THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

MODULE 1 | GENETIC MAPPING

Session 5. Trouble Shooting

July 2026

Slides, practicals & data can be downloaded from the cluster:
`/data/module1/downloadsTuesPM.zip`

Slides can be downloaded from the website:
https://cnsgenomics.com/data/teaching/GNGWS26/module1/

Outline

GWAS is a simple concept but requires careful thought

lots can go wrong!

Common issues and suggested fixed

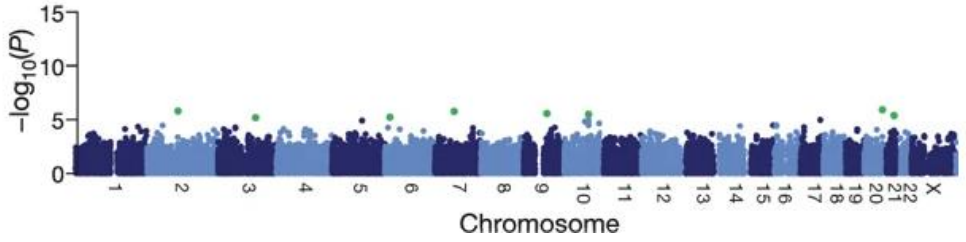
Reality check

GWAS diagnostics

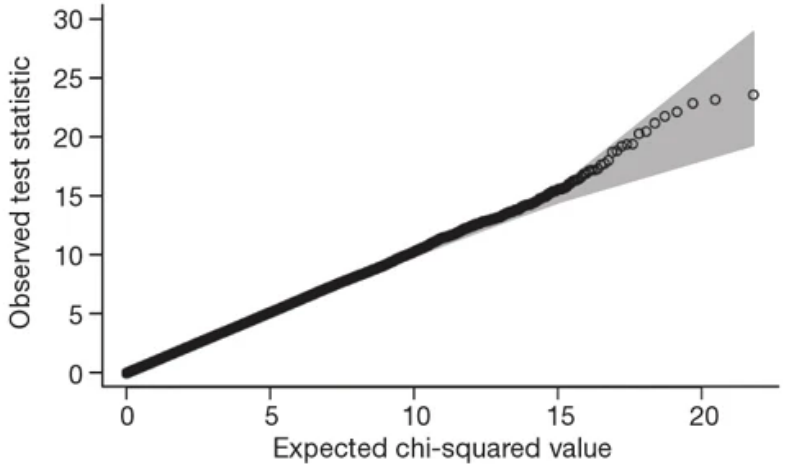
GWAS performs millions to statistical tests – do you believe your results?

How can you tell?

Manhattan plot



QQ plot



λ_{GC}*

$$\lambda_{GC} = \frac{median(\chi^2)}{0.454}$$

Inflated: $\lambda_{GC} >> 1$

Deflated: $\lambda_{GC} << 1$

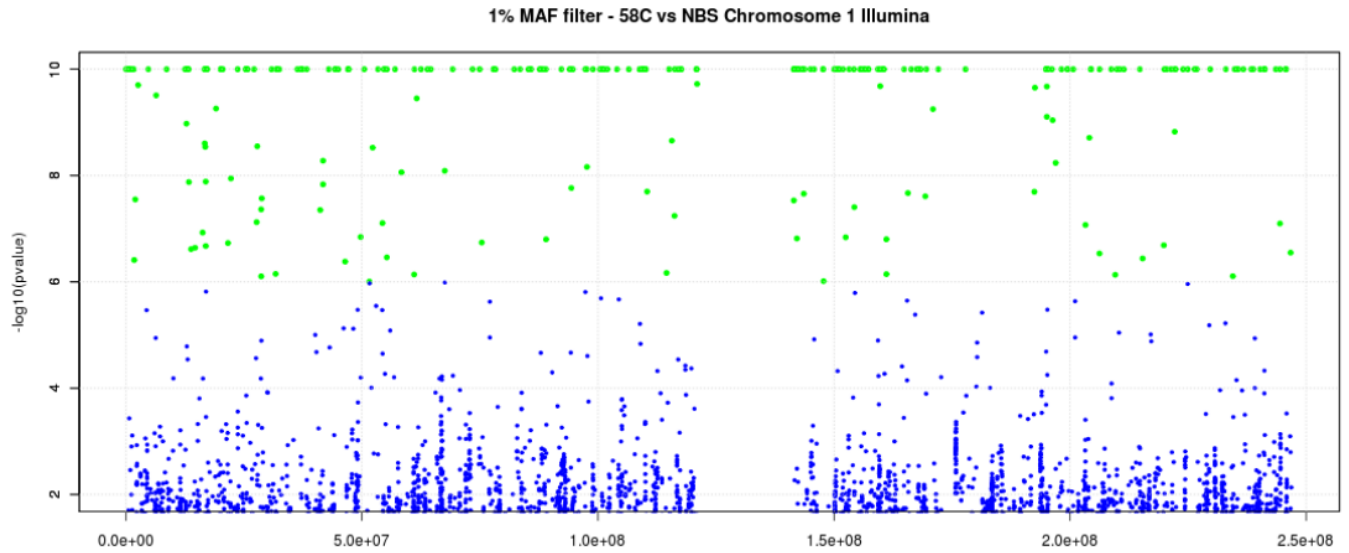
*scales with sample size

Issue : Technical artifacts

False-positives caused by genotyping errors / technical artifacts

How to spot it?

Single isolated ‘significant’ SNPs, strange patterns in Manhattan plots, inflated test statistics



Issue : Technical artifacts

False-positives caused by genotyping errors / technical artifacts

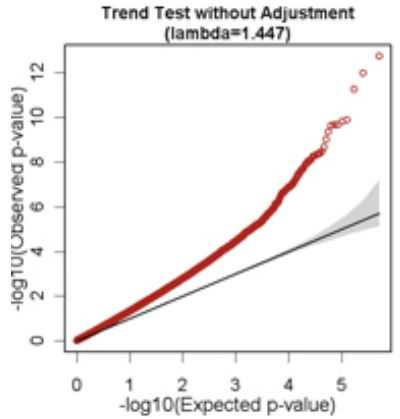
How to spot it?

Single isolated ‘significant’ SNPs, strange patterns in Manhattan plots, inflated test statistics

*How to fix it. G*enotype QC at sample & SNP level, check phenotypes

e.g. exclude SNP with high missingness, remove rare variants, SNP out of HWE, samples with lots of missingness or heterozygosity, careful handling of batch effects + randomisation of cases/controls

Issue : Inflated test statistics



What's the issue:

"I've carefully QC'd my GWAS but my QQ-plot looks inflated. Help."

First response: check your GWAS model

(1) Normality. Are your residuals normally distributed? Outliers

(2) Model specification. Have you included relevant covariates? e.g. age, sex, contemporary or socioeconomic group

(3) Independence. Is there population structure?

Issue : Inflated test statistics

How homogeneous is your population?

How did you make your PCs?

Are you overfitting?

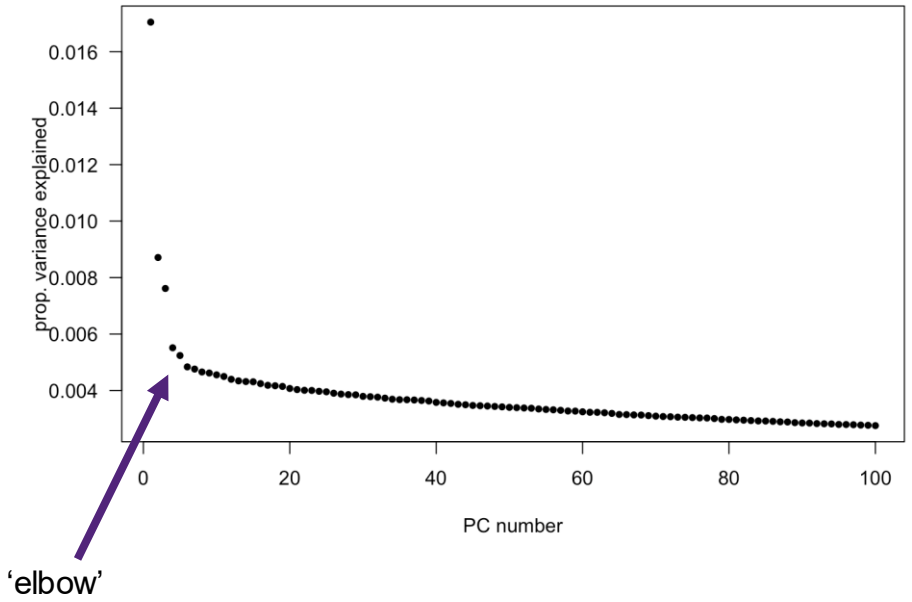
Population stratification fix: remove ancestry outliers, fit PCs or GRM.

How many PCs?

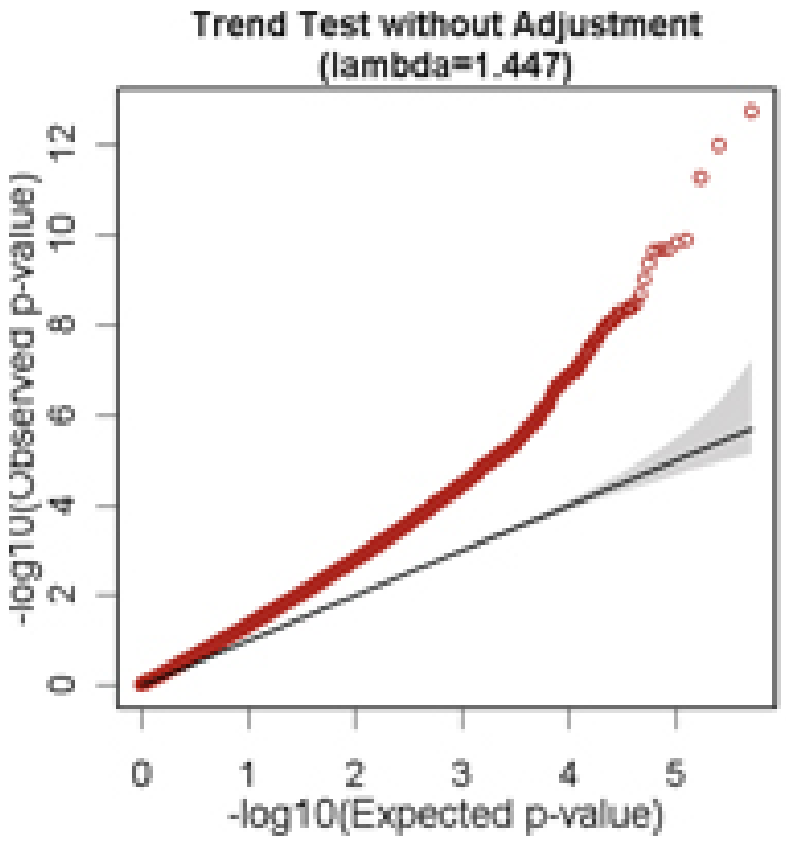
- In humans, typically 10-20 PCs
- OPTION 1: visual inspection of eigenvalues
- OPTION 2: test PCs against phenotype

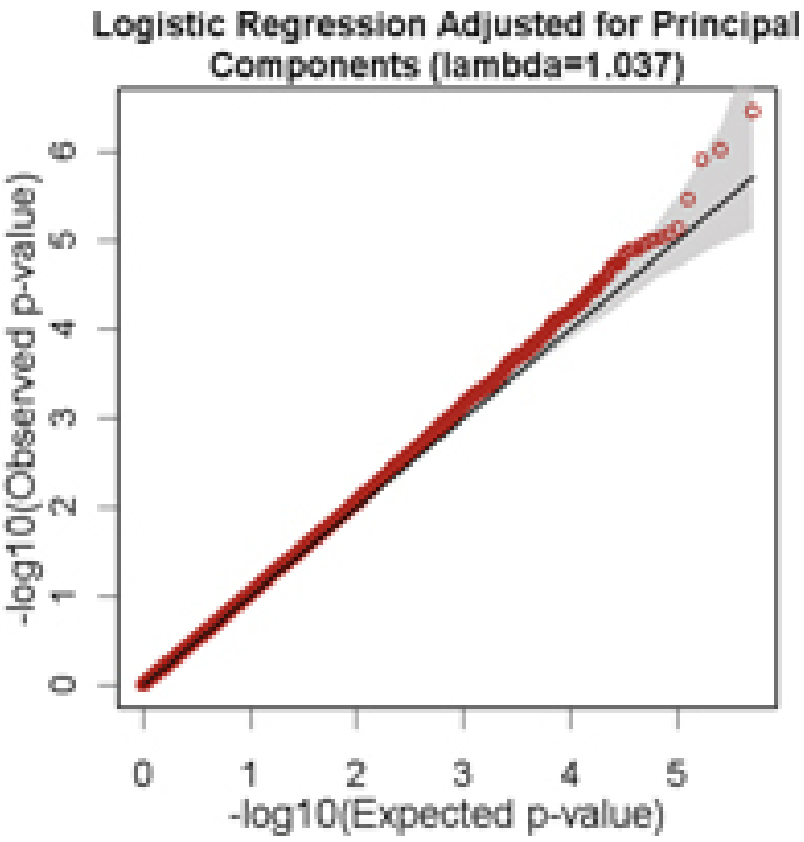
Or use a GRM?

- Alternative to PCs, more reliable particularly for heterogeneous or highly structured populations



Issue : Inflated test statistics





Issue : Inflated test statistics

The (naïve) inclusion of relatives means that:

1. test statistics are inflated, s.e. are too small ~ samples are not independent!
2. effects might also be biased. e.g. in humans, family members also share common ‘E’

How to spot it?

Genomic relationship matrix, e.g. off-diagonals > 0.05 (2^{nd} cousins)

How to fix it?

Lots of relatives: use MLM (fast-GWA, BOLT-LMM) for association testing

Only a few: remove one member of each pair

Issue : Inflated test statistics

“I’ve carefully QC’d my GWAS AND accounted for population structure but my QQ-plot still looks inflated. Argh.”

Polygenicity can cause inflated test statics in the absence of other causes

Worse for extremely well powered studies

How to spot it?

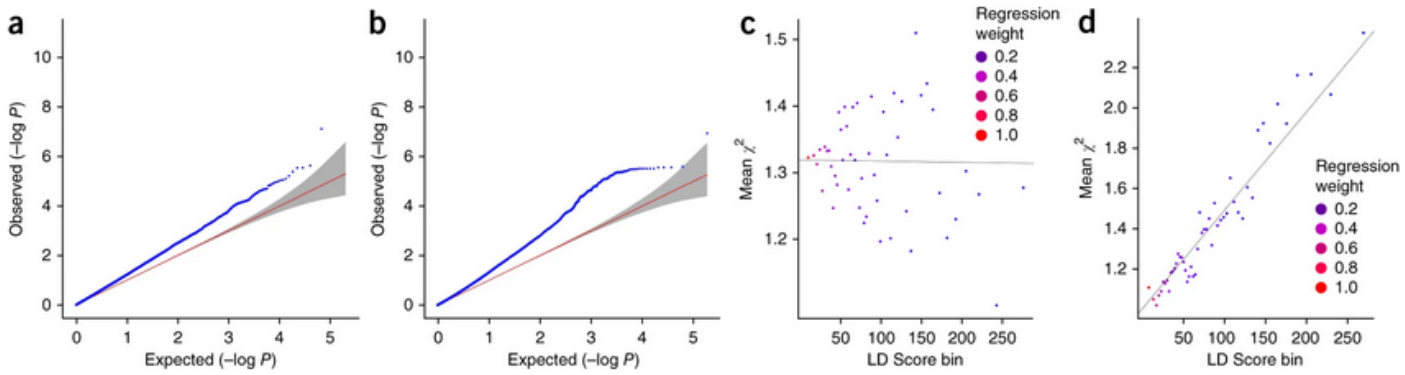
LD score regression

Issue : Inflated test statistics

LD score regression to distinguish between stratification and polygenicity

intercept ~ stratification

slope ~ polygenicity



(**a**) Quantile-quantile plot with population stratification ($\lambda_{GC} = 1.32$, LD Score regression intercept = 1.30). (**b**) Quantile-quantile plot with a polygenic genetic architecture where 0.1% of SNPs are causal ($\lambda_{GC} = 1.32$, LD Score regression intercept = 1.006). (**c**) LD Score plot with population stratification. Each point represents an LD Score quantile, where the *x* coordinate of the point is the mean LD Score of variants in that quantile and the *y* coordinate is the mean χ^2 statistic of variants in that quantile. Colors correspond to regression weights, with red indicating large weight. The black line is the LD Score regression line. (**d**) LD Score plot as in **c** but with polygenic genetic architecture.

Bulik-Sullivan et al. Nature Genetics (2015)

Issue : Inflated test statistics

Inflated: $\lambda_{GC} >> 1$

Underfitting

- increased false-positives i.e. sig loci that are not real (type I error)

- more typical situation

Sweet Spot

Deflated: $\lambda_{GC} << 1$

Overfitting

- increased false-negatives i.e. true loci are not detected (type II error)

- deflated test statistic, unusual. Overfitting PCs or GRM in homogenous population with low power.

Issue : Genotype-Environment confounding

G-E confounding occurs when genotypes are not randomly distributed across environments, i.e. your genotype predicts your environment

e.g. British individuals tend to share ancestry & cultural habits OR indirect genetic effects

How to spot it?

...did you measure ‘E’? -> yes, plot phenotype vs. E, F_{ST} by E, GWAS with E as phenotype, etc.

...is your ‘E’ heritable? -> yes, compare population vs. within family estimates

How to fix it.

adjust for ‘E’ if you can, a heritable ‘E’ probably needs a different design

Issue : Genotype-Environment confounding

e.g. Educational attainment

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01062-7

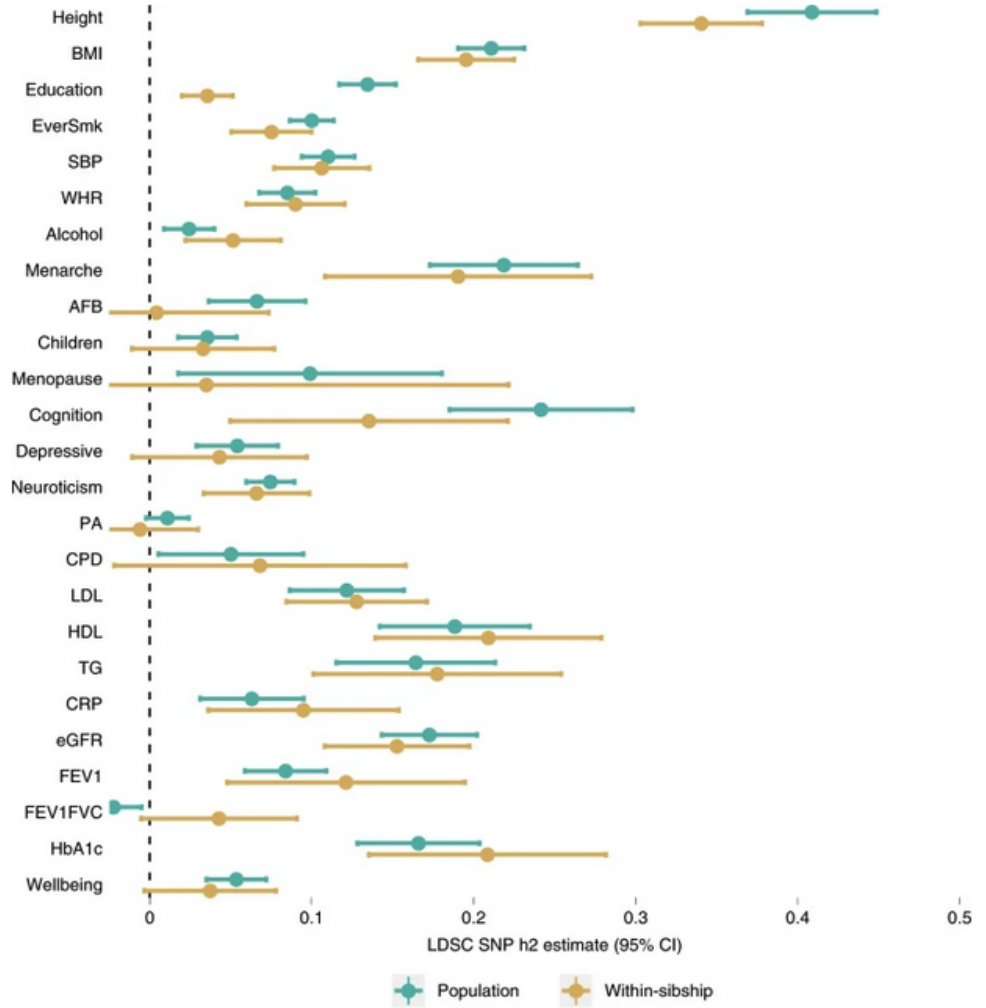
Check for updates

OPEN

Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects

Estimates from genome-wide association studies (GWAS) of unrelated individuals capture effects of inherited variation (direct effects), demography (population stratification, assortative mating) and relatives (indirect genetic effects). Family-based GWAS designs can control for demographic and indirect genetic effects, but large-scale family datasets have been lacking. We combined data from 178,086 siblings from 19 cohorts to generate population (between-family) and within-sibship (within-family) GWAS estimates for 25 phenotypes. Within-sibship GWAS estimates were smaller than

Fig. 5: LDSC SNP h^2 estimates for 25 phenotypes using population and within-sibship meta-analysis data with corresponding 95% CIs.



Other issues... – reality check!

With huge genetic studies, genetic signatures of many different things are being found.

e.g. *assortative mating* (tall people tend to pair with other tall people)

participation bias (UK Biobank has healthy volunteer bias)

survival bias (variants causing early mortality absent from older cohorts)

Often, we don't know about these factors and/or can't do much about them unless we specifically start looking their effects

Summary

GWAS are powerful but it can be difficult to be confident in your results

Try to avoid false-positives

e.g. cause by population stratification, poor QC, naïve inclusion of relatives, multiple testing

However, real genetic signals can also be confounded in GWAS and these are more difficult to sort out

e.g. indirect genetic effects