THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

UK Biobank Workshop: Introduction to Genome Wide Association Study

Monday 9th February 2024

Kathryn Kemper

Institute for Molecular Bioscience @UQ

biobank uk

Enabling scientific discoveries that improve human health

Introduction to Genome Wide Association Studies (GWAS)

Outcome: Participants are familiar with the concepts, terminology and outputs of a GWAS study.

Target Audience: individuals from other fields with basic understanding of genetics and statistics

Outline:

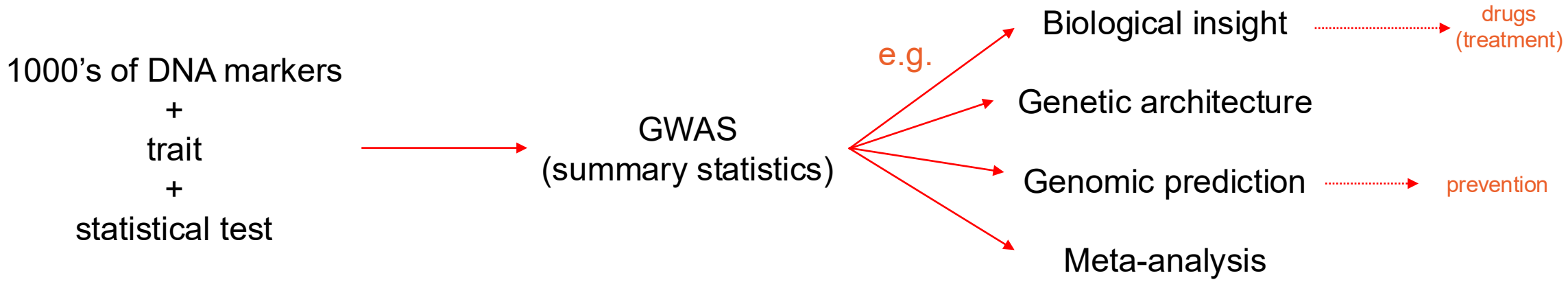
(1) context & motivation for GWAS

(2) inputs, output, and methodologies

(3) quality control, diagnostics & software

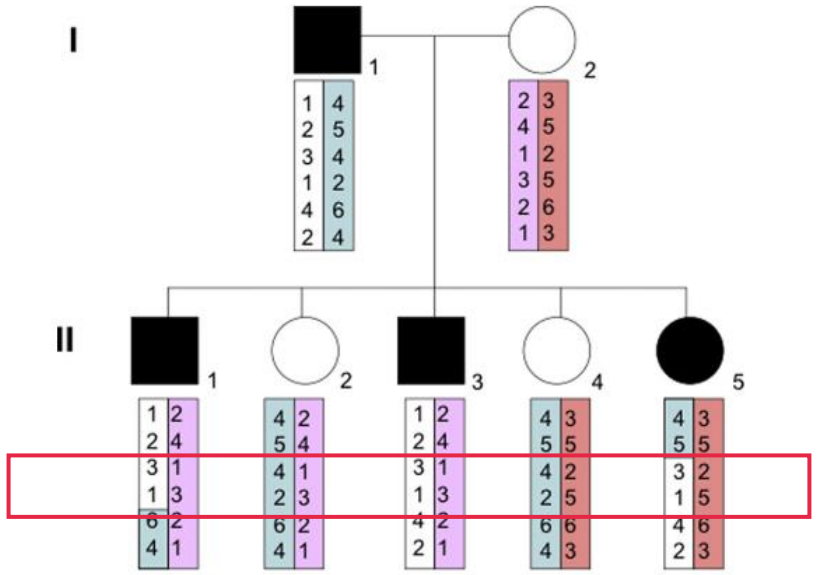
Motivation: why conduct a GWAS?

- GWAS scan the genome for associations between DNA markers (e.g. SNPs) and a trait of interest (e.g. height or heart disease) to identify genomic regions associated with the trait
  - identify genes involved in disease, i.e. biological insight
  - starting point for many downstream analyses



Context: why GWAS and what came before?

- Prior to GWAS, researchers were limited to linkage analysis or investigated ‘candidate’ genes
- linkage analysis = co-segregation of alleles & disease within families
  - i.e. rare alleles with large effects
- Not feasible to conduct population-level studies



Korf and Liu (2012)
Principals and Practice of Clinical Research

Context: human genome project & SNP chips

‘SNP chips’ (developed in late 1990’s & early 2000’s) have enabled the GWAS explosion

TIMELINE

1990-2003
Human Genome Project

2002-2005
HapMap Project

2005
First GWAS published

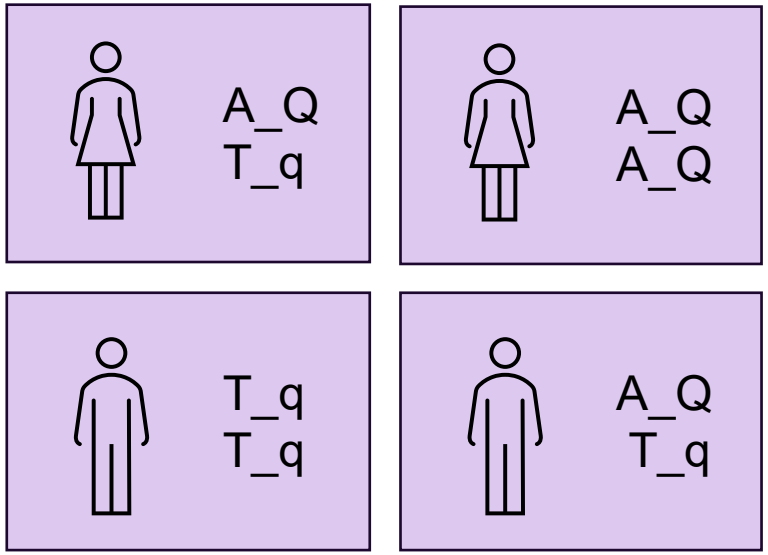
2007
WTCCC published

Context: basic principle of a GWAS

‘SNP chips’ measure 1000’s of markers all over the genome at low cost

- they allow population-level exploitation of LD (linkage disequilibrium) between SNP and a causal variant

e.g. an observed SNP marker with an ‘A’ allele is in LD with an unobserved (causal) variant ‘Q’

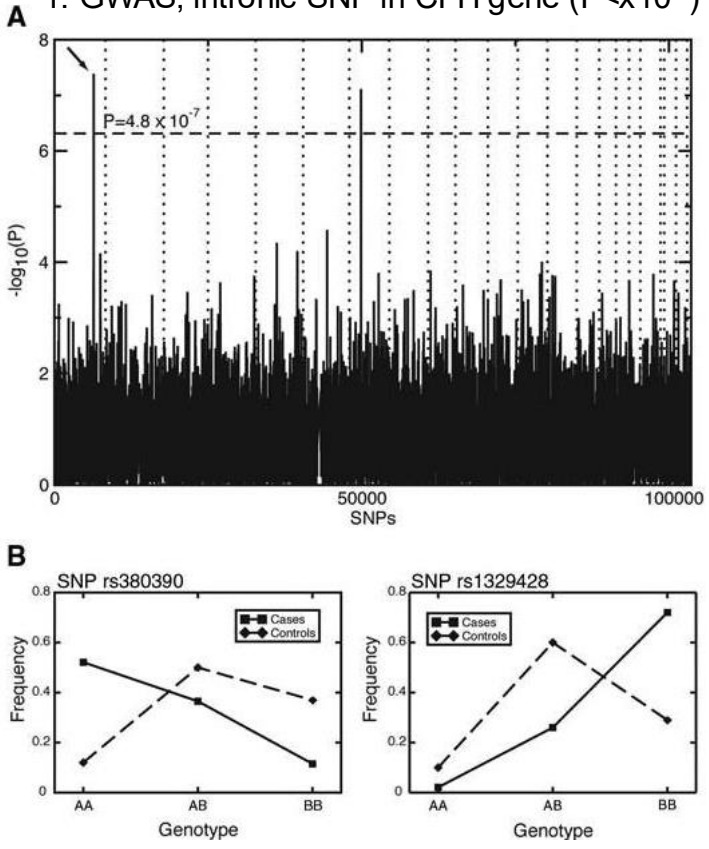


The first GWAS

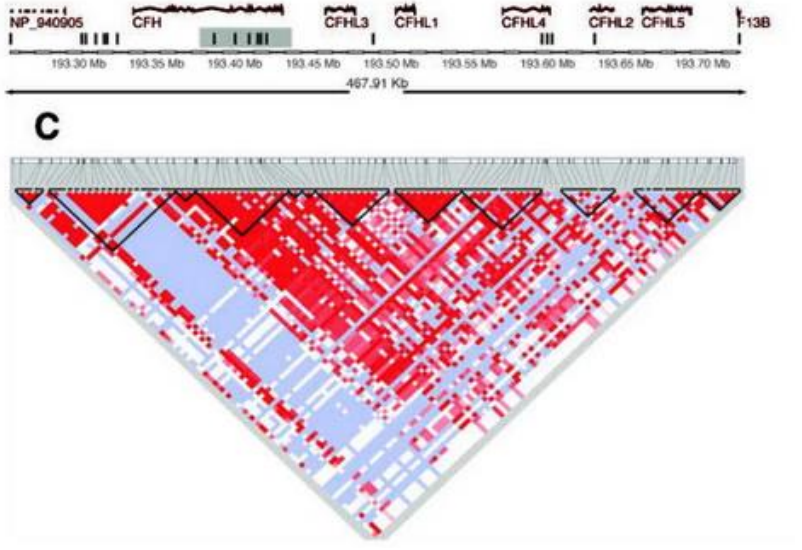
2005: one of the first GWAS, age-related macular degeneration

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

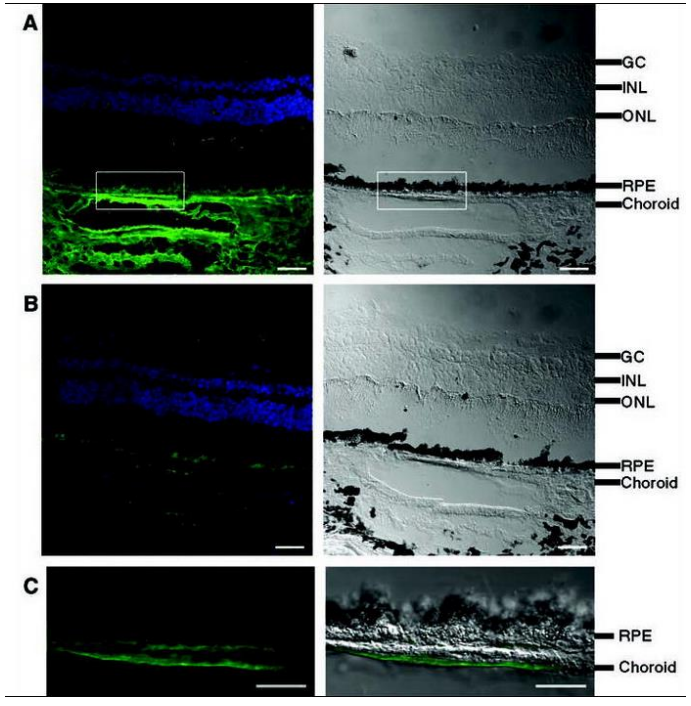
1. GWAS, intronic SNP in CFH gene ($P<x10^{-7}$)



2. re-sequencing region to identify missense variant



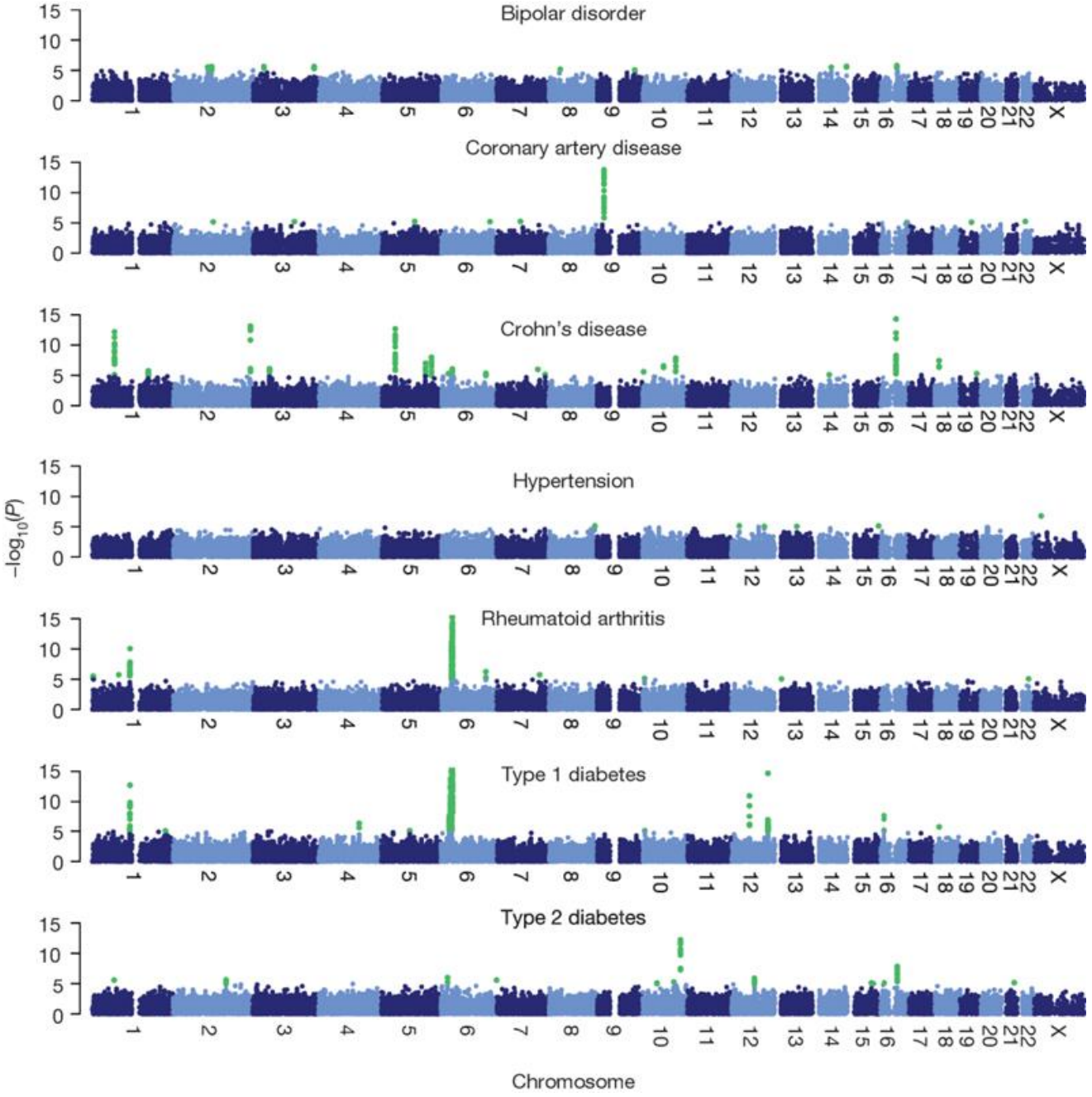
3. functional follow-up of CFH gene in retina



WTCCC

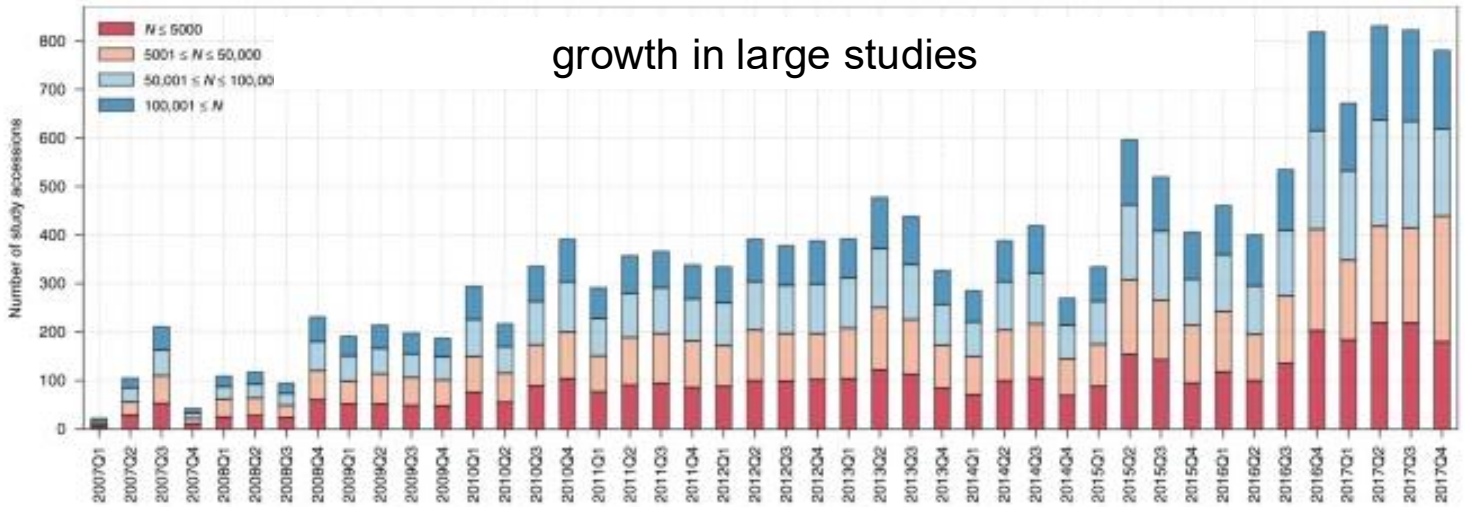
Wellcome Trust Case-Control Consortium

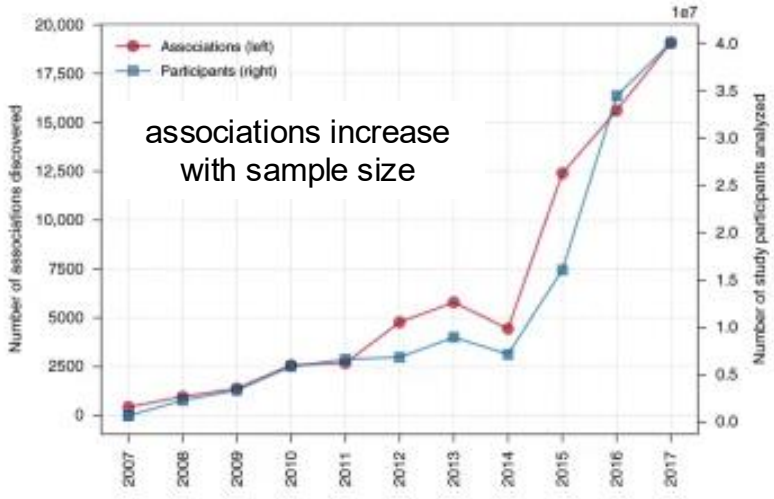
- First large scale GWAS (2007)
- 14,000 cases over 7 diseases
- 3,000 shared controls
- 500K Affymetrix GeneChip

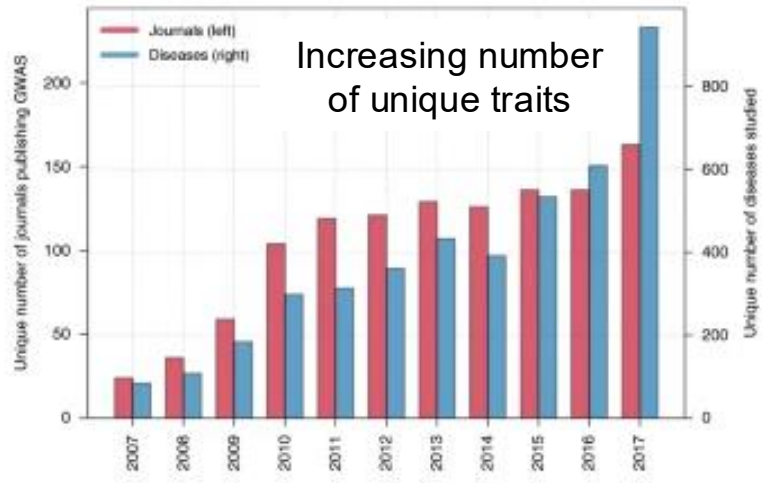


Trends in GWAS

- More markers
- Bigger sample sizes
- New traits & diseases
- Lots of discoveries + insights
- mostly EUR ancestry







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. (2022) *Nature*

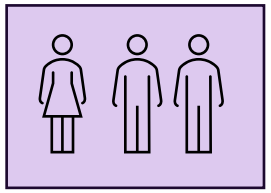
GWAS methodology

QC

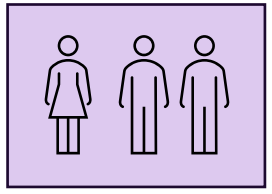
phenotype of interest → genotypes → Association between genotype & phenotype

Binary trait:

case:



control:



Binary trait: e.g. chi-square test

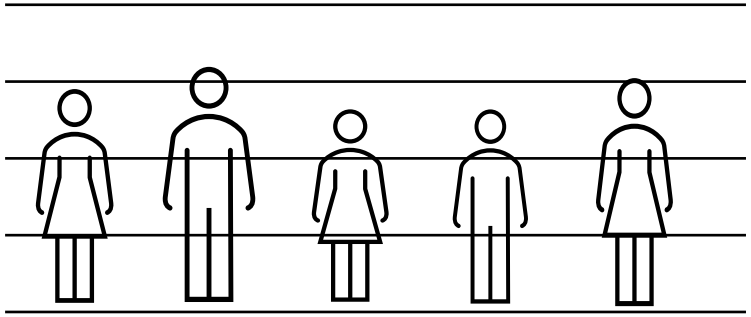
Alleles

	1	2	Total
Case	n_1	n_2	2N
Ctrl	m_1	m_2	2M
Total	T_1	T_2	2(N+M)

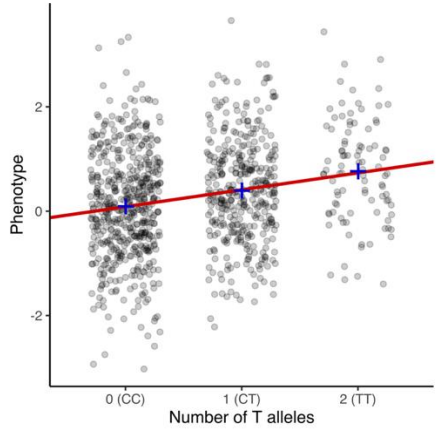
2x2 contingency table

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

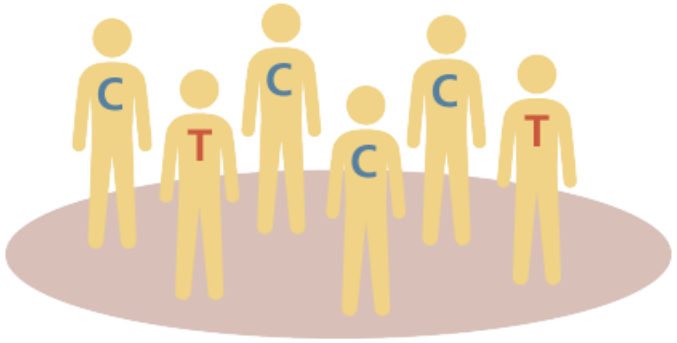
Quantitative trait:



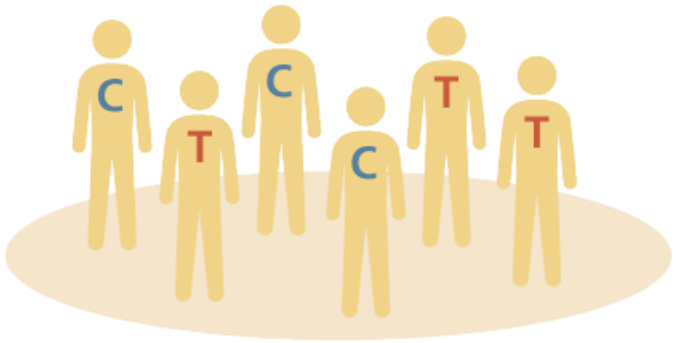
Quantitative trait: linear regression



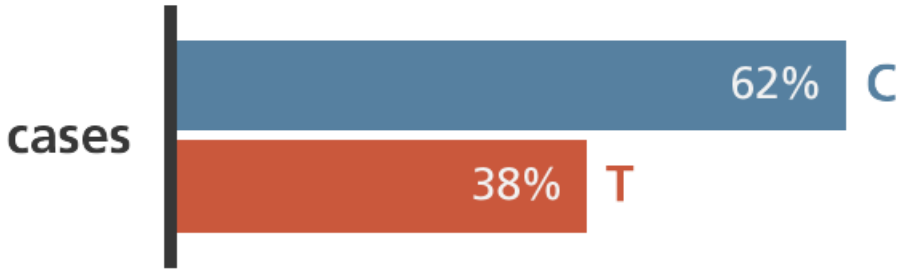
GWAS methodology - binary trait



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



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

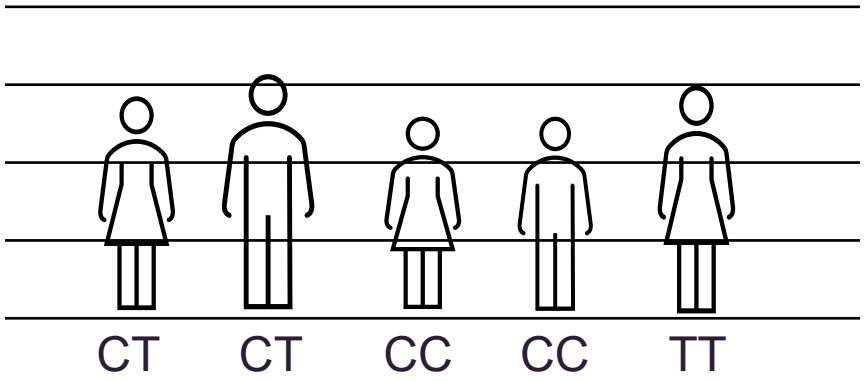


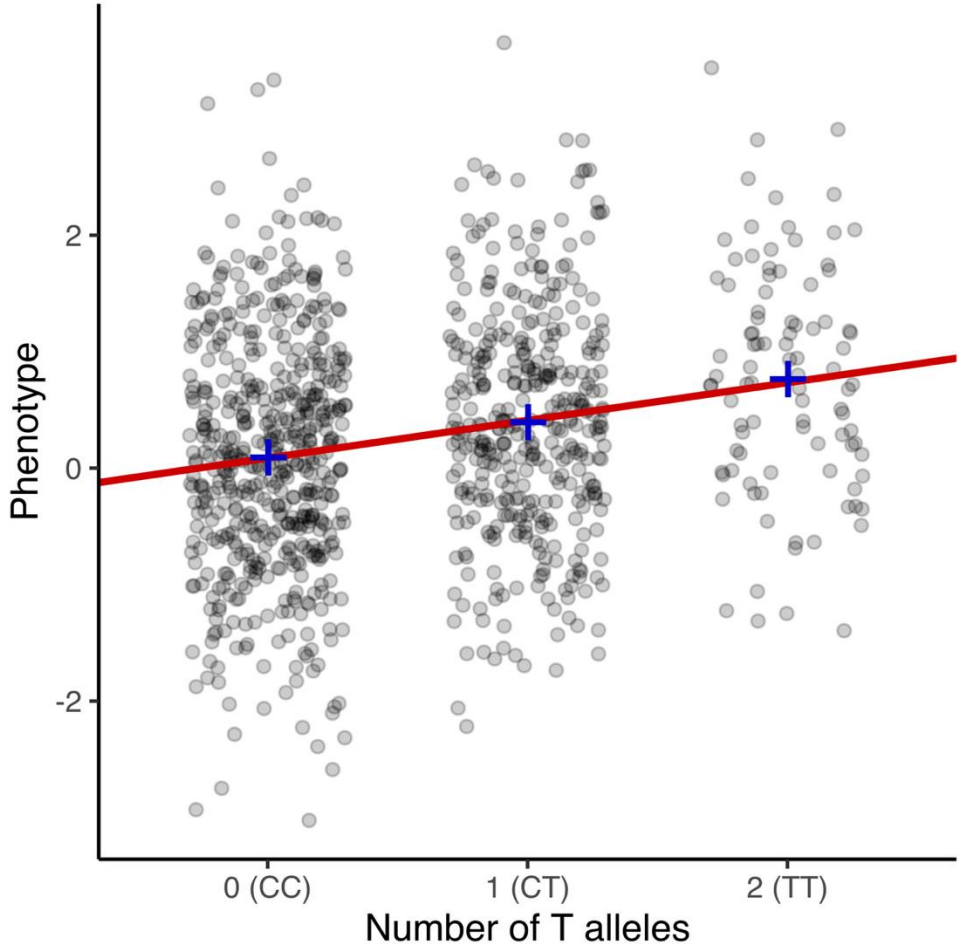
Test of association:

"Is the frequency of the 'C' allele different in cases vs. controls"

P = 0.0012

GWAS methodology - quantitative trait





- Linear model:

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

phenotypes

intercept

genotypes

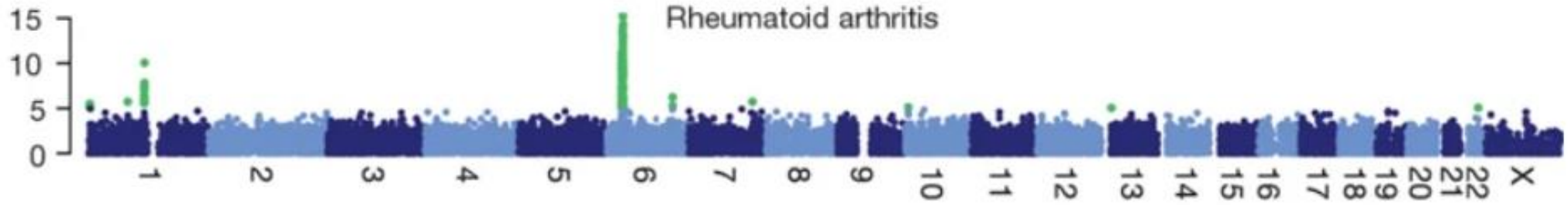
SNP effect

error

Output:

GWAS results are typically visualised as a ‘Manhattan plot’

y-axis: -$\log_{10}$P



x-axis: genome location of markers

- SNPs/markers with the strongest associations will have the greatest negative logarithms, and will tower over the background of unassociated SNPs
  - like skyscrapers in Manhattan →

P-value	-$\log_{10}$(P)
0.5	0.30
0.01	2
0.001	3
0.0001	4
... etc.	

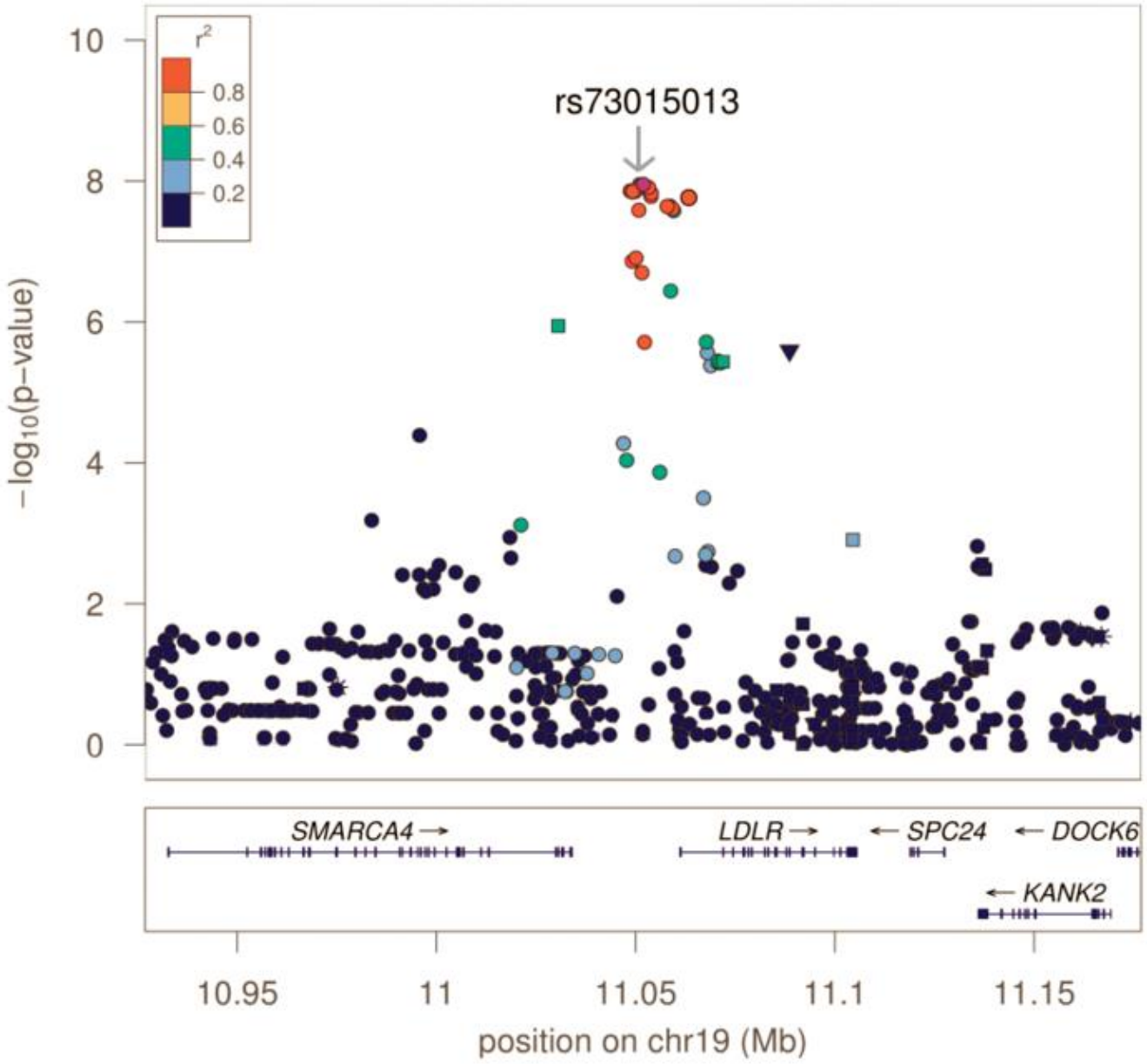
Output:

Specific regions are sometimes highlighted using ‘locus zoom’

Multiple testing means many tests will be ‘significant’ (P < 0.05) by chance,

e.g. 1M tests = 50,000 sig SNP

The accepted significance threshold for GWAS is usually $P < 5x10^{-8}$ $[-\log_{10}(P) > 7.3]$



Output:

Usually, GWAS ‘summary statistics’ are required (by journals) to be made public

```
SNP A1 A2 freq b se p N
rs1001 A G 0.8493 0.0024 0.0055 0.6653 129850
rs1002 C G 0.0306 0.0034 0.0115 0.7659 129799
rs1003 A C 0.5128 0.0045 0.0038 0.2319 129830
```

Cell Genomics

CellPress
OPEN ACCESS

Perspective

Workshop proceedings: GWAS summary statistics standards and sharing

Jacqueline A.L. MacArthur,[1,2,*] Annalisa Buniello,[1] Laura W. Harris,[1] James Hayhurst,[1] Aoife McMahon,[1] Elliot Sollis,[1] Maria Cerezo,[1] Peggy Hall,[3] Elizabeth Lewis,[1] Patricia L. Whetzel,[1] Orli G. Bahcall,[4] Inês Barroso,[5] Robert J. Carroll,[6] Michael Inouye,[7,8,9] Teri A. Manolio,[3] Stephen S. Rich,[10] Lucia A. Hindorff,[3] Ken Wiley,[3] and Helen Parkinson[1,*]

Table 1. Recommended standard reporting elements for GWAS SumStats

Data element	Column header	Mandatory/Optional
variant id	variant_id	One form of variant ID is mandatory, either rsID or chromosome, base pair location, and genome build[a]
chromosome	chromosome	
base pair location	base_pair_location	
p value	p_value	Mandatory
effect allele	effect_allele	Mandatory
other allele	other_allele	Mandatory
effect allele frequency	effect_allele_frequency	Mandatory
effect (odds ratio or beta)	odds_ratio or beta	Mandatory
standard error	standard_error	Mandatory
upper confidence interval	ci_upper	Optional
lower confidence interval	ci_lower	Optional

QC considerations:

The basic concept of a GWAS is simple

However, LOTS can go wrong!

The multiple testing burden is high, and small biases can add up

Most of the time required for GWAS is spent on QC

Diagnostics - how to tell if your results are biased?

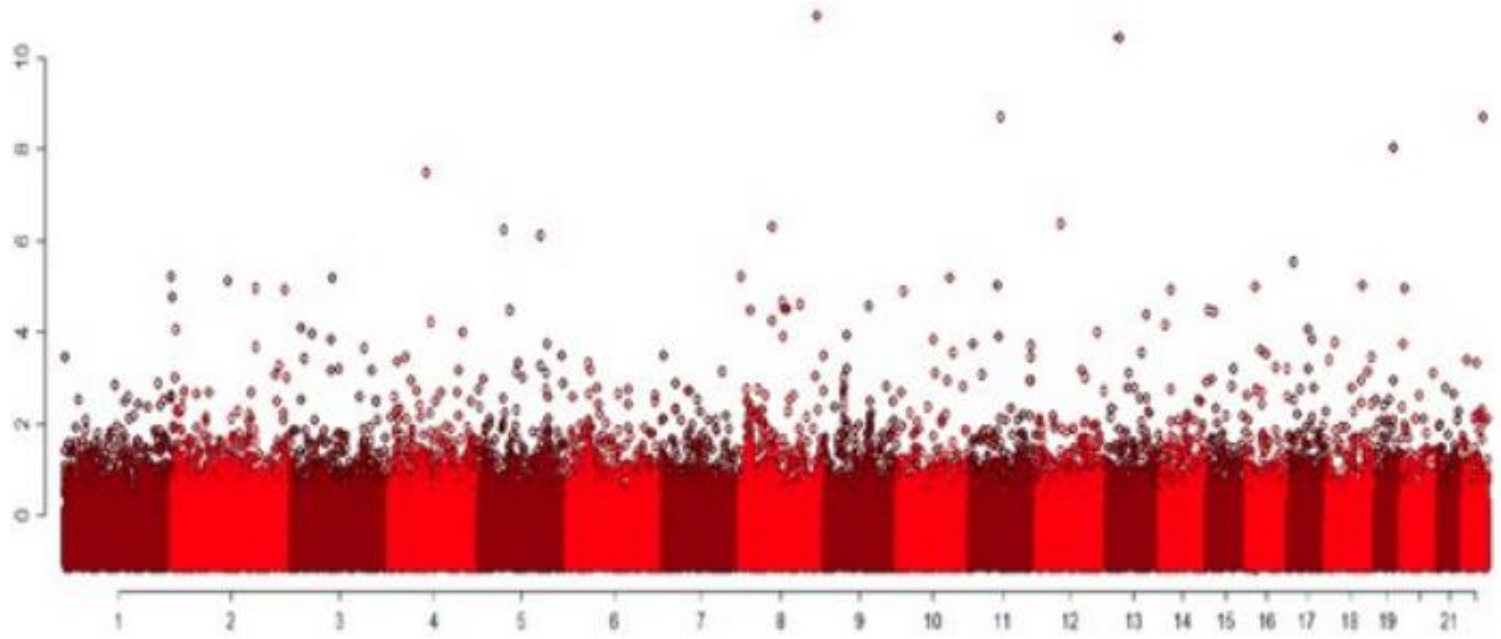
Examples, two common causes of bias and false-positive results

QC considerations - diagnostics

There are many ways to check your GWAS results

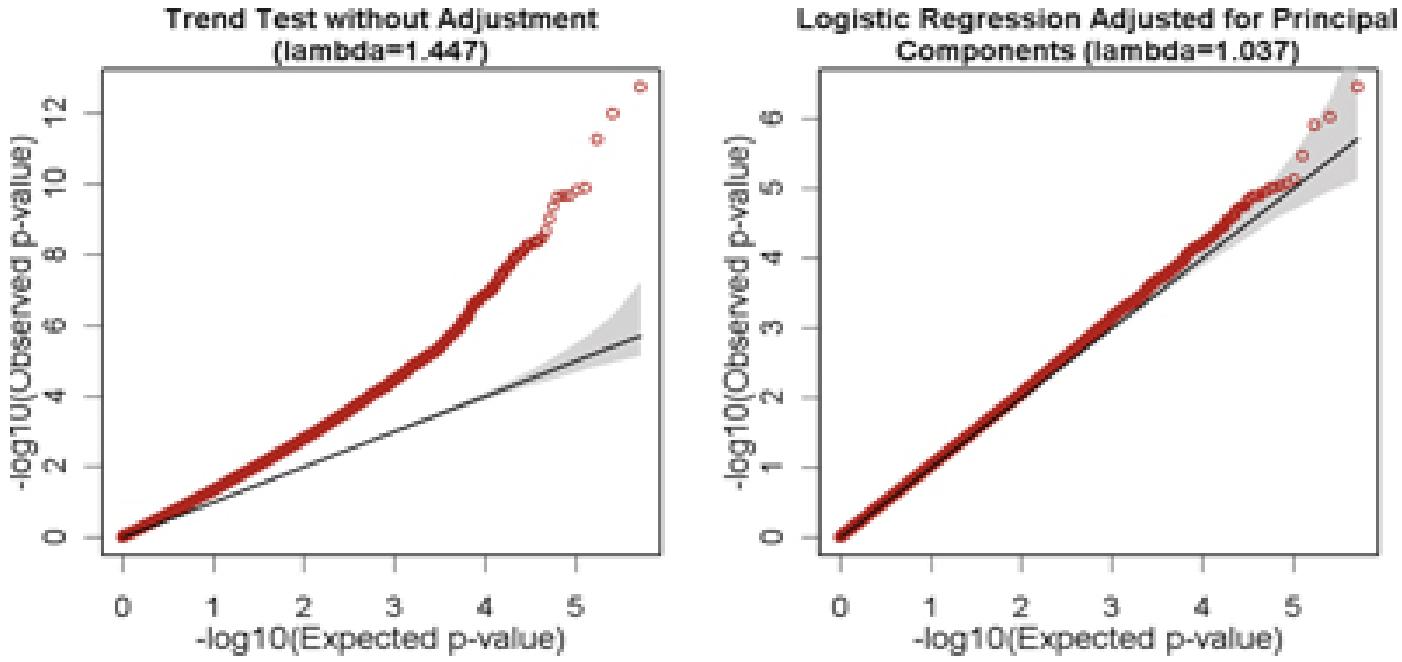
QC considerations - diagnostics

- Inspect your Manhattan plot
- e.g. a **bad** Manhattan plot - published but then retracted (!) for poor QC



QC considerations - diagnostics

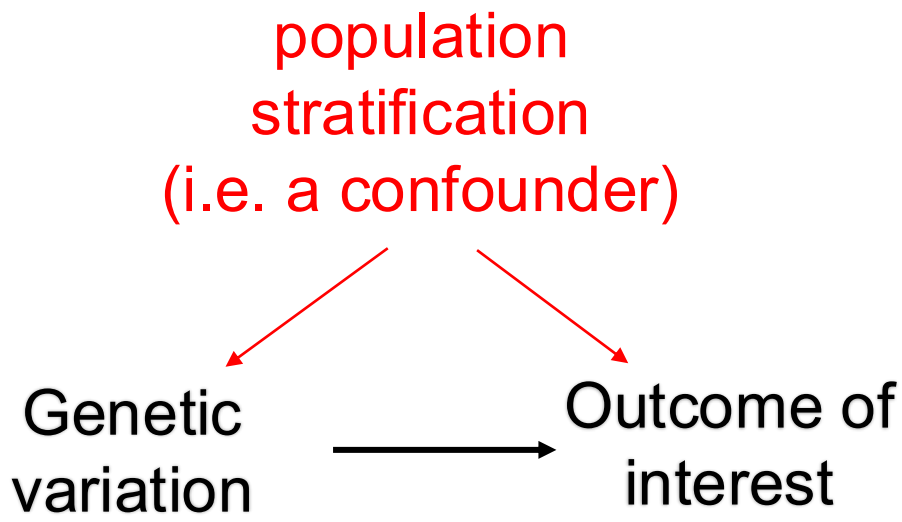
- compare your tests statistics to the null distribution,
- e.g. a QQ-plot



QC considerations: (1) Population stratification

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

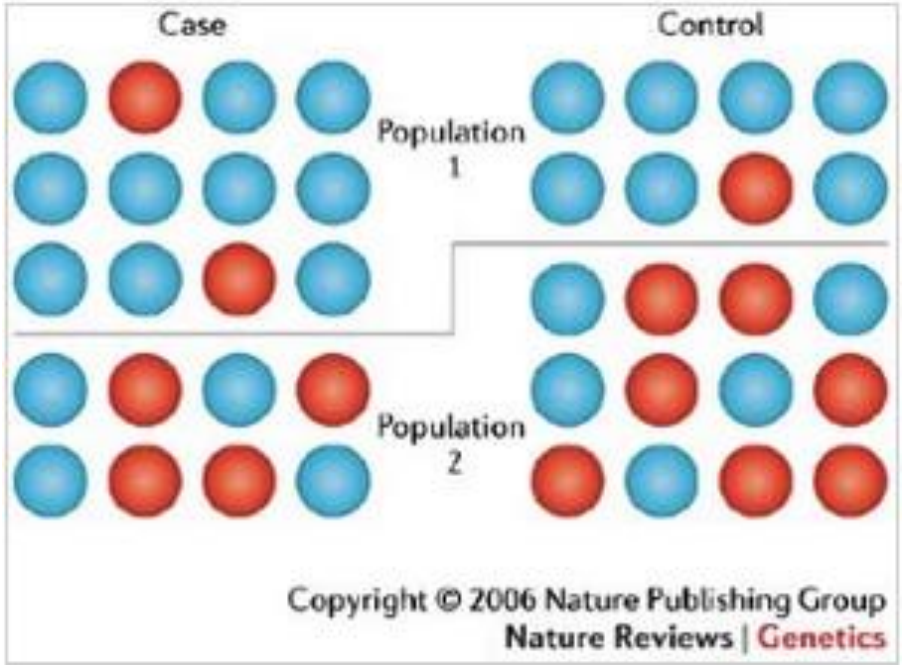
- Occurs when there are (unknown) subpopulations within the study sample which have *systematic differences in both ancestry and phenotypes*



QC considerations: (1) 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



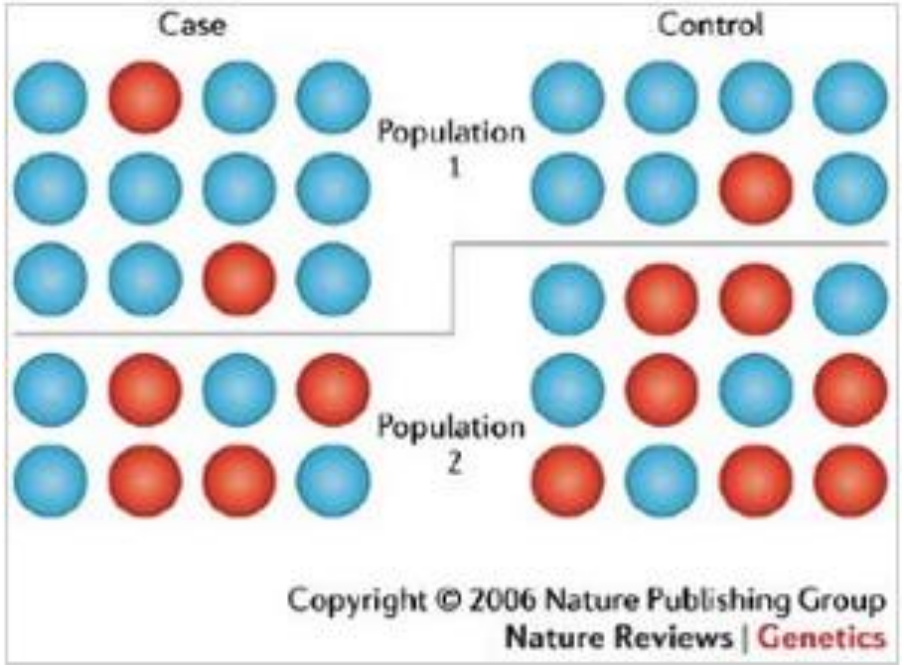
	Case	Control
ALL	14/20 = 0.7	12/20 = 0.6

Balding (2006) *Nature Rev. Genetics*

QC considerations: (1) 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*

collider bias / participation bias / G-E correlation

- Postal invitations were sent to 9.2 million individuals living near an assessment center, but 'only' 5.2% of those people joined the UK Biobank
- Those who joined were not a random sample of the UK population. Commonly referred to as 'healthy volunteer' bias. That is UKB participants tend to be, e.g.
  - from more affluent areas
  - non-smokers
  - use vitamin supplements
  - have lower rates of disease
  - etc.
- Be aware of how this selection bias may influence your results Various ways to address this issue:
  - sensitivity analysis
  - probability weightings
  - covariates
  - simulations

Technical Report | Open access | Published: 13 July 2023

Studying the genetics of participation using footprints left on the ascertained genotypes

Stefania Benonisdottir & Augustine Kong

Nature Genetics **55**, 1413–1420 (2023) | Cite this article

8923 Accesses | **4** Citations | **1248** Altmetric | Metrics

Abstract

The trait of participating in a genetic study probably has a genetic component. Identifying this component is difficult as we cannot compare genetic information of participants with nonparticipants directly, the latter being unavailable. Here, we show that alleles that are more common in participants than nonparticipants

QC considerations: (2) Genotype cleaning

- *WHY?*

Poor quality data = false positives / false negatives

Remove genotyping errors caused by e.g. low quality or quantity of DNA, contaminated DNA, chemical or machinery failure, human error, failure in clustering intensities

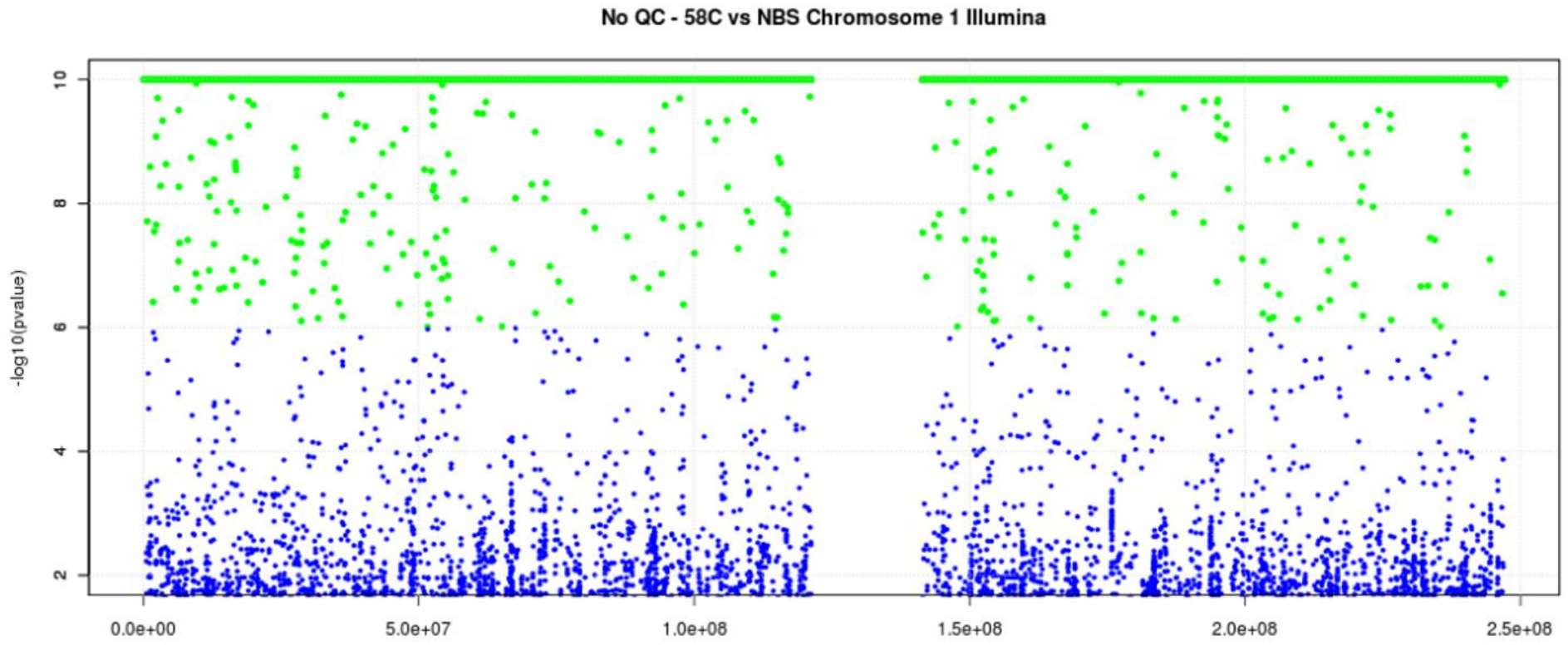
Example: Importance of Good Cleaning

- The WTCCC study used controls from two populations:

  - 1,500 from the 1958 British Birth Cohort (58C)
  - 1,500 from the National Blood Service (NBS)

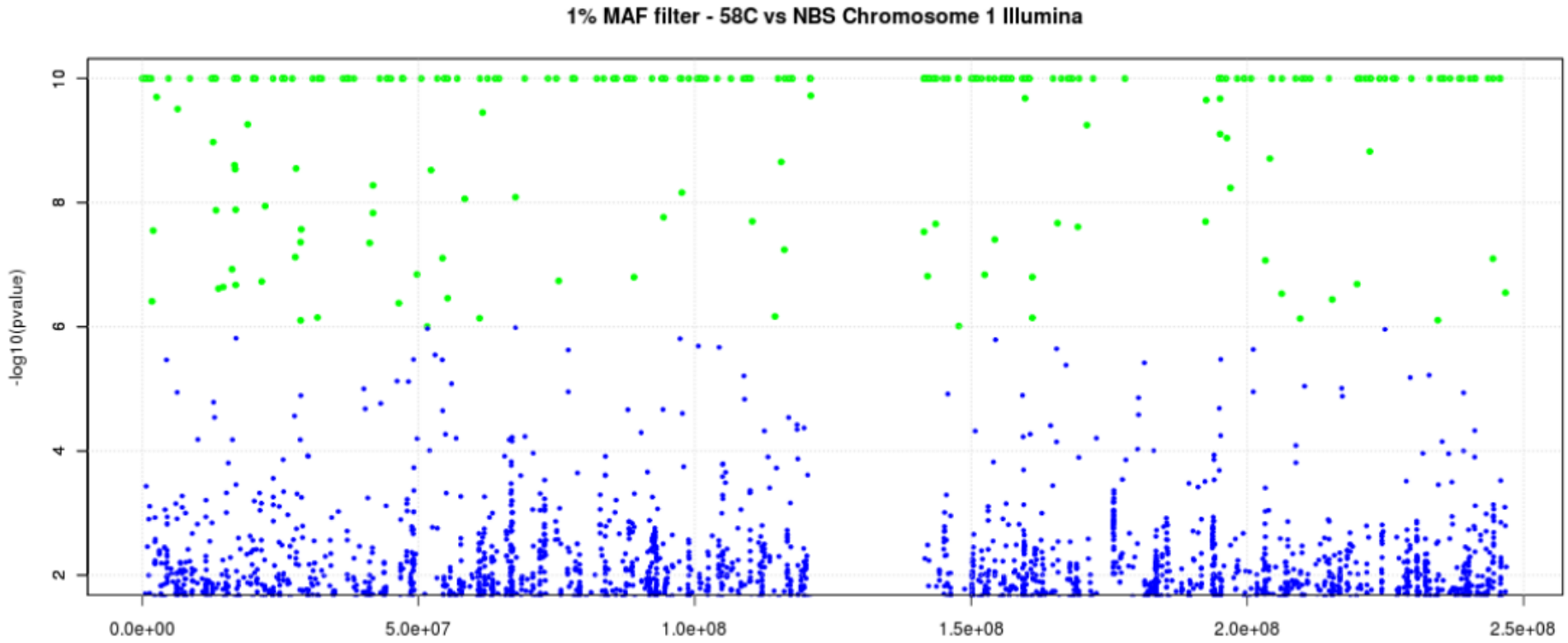
- Both these are unselected population cohorts, so performing a “case-control” study between these populations should find no significant differences

Importance of Good Cleaning



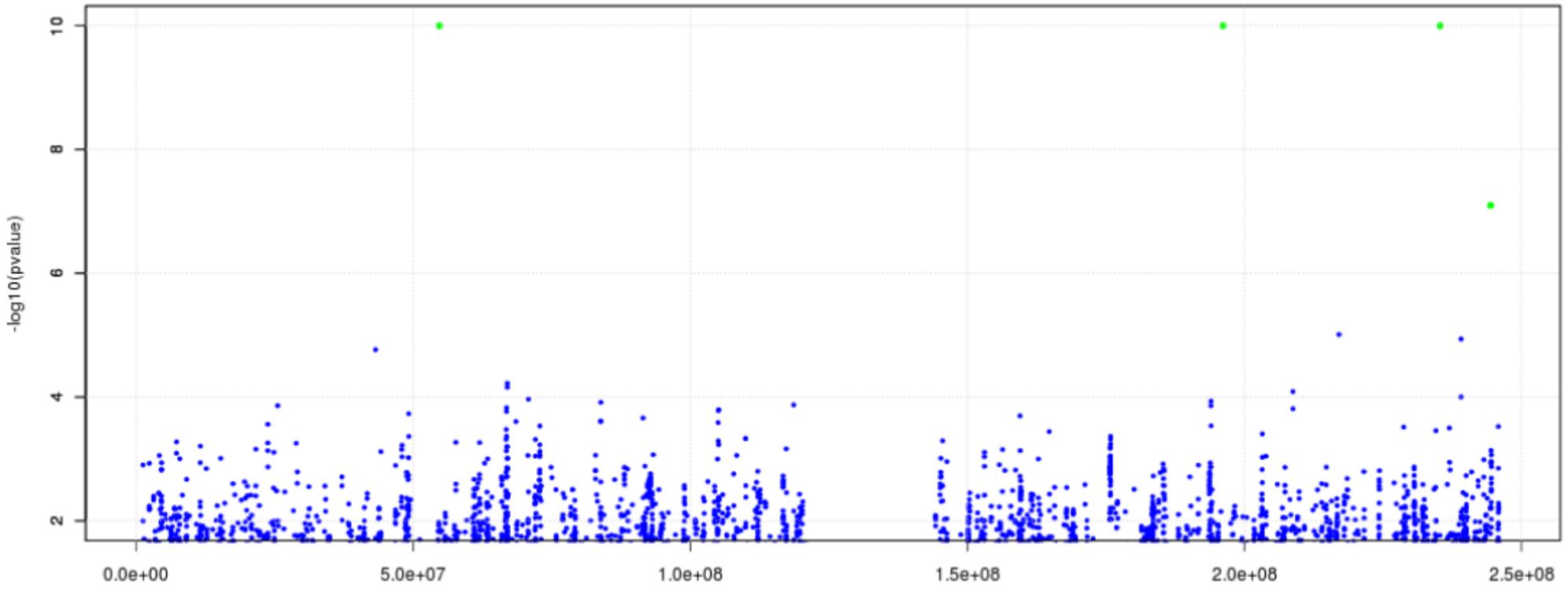
100% of SNPs

Importance of Good Cleaning



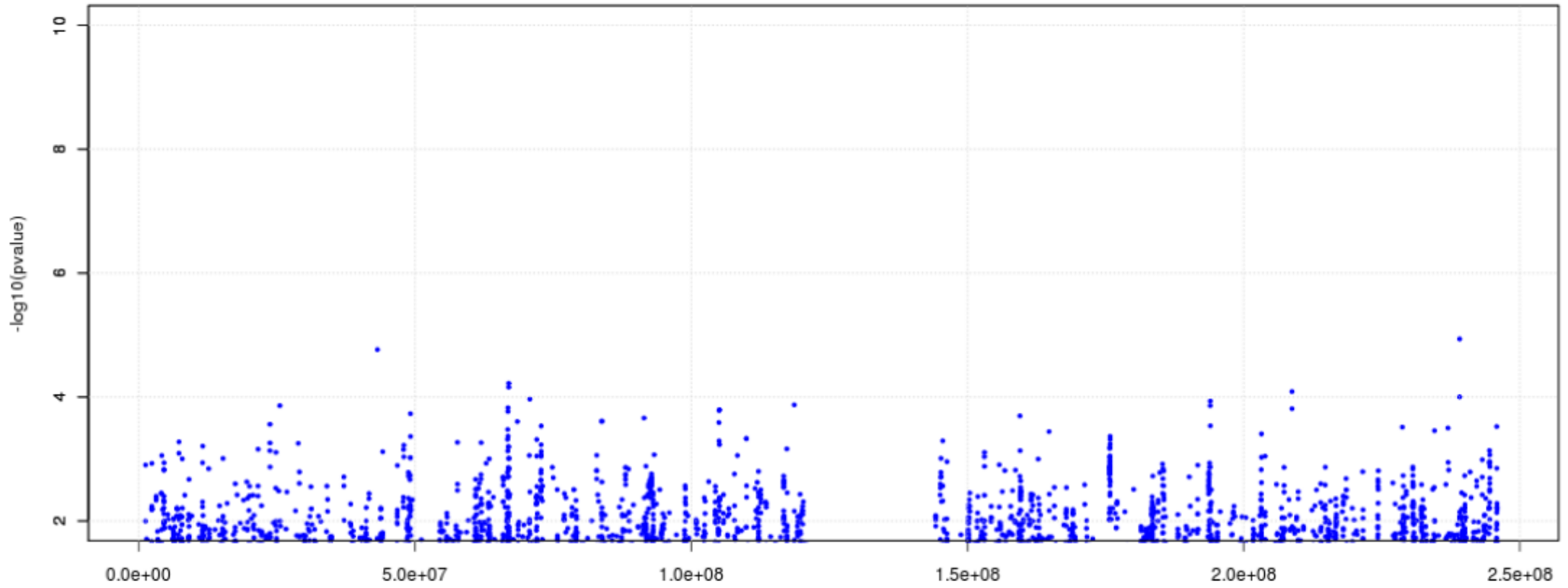
80.69% of SNPS
Filtering: minor allele frequency (MAF)

Importance of Good Cleaning



78.36% of SNPs
Filtering: MAF + Hardy Weinberg Equilibrium (HWE)

Importance of Good Cleaning



77.92% of SNPs
Filtering: MAF + HWE + Missingness

Software considerations

- There are many (command line) based programs to do GWAS
- I use/recommend PLINK and GCTA

PLINK is a free, open-source whole genome association analysis toolset

- Efficiently store, manipulate and analyse large datasets
- Many options, good for data manipulation & cleaning
- Widely used, v2.0 very efficient

https://www.cog-genomics.org/plink/

GCTA is also free

- Primarily used for REML, but has some useful GWAS capabilities
- Similar to PLINK, interchangeable files

https://yanglab.westlake.edu.cn/software/gcta/#Overview

Summary

- GWAS scan the genome for associations between DNA markers (e.g. SNPs) and a trait of interest (e.g. height or heart disease) to identify genomic regions associated with the trait
  - rely on linkage disequilibrium between markers & causal variant
  - summary stats are starting point for many downstream analyses

- Multiple testing burden magnifies any biases present
  - QC is very important, e.g. cleaning genotypes & population stratification