Bioinformatics, 2026, **42**(3), btag032
https://doi.org/10.1093/bioinformatics/btag032
Published: 22 January 2026
Applications Note | Genetics and population analysis

OXFORD

SAIGE-GPU: accelerating genome- and phenome-wide association studies using GPUs

Alex Rodriguez[1,*,†], Youngdae Kim[2,†], Tarak Nath Nandi[1], Karl Keat[3], Rachit Kumar[3], Mitchell Conery[1,4], Rohan Bhukar[5,6], Molei Liu[7], John Hessington[8], Ketan Maheshwari[9], VA Million Veteran Program[10], Edmon Begoli[9], Georgia Tourassi[11], Sumitra Muralidhar[12], Pradeep Natarajan[6,13,14,15], Benjamin F. Voight[4,16,17,18], Kelly Cho[13,19,20], John Michael Gaziano[13,19,20], Scott M. Damrauer[16,17,21,22], Katherine P. Liao[13,23,24,25,26], Wei Zhou[5,27,28], Jennifer E. Huffman[13,23,29], Anurag Verma[3,16,30,‡], Ravi K. Madduri[1,‡]

[1]Data Science and Learning, Argonne National Laboratory, Lemont, IL 60439, United States
[2]Department of Industrial Engineering, Artificial Intelligence Graduate School, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea
[3]Institute for Biomedical Informatics, University of Pennsylvania – Perelman School of Medicine, Philadelphia, PA 19104, United States
[4]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania – Perelman School of Medicine, Philadelphia, PA 19104, United States
[5]Program in Medical and Population Genetics, Cambridge, MA 02142, United States
[6]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA 02114, United States
[7]Department of Biostatistics, Columbia University's Mailman School of Public Health, New York, NY 10032, United States
[8]Information Systems, University of Pennsylvania, Philadelphia, PA 19104, United States
[9]Oak Ridge National Laboratory, Oak Ridge, TN 37830, United States
[10]See Supplement for a List of MVP Contributors
[11]Computing and Computational Sciences Directorate, Oak Ridge National Laboratory, Oak Ridge, TN 37830, United States
[12]Department of Veterans Affairs, Office of Research and Development, Washington, DC 20420, United States
[13]Department of Medicine, Harvard Medical School, Boston, MA 02115, United States
[14]Program in Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Cambridge, MA 02142, United States
[15]Cardiology Division, Massachusetts General Hospital, Boston, MA 02114, United States
[16]Corporal Michael Crescenz VA Medical Center, Philadelphia, PA 19104, United States
[17]Department of Genetics, University of Pennsylvania – Perelman School of Medicine, Philadelphia, PA 19104, United States
[18]Institute for Translational Medicine and Therapeutics, University of Pennsylvania – Perelman School of Medicine, Philadelphia, PA 19104, United States
[19]MVP Boston Coordinating Center, VA Boston Healthcare System, Boston, MA 02111, United States
[20]Division of Aging, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, United States
[21]Department of Surgery, University of Pennsylvania – Perelman School of Medicine, Philadelphia, PA 19104, United States
[22]Cardiovascular Institute, University of Pennsylvania – Perelman School of Medicine, Philadelphia, PA 19104, United States
[23]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA 02130, United States
[24]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, United States
[25]Medicine, Rheumatology, VA Boston Healthcare System, Boston, MA 02130, United States
[26]Division of Rheumatology, Department of Medicine, Inflammation, and Immunity, Brigham and Women's Hospital, Boston, MA 02115, United States
[27]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, United States
[28]Stanley Center for Psychiatric Research, Cambridge, MA 02142, United States
[29]Palo Alto Veterans Institute for Research (PAVIR), Palo Alto Health Care System, Palo Alto, CA 94304, United States
[30]Division of Translational Medicine and Human Genetics, Department of Medicine, University of Pennsylvania – Perelman School of Medicine, Philadelphia, PA 19104, United States
*Corresponding author. Data Science and Learning (DSL) Division, Argonne National Laboratory, 9700 S. Cass Ave, Lemont, IL, 60439, USA. E-mail: a.rodriguez@anl.gov
[†]These authors contributed equally to this work.
[‡]These authors supervised equally to this work.
Associate Editor: Russell Schwartz

Abstract

Motivation: **Genome-wide association studies (GWAS) at biobank scale are computationally intensive, especially for admixed populations requiring robust statistical models. SAIGE is a widely used method for generalized linear**

Received: 6 August 2025. **Revised:** 10 December 2025. **Accepted:** 5 January 2026

mixed-model GWAS but is limited by its CPU-based implementation, making phenome-wide association studies impractical for many research groups.

Results: **We developed SAIGE-GPU, a GPU-accelerated version of SAIGE that replaces CPU-intensive matrix operations with GPU-optimized kernels. The core innovation is distributing genetic relationship matrix calculations across GPUs and communication layers. Applied to 2068 phenotypes from 635 969 participants in the Million Veteran Program, including diverse and admixed populations, SAIGE-GPU achieved a 5-fold speedup in mixed model fitting on supercomputing infrastructure and cloud platforms. We further optimized the variant association testing step through multi-core and multi-trait parallelization. Deployed on Google Cloud Platform and Azure, the method provided substantial cost and time savings.**

Availability and implementation: **Source code and binaries are available for download at https://github.com/saigegit/SAIGE/tree/SAIGE-GPU-1.3.3. A code snapshot is archived at Zenodo for reproducibility (DOI: [10.5281/zenodo.17642591]). SAIGE-GPU is available in a containerized format for use across HPC and cloud environments and is implemented in R/C++ and runs on Linux systems.**

1 Introduction

Biobanks linked to electronic health records are powering large-scale genome-wide association studies (GWAS) and other translational research (Gottesman *et al.* 2013, Sudlow *et al.* 2015, Wolford *et al*. 2018, Verma *et al.* 2022, Kurki *et al.* 2023, Zawistowski *et al.* 2023, Bick *et al.* 2024). However, the size of modern cohorts presents major computational challenges. Mixed-model GWAS rely on generalized linear mixed models (GLMMs) to control for population structure and relatedness (Uffelmann *et al.* 2021), and their cost scales with both sample and number of variants (Chang *et al.* 2015). Broad phenotyping in biobanks has further driven growth of multi-trait and phenome-wide analyses, amplifying the need for faster and more scalable computations (Sakaue *et al.* 2021, Karczewski *et al.* 2024, Verma *et al*. 2024).

To meet these demands, we optimized SAIGE (Zhou *et al.* 2018), a leading mixed-model GWAS method, by integrating Graphics Processing Units (GPUs) acceleration and distributed computing. SAIGE accounts for relatedness using a genetic relationship matrix (GRM); although full GRMs offer improved control of confounding (Lee *et al.* 2012), they are computationally intensive. SAIGE uses a conjugate-gradient solver (Kaasschieter 1988) that avoids explicit GRM construction, an approach shared with methods like Bolt-LMM (Loh *et al.* 2015), and is well suited to GPU architectures with high memory bandwidth and massive parallelism (Sunitha *et al*. 2017, Baji 2018, Cook).

The computational cost of phenome-wide GWAS remains a key bottleneck preventing many research groups from conducting such analyses. In prior work, our SAIGE-GPU implementation enabled a phenome-wide GWAS in the Million Veteran Program (MVP) (Verma *et al*. 2024), analyzing 3.5 billion genetic variants, over 2000 traits, and 635 969 MVP participants while achieving a 5-fold speedup in GLMM fitting. Standard CPU-based SAIGE would have required approximately 481 186 CPU-core hours—equivalent to 55 years of single-core time or 126 days on a 16-core workstation.

Large-scale resources such as UK Biobank, All of Us, and other national cohorts make phenome-wide GWAS increasingly common, yet computational cost remains a barrier. GPU acceleration substantially reduces this burden: cloud benchmarks show that a single-trait analysis costing $3.88 on CPUs costs $1.45 on one GPU—a 2.7× reduction. For 1000 traits, this saves approximately $2430 in compute cost ($3880 versus $1450) enabling iterative analyses that would otherwise exceed typical budgets or require national supercomputing facilities.

To confirm that GPU acceleration addresses a true bottleneck rather than providing marginal improvements, we directly compared SAIGE-GPU and REGENIE v4.1 (Mbatchou *et al.* 2021)—a widely used alternative—using identical datasets and hardware (400 000 samples, 121 587 variants). SAIGE-GPU completed model fitting step in approximately 15 min using three GPUs, while REGENIE required 31–34 min using 16–32 optimized CPU threads, a 2–2.3× speedup. In a phenome-wide analysis of 2068 traits, this difference corresponds to approximately 23 days of compute saved. Although SAIGE and REGENIE use distinct statistical frameworks (exact versus approximate), these benchmarks show that GPU acceleration of SAIGE's model fitting step—the dominant computational cost—offers practical advantages over CPU-based implementations and alternative approximate methods. Full comparative results appear in Tables 6 and 7, available as supplementary data at *Bioinformatics* online.

2 Materials and methods

2.1 Optimizations for SAIGE linear mixed model fitting (Step 1)

SAIGE fits a GLMM for each phenotype (Y) to account for fixed covariates (X), genotypes (G) and sample relatedness through random genetic effects ($\boldsymbol{b}$). Specifically, the model uses the logistic mixed model and can be written as:

$$logit(\mu_i) = \boldsymbol{X_i}\alpha + \boldsymbol{G_i}\beta + b_i + e \quad (1)$$

where, $\mu_i = P(y_i = 1|X_i, \boldsymbol{G}_i, b_i)$, α and β are vectors of fixed-effect and genetic-effect coefficients, and $b \sim N(0, \tau\boldsymbol{\psi})$. The genetic relationship matrix (GRM) $\boldsymbol{\psi}$ is defined as:

$$\boldsymbol{\psi} = \frac{1}{M}\boldsymbol{A}\boldsymbol{A}^T \quad (2)$$

with genotype matrix $\boldsymbol{A} \in \mathbb{R}^{NxM}$. Fitting the model under the null hypothesis ($\beta = 0$) is an iterative process, and each iteration involves solving the linear system $\boldsymbol{\Sigma x} = \boldsymbol{b}$, where $\boldsymbol{\Sigma} = \boldsymbol{W}^{-1} + \tau\boldsymbol{\Psi}$

with $\boldsymbol{W}$ denoting the diagonal weight matrix. Constructing and storing an explicit $N \times N$ GRM becomes infeasible as sample size and variant count grow. SAIGE addresses this by never materializing the GRM; instead, it encodes genetic relatedness through efficient matrix-vector operations used within the preconditioned conjugate gradient (PCG) solver. This implicit formulation dramatically reduces memory demands and removes the need to store the full GRM.

The main computational cost is the PCG iteration. Each iteration requires two matrix-vector multiplications with $\boldsymbol{A}$ and $\boldsymbol{A}^T$, an $O(MN)$ operation. Empirically, it converges in $O(N^{0.5})$ iterations (Loh *et al.* 2015), giving a total complexity $O(MN^{1.5})$. Convergence depends on the eigenvalue distribution of $\boldsymbol{\Sigma}$, and clustering via preconditioning can accelerate the convergence (Nocedal and Wright 2006). The current implementation uses a diagonal preconditioning. A more sophisticated preconditioning is reserved for future work.

The current version of SAIGE relies on Intel's Threading Building Blocks (TBB) library (Reinders 2007) for CPU parallelism, which is incompatible with some architectures such as IBM POWER9. This prevented deployment of native SAIGE on the DOE Oak Ridge Summit supercomputer used for MVP analyses.

2.2 GPU-accelerated and distributed implementation

To overcome scaling and hardware limitations, we developed a GPU-accelerated, distributed version of SAIGE to fit the model. The genotype matrix $\boldsymbol{A}$ is column-partitioned across nodes using the Message Passing Interface (MPI) (Nielsen 2016), so node i stores $A_{:,s_i:e_i}$. During each PCG iteration, the GRM-vector product as:

$$\boldsymbol{\psi v} = \frac{1}{\boldsymbol{M}} \sum_{i} \boldsymbol{A}_{:,s_i:e_i} (\boldsymbol{A}^{\boldsymbol{T}}_{:,s_i:e_i} \boldsymbol{v}) \tag{3}$$

Each node performs both multiplications on its GPUs using NVIDIA's *cuBLAS* library *cublasgemv* (clMathLibraries/clBLAS). Partial results are summed and broadcast using MPI. This hybrid GPU-MPI design reduces both runtime and memory overhead, enabling scalable mixed-model inference across heterogeneous clusters (Fig. 1A).

Because the GPUs have limited memory relative to CPUs, very large matrices often require multi-GPU distribution. The number of GPUs needed to store and process the GRM implicitly is:

$$n_{gpu} = \left\lceil \frac{4MN}{GPU_{mem} \times 10^9} \right\rceil \tag{4}$$

reflecting 4 bytes per floating-point value and unit conversion from bytes to gigabytes.

2.3 Optimizations for SAIGE associations test (Step 2)

SAIGE's association testing includes two key corrections that ensure statistical robustness at biobank scale. Saddlepoint Approximation (SPA) (Daniels 1954) provides accurate values under extreme case-control imbalance, and optional Firth (1993) penalized likelihood mitigates small-sample bias and separation issues. Together, these corrections preserve type I error control across a wide range of allele frequencies and phenotype distributions.

Building on this, we optimized SAIGE's second step by parallelizing both across phenotypes and within each phenotype. When the number of threads is set to >1, the sample list is split into equal chunks processed in parallel via R's *mclapply* and UNIX multiprocessing utilities.

We added support for a trait manifest file, allowing multiple phenotypes to be processed in a single batch by listing all required GMMAT model files, variance ratio files, and output destinations. This approach maximizes CPU utilization, reduces redundant file I/O, and allows batching as many traits as memory permits.

2.4 Biobank-scale analyses across population groups

We applied SAIGE-GPU to a phenome-wide GWAS of the 635 969 MVP participants (Gaziano *et al.* 2016, Verma *et al.* 2024) across four population groups: African, Admixed Americans, East Asian, and European (Table 1, available as supplementary data at *Bioinformatics* online). This required 4045 individual SAIGE runs. Model fitting used LD-pruned directly genotyped variants; association testing used imputed dosages. All analyses adjusted for age, sex, and population-specific genetic principal components. In total, over 350 billion variant-trait tests were detected (Fig. 1, available as supplementary data at *Bioinformatics* online).

2.5 Computational infrastructures

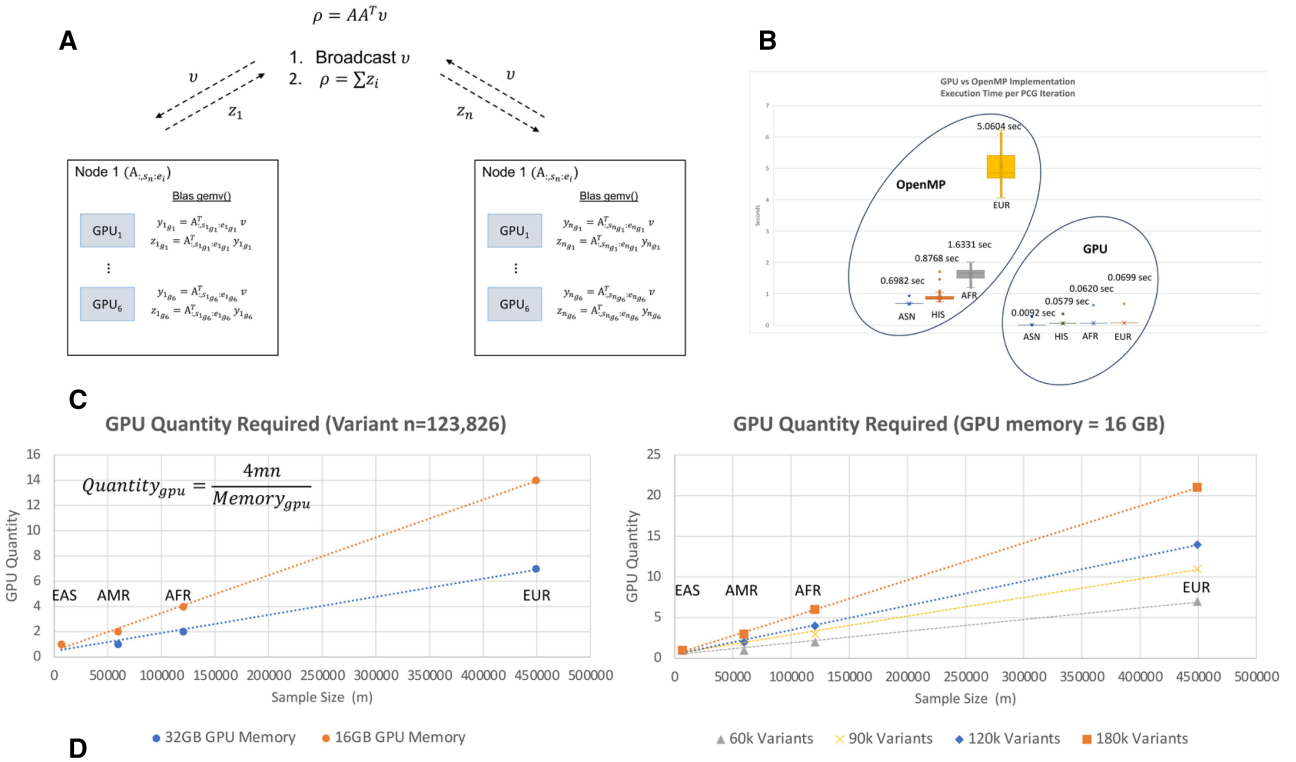
Initial analyses were run on the OLCF Summit supercomputer, equipped with two IBM POWER9 CPUs (512 GB DDR4) and six NVIDIA V100 GPUs (96 GB HBM2). Since POWER9—used mainly in national labs—is incompatible with Intel TBB and not representative of the x86 systems common in genomics research, we developed a custom CPU baseline for comparison.

To benchmark on standard x86 architectures, we used OLCF Frontier. Each node includes an AMD EPYC processor with 56 cores, four AMD MI250X GPUs (64 GB HBM2E each; 8 logical GPUs), and 512 GB DDR4 memory. We used three physical GPUs per trait to match Summit's memory footprint and ensure fair comparison. SAIGE v1.4.4.1 was containerized and executed using all 56 CPU cores. Docker and Singularity containers were also tested on Google Cloud Platform (GCP) and Microsoft Azure.

3 Results

3.1 GPU acceleration for SAIGE step 1 (model fitting)

Standard SAIGE was not tractable at MVP scale or deployable on OLCF's Summit architecture, motivating development of SAIGE-GPU. In benchmarking on Summit, GPU acceleration dramatically reduced the time per PCG iterations—from ~5.06 s



Category	All of Us European+	All of Us African American+	UK Biobank European+	UK Biobank African+
Variant Size	100,000	100,000	100,000	100,000
Sample Size	133,000	55,000	420,500	6,600
SAIGE GPU Analysis Time (hours)*	0.16	0.1	0.55	0.02
SAIGE GPU Analysis Cost	$0.42	$0.26	$1.45	$0.05
SAIGE CPU Analysis Time (hours)**	0.8	0.17	0.98	0.25
SAIGE CPU Analysis Cost	$3.17	$0.67	$3.88	$0.99

* Google Cloud - A100 GPU, 85 GB RAM, $2.64/hour

** Google Cloud - 96 Core VM, $3.96/hour

\+ Phenotype used was Type 2 Diabetes

Figure 1 (A) Distributed BLAS *gemv()*, matrix-vector multiplication, using GPUs on the cluster. The columns of matrix A are distributed and preloaded on GPUs, with node i having columns with indices from s_i to e_i, and these columns are distributed on GPUs on that node. The distributed matrix-vector multiplication computes $p = AA^T v$ through partial products on each node, avoiding explicit formation of the full GRM matrix. These partial products are aggregated to compute a solution p. (B) Demonstration of the time required for a single PCG iteration on a GPU, showcasing the efficient parallelization within the GPU. (C) GPU node requirements scale linearly with genotype matrix size (m samples × n variants). Left panel shows the effect of GPU memory size (16 GB versus 32 GB) with fixed variant count (n = 123 826). Right panel shows the effect of variant count with fixed GPU memory (16 GB). The relationship follows Quantity_gpu = 4 mn/Memory_gpu, where 4 is the byte size of a single-precision floating-point number and Memory_gpu is in bytes (e.g. 16 GB = 16×10^9 bytes). (D) Cost and time execution comparison using All of Us and UK Biobank data on Google Cloud Platform for SAIGE-GPU version versus the native SAIGE version.

on a 42-core OpenMP implementation to ∼0.069 s on GPUs (72× faster; Fig. 1B). Overall, model fitting for a representative trait completed in ∼30 min with GPUs versus 4 h 8 min with OpenMP (3× faster; Table 2, available as supplementary data at *Bioinformatics* online), and the required number of GPUs followed the predicted scaling relationship (Fig. 1C).

Performance varied with hardware. On Summit's higher-memory 32 GB GPUs, SAIGE-GPU achieved a 5× improvement relative to projected OpenMP performance (Table 3, available as supplementary data at *Bioinformatics* online). On Frontier—where native SAIGE could use 56 CPU cores per node—SAIGE-GPU with 3 × 64 GB GPUs achieved a more modest 2× advantage (Fig. 2, available as supplementary data at *Bioinformatics* online). These comparisons highlight that relative GPU speed-ups depend on CPU/GPU balance, memory capacity, and system architecture.

3.2 SAIGE-GPU container for cloud environment

Cloud deployment dramatically lowers barriers to entry for large-scale genomic analyses. To support cloud-based biobanks (UK Biobank, All of Us), we developed a portable SAIGE-GPU container. Across Google Cloud Platform and Microsoft Azure GPU acceleration consistently outperformed CPU-based SAIGE. For example, Type 2 Diabetes model fitting in All of Us (EUR population group) completed in ∼10 min on a single GPU versus 45 min on a 64-core VM (5× faster) at substantially lower cost ($0.42 versus $3.17) (Fig. 1D, Fig. 3, available as supplementary data at *Bioinformatics* online). UK Biobank analyses showed similar gains: ∼30 min on an A100 GPU ($1.45) versus 58 min on a 96-core CPU VM ($3.88).

3.3 SAIGE step 2 (association testing) parallelization

We incorporated multi-core and multi-trait parallelization into Step 2 by partitioning SNPs within traits and enabling concurrent multi-trait execution. This approach scales particularly well when many traits are processed together (Table 4, available as supplementary data at *Bioinformatics* online). In a test of 15 traits for a simulated 400 000-sample cohort, parallelized SAIGE-GPU reduced runtime by ∼11.5% relative to the native implementation (Table 5, available as supplementary data at *Bioinformatics* online). Larger gains are expected once the association testing step transitions to GPU-based kernels.

3.4 Future optimization directions

Remaining bottlenecks derive from large $\boldsymbol{AA}^T$operations, not from fixed-effects design matrices (which typically have $<$50 covariates), so GPU acceleration of fixed-effect terms would yield only minor ($<$5%) improvements. We are actively developing a GPU-optimized association-testing module; full end-to-end GPU acceleration of both SAIGE steps will provide additional speedups and further expand applicability across diverse computation environments.

4 Conclusion

We optimized SAIGE to fully leverage GPU acceleration, transforming previously impractical phenome-wide analyses on biobank-scale data into workloads that can be completed in under a month. These improvements substantially reduce runtime and node-hour consumption, addressing a major computational barrier as whole-genome sequencing continues to expand in scale and accessibility.

There is a growing need for mixed-model methods that operate efficiently across both HPC systems and commercial cloud platforms. SAIGE-GPU meets this need by providing scalable, portable, and cost-effective GLMM-based association testing across diverse computing environments. The tool is available in both source and containerized form for deployment on GPU-enabled systems. Source code and documentation are accessible at: https://github.com/saigegit/SAIGE/tree/SAIGE-GPU-1.3.3 and https://exascale-genomics.github.io/SAIGE-GPU. A code snapshot is archived at Zenodo for reproducibility (DOI: [10.5281/zenodo.17642591]).

Acknowledgements

We thank the Million Veteran Program, Office of Research and Development, and Veterans Health Administration for supporting this work. We would like to sincerely thank Dr Thomas Zacharia for providing access to the supercomputers at the Oak Ridge National Laboratory Leadership Computing Facility and Dr Dimitri Kusenov, the previous DOE Headquarters lead for the VA-DOE partnership, for his invaluable guidance and support. Their contributions have been instrumental in the successful completion of this study. Finally, we thank former staff members, and volunteers, who have contributed to MVP and, most of all, MVP participants for their service and their continued contributions to our nation through participation in this study. This publication does not represent the views of the Department of Veteran Affairs or the United States Government.

Author contributions

Alex Rodriguez (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [lead], Supervision [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Youngdae Kim (Conceptualization [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Resources [equal], Software [lead], Supervision [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Tarak Nath Nandi (Data curation [equal], Formal analysis [equal], Methodology [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Karl Keat (Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Mitch Conery (Formal analysis [equal], Methodology [equal], Resources [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Rachit Kumar (Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal]), Rohan Bhukar (Methodology [equal], Software [equal]), Molei Liu (Methodology [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal]), John Hesignton (Formal analysis [equal], Investigation [equal], Methodology [equal], Resources [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal]), Ketan Maheshwari (Methodology [equal], Resources [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal]), Edmon Begoli (Conceptualization [equal], Methodology

[equal], Resources [equal], Writing—original draft [equal], Writing—review & editing [equal]), Georgia Tourassi (Conceptualization [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), Sumitra Muralidhar (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Pradeep Natarajan (Project administration [equal], Resources [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), Benjamin Voight (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Kelly Cho (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), J. Michael Gaziano (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Scott Damrauer (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Katherine P. Liao (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Wei Zhou (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Jennifer E. Huffman (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Anurag Verma (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Funding acquisition [lead], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [lead], Supervision [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), and Madduri Ravi (Conceptualization [lead], Data curation [equal], Formal analysis [lead], Funding acquisition [equal], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [lead], Supervision [lead], Validation [equal], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

K.P.L. received a one-time consulting fee from UCB. S.M.D. receives research support from RenalytixAI and Novo Nordisk, outside the scope of the current research, and is named as a co-inventor on a Government-owned US Patent application related to the use of genetic risk prediction for venous thromboembolic disease and for the use of PDE3B inhibition for preventing cardiovascular disease, both filed by the US Department of Veterans Affairs in accordance with Federal regulatory requirements. All other authors declare that they have no competing interests.

Funding

The work was supported by the Million Veteran Program award #MVP000. This research used resources from the Knowledge Discovery Infrastructure at the Oak Ridge National Laboratory, supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725 and the Department of Veterans Affairs Office of Information Technology Inter-Agency Agreement with the Department of Energy under IAA No. VA118-16-M-1062. Other support by the National Institute of General Medical Sciences [R01GM138597 to A.V.]; National Institute Health [T32 AA028259 to J.D.D.]; National Library of Medicine [5R01LM010685 to (R.J.C.]; National Human Genome Research Institute [K99HG012222 to W.Z.]; National Institute of Arthritis and Musculoskeletal and Skin Diseases [P30AR072577 to K.P.L.]; National Institute of Diabetes and Digestive and Kidney Diseases [DK126194 to B.F.V.]; National Institute of Health [NIR01AG067025, K08MH122911 to G.V.]; National Institute of Health [BX004189, R01AG065582, R01AG067025 to P.R.]; Office of Research and Development, Veterans Health Administration award [I01CX001849-01 to J.G.]; Office of Research and Development, Veterans Health Administration awards [BX004821, CX001737, BX005831 to Y.S. V.]; Veterans Health Administration awards [IK2-CX001780 to S. M.D.]. National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00520902) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2020-II201336, Artificial Intelligence graduate school support (UNIST) to Y.K.].

Data availability

The tool is available in both source and containerized form for deployment on GPU-enabled systems. Source code and documentation are accessible at: https://github.com/saigegit/SAIGE/tree/SAIGE-GPU-1.3.3 and https://exascale-genomics.github.io/SAIGE-GPU. A code snapshot is archived at Zenodo for reproducibility (DOI: [10.5281/zenodo.17642591]).

References

Baji T. Evolution of the GPU device widely used in AI and massive parallel processing. In: *2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM)*. Kobe: IEEE, 2018, 7–9.

Bick AG, Metcalf GA, Mayo KR *et al.* Genomic data in the all of us research program. *Nature* 2024;**627**:340–6.

Chang CC, Chow CC, Tellier LC *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**:7–8.

clMathLibraries/clBLAS. https://github.com/clMathLibraries/clBLAS (22 July 2025, date last accessed).

Cook S. Chapter 9: optimizing your application. In: *CUDA Programming. Applications of GPU Computing Series,* **2013**, 305–440.

Daniels HE. Saddlepoint approximations in statistics. *Ann Math Stat* 1954;**25**:631–50.

Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;**80**:27–38.

Gaziano JM, Concato J, Brophy M *et al.* Million veteran program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* 2016;**70**:214–23.

Gottesman O, Kuivaniemi H, Tromp G *et al.*; eMERGE Network. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet Med* 2013;**15**:761–71.

Kaasschieter EF. Preconditioned conjugate gradients for solving singular systems. *J Comput Appl Math* 1988;**24**:265–75.

Karczewski KJ, Gupta R, Kanai M *et al.* Pan-UK Biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. *Nat Genet* 2025;**57**:2408–17. https://doi.org/10.1038/s41588-025-02335-7

Kurki MI, Karjalainen J, Palta P *et al.*, FinnGen. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* 2023;**613**:508–18.

Lee SH, Yang J, Goddard ME *et al.* Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 2012;**28**:2540–2.

Loh P-R, Tucker G, Bulik-Sullivan BK *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015;**47**:284–90.

Mbatchou J, Barnard L, Backman J *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* 2021;**53**:1097–103.

Nielsen F. Introduction to HPC with MPI for data science. Undergraduate Topics in Computer Science, **2016**, 21–62.

Nocedal J, Wright S. *Numerical Optimizations*. 2nd edn. New York, NY, USA: Springer, 2006.

Reinders J. *Intel Threading Building Blocks: Outfitting C++ for Multi-Core Processor Parallelism*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2007.

Sakaue S, Kanai M, Tanigawa Y, FinnGen *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* 2021;**53**:1415–24.

Sudlow C, Gallacher J, Allen N *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**:e1001779.

Sunitha NV, Raju K, Chiplunkar NN. Performance improvement of CUDA applications by reducing CPU-GPU data transfer overhead. In: *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*. Coimbatore, India: IEEE, 2017, 211–5.

Uffelmann E, Huang QQ, Munung NS *et al.* Genome-wide association studies. *Nat Rev Methods Primers* 2021;**1**:59.

Verma A, Damrauer SM, Naseer N *et al.* The Penn medicine BioBank: towards a genomics-enabled learning healthcare system to accelerate precision medicine in a diverse population. *J Pers Med* 2022;**12**. https://doi.org/10.3390/jpm12121974

Verma A, Huffman JE, Rodriguez A *et al.* Diversity and scale: genetic architecture of 2068 traits in the VA million veteran program. *Science*, 2024**385**:eadj1182.

Wolford BN, Willer CJ, Surakka I. Electronic health records: the next wave of complex disease genetics. *Hum Mol Genet* 2018;**27**:R14–21.

Zawistowski M, Fritsche LG, Pandit A *et al.* The Michigan genomics initiative: a biobank linking genotypes and electronic clinical records in Michigan medicine patients. *Cell Genom* 2023;**3**:100257. https://doi.org/10.1016/j.xgen.2023.100257

Zhou W, Nielsen JB, Fritsche LG *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018;**50**:1335–41.

Bioinformatics, 2026, 42, 1–7
https://doi.org/10.1093/bioinformatics/btag032
Applications Note