occurred before 2007. Each panel contains full calendar years only. Source: NHGRI-EBI GWAS Catalog

- More genetic markers
- Bigger sample sizes + meta-analysis
- New traits & diseases
- More discoveries + insights
- Predominately EUR ancestry recruited from US, UK & Iceland

Mills & Rahal (2019) *Communications Biology*

Latest (published) GWAS has 5.4M people!



Article

A saturated map of common genetic variants associated with human height

https://doi.org/10.1038/s41586-022-05275-y

Received: 19 December 2021

Accepted: 24 August 2022

Published online: 12 October 2022

Open access

Check for updates

Common single-nucleotide polymorphisms (SNPs) are predicted to collectively explain 40–50% of phenotypic variation in human height, but identifying the specific variants and associated regions requires huge sample sizes[1]. Here, using data from a genome-wide association study of 5.4 million individuals of diverse ancestries, we show that 12,111 independent SNPs that are significantly associated with height account for nearly all of the common SNP-based heritability. These SNPs are clustered within 7,209 non-overlapping genomic segments with a mean size of around 90 kb, covering about 21% of the genome. The density of independent associations varies across the genome and the regions of increased density are enriched for biologically relevant genes. In out-of-sample estimation and prediction, the 12,111 SNPs (or all SNPs in the HapMap 3 panel[2]) account for 40% (45%) of phenotypic variance in

Yengo et al. (2023) *Nature*

A word of warning…

- Conceptually, a GWAS is straight forward. However, LOTS can go wrong!

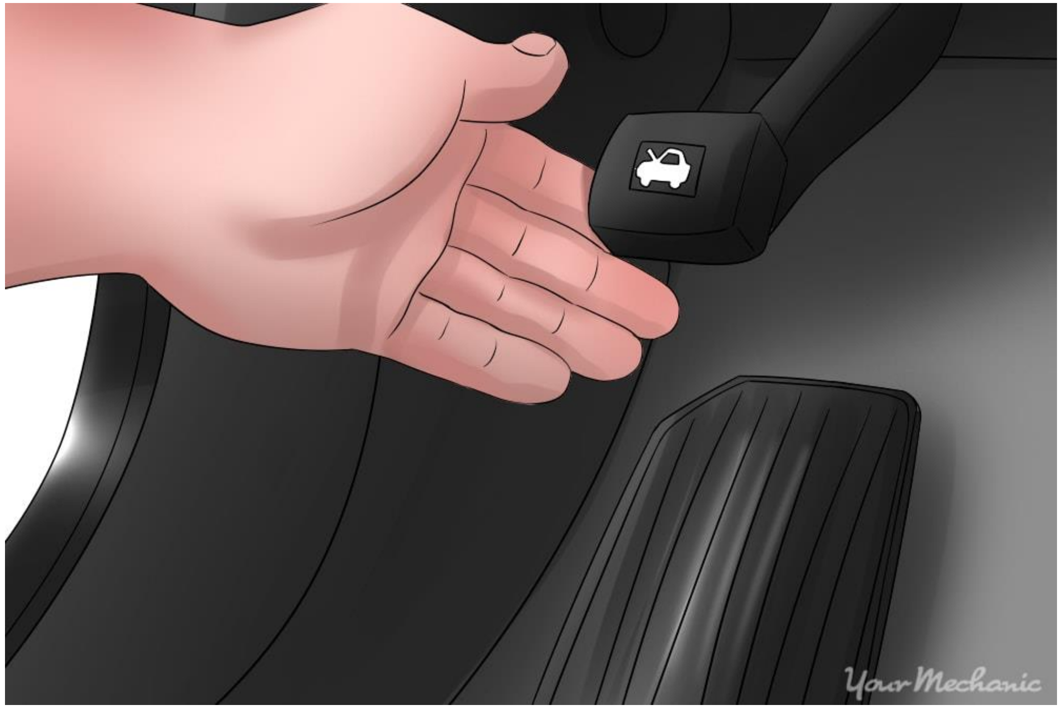
- **Most of the time in GWAS is spent in preparing the data to avoid various pitfalls**

Objective for the Module: Genetic Mapping

- during the module you will *not* find recipes and/or pipelines

- We hope you will *understand* what your doing & be able to critique others

- You may not complete all the practicals but hopefully you will have resources to come back to
  - everyone brings different skills



GWAS Experimental Design: phenotypes

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

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?

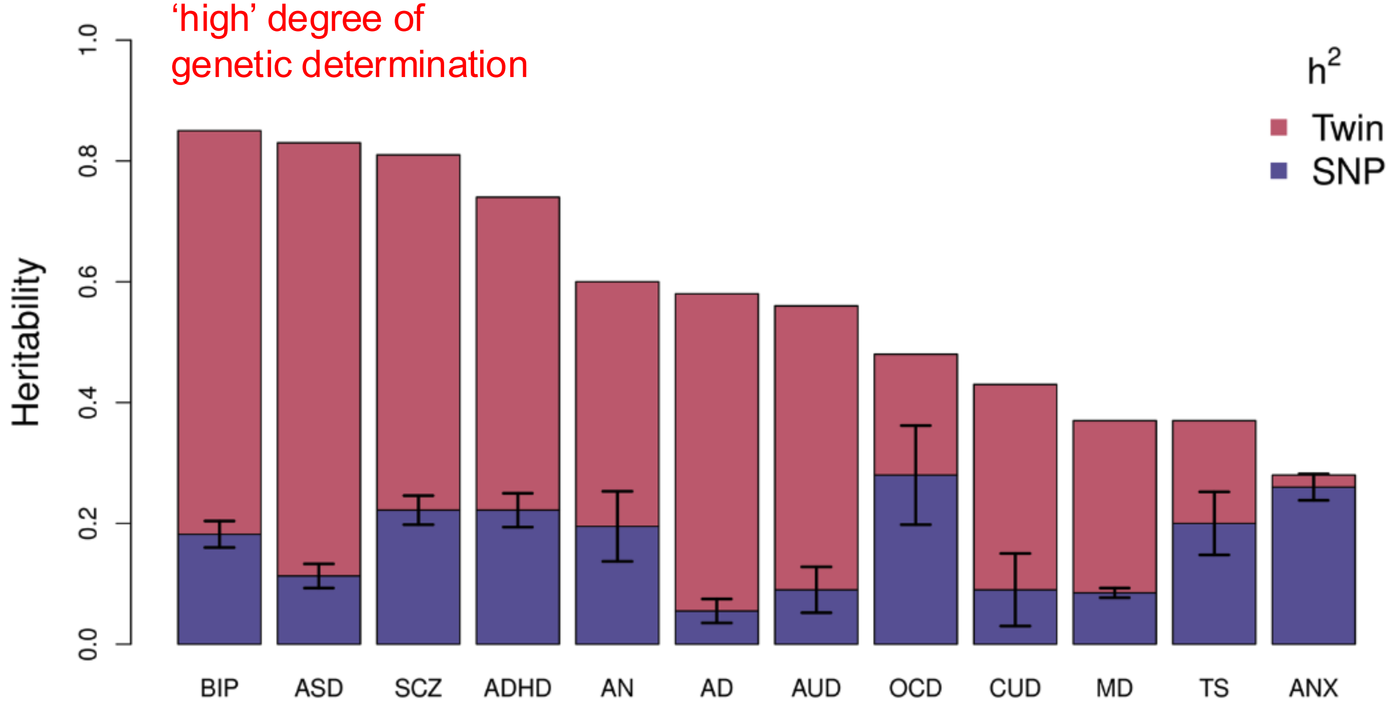
- A **phenotype** is an observable trait
  - it is influenced by both genetic and environmental (non-genetic) factors

- 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

'high' degree of genetic determination



- *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).

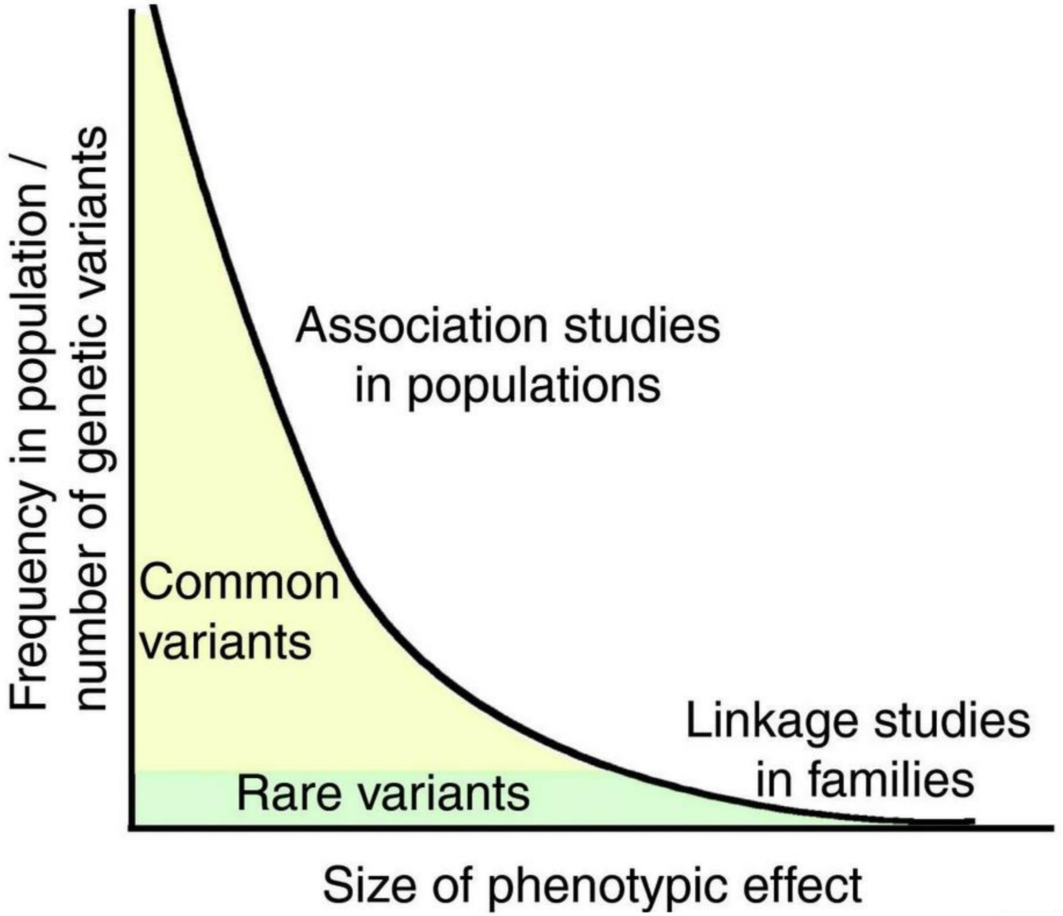
O'Connell & Coombes (2021) *Psychol Med.*

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%

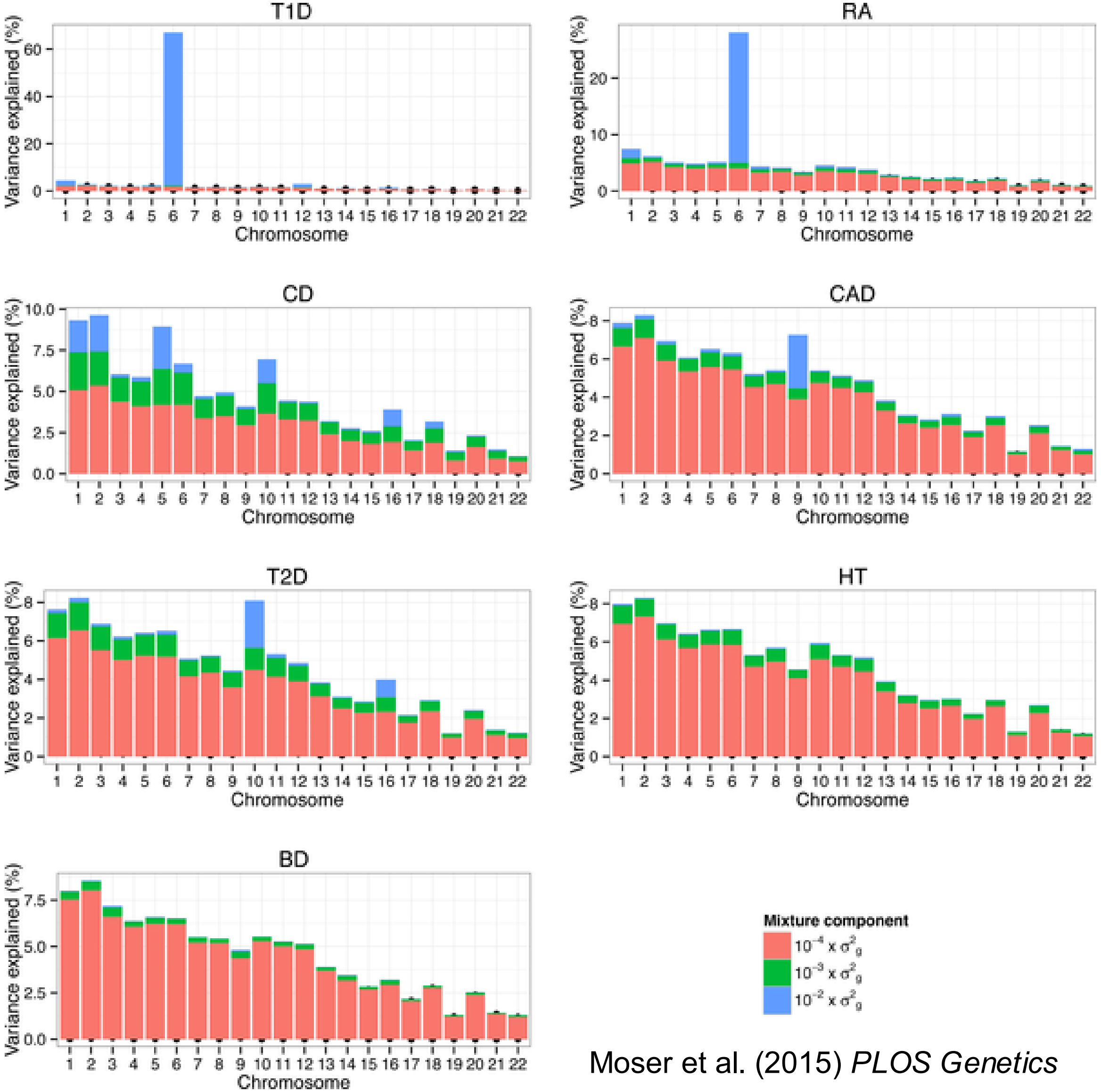
- *Common variants* tend to have smaller effect sizes



Rahim et al. (2008) *Genome Biology*

Genetic architecture

- **Genetic architecture** can differ between traits, even when the 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

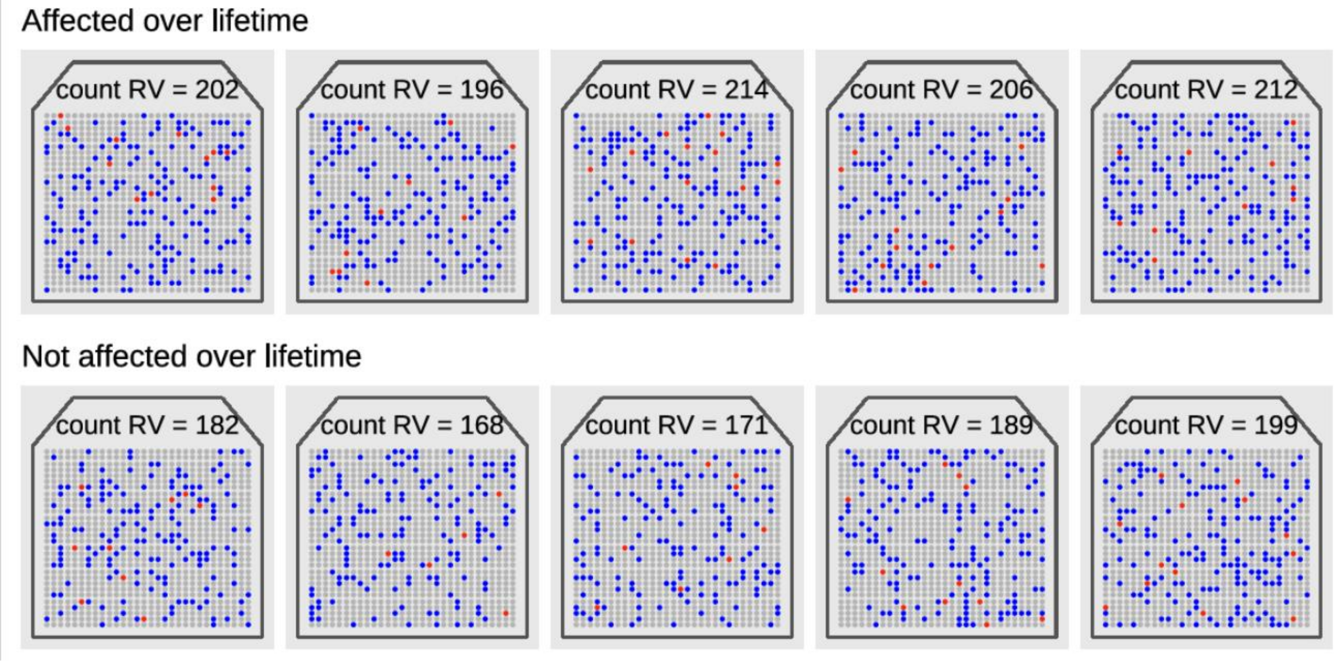


Moser et al. (2015) *PLOS Genetics*

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

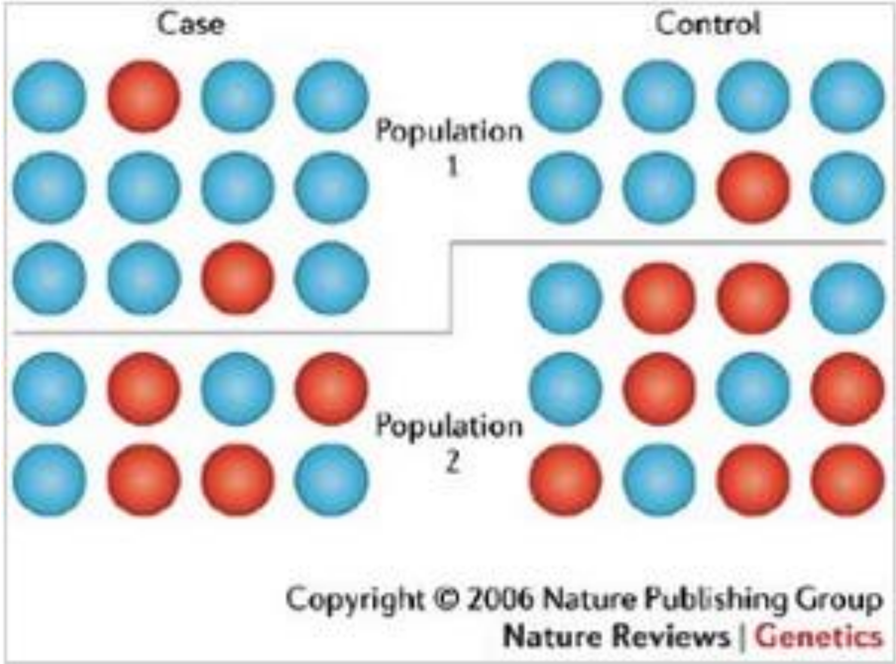


RV = risk variant

Slide adapted from Prof Naomi Wray

Phenotype QC - 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



Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

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

Balding (2006) *Nature Rev. Genetics*

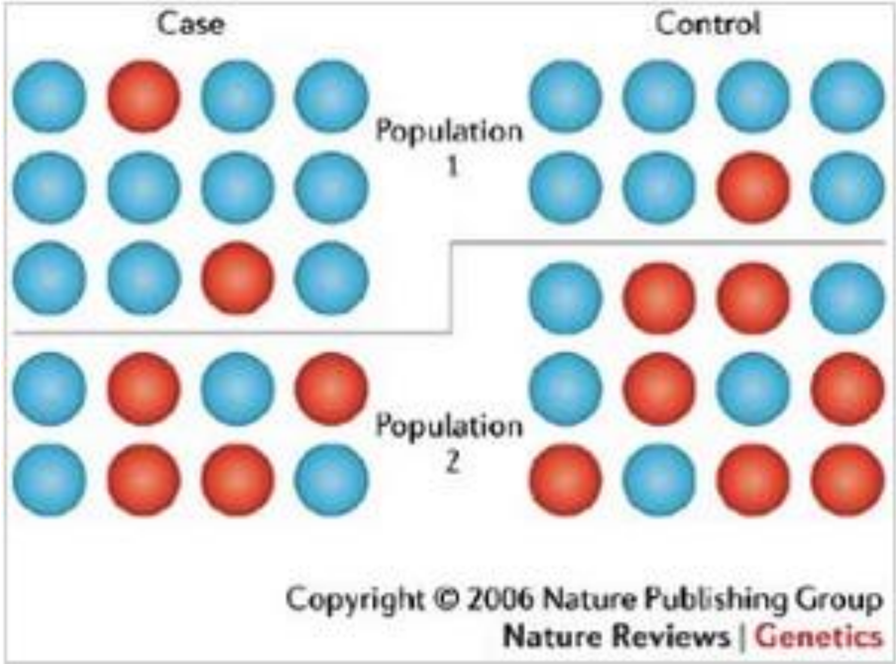
Phenotype QC - Population stratification

- **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



Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

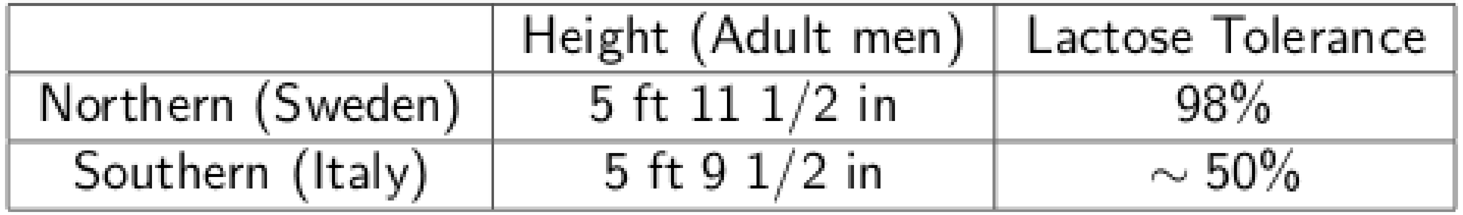
	Case	Control
Pop 1	10/12 = 0.83	7/8 = 0.87
Pop 2	4/8 = 0.5	5/12 = 0.41
ALL	14/20 = 0.7	12/20 = 0.6

Balding (2006) *Nature Rev. Genetics*

Phenotype QC - 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	∼ 50%

Campbell et al. (2005) *Nature Genetics*

Phenotype QC - Close relatives

- 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 & relatives

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.

a) fitting PCs as covariates to account for ancestry differences, or

a) use a mixed model (with a genomic relationship covariance matrix) to account for close relatives

The genomic relationship matrix

Many approach to deal with population structure rely on a genomic relationship matrix or GRM

$$\begin{pmatrix} 1.1 & 0.22 & 0.12 & -0.01 \\ 0.22 & 0.95 & 0.12 & 0.01 \\ 0.12 & 0.12 & 1.05 & 0.52 \\ -0.01 & 0.01 & 0.52 & 1.00 \end{pmatrix}$$

off-diagonal elements of **A** estimate the genomic relationship (π) between pairs [i.e. average allele sharing]

The genomic relationship matrix

Many approach to deal with population structure rely on a genomic relationship matrix or GRM

Square symmetric matrix

individuals →

individuals ↓

$$\begin{pmatrix} 1.1 & 0.22 & 0.12 & -0.01 \\ 0.22 & 0.95 & 0.12 & 0.01 \\ 0.12 & 0.12 & 1.05 & 0.52 \\ -0.01 & 0.01 & 0.52 & 1.00 \end{pmatrix}$$

off-diagonal elements of **A** estimate the genomic relationship (π) between pairs [i.e. average allele sharing]

The genomic relationship matrix

Many approach to deal with population structure rely on a genomic relationship matrix or GRM

Square symmetric matrix

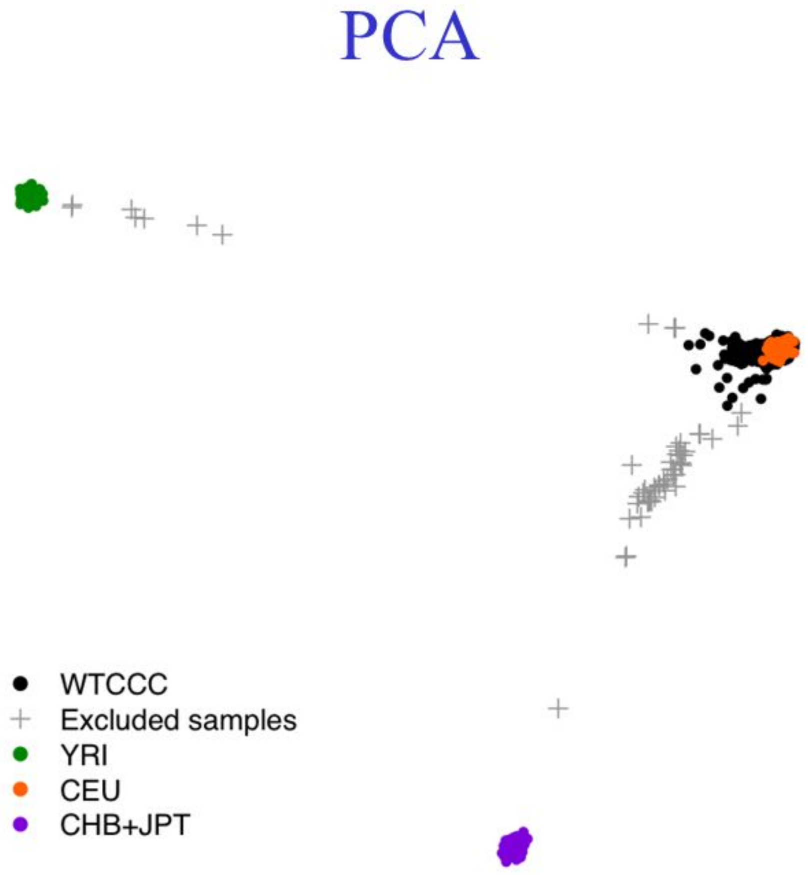
$$\begin{array}{c|c} & \text{individuals} \rightarrow \\ \hline \text{individuals} \downarrow & \begin{pmatrix} 1.1 & 0.22 & 0.12 & -0.01 \\ 0.22 & 0.95 & 0.12 & 0.01 \\ 0.12 & 0.12 & 1.05 & 0.52 \\ -0.01 & 0.01 & 0.52 & 1.00 \end{pmatrix} \end{array}$$

off-diagonal elements of **A** estimate the genomic relationship (π) between pairs [i.e. average allele sharing]

Dealing with population structure & relatives (1)

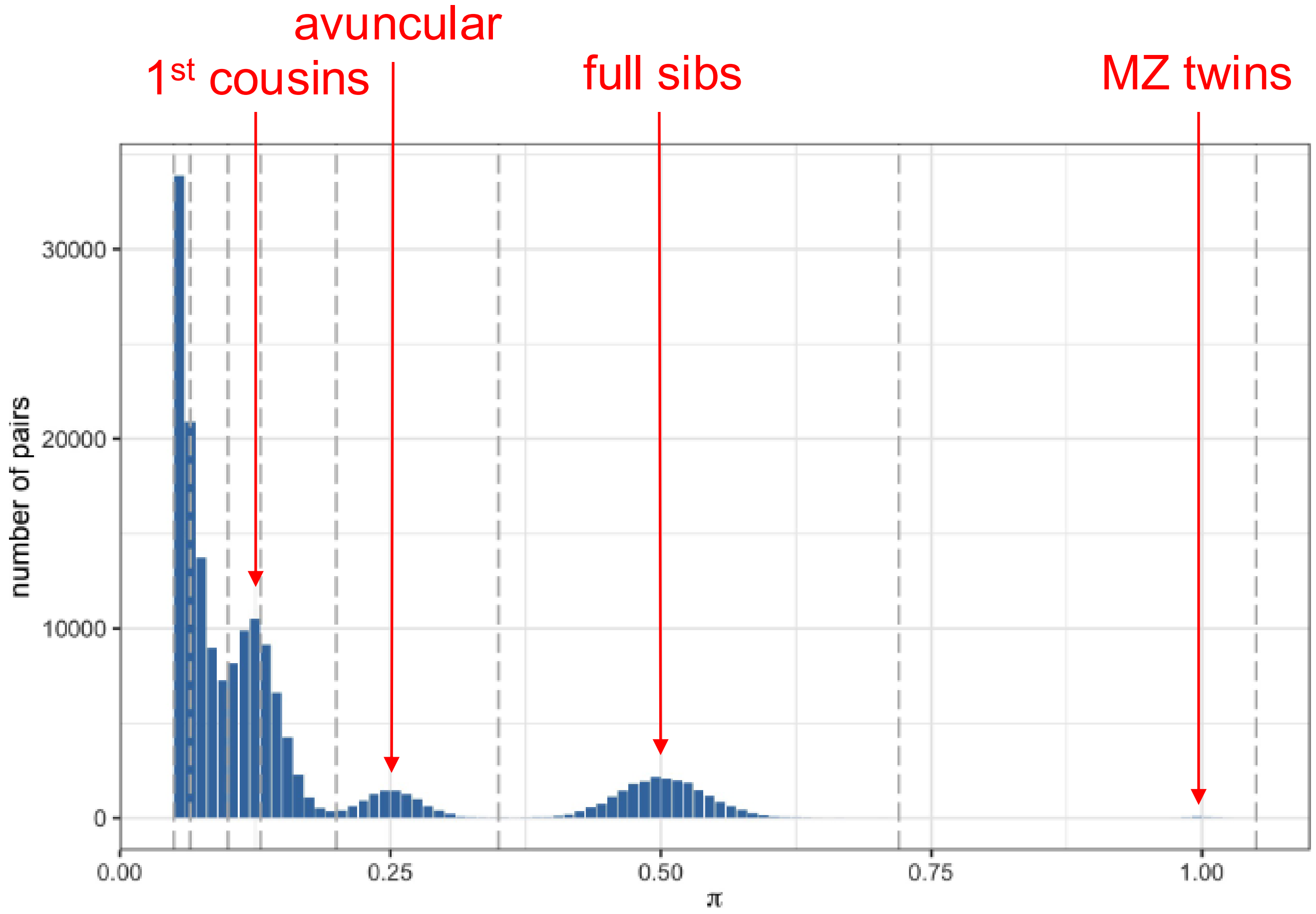
remove ancestry outliers

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



Dealing with population structure & relatives (2)

remove one member from each pair of relatives with $\pi > 0.05$



Genomic relationship among each pair in UK Biobank (π)

Kemper et al. (2021) *Nature Communications*

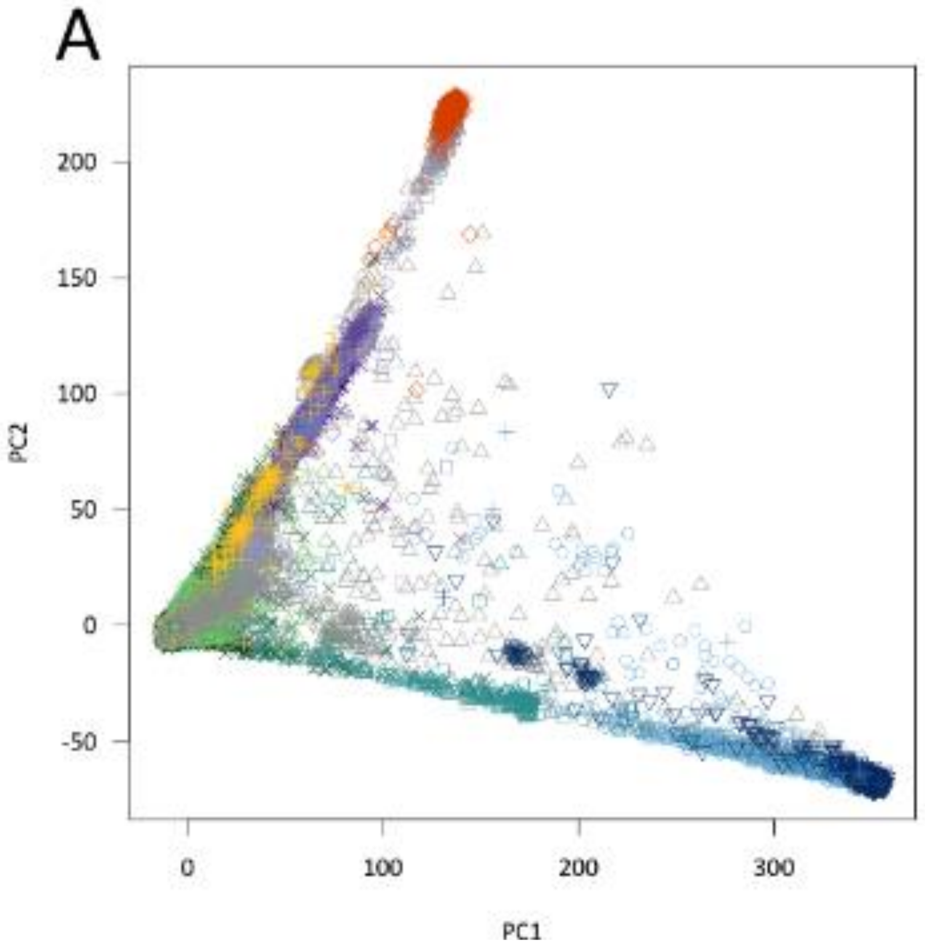
Dealing with population structure & relatives (3)

account for structure by fitting PCs as covariates

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

e.g. file of covariates:

ID	PC1	PC2
456859	-10	0
456185	150	-10
523014	323	-47
…	…	…



1st and 2nd principal components in UK Biobank, coloured by self-reported ancestry

Bycroft et al. (2021) *Nature Genetics*

Dealing with population structure & relatives (4)

account for close relatives using a mixed model

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

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}\beta + \mathbf{W}\mathbf{g} + \mathbf{e}$$

phenotypes

intercept

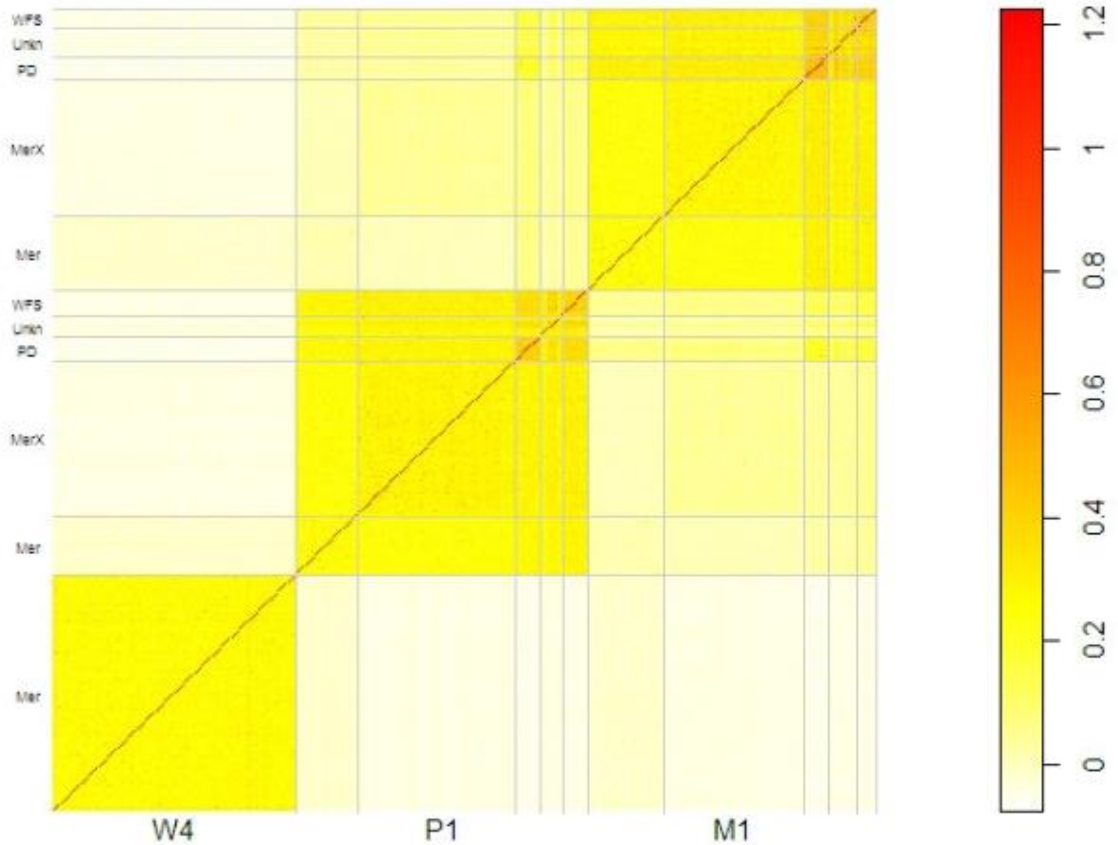
genotypes

SNP effect

(random) additive genetic effect

error

Example GRM from sheep



Kemper et al. (2011) *Genetics Research*

Pre-correcting (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

- *Why?* Pre-correcting results in some loss of 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 & relatives are a **major** source bias in GWAS

  - Best addressed during study design

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

Practical Session

Choose Part 1 or 2

Part 1: pre-adjusting (quantitative) phenotype

Part 2: simple PCA analysis

- download practical notes & slides from
  https://cnsgenomics.com/data/teaching/GNGWS25/module1/
- or on the cluster `/data/module1/downloadsWebsite.zip`

- On the cluster please work in your own folder, `/scratch/username/`

- Some data can be downloaded for the pracs from:
  `/data/module1/downloadsDataMonPM.zip`

Practical Session – cluster login

Username and password to cluster sent via email

Wifi access

- eduroam (university affiliates)
- UQ Guest/hotspot to set up UQ visitor account (non-university), see Instructions to Computing Resources sent via email

Mac machine

Open terminal: type e.g. ‘`ssh username@203.101.229.126`’
Answer ‘Yes’
Type in your password

Windows machine

PowerShell: as above
Command prompt: as above
Putty with user credentials

Practical Session - downloading

Mac machine

- Open new terminal: type e.g.

‘`scp username@203.101.229.126:/scratch/username/<file> <yourLocalDirectory>`’

- type your password

Windows machine

- Open new PowerShell: as above
- Open new command prompt: type e.g.

1. ‘`sftp username@203.101.229.126`’
2. type your password
3. ‘`pwd`’ and/or ‘`cd`’ to navigate to `/scratch/username/`
4. ‘`get <file>`’

- WinSCP or FileZila with user credentials

Practical Session – unix for new users

`cd <directory>` = change directories
`pwd` = present working directory
`ls` = list files in current directory
`Ctrl+c` = kill a process (e.g. PLINK when it’s taking too long)
`mkdir <name>` = make a new directory

`R` = start an interactive R session
`quit()` within R = exit an interactive R session

The data is located in `/data/module1/`
Your working directory `/scratch/username/`