# Genome-wide Association Studies

**Practical 1: Cleaning genotype data and intro to software**

Genetics & Genomics Winter School
Alesha Hatton
8/7/2025

# Why clean genotype data?

**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
- 3 main versions of PLINK
  - PLINK v1.07 2007 https://zzz.bwh.Harvard.edu/plink/ < good website on basics >
  - PLINK v1.90 2015 https://www.cog-genomics.org/plink/1.9/ < major upgrade of v1.07 >
  - PLINK v2.0 2017 https://www.cog-genomics.org/plink/2.0/ < under development? >

# PLINK v1.90 website



many commands
& info on their use

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

PLINK 1.9 home | plink2-users | GitHub | File formats | PLINK 1.9 index | PLINK 2.0

Introduction, downloads
S: 11 Dec 2023 (b7.2)
D: 11 Dec 2023
Recent version history
What's new?
Future development
Limitations
Note to testers
[Jump to search box]
General usage
Getting started
Citation instructions
Standard data input
PLINK 1 binary (.bed)
Autoconversion behavior
PLINK text (.ped, .tped...)
VCF (.vcf[.gz], .bcf)
Oxford (.gen[.gz], .bgen)
23andMe text
Generate random
Unusual chromosome IDs
Recombination map
Allele frequencies
Phenotypes
Covariates
Clusters of samples
Variant sets
Binary distance matrix
IBD report (.genome)
Input filtering
Sample ID file
Variant ID file
Positional ranges file

## PLINK 1.90 beta

This is a comprehensive update to Shaun Purcell's PLINK command-line program, developed by Christopher Chang with support from the NIH-NIDDK's Laboratory of Biological Modeling, the Purcell Lab, and others. (What's new?) (Credits.) (Methods paper.) (Usage questions should be sent to the **plink2-users Google group**, not Christopher's email.)

### Binary downloads

| | Build | | |
|---|---|---|---|
| Operating system[1] | Stable (beta 7.2, 11 Dec 2023) | Development (11 Dec 2023) | Old[2] (v1.07) |
| Linux 64-bit | **download** | download | download |
| Linux 32-bit | **download** | download | download |
| macOS (64-bit) | **download** | [illegible] | [illegible] |
| Windows 64-bit | **download** | download | download |
| Windows 32-bit | **download** | download | |

1: Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.
2: These are just mirrors of the binaries posted at https://zzz.bwh.harvard.edu/plink/download.shtml.

Source code, compilation instructions, and the like are on the developer page.

**The following documented PLINK 1.07 flags are not supported by 1.90 beta 7:**

- --qual-geno-scores[3]
- --segment[4]
- --p2, --genedrop

download the right version

Index!

plink2-dev
Credits
File formats
Quick index search

addition, several key sample x sample and variant x variant matrix computations (including the GRM mentioned below) can be cleanly split across computing clusters (or serially handled in manageable chunks by a single computer).

**Command-line interface improvements**
We've standardized how the command-line parser works, migrated from the original "everything is a flag" design toward a more organized flags + modifiers approach (while retaining backwards compatibility), and added a thorough command-line help facility.

**Additional functions**

# PLINK v1.90

- Need to run PLINK via command line, e.g.

```
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,
--rel-cutoff, --cluster, --pca, --neighbour, --ibs-test, --regress-distance,
--model, --bd, --gxe, --logistic, --dosage, --lasso, --test-missing,
--make-perm-pheno, --tdt, --qfam, --annotate, --clump, --gene-report,
--meta-analysis, --epistasis, --fast-epistasis, and --score.

"plink --help | more" describes all functions (warning: long).
```

- `plink --bfile` *`filename`* `--missing --out` *`newfilename`*
- If you have downloaded PLINK into your local directory, could be: ./plink

# 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

# 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

AJHG

Volume 88, Issue 1, 7 January 2011, Pages 76-82

Report

## GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang [1], S. Hong Lee [1], Michael E. Goddard [2 3], Peter M. Visscher [1]

Show more

+ Add to Mendeley  Share  Cite

https://doi.org/10.1016/j.ajhg.2010.11.011

Get rights and content



For most human complex diseases and traits, SNPs identified by genome-wide association studies (GWAS) explain only a small fraction of the heritability. Here we report a user-friendly software tool called genome-wide complex trait analysis (GCTA),

# 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

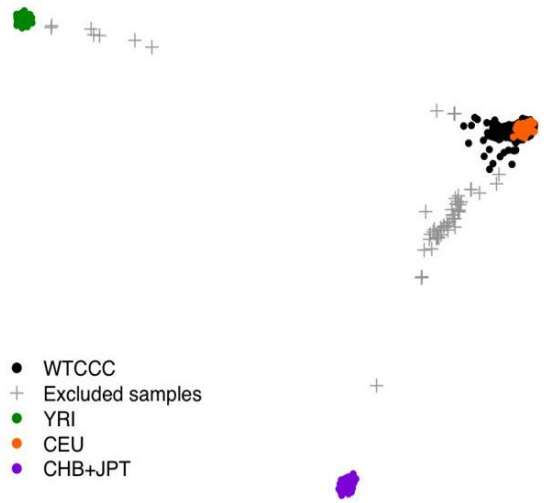
1. Removal of individuals with excess missing genotypes
2. Removal of individuals with outlying homozygosity values
3. Remove of samples showing a discordant sex

more details in the prac

Ensure data suitable for the analyses

4. Removing of related or duplicate samples, and
5. Population structure - removing of ancestry outliers

# 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 Marker 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)

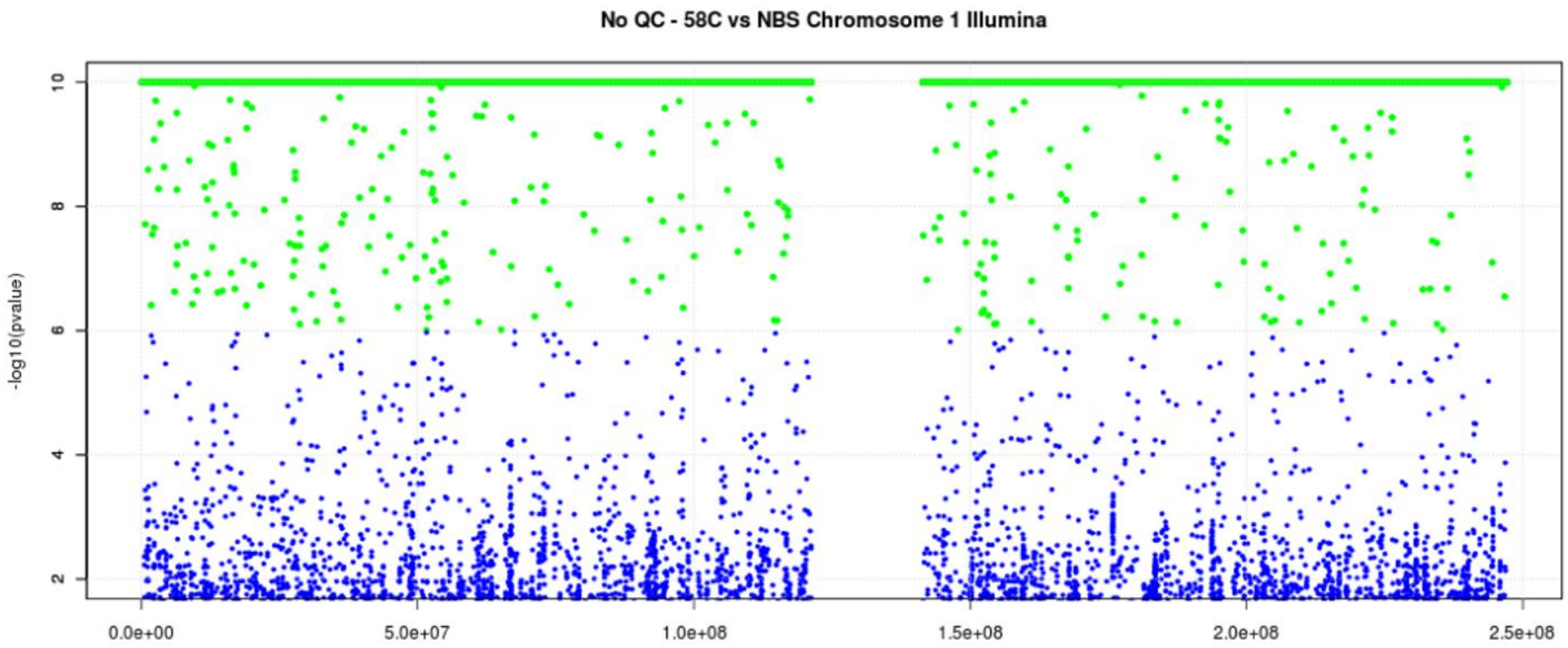
# Example: Importance of Good Cleaning

- 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)

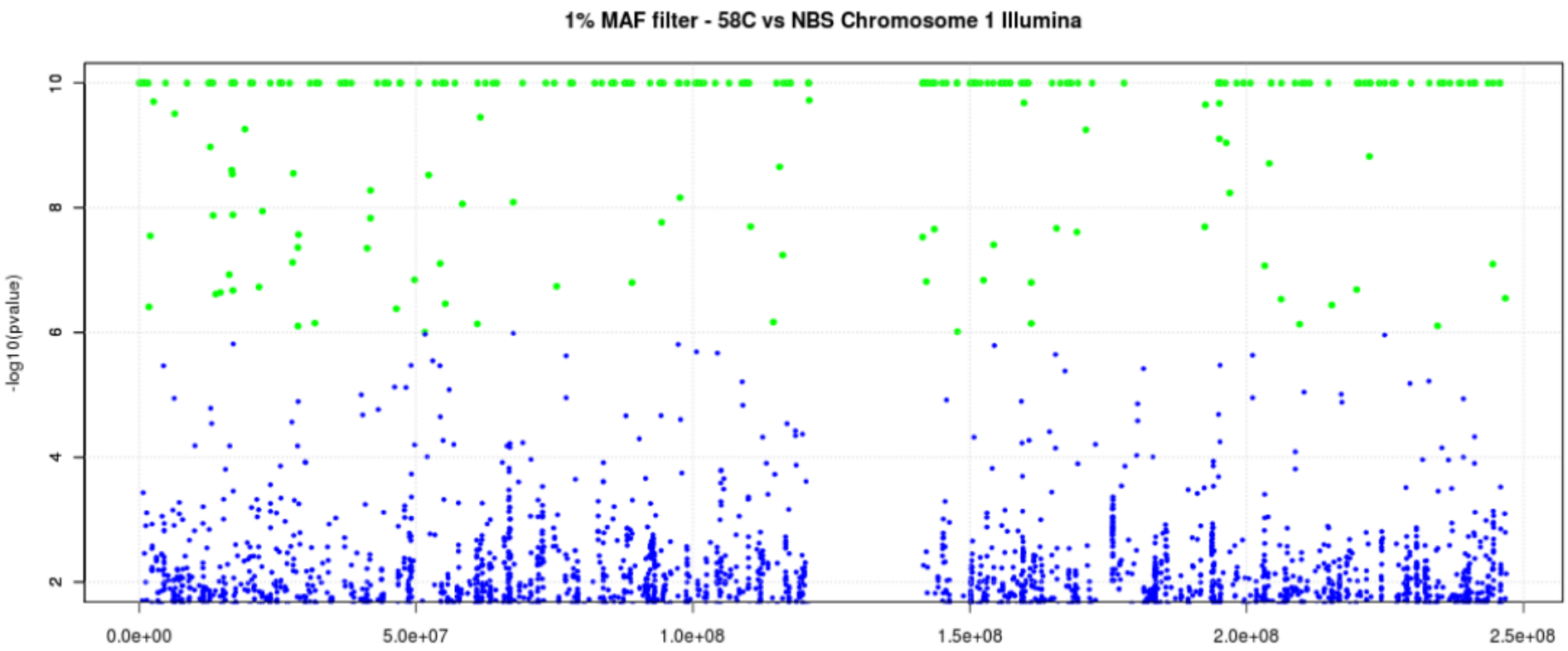
- Both these are unselected population cohorts, so performing a “case-control” study between these populations should find no significant differences

# Importance of Good Cleaning



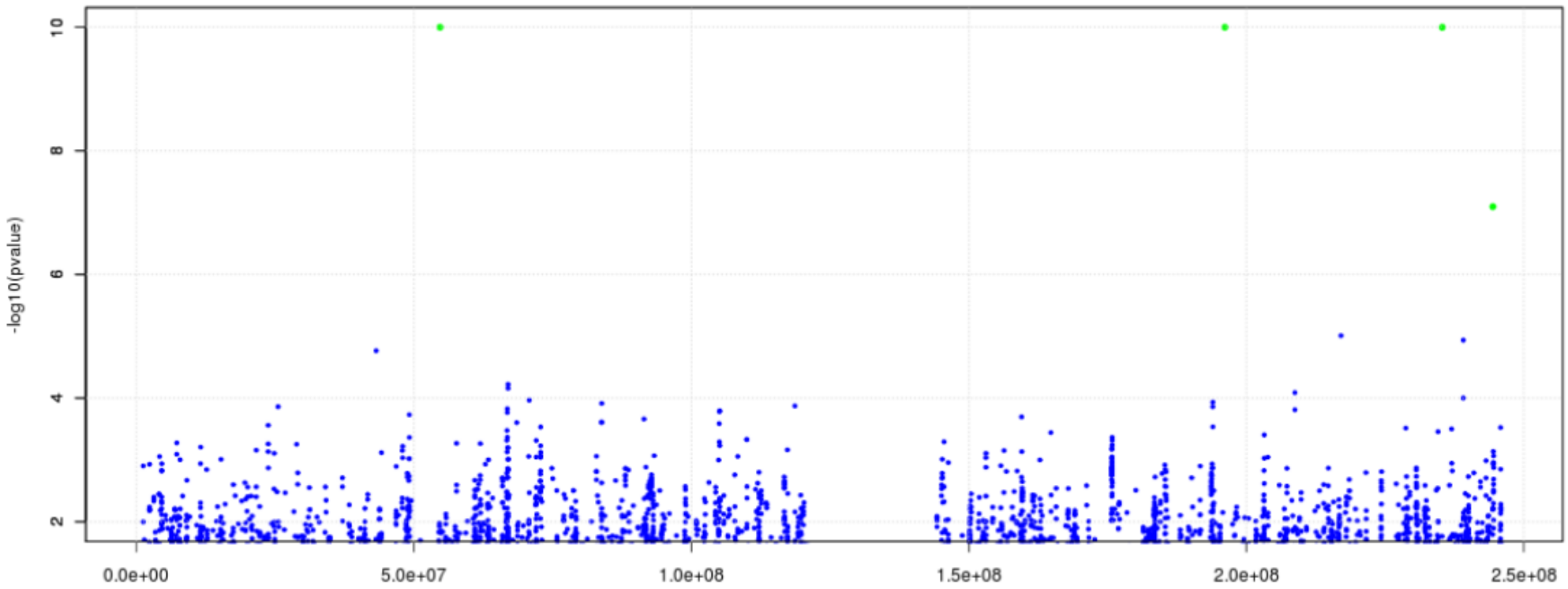

100% of SNPs

# Importance of Good Cleaning



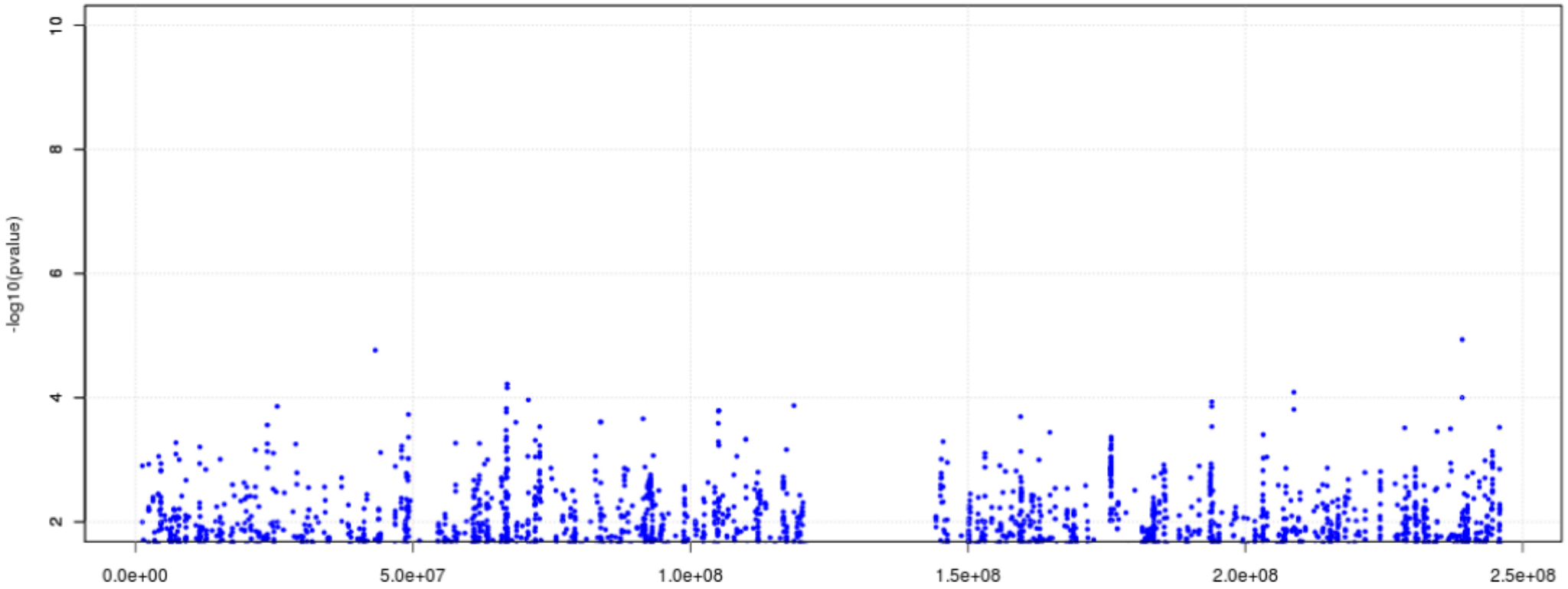

80.69% of SNPS
Filtering: MAF

# Importance of Good Cleaning




78.36% of SNPs
Filtering: MAF + HWE

# Importance of Good Cleaning




77.92% of SNPs
Filtering: MAF + HWE + Missingness

# In the prac - we will use PLINK to do the 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<br>--rel-cutoff | 4) Compare to known allele freq | --freq |