Quality Control of Genotype Data

KidneyGenAfrica Course – January 2026

Jean-Tristan Brandenburg

January 21, 2026

Contents

1 GWAS Quality Control (QC): Context and Objectives

The objective of this practical is to understand the **basics of genotype quality control (QC)** and how QC affects **bias in genome-wide association studies (GWAS)** results. Most genotype QC steps are performed **before imputation** to ensure that only high-quality data are carried forward for analysis.

In this session, we will focus on:

- Missingness
- Minor allele frequency
- Hardy–Weinberg equilibrium (HWE)
- Relatedness

Exercise 1: Genotype Quality Control (MAF, Missingness, HWE)

1.1 Missingness

Missingness can reflect **genotyping problems**. Variants or individuals with a high proportion of missing genotype calls often indicate:

- Poor DNA quality or contamination
- Technical issues during array hybridization or scanning
- Systematic batch effects

Filtering out SNPs and individuals with high missingness rates (e.g., $> 5\%$) reduces technical noise and improves the reliability of downstream analyses.

1.2 Minor Allele Frequency (MAF)

Purpose: remove unreliable or non-informative variants.

Variants with very low frequency ($< 1\%$):

- Have higher genotyping error rates
- Often fail imputation or have low INFO/R^2 scores
- Are poorly represented in reference panels

Filtering variants with MAF < 0.01 improves imputation accuracy by retaining variants with stable LD patterns.

1.3 Hardy–Weinberg Equilibrium (HWE)

Deviation from Hardy–Weinberg expectations can indicate:

- Allele miscalls (e.g., strand flips)
- Cluster calling errors
- Batch effects
- Population stratification
- Sample contamination or sex mislabeling

SNPs with strong deviations from HWE are excluded prior to imputation.

1.4 Exercise 1: Filtering Criteria

Apply the following QC thresholds:

- SNP MAF < 0.01
- SNP missingness > 0.05
- Individual missingness > 0.005
- HWE $p < 1 \times 10^{-6}$

All filters can be applied in a single **PLINK** command.

1.5 Common PLINK Arguments

- `--bfile`: input PLINK binary files
- `--maf`: minor allele frequency filter
- `--geno`: SNP missingness filter
- `--mind`: individual missingness filter
- `--hwe`: Hardy–Weinberg equilibrium filter
- `--make-bed`: create a new dataset
- `--out`: output prefix

To obtain help for a specific argument:

```
plink --help bfile
```

1.6 Exercise 1 – Questions

- How many SNPs are removed by each filter?
- How many individuals are excluded?
- Which QC criterion has the largest impact on the dataset?

Exercise 2: Relatedness Filtering

GWAS assumes that samples are **independent**. Closely related individuals (e.g., siblings or parent–child pairs) can inflate association signals.

1.7 Concept of Relatedness

Relatedness is quantified using the **pi-hat** statistic (identity-by-descent, IBD). PLINK 2 uses the **KING** algorithm to efficiently detect related individuals.

More information: https://www.kingrelatedness.com/

Relationship	pi-hat range	Action
Duplicate samples	> 0.9	Remove one
First-degree relatives	$0.35 - 0.9$	Remove one per pair
Second-degree relatives	$0.125 - 0.35$	Often acceptable

1.8 Typical pi-hat Thresholds

1.9 Key PLINK 2 Arguments

- `--bfile`: input dataset
- `--king-cutoff`: relatedness threshold
- `--make-bed`: create filtered dataset
- `--out`: output prefix

1.10 Exercise 2 – Question

- How many individuals were removed due to relatedness filtering?