MODULE 1 | GENETIC MAPPING

Session 4. GWAS in practice I – Quality Control

July 2026

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

GWAS in practice

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Part 1: Quality control

- Genotype QC - SNP and sample level
- Relatedness / ancestry outliers
- Phenotype QC and covariates
- Practical: genotype QC with PLINK

Part 2: GWAS Models

- GWAS models
- Quantitative traits
- Binary traits
- GWAS with relatives
- Practical: Run a GWAS using each of these models

Why do we need to do quality control?

Poor quality data → false positives / negatives

- To remove genotyping errors
  - Low quality or quantity of DNA
  - Contaminated DNA
  - Chemical or machinery failure
  - Human error
  - Failure in clustering of intensities
- To ensure data suitable for the analyses
  - Relatedness
  - Population structure

PLINK

- PLINK is a free, open-source whole genome association analysis toolset
  - Efficiently store, manipulate and analyse large datasets
  - Widely used
- Run PLINK via command line
- `plink --bfile` *`filename`* `--missing --out` *`newfilename`*
- If you have downloaded PLINK into your local directory, could be: ./plink

```
delta2:~/60days/UQWS_2023$ plink
PLINK v1.90b6.22 64-bit (16 Apr 2021)          www.cog-genomics.org/plink/1.9/
(C) 2005-2021 Shaun Purcell, Christopher Chang   GNU General Public License v3

  plink <input flag(s)...> [command flag(s)...] [other flag(s)...]
  plink --help [flag name(s)...]

Commands include --make-bed, --recode, --flip-scan, --merge-list,
--write-snplist, --list-duplicate-vars, --freqx, --missing, --test-mishap,
--hardy, --mendel, --ibc, --impute-sex, --indep-pairphase, --r2, --show-tags,
--blocks, --distance, --genome, --homozyg, --make-rel, --make-grm-gz,
```

PLINK data format

- Three files:
  - gwas.bim → information about SNP markers
  - gwas.fam → information about individuals
  - gwas.bed → binary file containing all genotypes

```
[allan@analysis1 ~]$ head /data/module1/gwas/part1/gwas.bim
1       rs3131972       0       752721  1       2
1       rs3115850       0       761147  1       2
1       rs12562034      0       768448  1       2
1       rs4040617       0       779322  2       1
1       rs4970383       0       838555  1       2
1       rs950122        0       846864  1       2
1       rs6657440       0       850780  2       1
1       rs13303101      0       862124  1       2
1       rs1110052       0       873558  2       1
1       rs3748592       0       880238  1       2
```

chromosome

SNP ID

(dummy variable) position, cM

position, bp

allele 1

allele 2

```
[allan@analysis1 ~]$ head /data/module1/gwas/part1/gwas.fam
7653762 7653762 0 0 2 -9
8144519 8144519 0 0 2 -9
2337680 2337680 0 0 2 -9
5219864 5219864 0 0 1 -9
1417721 1417721 0 0 1 -9
2371103 2371103 0 0 2 -9
472262 472262 0 0 1 -9
566177 566177 0 0 2 -9
8097907 8097907 0 0 2 -9
8738370 8738370 0 0 2 -9
```

FID (Family ID)

IID (individual ID)

PID/MID/sex/phenotype
-> not necessary to specify, but “0 0 0 -9” is needed

- Other input formats also specified on the PLINK website

3 main versions of PLINK

Feature	PLINK 1 (1.07)	PLINK 1.9	PLINK 2
Status	Legacy / Retired	Fully stable and complete	Advanced alpha (active development)
Speed	Slow (often limited by memory)	1-4 orders of magnitude faster than PLINK 1	Even faster, highly parallelized multithreading
Native File Formats	`.bed`, `.bim`, `.fam`	`.bed`, `.bim`, `.fam`	`.pgen`, `.pvar`, `.psam` (faster I/O)
Allele Definitions	Biallelic (A1/A2, determined by frequency)	Biallelic (A1/A2, determined by frequency)	Multiallelic natively supported (REF/ALT)
Missing Data / Dosage	Lossy import of dosage/probabilities	Lossy import of dosage/probabilities	Retains genotype likelihoods & phased states

Other file formats

Variant Call Formats (`.vcf, .bcf`) for sequencing data

Oxford format (`.gen, .bgen, .sample`)

GCTA

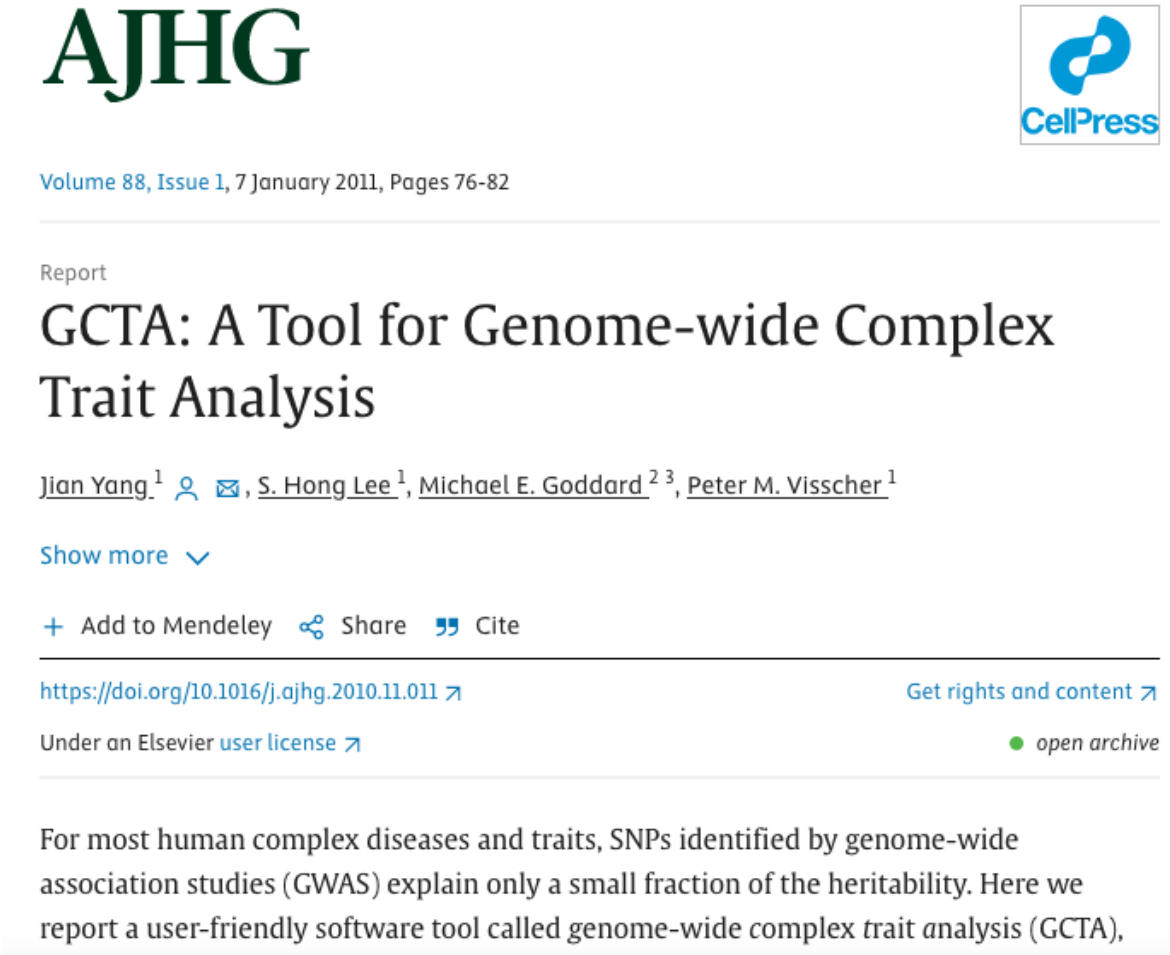
- We will also use GCTA

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

- Runs like PLINK, same command format and input format

gcta64 --bfile <data prefix> --command

- Primarily for variance component estimation via REML (StatGen2 module) but has expanded to include other useful features



Quality control for genotype data

We divide the cleaning of genotype data into two steps

STEP 1) removing any individuals with poor quality genotype data

STEP 2) removing SNP markers that have substandard genotyping performance

- Performing the per-individual steps first prevents individuals with poor quality genotypes having an undue influence on the removal of SNP markers in the later step.
- We use on statistical measures to detect bad quality data and remove it

```
plink --bfile filename --maf 0.01
```

Per Individual Quality Control

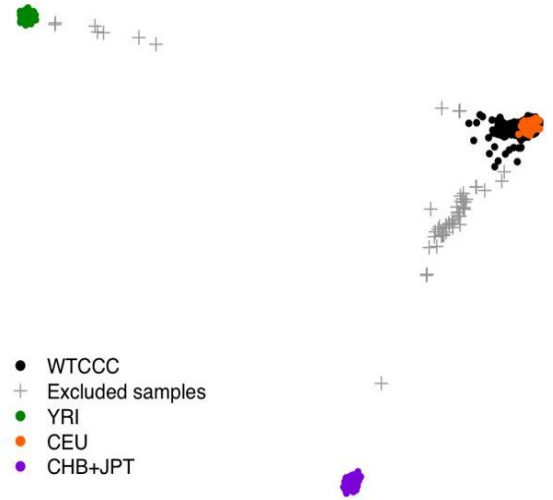
Suggestions for removing individuals with ‘poor quality’ genotypes

1. Removal of individuals with excess missing genotypes
2. Removal of individuals with outlying homozygosity values
3. Remove of samples showing a discordant sex (self reported sex and sex from X chromosome heterozygosity)

Ensure data suitable for the analyses

4. Removing of related or duplicate samples, and
5. Population structure - participants in the GWAS do not represent multiple subgroups with different allele distributions (e.g., genetic ancestry groups)

Per Individual Quality Control - removal of ancestry outliers



This can take a LONG time to run!

1. Download and perform PCA on diverse individuals with known ancestry, e.g. 1000 Genomes
2. Project your samples onto PCs
3. Exclude ‘outliers’ from further analysis

e.g. with GCTA

Example

REF: SNP genotype data of the reference sample; TAR: SNP genotype data of the target sample;

```
# To make a GRM
gcta64 --bfile REF --maf 0.01 --autosome --make-grm --out REF
# PCA analysis
gcta64 --grm REF --pca 20 --out REF_pca20

# To use the PCs generated above to produce PC loadings of each SNP
gcta64 --bfile REF --pc-loading REF_pca20 --out REF_snp_loading

# To compute the PCs of the target sample using the PC loading generated above
# Note that the analysis can be performed with one chromosome at a time
gcta64 --bfile TAR --project-loading REF_snp_loading 20 --out TAR_pca20
```

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

Per SNP Quality Control

Suggestions for removing ‘bad’ SNPs,

1. Removal of SNPs with excess missing genotypes
2. Removal of SNPs that deviate from Hardy-Weinberg equilibrium
3. Remove of SNPs with low minor allele frequency
4. Comparing allele frequency to known values (from reference dataset)

Post imputation QC

Further QC steps after imputation

- Removal of SNPs with low imputation quality - INFO score < 0.8
- Removal of SNPs that deviate from Hardy-Weinberg equilibrium
- Remove of SNPs with low minor allele frequency

Phenotypic data preparation

- Inclusion / exclusion criteria
  - Data collection details

- File Format
  - File type (PLINK or other)
  - Header, delimiter, missing-value codes

- Continuous data – is it normally distributed
- Binary data – case/control ratio, cases *should* be the same as controls in all aspects except for the case status

Covariates

- Account for variables that influence the outcome (or genotyping data quality)
- Classic Covariates- age, age^2, biological sex
- Laboratory/Study Design-Related Concerns
  - Processing/isolation of DNA
  - Genotyping batch effects
  - Project study site
  - Multiple genetic ancestry (sub)groups

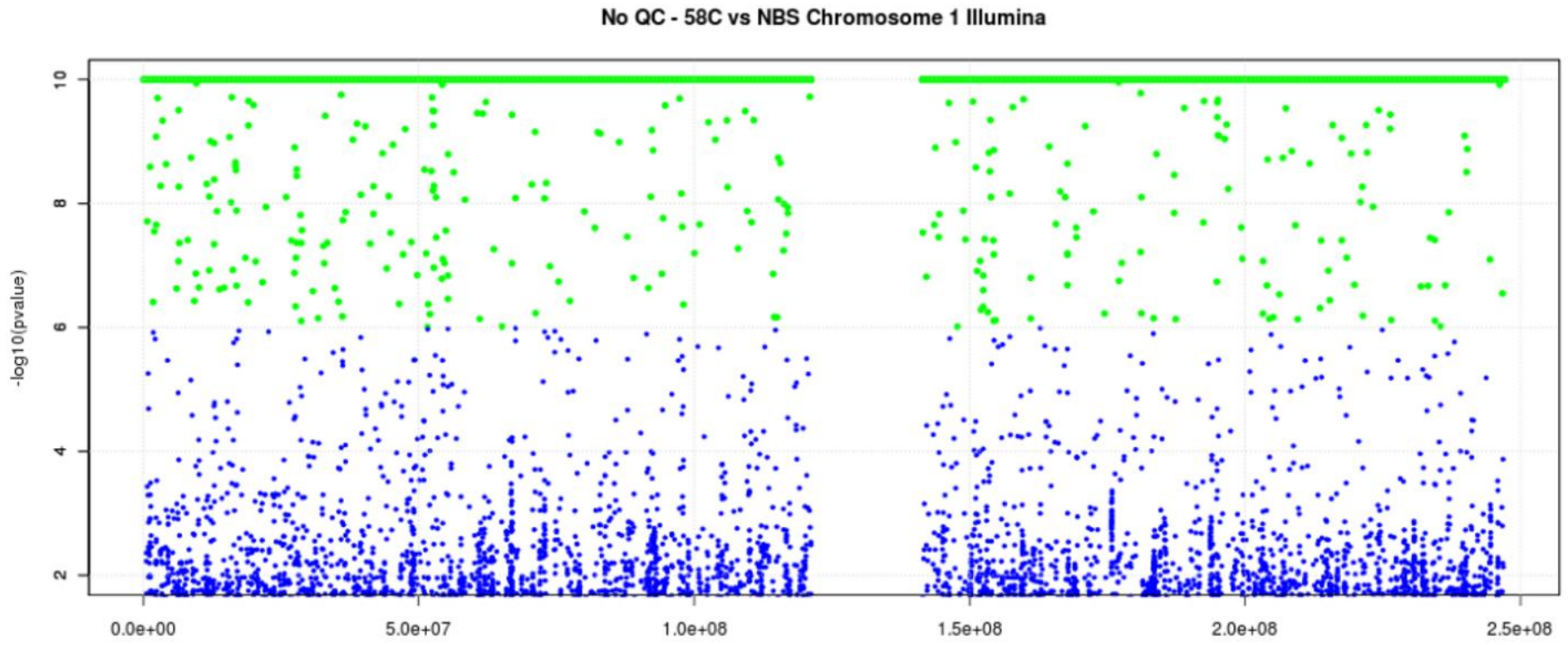
Example: Importance of Good Cleaning

- 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)

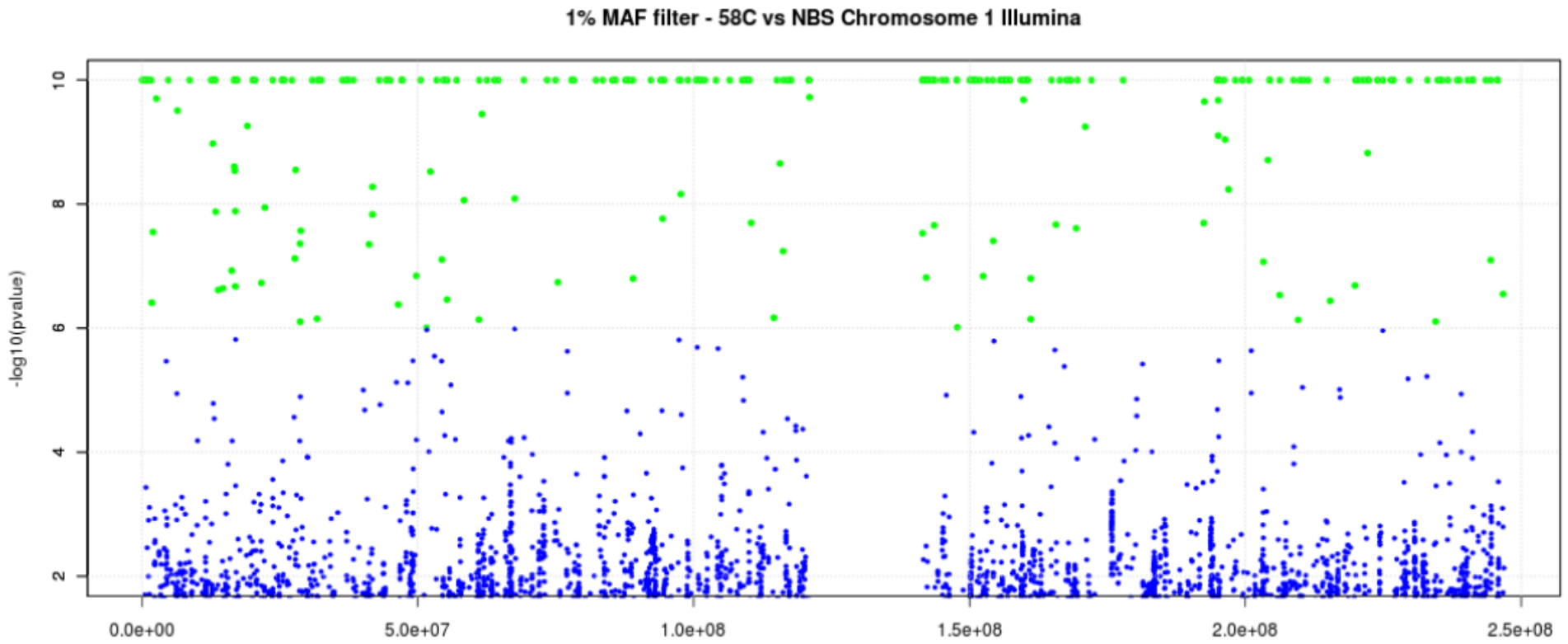
- Both these are unselected population cohorts, so performing a “case-control” study between these populations should find no significant differences

Importance of Good Cleaning



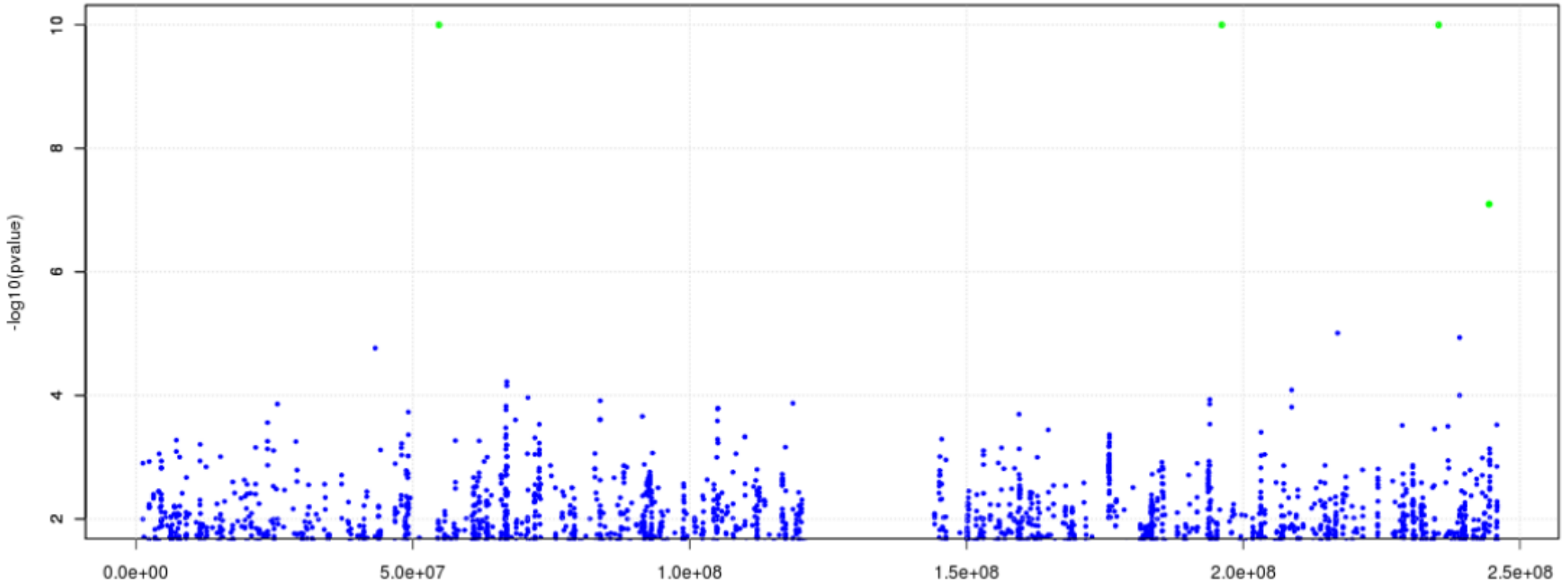
100% of SNPs

Importance of Good Cleaning



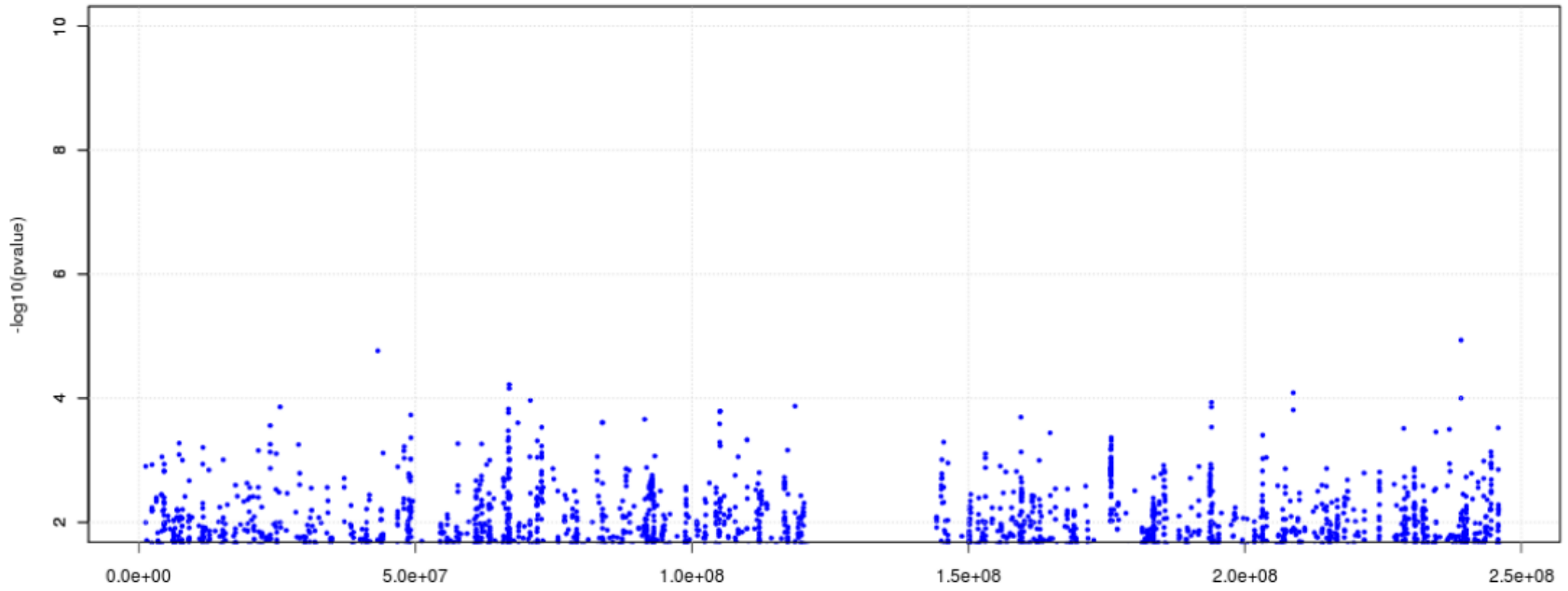
80.69% of SNPS
Filtering: MAF

Importance of Good Cleaning



78.36% of SNPs
Filtering: MAF + HWE

Importance of Good Cleaning



77.92% of SNPs

Filtering: MAF + HWE + Missingness

Practical - use PLINK for genotype QC

- Summary of PLINK commands
  - the commands can be run individually to help visualise what you're doing, and for trouble shooting
  - In practice, they are usually grouped & several commands run in a single step where appropriate

Individual QC	command	SNP QC	command
1) Excess missing genotypes	--missing	1) Excess missingness	--missing
2) Outlying homozygosity	--het	2) Hardy-Weinberg equilibrium	--hardy
3) Discordant Sex	--check-sex	3) MAF	--maf
4) Remove relatives	--genome --rel-cutoff	4) Compare to known allele freq	--freq

Genotype QC

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

QC step	Check	Rationale	Analysis
Set up	Genome build	Ensure proper SNP location alignment	Liftover (if needed)
Sample-level QC	Missingness	High levels of SNP genotypes missingness can indicate poor quality DNA or technical problems	Remove if >5%
	Heterozygosity	High individual-level SNP heterozygosity might be due to low sample quality. Low levels of heterozygosity may result from inbreeding	Remove if >3SD
	Sex Check	A discrepancy may reflect sample mix-ups in the lab	Drop sample if sex error
	Relatedness	Without appropriate correction, the inclusion of data from relatives could bias estimations of SNP effect sizes and standard errors	Remove one individual from each pair with GRM value >0.05
	Ancestry	Allele frequencies can differ between genetic ancestry subpopulations. Without appropriate correction, population stratification can lead to false positives and/or mask true associations	PCA
SNP-level QC	Missingness	Ensure small proportion of SNPs that are missing	Remove if >5%
	Minor allele frequency	Most GWAS are underpowered to detect associations for SNPs with low MAF	Remove if <1% (0.5%)
	Hardy Weinberg Equilibrium	Ensure use of SNPs with genotype and allele frequences that are constant over generations	Remove if p<1e-6, threshold may differ
	Ambiguous SNPs	SNPs with indeterminant calls may result in incorrect interpretations	Remove strand-ambiguous SNPs

Phenotype QC

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

QC step	Check	Rationale	Analysis
Phenotype set-up	Verify phenotype coding	May result in incorrect interpretations or challenges when aligning with other studies (i.e meta-analysis)	Check units for continuous traits Check case-control definitions and coding Inclusion/exclusion criteria
	File format		Header, delimiter, missing-value codes
	Duplicates or repeated measures		Apply a rule (e.g. latest-visit, case-status)
	Related samples		Use appropriate GWAS analysis method
Binary traits	Unbalanced case/control ratio	Extremely unbalanced ratios may increase Type I error (false positive)	Use appropriate GWAS analysis method
Continuous traits	Distribution	Failure to assess and address may reduce sensitivity of measurement and subsequent power to detect real effect of genetic association	Use histogram/box plots to check for outliers and skewness Check for normality
	Transformation		Transform if required Ensure all cohorts use the same transformation to keep the same scale
	Outliers		Remove implausible values e.g. >3SD or winsorize