Post-Association Analyses

KidneyGenAfrica Course – January 2026

January 21, 2026

Contents

1 Objective

The objective of this practical is to interpret GWAS results through standard post-association analyses. Specifically, you will:

- Generate QQ and Manhattan plots
- Compute the genomic inflation factor (λ)
- Identify independent association signals using LD-based clumping
- Visualize association signals using regional (locus) plots

2 Data and Files

The following files are used:

- `5_association/egfr_covar_afterqc.egfr.glm.linear` Association results after QC
- `5_association/egfr_covar_beforeqc.egfr.glm.linear` Association results before QC
- `4_Data_qc_admixture_pheno/genotyped_qc` PLINK binary fileset used to compute LD

Exercise 14: QQ Plot and Genomic Inflation Factor

Compute the genomic inflation factor (λ) for each set of summary statistics and compare them.

What is the Genomic Inflation Factor?

The genomic inflation factor (λ) evaluates whether GWAS test statistics are inflated relative to the null hypothesis:

$$\lambda = \frac{\text{median of observed } \chi^2}{\text{expected median under the null } (0.456)}$$

Interpretation

- $\lambda \approx 1.0$: no inflation
- $\lambda > 1.1$: evidence of inflation, potentially due to
  - population stratification
  - cryptic relatedness
  - genotyping or imputation errors

R Example: QQ Plot and λ

```
library(qqman)

sumstats <- read.table(
  "5_association/egfr_covar_afterqc.egfr.glm.linear",
  header=TRUE
)

chisq <- qchisq(1 - sumstats$P, 1)
```

```
lambda <- median(chisq, na.rm=TRUE) / qchisq(0.5, 1)
print(paste("Lambda:", round(lambda, 3)))

qq(sumstats$P, main="QQ plot -- After QC")
```

Exercise 14 – Questions

- Compare λ between the two models.
- Which model shows less inflation?
- What factors could explain the observed differences?

Exercise 15: Manhattan Plot

Generate Manhattan plots for the association results using `qqman`.

R Example

```
library(qqman)

sumstats <- read.table(
  "5_association/egfr_covar_afterqc.egfr.glm.linear",
  header=TRUE
)

manhattan(
  sumstats,
  chr = "CHR",
  bp  = "BP",
  snp = "ID",
  p   = "P",
  genomewideline = -log10(5e-8),
  main = "Manhattan Plot -- After QC"
)
```

Exercise 15 – Questions

- Are there genome-wide significant loci?
- Do the strongest signals appear biologically plausible?

Exercise 16: Identification of Independent SNPs (Clumping)

Identify independent lead SNPs using LD-based clumping with **PLINK 2**.

Example PLINK 2 Command

```
../bin/plink2 \
  --bfile 4_Data_qc_admixture_pheno/genotyped_qc \
  --clump 5_association/egfr_covar_afterqc.egfr.glm.linear \
  --clump-p1 5e-8 \
  --clump-p2 0.1 \
  --clump-r2 0.1 \
```

```
  --clump-kb 1000 \
  --out 5_association/egfr_afterqc_clumped
```

Notes

- `--clump-p1`: primary significance threshold
- `--clump-p2`: secondary threshold for nearby SNPs
- `--clump-r2`: LD threshold
- `--clump-kb`: window size (1000 kb = 1 Mb)

Exercise 16 – Questions

- How many independent SNPs are identified?
- Have these loci been previously reported in the GWAS Catalog?

Exercise 17: Regional (Locus) Plots

Regional plots allow visualization of association signals and local LD structure around lead SNPs.

R Example Using `locusplotr`

```
library(locusplotr)

gg_locusplot(
  df = Data,
  lead_snp = "rs1719245",
  rsid = "ID",
  chrom = "CHR",
  pos = "POS",
  ref = "REF",
  alt = "ALT",
  p_value = P,
  plot_genes = TRUE,
  genome_build = "GRCh38",
  plink = "../../bin/plink",
  bfile = "../Data_qc/genotyped_qc",
  compute_ld = TRUE
)
```

Exercise 17 – Questions

- Does the association signal span multiple SNPs in LD?
- Are there plausible candidate genes in the region?

Summary Questions

- How does genomic inflation change before versus after QC and ancestry adjustment?
- How many independent association signals remain after clumping?