SAGA (Simplified Association Genomewide Analyses): a user-friendly Pipeline to Democratize Genome-Wide Association Studies

Basilio Cieza[1], Neetesh Pandey[1], Vivek Ruhela[1], Sarwan Ali[1], Giuseppe Tosto*[1,2,3]

1. Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Vagelos College of Physicians and Surgeons, Columbia University. 630 West 168th Street, New York, NY 10032, USA.
2. The Gertrude H. Sergievsky Center, Vagelos College of Physicians and Surgeons, Columbia University. 630 West 168th Street, New York, NY 10032, USA.
3. Department of Neurology, Vagelos College of Physicians and Surgeons, Columbia University, and the New York Presbyterian Hospital. 710 West 168th Street, New York, NY 10032, USA.

*Corresponding author:

Giuseppe Tosto, M.D., Ph.D.

G.H.Sergievsky Center, The Taub Institute for Research on Alzheimer's Disease and the Aging Brain

Department of Neurology, Columbia University Irving Medical Center

630W 168th street, room PH19-314

New York City, New York, 10032

Tel. 212-305-9274

gt2260@cumc.columbia.edu

Abstract

Genome-wide association studies (GWAS) have enabled clinicians and researchers to identify genetic variants linked to complex traits and diseases(1). However, conducting GWAS remains technically challenging without bioinformatics expertise due to required data preprocessing, software installation, and analysis scripting (2,3). SAGA is a BASH-based, open-source, fully automated pipeline that integrates three widely adopted tools—PLINK(4), GMMAT(5), and SAIGE(6)—for accessible, robust, and reproducible GWAS. After installation, users only provide standard genotype and phenotype files. The pipeline automates preprocessing, association testing, and visualization, outputting summary statistics, Manhattan plots, and quantile-quantile plot. SAGA enables robust GWAS for users with no scripting experience, expanding access to complex genetic analyses.

Keywords: GWAS, SAIGE, GMMAT, PLINK, pipeline, bash, summary statistics, user-friendly, Manhattan plot, quantile-quantile plot.

1. Introduction

Genome-wide association studies (GWAS) have transformed discovery of gene-variant associations for complex traits and diseases in human populations (1). Yet, the broader deployment of GWAS is limited by technical barriers: multi-step software installations, data formatting, and intricate analytic workflows (7–9) SAGA addresses this gap for non-bioinformaticians with a modular, bash-based pipeline that integrates three popular platforms: **PLINK** for standard linear/logistic regression (4), **GMMAT** for generalized mixed model association testing, especially with related or family-based samples (5), and **SAIGE** for very large, imbalanced case-control studies and rare variant analysis (6). SAGA offers an automated, reproducible solution for GWAS analysis on mainstream genomic data, providing ready-to-use outputs for downstream exploration and publication.

2. Methods

Tool Selection and Rationale

We chose PLINK, GMMAT, and SAIGE for their complementary strengths (4–6) (**Table 1**).

Tool	Key Strengths	Ideal For	Model Type	Handles Relatedness	Handles Case-Control Imbalance
PLINK	Fast, user-friendly	Unrelated individuals	Linear/logistic regression	☐	☐
GMMAT	Mixed models, kinship-aware	Family/related samples	Linear/logistic mixed model	☐	☐
SAIGE	Scalability, rare variants	Biobank, unbalanced traits	Saddlepoint Approximation (SPA)	☐	☐

Table 1. Software included in SAGA and their features.

Pipeline Workflow

The user has to shape input genetic files in popular PLINK's binary format (.bed/.bim/.fam) along with a structured phenotype file containing family and individual IDs ("FID" and "IID", respectively), outcome measures (binary or quantitative), and optional covariates. All input files undergo automated validation to ensure completeness and correct formatting.

In the preprocessing stage, SAGA performs quality control (QC) to filter variants and samples based on missingness, Hardy–Weinberg equilibrium, and minor allele frequency thresholds. Population structure is assessed via principal component analysis (PCA), generating the top principal components for covariate adjustment. A genetic relationship matrix (GRM) is generated for use with GMMAT and SAIGE by providing dedicated files again in PLINK format. Each preprocessing step is standardized and error-checked to minimize user intervention.

Association analysis is conducted using one of three backends, selected automatically according to user-defined parameters or data characteristics: PLINK for fast linear or logistic regression, GMMAT for generalized linear mixed models in related populations, and SAIGE which uses a saddlepoint approximation (SPA)-adjusted logistic mixed model designed to handle large-scale datasets (e.g. biobanks) featuring highly unbalanced case-control ratios.

Finally, post-analysis procedures standardize summary statistics across backends and produce publication-ready Manhattan and quantile-quantile (QQ) plots in both high-resolution .png and .pdf formats using automated R scripts (10). This fully integrated workflow ensures consistent, reproducible results from raw genotype data to final visualization.

Technical Implementation

SAGA is implemented entirely in BASH and requires only a UNIX/Linux environment for execution. The pipeline depends on PLINK for genotype processing, R 4.2.2 for running GMMAT and generating plots, and SAIGE (version 1.3.0 or later) for scalable mixed-model association testing. Installation is straightforward via a single command:

git clone https://github.com/bciezah1/SAGA.git

Upon completion of the analysis, SAGA generates all results in an organized output/ directory. This includes standardized summary statistics ("*summary_stats.txt*") and high-quality Manhattan and QQ plots ("*manhattan.png*" and "*qqplot.png*"). Examples of the results are shown in **Figure 1**.

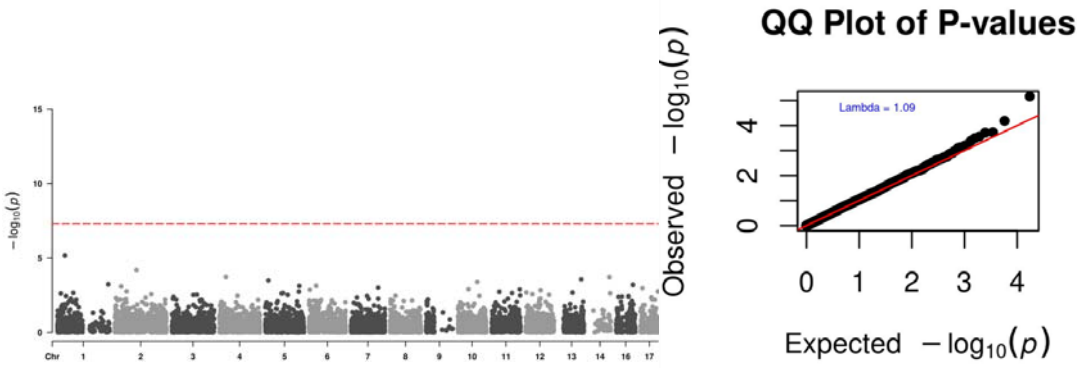

Figure 1. Example output. Manhattan plot and QQ plot with lambda value for inflation.

Automated Preprocessing Details

Once installed, the package can be launched via a single command. For example, employing PLINK with a quantitative phenotype and three covariates:

./run_pipeline_plink.sh *{user's path}***/toy_data/input_dosage **
*{user's path}***/toy_data/pheno_quantitative.txt **
**COV1,COV2,COV3 **
**PHENO **
quantitative

In this example, the first argument specifies the location of the PLINK genotype data, followed by the phenotype file location, a comma-separated list of covariates, the target phenotype name, and the phenotype type (quantitative or binary). Upon completion, SAGA generates the results in an organized output/ directory, including standardized summary statistics (*summary_stats.txt*) and high-quality Manhattan and QQ plots (*manhattan.png and qqplot.png*).

3. Benchmarking and Experimental Design

We benchmarked the performance and usability of SAGA, using simulated datasets that varied in complexity in a controlled manner. Specifically, we simulated datasets with three different sample sizes (N=1,000, 5,000, and 10,000 individuals), and three variant counts (N=1,000, 10,000, and 100,000 single nucleotide polymorphisms (SNPs; **Figure 1**, **Table S1**). Phenotypes and covariates were simulated to mimic binary traits with typical covariate structures encountered in common GWAS studies. Each dataset was analyzed separately using the three GWAS analysis backends integrated within SAGA.

Benchmarking metrics included runtime efficiency, error rates, reproducibility, and ease of use.

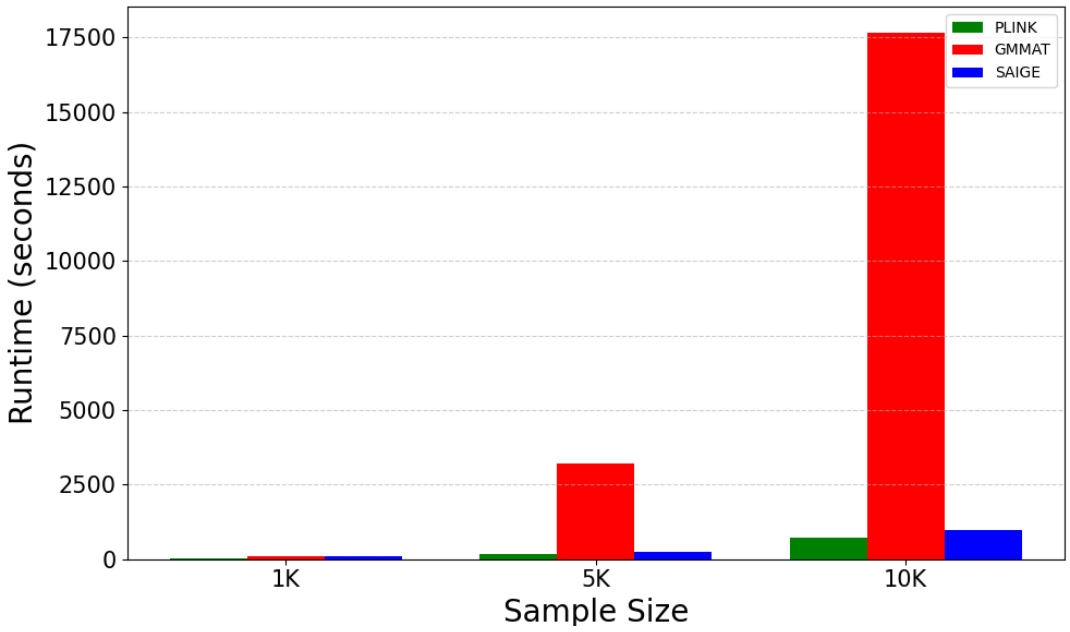

Figure 1. Shows the benchmark for three different samples sizes for PLINK, GMMAT, and SAIGE (green, red and blue color, respectively).

4. Discussion and conclusions

SAGA streamlines the GWAS workflow for users without scripting or programming expertise by providing an intuitive and automated framework. One of its major advantages is ease of access, as users do not need to edit code and only require minimal familiarity with command-line operations. The pipeline covers all critical steps of genome-wide association studies, from quality control through analysis to the generation of publication-ready visualizations, thereby significantly reducing the technical burden on researchers. By employing consistent, standardized processing methods compliant with leading best practices, SAGA ensures reproducible results across multiple runs. Importantly, it includes rigorous quality control measures that account for population structure and relatedness among samples, facilitating robust association testing.

Our benchmarking results highlight the strengths of SAGA in terms of runtime efficiency, error control, reproducibility, and usability. Importantly, the framework demonstrated the ability to scale with increasing dataset sizes and variant complexity while preserving automated and reproducible workflows across multiple statistical models and software implementations (See Figure 1). These findings underscore the utility of SAGA for a wide range of GWAS applications and emphasize its potential to broaden access to advanced genetic association analyses for researchers without extensive computational expertise.

However, the pipeline has some limitations. It assumes that users provide valid and properly formatted phenotype files, as it does not incorporate functionality to check or compose phenotypic data. Additionally, SAGA currently lacks native support for meta-analysis or the merging of multiple cohorts, which are important considerations for large-scale genetic studies. While the pipeline can handle moderately large datasets, analyzing very large datasets—such as those exceeding 50,000 individuals—may still require high-performance computing clusters or cloud resources due to memory and processing demands.

Looking forward, plans for future development include the addition of modules for gene-based and rare variant analyses, which would extend the pipeline's scope to capture a broader spectrum of genetic variation. The developers also intend to implement advanced post-GWAS annotation functionalities, including compatibility with tools such as FUMA, which can facilitate the biological interpretation of significant loci. Another anticipated feature is support for polygenic risk scoring, enabling the translation of GWAS findings into quantitative measures of genetic risk for complex diseases. These enhancements aim to further empower researchers with limited computational backgrounds to perform comprehensive and insightful genetic association studies.

Data and Software Availability

All code, documentation, and test data are available at:

https://github.com/bciezah1/SAGA.git

Author Contributions

BC, GT: Conceptualization, manuscript writing

BC: software development

BC, NP, VR, SA, GT: documentation, testing

Ethical Statement

No human or animal subjects were involved in this work. All testing datasets were simulated and fully anonymized.

Acknowledgements

We thank the GiusTo Lab at Columbia University for cohort data access and deployment support. This work is supported by NIA/NIH award U01AG081817, RF1AG082009 and U19AG074865.

References

1. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Vol. 1, Nature Reviews Methods Primers. Springer Nature; 2021.
2. Schönherr S, Schachtl-Riess JF, Di Maio S, Filosi M, Mark M, Lamina C, et al. Performing highly parallelized and reproducible GWAS analysis on biobank-scale data. NAR Genom Bioinform. 2024 Mar 1;6(1).
3. Song Z, Gurinovich A, Federico A, Monti S, Sebastiani P. nf-gwas-pipeline: A Nextflow Genome-Wide Association Study Pipeline. J Open Source Softw. 2021 Mar 2;6(59):2957.
4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
5. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am J Hum Genet. 2016 Apr 7;98(4):653–66.
6. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018 Sep 1;50(9):1335–41.
7. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Vol. 26, Bioinformatics. 2010. p. 445–55.
8. Akdeniz BC, Frei O, Hagen E, Filiz TT, Karthikeyan S, Pasman J, et al. COSGAP: COntainerized Statistical Genetics Analysis Pipelines. Bioinformatics Advances. 2024;4(1).
9. Lück S, Scholz U, Douchkov D. Introducing GWAStic: a user-friendly, cross-platform solution for genome-wide association studies and genomic prediction. Bioinformatics Advances. 2024;4(1).
10. Stephen Turner A, Stephen Turner M. Package “qqman” Title Q-Q and Manhattan Plots for GWAS Data. 2025.