Training Exercise: Post-GWAS Analysis KidneyGenAfrica Training Programme

Jean-Tristan Brandenburg

January 2026

Contents

1 Post-GWAS Analysis – Context **2**

2 Examples of Post-GWAS Analyses **2**

3 Software, Platforms, and Technical Approaches **2**
- 3.1 Manual Analysis . . . 2
- 3.2 Pipelines and Workflows . . . 3
- 3.3 Web-Based Platforms . . . 3

4 Annotation and Post-GWAS Interpretation Using FUMA **3**
- 4.1 General Description . . . 3
- 4.2 FUMA Modules . . . 4
- 4.3 SNP2GENE: Definition of SNPs and Loci . . . 4
  - 4.3.1 Independent Significant SNPs . . . 4
  - 4.3.2 Lead SNPs . . . 4
  - 4.3.3 Genomic Risk Loci . . . 4
  - 4.3.4 Candidate SNPs . . . 4
- 4.4 Functional Annotation Tables . . . 5
- 4.5 Gene-Based Analysis Using MAGMA . . . 5
- 4.6 Visualization . . . 5
  - 4.6.1 Genome-Wide Plots . . . 5
  - 4.6.2 Gene-Based Plots . . . 5
  - 4.6.3 Regional Association Plots . . . 5
- 4.7 Interpretation and Limitations . . . 5

5 Exercise **6**
- 5.1 Dataset Description . . . 6
- 5.2 FUMA Parameters . . . 6
- 5.3 Questions . . . 6

1 Post-GWAS Analysis – Context

Post-GWAS analysis refers to the set of analyses performed **after a genome-wide association study (GWAS)** to interpret and translate GWAS signals into biological insights.

Post-GWAS analyses aim to **extract, understand, and interpret association results at multiple levels**, including significant genomic regions, variants, and genes. These analyses typically rely on **GWAS summary statistics**, **linkage disequilibrium (LD) information**, and/or **genotype data**, combined with external biological databases and specialized software.

2 Examples of Post-GWAS Analyses

(non-exhaustive list)

- Assessment of association quality and bias using genomic inflation factors (λ), QQ plots, and Manhattan plots
- Extraction of significant SNPs and fine-mapping to identify putative causal variants
- Functional annotation and visualization of variants
- Replication of known associations and validation of novel signals
- Gene-based and gene-set association analyses
- Variant-to-gene mapping using eQTLs and chromatin interaction data
- Pathway and tissue-specific enrichment analyses

More advanced analyses

- Polygenic risk score (PRS) construction and validation
- Mendelian randomization
- Multi-omics integration (transcriptomics, epigenomics, proteomics)

3 Software, Platforms, and Technical Approaches

Post-GWAS analyses rely on a wide range of tools and strategies. The choice depends on the research question, available data, and computational resources.

3.1 Manual Analysis

Manual analysis involves running each step independently. This approach is **flexible and transparent** but **time-consuming** and prone to human error.

Common tools

- **PLINK** – SNP filtering, clumping
- **GCTA** – conditional analysis, heritability
- **LocusZoom** – regional association plots
- **FINEMAP**, **SuSiE**, **CAVIAR** – fine-mapping
- **ANNOVAR**, **VEP** – variant annotation
- **MAGMA** – gene and gene-set analysis
- **R / Python** – custom analyses

3.2 Pipelines and Workflows

Automated workflows deployed on local machines, HPC clusters, or cloud infrastructure. These are **reproducible and scalable** but require maintenance.

Example

- **H3AGWAS pipeline** – GWAS and post-GWAS workflows for African genomics

3.3 Web-Based Platforms

Web platforms provide predefined pipelines with minimal computational burden but reduced flexibility.

Example

- **FUMA** – functional mapping, annotation, and gene prioritization

4 Annotation and Post-GWAS Interpretation Using FUMA

4.1 General Description

Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) is a web-based platform for annotating, prioritizing, visualizing, and interpreting GWAS results using summary statistics.

FUMA integrates LD information, functional annotations, eQTLs, chromatin interactions, and gene-based association analyses. User registration is required to upload private datasets, while public results can be explored freely.

Tutorial: `https://fuma.ctglab.nl/tutorial`

4.2 FUMA Modules

- **SNP2GENE**: identification of significant SNPs, locus definition, functional annotation, and MAGMA analyses
- **GENE2FUNC**: biological interpretation of gene lists (expression and pathways)
- **Cell Type**: identification of relevant cell types using MAGMA results

4.3 SNP2GENE: Definition of SNPs and Loci

4.3.1 Independent Significant SNPs

Independent significant SNPs:

- Have $P \leq 5 \times 10^{-8}$ (default)
- Are independent at $r^2 \leq 0.6$

These SNPs correspond to PLINK-clumped SNPs and serve as anchors for downstream analyses.

4.3.2 Lead SNPs

Lead SNPs are independent significant SNPs that are independent at a stricter threshold ($r^2 < 0.1$). They represent the most independent association signals.

4.3.3 Genomic Risk Loci

Lead SNPs are grouped into genomic risk loci:

1. SNPs in LD ($r^2 \geq 0.1$) are grouped
2. Nearby loci (default: 250 kb) are merged
3. The SNP with the smallest P-value represents the locus

4.3.4 Candidate SNPs

Candidate SNPs:

- Are in LD with independent significant SNPs
- Pass an optional GWAS threshold (default $P \leq 0.05$)

They are fully annotated and used for gene mapping.

4.4 Functional Annotation Tables

- **SNPs annotation**: all candidate SNPs with functional information
- **ANNOVAR**: predicted functional consequences
- **Mapped genes**: genes mapped by positional, eQTL, or chromatin interaction mapping
- **eQTL mapping**: SNP–gene–tissue combinations
- **GWAS Catalog**: overlap with published GWAS associations

4.5 Gene-Based Analysis Using MAGMA

FUMA performs MAGMA gene-based, gene-set, and gene-property analyses using default parameters (SNP-wise mean model).

SNPs are assigned to genes based on physical position, and gene-level association statistics are computed using the selected reference panel.

4.6 Visualization

4.6.1 Genome-Wide Plots

Manhattan plots are generated from GWAS summary statistics. SNPs with $P \geq 1 \times 10^{-5}$ are excluded, and overlapping points are filtered for clarity.

4.6.2 Gene-Based Plots

Gene-level Manhattan plots are generated using MAGMA (v1.10) considering protein-coding genes from Ensembl build 85.

4.6.3 Regional Association Plots

Regional plots can be generated for independent significant SNPs, lead SNPs, or genomic risk loci.

Alternative tools:

- LocusZoom v2: `http://locuszoom.org/`

4.7 Interpretation and Limitations

LD-based definitions provide approximations of association signals. Lead SNPs are not necessarily causal, and multiple biological signals may be merged into a single locus.

Bayesian fine-mapping approaches offer a more precise framework to identify causal variants and define credible sets.

5 Exercise

5.1 Dataset Description

The dataset includes approximately **67,000 African American individuals** from a meta-analysis of eight cohorts generated using METAL.

Data access: `https://susztaklab.com/GWAS2M/Download.php`

Reference publication: `https://www.science.org/doi/10.1126/science.adp4753`

5.2 FUMA Parameters

- Reference panel: African populations (1000 Genomes)
- LD threshold: $r^2 > 0.6$
- Minor allele frequency: > 0.01
- eQTL mapping: enabled
- MAGMA analysis: enabled

FUMA results: `https://fuma.ctglab.nl/browse/697303`

5.3 Questions

- How many putative genomic risk loci are identified using a genome-wide significance threshold of 5×10^{-8}?
- Using regional association plots and SNPs in linkage disequilibrium (LD), how is each lead SNP supported by neighboring variants?
- What differences do you observe between loci in terms of effect size, LD structure, and number of supporting SNPs?
- Which loci or genomic regions have been previously reported in GWAS catalogs?
  - Which loci appear novel?
  - What are the limitations of assessing novelty using GWAS catalog overlap alone?
  - What additional analyses or external resources could be used to validate novelty?
- Using functional annotations and pathogenicity scores (e.g. CADD), which SNPs are predicted to be potentially pathogenic and in which genes?
- Which genes are identified as significant in the MAGMA gene-based analysis?
  - Which of these genes have been previously described in kidney-related traits?
  - How do gene-based results compare with SNP-level (locus-based) findings?
  - How would you interpret discrepancies between the two approaches?
- **African-ancestry considerations**

- Is the conventional genome-wide significance threshold (5×10^{-8}) appropriate for African-ancestry GWAS?
- How do higher genetic diversity and shorter LD blocks in African populations affect locus discovery and interpretation?
- How should significance thresholds be handled in trans-ethnic or multi-ancestry GWAS?
- What alternative strategies could be considered (e.g. ancestry-specific thresholds, empirical thresholds, replication-based validation)?