Accelerating Genome- and Phenome-Wide Association Studies using GPUs – A case study using data from the Million Veteran Program

Alex Rodriguez[†1], Youngdae Kim[†2], Tarak Nath Nandi[1], Karl Keat[3], Rachit Kumar[3], Rohan Bhukar[4,5], Mitchell Conery[6], Molei Liu[7], John Hessington[8], Ketan Maheshwari[9], Drew Schmidt[9], *VA Million Veteran Program*[10], Edmon Begoli[9], Georgia Tourassi[11], Sumitra Muralidhar[12], Pradeep Natarajan[5,13,14,15], Benjamin F Voight1[16,6,17,18], Kelly Cho[19,13,20], J Michael Gaziano[19,13,20], Scott M Damrauer[16,17,21,22], Katherine P Liao[23,13,24,25,26], Wei Zhou[4,27,28], Jennifer E Huffman[23,13,29], Anurag Verma[‡16,3,30], Ravi K Madduri[‡1]

[1]Data Science and Learning, Argonne National Laboratory, Lemont, IL, 60439, USA
[2]Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, 60439, USA
[3]Institute for Biomedical Informatics, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA
[4]Program in Medical and Population Genetics, Cambridge, MA, 02142, USA
[5]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, 02114, USA
[6]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA
[7]Department of Biostatistics, Columbia University's Mailman School of Public Health, New York, NY, 10032, USA
[8]Information systems, University of Pennsylvania, Philadelphia, PA, 19104, USA
[9]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[10]See Supplement for a list of MVP contributors
[11]Computing and Computational Sciences Directorate, Oak Ridge National Laboratory, Oak Ridge, TN, 37830, USA
[12]Office of Research and Development, Department of Veterans Affairs, Washington, DC, 20420, USA
[13]Department of Medicine, Harvard Medical School, Boston, MA, 02115, USA
[14]Program in Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Cambridge, MA, USA
[15]Cardiology Division, Massachusetts General Hospital, Boston, MA, 02114, USA
[16]Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, 19104, USA
[17]Department of Genetics, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA
[18]Institute of Translational Medicine and Therapeutics, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA
[19]MVP Boston Coordinating Center, VA Boston Healthcare System, Boston, MA, 02111, USA
[20]Department of Medicine, Division of Aging, Brigham and Women's Hospital, Boston, MA, 02115, USA

[21]Department of Surgery, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA
[22]Cardiovascular Institute, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA
[23]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA, 02130, USA
[24]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115, USA
[25]Medicine, Rheumatology, VA Boston Healthcare System, Boston, MA, 02130, USA
[26]Department of Medicine, Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Boston, MA, 02115, USA
[27]Department of Medicine, Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, 02114, USA
[28]Stanley Center for Psychiatric Research, Cambridge, MA, 02142, USA
[29]Palo Alto Veterans Institute for Research (PAVIR), Palo Alto Health Care System, Palo Alto, CA, 94304, USA
[30]Department of Medicine, Division of Translational Medicine and Human Genetics, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA

Joint Authorship

†These authors contributed equally to this work
‡These authors supervised equally to this work
*Corresponding author: Ravi K. Madduri: madduri@anl.gov

Abstract

The expansion of biobanks has significantly propelled genomic discoveries yet the sheer scale of data within these repositories poses formidable computational hurdles, particularly in handling extensive matrix operations required by prevailing statistical frameworks. In this work, we introduce computational optimizations to the SAIGE (Scalable and Accurate Implementation of Generalized Mixed Model) algorithm, notably employing a GPU-based distributed computing approach to tackle these challenges. We applied these optimizations to conduct a large-scale genome-wide association study (GWAS) across 2,068 phenotypes derived from electronic health records of 635,969 diverse participants from the Veterans Affairs (VA) Million Veteran Program (MVP). Our strategies enabled scaling up the analysis to over 6,000 nodes on the Department of Energy (DOE) Oak Ridge Leadership Computing Facility (OLCF) Summit High-Performance Computer (HPC), resulting in a 20-fold acceleration compared to the baseline model. We also provide a Docker container with our optimizations that was successfully used on multiple cloud infrastructures on UK Biobank and All of Us datasets where we showed significant time and cost benefits over the baseline SAIGE model.

Introduction

The rapid expansion of biobanks has significantly advanced genomic discoveries, facilitating studies on the genetic basis of disease and catalyzing studies in personalized medicine. The increasing number of newly formed biobanks, alongside the continued growth of established ones, enables researchers to conduct studies with increasingly larger sample sizes, yielding more robust and generalizable findings. Furthermore, biobanks linked to electronic health records (EHR) have been instrumental in translational studies, providing data on both the genome and phenome in large populations (1-7).

However, the unprecedented size of the data accessible via biobanks require researchers to consider the computational challenges arising from data complexity, analysis methodologies, statistical frameworks, and infrastructure constraints. Addressing the computational limitations requires development of innovative algorithms, optimization strategies, and adaptable computing architectures tailored to the unique requirements of biomedical research. Additionally, fostering collaboration among computational scientists, statisticians, and domain experts proves indispensable in crafting resilient computational tools and workflows capable of facilitating the efficient analysis of burgeoning biobank data. Leveraging the enhanced computational infrastructures afforded by high-performance computing (HPC) and cloud environments further augments the capacity for comprehensive analysis within the biomedical domain.

One routine statistical analysis that utilizes genetic and phenotypic data available from large biobanks is the genome-wide association study (GWAS). The purpose of GWAS is to identify association between polymorphic DNA variants in the genome among biobank participants and a

phenotypic trait or disease of interest, typically extracted from EHR or clinical data from participants where the underlying computation involves millions of iterations of a generalized linear model over the available genetic exposures (8). Furthermore, the complexity of the analysis increases when state-of-the-art approaches extend the analysis to use multi-level models to better account for population architecture and relatedness, in addition to the volume of the data resulting in huge, dense matrix-matrix and matrix-vector operations that are performed as a part of statistical scaffolding (9-12). Compounding this computational complexity are data scaling challenges, including the desire to analyze the entire catalog of traits extracted from EHR data (i.e., the "phenome", 100s-1000s of traits) and doing so across all population groups represented in order to capture the diverse representation that is increasingly available. The scale of this undertaking requires large amounts of disk space, fast processors, and innovative techniques to take advantage of the resources available.

The U.S. Department of Veterans Affairs (VA) Million Veteran Program (MVP) stands as a pioneering research endeavor, continually expanding in scope and offering researchers a platform to tackle the aforementioned challenges. This biobank, aimed at advancing precision medicine and improving healthcare outcomes for Veterans, includes a large number of individuals from underrepresented populations. The VA and the Department of Energy (DOE) established an Interagency Agreement (IAA) to combine VA's vast array of clinical and genomic data with DOE's national computing capabilities, including the most powerful supercomputer in the Nation, to push the frontiers of precision medicine and computing with vision to improve the lives of Veterans and all Americans and support the national Precision Medicine Initiative. Its primary mission revolves around furnishing comprehensive genotype-phenotype insights into prevalent and significant health outcomes, leveraging the most extensive EHR-linked biobank in the United States. Among the formidable computational challenges encountered in this research is the task of conducting GWAS across a staggering 3.5 billion genetic variants, spanning thousands of traits gleaned from the electronic health records of 635,969 MVP participants.

SAIGE (the Scalable and Accurate Implementation of Generalized Mixed Model algorithm) (10) is one such state-of-the-art multi-level modeling based GWAS approach designed to accommodate sample relatedness and manage unbalanced case-control ratios typical in biobanks like MVP. A crucial aspect of GWAS analysis entails constructing a Genetic Relationship Matrix (GRM), which measures the genetic relatedness or similarity among individuals within the study cohort. SAIGE offers users the option to generate either a sparse or a full GRM. While a sparse GRM boasts faster executions and lower memory demands, opting for a full GRM provides a more precise assessment of pairwise relatedness among all individuals, enhancing the depiction of genetic relationships. This proves invaluable for various downstream analyses, including estimating heritability, evaluating genetic correlations, and achieving a deeper understanding of the genetic architecture of the trait in question (14). To mitigate the computational burden associated with storing and inverting the full GRM, SAIGE employs the pre-conditioned

conjugate gradient (15) method to iteratively solve linear equations. Nevertheless, it still faces substantial computation challenges due to extensive matrix and vector multiplications across numerous iterations.

In high-performance computing, processors face performance bottlenecks due to memory and disk input/output (I/O) operations. While processors execute computations rapidly, they depend on memory for data storage. Limited memory capacity necessitates frequent reads and writes to disk storage, slowing overall execution. Parallel processing distributes workloads across multiple compute processor units (CPU) but memory limitations persist. These bottlenecks are especially pronounced in large matrix operations. Graphics Processing Units (GPUs) provide an optimized architecture through high memory bandwidth and capacity (16), massive parallelism (17), and reduced data transfer between CPU and memory (18), particularly for large matrix operations. By leveraging these GPU attributes, computationally intensive tasks, like large matrix operations, can experience significance performance enhancements compared to CPU processing alone, mitigating memory and disk I/O bottlenecks.

We adapted SAIGE, initially tailored for CPU infrastructure, to utilize both CPUs and GPUs at the DOE Oak Ridge Leadership Computing Facility (OLCF) Summit HPC. This adaptation markedly expedited the analysis process, resulting in a more than 20-fold acceleration, far surpassing what would have been attainable on a CPU-based cluster. Furthermore, our work provided a *generic optimization framework* for other analytical tools based on generalized linear mixed models using a full GRM. We also created a Docker container for deployment on various cloud infrastructures. We present a comparison of the time and cost between the SAIGE GPU and CPU versions.

Materials and Methods

Study Design, Population Groups, and Phenotypic Definitions

The analysis involved a series of GWAS across 2,068 traits, covering a deep catalog of phenotypes extracted from EHR-derived diagnosis codes, clinical laboratory tests, vital signs, and survey responses. As previously described (13, 19, 20, 21, 22), the analysis was performed using data from 635,969 participants from MVP Genomics Release 4 (23) (**Table 1**) classified into four population groups based on genetic similarity (GIA) to the 1000 Genomes Project (24, 25) African (AFR, n = 121,177), Admixed Americans (AMR, n = 59,048), East Asian (EAS, n = 6,702), and European (EUR, n = 449,042) superpopulations. After imputation and quality control (QC) filtering, > 44.3M variants (minor allele count (MAC) ≥ 40) were included for analysis. For a visual representation of the analysis, please refer to **Figure 1a**, which illustrates the different quantities for which the analysis was conducted. After trait quality control, 1,854 binary and 214 quantitative traits were included in the downstream analysis in at least one population group (**Figure 1b**).

Additionally, genotype data from the UK Biobank (3) and the All of Us Research Program (7) were utilized to test the software capabilities on a cloud environment. Within the UK Biobank data, we employed the European (EUR, n = 420,500) and African (AFR, n = 6,600) cohorts where PCA was used to measure the population structure (26). The All of Us cohorts included European (EUR, n = 133,000) and African (AFR, n = 55,000) population groups where PCA was used to measure the population structure (7).

Biobank-scale genomic analysis across population groups

In total, 4,045 GWAS SAIGE runs were needed for the GW-PheWAS analysis which resulted in over 350 billion variant-trait associations across population groups. The current implementation of the SAIGE algorithm was not analytically tractable at this scale of computation. SAIGE uses R/C++ based tools developed for CPU environments and uses the Intel Threading Building Blocks (TBB) (27) library to enable parallelization. The SAIGE method comprises two primary steps. Step one involves fitting a linear/logistic mixed model with a GRM included under the null hypothesis that no genetic variants are associated with the phenotype of interest. We note that fitting the null mixed model involves thousands of matrix-matrix and matrix-vector operations, which are best suited for a GPU environment. Step two tests each SNP at a time across the genome for their association with the phenotype with a score test using the saddlepoint approximation (SPA) (28) and Firth regression methods (29) to account for unbalanced case-control sample sizes.

Directly genotyped variants were used for step one of SAIGE and were filtered for pairwise correlation with a window size, number of SNPs and VIF threshold of 50, 5, and 0.2 respectively using Plink1.9 (9). Imputed genetic dosages were used for step two of SAIGE. Only variants with an imputation quality > 0.3 and MAC ≥ 40 within the relevant population groups were included in the GWAS execution. Analyses were adjusted for age, sex, and the top ten population specific genetic principal components (PC) estimated by Principal component analysis (13).

Computational Infrastructures

All GWAS analysis was conducted on the Summit HPC, located at DOE's OLCF. It consists of 4,608 nodes, with each node featuring two IBM POWER9 processors and six NVIDIA Tesla V100 GPUs of 16 GB memory. All but 54 of the Summit nodes are equipped with 512 GB of DDR4 memory for the POWER9 processors and 96 GB of high-bandwidth memory (HBM2) for the V100 GPUs. The remaining 54 nodes in Summit HPC are high-memory nodes equipped with 2 TB of DDR4 memory for the CPUs and 192 GB of HBM2 for the GPUs with 32 GB of memory per GPU. Specifically, for the GW-PheWAS analysis, we utilized the nodes on Summit with 512 GB DDR4 memory and 96 GB HBM2 memory.

To further advance the use of SAIGE-GPU in various research environments, we generated Docker and Singularity containers. We conducted extensive testing of this containerized solution on the GPUs available on Google Cloud Platform (GCP) and Azure Cloud Platforms. Our code for containers and optimizations is provided (**Data and materials availability**).

Results

We initially focused on optimizing step one for SAIGE as it can largely benefit by using GPUs for matrix-vector operations in the calculation of the GRM on the fly and employed MPI to distribute the data across multiple GPUs. While the standard SAIGE method on CPU-based machines was suitable for relatively small cohorts, it became impractical for larger population groups (e.g., groups similar to 1KG-European and 1KG-Africa in MVP) due to the substantial size of the matrix-vector operations involved. Employing the GPU-modified SAIGE framework, we successfully conducted a total of 4,045 independent GWAS runs. The GWAS analysis was accomplished within 14,286 GPU hours for step one, equivalent to 5 days of wall time, resulting in a 160-fold reduction in required core CPU hours in a CPU environment cluster (**Table 2**). Step two in SAIGE presented distinct challenges due to the need for millions of association tests for each trait, totaling over 3 billion association tests.

Optimizations for SAIGE Step One Using GPUs

The primary challenge we encountered when using the SAIGE algorithm on the DOE OLCF Summit HPC was the IBM POWER9 processors incompatibility with the Intel threading building block (TBB) library, which is instrumental for parallelization within step one. This issue prevented us from installing the native SAIGE version, prompting us to find alternative solutions. In addition, solving the logistic mixed model using the PCG algorithm posed challenges due to the numerous iterations and the time-consuming nature of the process. Step one's time complexity (O) is $O(MN^{1.5})$, where N is the sample size, and M is the number of genetic markers per sample, making the calculation of the GRM a substantial contributor to the overall computational time for this step, particularly when dealing with large sample sizes (as indicated in equations 1, 2, 3) (10). Building and storing the GRM demanded substantial memory and computational resources. SAIGE's approach addressed the memory issue by generating GRM segments on-demand, albeit at the cost of increased time requirements and the need for extensive parallelization using multiple CPUs.

SAIGE models the relationship between traits ($\boldsymbol{Y}$) and genotypes ($\boldsymbol{G}$) while adjusting for other covariates ($\boldsymbol{X}$) and random genetic effects ($\boldsymbol{b}$) accounting for unknown sample relatedness based on the linear and logistic mixed models (9) (equation 1):

$$logit(\boldsymbol{Y}) = \boldsymbol{X}\alpha + \boldsymbol{G}\beta + \boldsymbol{b} + e \quad (1)$$

α and β are the coefficient vectors of fixed effects and genetic effects, respectively and e is a random effect of residual errors. Each element in $\boldsymbol{Y}$ represents the probability for an individual being a case given the covariates and genotypes as well as the random effect. The variable $\boldsymbol{b}$ is assumed to be sampled from a normal distribution with a mean of zero, and a standard deviation of $\tau\boldsymbol{\psi}$, where $\boldsymbol{\psi}$ is the GRM calculated as

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

where $\boldsymbol{A}$ is the genotype matrix of size $N \times M$. Optimizing a linear system solution involving the GRM ($\boldsymbol{\psi}$) matrix on GPUs is our focus.

The model is fit under the null hypothesis of $\beta = 0$, in which the iterative PCG method is used to obtain a solution to a linear system of equations defined by $\boldsymbol{\psi x} = \boldsymbol{b}$ for a given vector b. This iterative process, central to step one, was time-consuming due to multiple matrix-vector operations involving the GRM. Furthermore, building the GRM itself is increasingly memory intensive as the number of individuals and marker panels increase in size. For example, the MVP release 4 population group similar to 1KG-Europeans (N = 445,444; M = 120,000) would produce a GRM of approximately 800 gigabytes.

To accelerate the computation time and reduce the memory footprint, we employed distributed computing techniques involving the use of Message Passing Interface (MPI) (30) and were able to successfully exploit the parallel computing capability of GPUs for matrix-vector multiplications. Specifically, we partition the columns of the matrix A that are used to form the GRM and distribute them into a set of nodes on a cluster. For example, node i contains columns $A_{:,s_i:e_i}$ with s_i and e_i denoting the start and end indices of the columns of A stored in node i. At each iteration of the PCG method, a matrix-vector multiplication $\boldsymbol{\psi v}$ for some vector $\boldsymbol{v}$ is performed. Using the fact that $\boldsymbol{\psi v} = \frac{\mathbf{1}}{\boldsymbol{M}}\sum_{\boldsymbol{i}} \boldsymbol{A}_{:,\boldsymbol{s_i}:\boldsymbol{e_i}}(\boldsymbol{A}^{\boldsymbol{T}}_{:,\boldsymbol{s_i}:\boldsymbol{e_i}}\boldsymbol{v})$, each node computes its summand in parallel on GPUs. The results of all nodes are summed and redistributed using MPI. NVIDIA's BLAS library *cublasgemv* (30) is used to compute the summand to further accelerate the two matrix-vector multiplications, $\boldsymbol{y_i} := \boldsymbol{A}^{\boldsymbol{T}}_{:,\boldsymbol{s_i}:\boldsymbol{e_i}}\boldsymbol{v}$ and $\boldsymbol{A}_{:,\boldsymbol{s_i}:\boldsymbol{e_i}}\boldsymbol{y_i}$, on GPUs (**Figure 2**).

To deal with the large memory requirement, SAIGE relied on the Intel TBB package to parallelize this step, which was incompatible with the Summit infrastructure. We initially replaced the TBB's parallelization method with OpenMP (32) for executing the matrix-vector operations. However, the primary benefit of accelerating step one lies in the considerably faster matrix computations achieved using GPUs compared to CPUs. We compared the SAIGE version that leveraged OpenMP API for parallelization with the GPU version (**Table 3**) using the Varicose Veins trait (454.1 ICD-9 code). In the OpenMP version, we utilized all 42 available cores on the compute node for parallelizing the matrix calculations to generate the GRM, while for the GPU version we utilized 16 GPUs each equipped with 16 GB of memory in each GPU to

distribute subsections of the matrix with dimensions of 8,256 by 445,444. On average, a single PCG iteration on a GPU required approximately 0.069 seconds for the group similar to 1KG-Europe in MVP. In contrast, the OpenMP SAIGE implementation took roughly 5.06 seconds, marking a substantial 72-fold improvement for PCG iterations to converge (**Figure 3**). It took 30 minutes (on 3 nodes with 6 GPUs each) using the GPU-SAIGE implementation to complete step one. Conversely, the same analysis conducted with the OpenMP implementation took 4 hours and 8 minutes in a single 42-core node, representing an overall 3-fold improvement and considers other processes within step one such as processing the input data. While the OpenMP implementation successfully executed the calculations in SAIGE step one on CPUs with a low memory footprint, it required numerous CPU parallel processes to achieve convergence. The advantage of using GPUs becomes readily apparent as the genotype matrix size grows, because it takes substantially longer for CPUs to parallelize the matrix operations. The GPU capitalizes on its inherent parallelization capabilities and pre-loading contents of the matrix into memory, offering a substantial performance boost for large-scale genetic analysis.

It is important to note that, due to computing the complete GRM in parallel GPUs, the memory footprint increased in comparison to the CPU-based approach which processes the GRM in small segments independent of one another. Thus, the number of nodes needed to cover the GPUs is increased per trait in larger population groups. The amount GPUs required for a run was calculated using the formula:

$$n_{gpu} = \text{ceil}(4 \times M \times N/(GPU_{mem} * 10\text{^}9)) \quad (3)$$

This calculation factored in GPU memory capacity (GPU_{mem}), the byte size of a single precision floating-point number (4 bytes), and the conversion between bytes and gigabytes (10^9). This formula can be used for any cohort in additional biobanks to determine the number of GPUs to be used on other computational environments (i.e., cloud infrastructure). This estimation considered the linear relationship between the genotype matrix size, GPU memory available and the required number of nodes which can be visualized (**Figures 4a, 4b**).

The optimizations made in step one effectively harnessed the speed of GPU matrix computation and parallelization, resulting in a significant reduction in analysis time. The GPU optimization of step one enabled the completion of the GWAS analysis for all traits and population groups within 2,381 node hours, representing a remarkable 20-fold improvement for step one in comparison to the initial native SAIGE implementation in a CPU-based cluster (as presented in **Table 2**). Consequently, step one was accomplished in less than 5 days through efficient utilization of node hours facilitated by high-memory Summit nodes for all MVP traits and population groups. Overall, an effective usage of 22,051 GPUs was needed to complete the analysis.

SAIGE Step two Job Management

In step two of the SAIGE algorithm, millions of variant association tests were conducted independently, given the highly parallel nature of these jobs. Execution times for both the SAIGE-GPU and SAIGE-OpenMP implementations incorporated these optimizations for step two which showed an improvement of 2 to 3-fold compared to initial tests (as summarized in **Table 4)**. To enable parallelization, the MVP genotype data files were partitioned into 219 files based on imputation analysis results. This data partitioning strategy facilitated the parallel execution of 219 jobs per trait and population group, totaling nearly 2 million independent jobs. The predominant challenge in step two revolved around managing the substantial number of jobs required for which we used the R library Tasktools (33), enabling successful submission and monitoring of the jobs.

SAIGE-GPU Container on Cloud Infrastructures

While the comprehensive analysis was conducted on the OLCF Summit HPC infrastructure, as the MVP cohort data was exclusively available on OLCF computational resources, we note that other cohorts, such as the United Kingdom Biobank (UKBB) (3) and All of Us Biobank (AoU) (7), can only be accessed through cloud infrastructures like the Google Cloud Platform (GCP) and Azure. In response to this demand, we have developed a specialized container image designed for versatile deployment across various cloud infrastructures.

To evaluate its performance, we conducted a comparative study that pitted SAIGE-GPU against SAIGE-CPU using data from the UK and AoU Biobanks. We employed the Type 2 Diabetes (T2D) trait to assess their precision, processing speed, and cost-effectiveness within the GCP cloud environment for two of the largest genetically inferred population groups, namely African and European (**Figure 5** and **Table 5**). For instance, a 5-fold improvement in execution time was seen when analyzing the T2D trait from AoU across the European population group (N = 133,000; M = 100,000). Step one completed in 10 minutes using 1 GPU (A100 GPU, 85 GB RAM), whereas the CPU-based SAIGE version consumed 45 minutes on a 64-core virtual machine. Furthermore, the cost of utilizing 1 GPU for the EUR cohort amounted to approximately $0.42, while the cost of the 64-core VM was $3.17. A similar trend in terms of cost and time is observed for the AFR population group, which would have a smaller memory footprint due to the matrix size.

This same pattern of advantages is evident when applied to UKBB traits, as exemplified in table 5. Specifically, we focused on the EUR population group, which consisted of 420,500 individuals, closely resembling the MVP EUR cohort in participant size. GCP infrastructure (NVIDIA Tesla A100 GPUs, 12 vCPUs, and 85GB of RAM) was employed to run the T2D trait and completed the analysis in just over 30 minutes, with an average cost of $1.45. In contrast, utilizing the CPU-based SAIGE version consumed 58 minutes and incurred a cost of $3.88 using a 96-core VM.

Conclusion

We leveraged the GPU computational resources of the DOE OLCF Summit HPC address major computational challenges posed by the increasingly large datasets utilized in genomics research. In this example, we demonstrate optimization of a highly used tool for genomic analyses designed for CPUs, SAIGE. Prior to optimizing step one with GPUs, the analysis would have spanned several years for all genetically inferred population groups. The optimizations have now condensed the completion time to under a month, reducing node hours by a substantial factor. Even though we largely focused on step one of SAIGE, in Step two we showed how we executed millions of variant association tests in parallel, a highly compute-intensive task. We intend to further improve this step by parallelizing this step instead of performing each association in serial mode.

In a recent article (34), the authors performed a large analysis on over 7,000 traits of the Pan-UK Biobank (35) data for multiple ancestries using hail-batch on the Google Cloud Platform. As previously mentioned, both the European population groups for UKBB and MVP are comparable in size, while the African, Admixed American and Eastern Asian population groups are larger for the MVP. The authors used the SAIGE-CPU implementation to perform close to 300 billion associations and required over 3.8 million CPU hours to complete both step one and step two in SAIGE. In comparison, the MVP analysis required 14,283 GPU hours for step one and approximately 2 million CPU hours for step two to perform over 350 billion associations.

The MVP has now expanded to a million individuals (36) and plans to collect whole-genome sequencing data, likely to increase the number of low-frequency variants that will be tested in the future. Thus, it is imperative to understand approaches to efficiently optimize software already developed for these data in HPC environments. Our primary focus lay in enhancing the efficiency of SAIGE's first step since it is iteratively employed in numerous downstream SAIGE-related analyses (e.g., SAIGE-GENE (37)). However, our ongoing efforts center on further streamlining SAIGE for GW-PheWAS studies across multiple biobanks such as All of Us, UK Biobank, Penn Medicine BioBank (2).

The continuous evolution of GPU technology in various implementations offers a promising outlook. The Summit infrastructure currently harnesses NVIDIA CUDA libraries for these operations, but future systems may incorporate different libraries, further accelerating execution times and lowering costs. These systems are expected to feature expanded memory and storage capacities. Additionally, our GPU-based SAIGE implementation can be readily adapted for Intel GPUs using the Intel oneAPI platform and AMD GPUs using their ROCm platform.

A container is available for deployment on cloud platforms equipped with GPU nodes. The code can be accessed at https://exascale-genomics.github.io/SAIGE-GPU. The significant improvements in efficiency achieved with SAIGE using GPUs demonstrate the potential for the

development of new and existing tools capable of performing population analysis at the exascale level by optimizing software for GPU usage.

References and Notes

1. B. N. Wolford, C. J. Willer, I. Surakka, Electronic health records: the next wave of complex disease genetics. *Hum. Mol. Genet.* **27**, R14–R21 (2018).
2. A. Verma, S. M. Damrauer, N. Naseer, J. Weaver, C. M. Kripke, L. Guare, G. Sirugo, R. L. Kember, T. G. Drivas, S. M. Dudek, Y. Bradford, A. Lucas, R. Judy, S. S. Verma, E. Meagher, K. L. Nathanson, M. Feldman, M. D. Ritchie, D. J. Rader, For The Penn Medicine BioBank, The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.* **12**, 1974 (2022).
3. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
4. M. Zawistowski, L. G. Fritsche, A. Pandit, B. Vanderwerff, S. Patil, E. M. Schmidt, P. VandeHaar, C. J. Willer, C. M. Brummett, S. Kheterpal, X. Zhou, M. Boehnke, G. R. Abecasis, S. Zöllner, The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genomics*. **3**, 100257 (2023).
5. M. I. Kurki, J. Karjalainen, P. Palta, T. P. Sipilä, K. Kristiansson, K. M. Donner, M. P. Reeve, H. Laivuori, M. Aavikko, M. A. Kaunisto, A. Loukola, E. Lahtela, H. Mattsson, P. Laiho, P. Della Briotta Parolo, A. A. Lehisto, M. Kanai, N. Mars, J. Rämö, T. Kiiskinen, H. O. Heyne, K. Veerapen, S. Rüeger, S. Lemmelä, W. Zhou, S. Ruotsalainen, K. Pärn, T. Hiekkalinna, S. Koskelainen, T. Paajanen, V. Llorens, J. Gracia-Tabuenca, H. Siirtola, K. Reis, A. G. Elnahas, B. Sun, C. N. Foley, K. Aalto-Setälä, K. Alasoo, M. Arvas, K. Auro, S. Biswas, A. Bizaki-Vallaskangas, O. Carpen, C.-Y. Chen, O. A. Dada, Z. Ding, M. G. Ehm, K. Eklund, M. Färkkilä, H. Finucane, A. Ganna, A. Ghazal, R. R. Graham, E. M. Green, A. Hakanen, M. Hautalahti, Å. K. Hedman, M. Hiltunen, R. Hinttala, I. Hovatta, X. Hu, A. Huertas-Vazquez, L. Huilaja, J. Hunkapiller, H. Jacob, J.-N. Jensen, H. Joensuu, S. John, V. Julkunen, M. Jung, J. Junttila, K. Kaarniranta, M. Kähönen, R. Kajanne, L. Kallio, R. Kälviäinen, J. Kaprio, FinnGen, N. Kerimov, J. Kettunen, E. Kilpeläinen, T. Kilpi, K. Klinger, V.-M. Kosma, T. Kuopio, V. Kurra, T. Laisk, J. Laukkanen, N. Lawless, A. Liu, S. Longerich, R. Mägi, J. Mäkelä, A. Mäkitie, A. Malarstig, A. Mannermaa, J. Maranville, A. Matakidou, T. Meretoja, S. V. Mozaffari, M. E. K. Niemi, M. Niemi, T. Niiranen, C. J. O´Donnell, M. Obeidat, G. Okafo, H. M. Ollila, A. Palomäki, T. Palotie, J. Partanen, D. S. Paul, M. Pelkonen, R. K. Pendergrass, S. Petrovski, A. Pitkäranta, A. Platt, D. Pulford, E. Punkka, P. Pussinen, N. Raghavan, F. Rahimov, D. Rajpal, N. A. Renaud, B. Riley-Gillis, R. Rodosthenous, E. Saarentaus, A. Salminen, E. Salminen, V. Salomaa, J. Schleutker, R. Serpi, H. Shen, R. Siegel, K. Silander, S. Siltanen, S. Soini, H. Soininen, J. H. Sul, I. Tachmazidou,

K. Tasanen, P. Tienari, S. Toppila-Salmi, T. Tukiainen, T. Tuomi, J. A. Turunen, J. C. Ulirsch, F. Vaura, P. Virolainen, J. Waring, D. Waterworth, R. Yang, M. Nelis, A. Reigo, A. Metspalu, L. Milani, T. Esko, C. Fox, A. S. Havulinna, M. Perola, S. Ripatti, A. Jalanko, T. Laitinen, T. P. Mäkelä, R. Plenge, M. McCarthy, H. Runz, M. J. Daly, A. Palotie, FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. **613**, 508–518 (2023).

6. O. Gottesman, H. Kuivaniemi, G. Tromp, W. A. Faucett, R. Li, T. A. Manolio, S. C. Sanderson, J. Kannry, R. Zinberg, M. A. Basford, M. Brilliant, D. J. Carey, R. L. Chisholm, C. G. Chute, J. J. Connolly, D. Crosslin, J. C. Denny, C. J. Gallego, J. L. Haines, H. Hakonarson, J. Harley, G. P. Jarvik, I. Kohane, I. J. Kullo, E. B. Larson, C. McCarty, M. D. Ritchie, D. M. Roden, M. E. Smith, E. P. Böttinger, M. S. Williams, The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* **15**, 761–771 (2013).
7. The All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. Nature (2024). https://doi.org/10.1038/s41586-023-06957-x.
8. Uffelmann, E., Huang, Q.Q., Munung, N.S. et al. Genome-wide association studies. Nat Rev Methods Primers 1, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9
9. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. doi:10.1186/s13742-015-0047-8
10. W. Zhou, J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. Gifford, L. A. Bastarache, W.-Q. Wei, J. C. Denny, M. Lin, K. Hveem, H. M. Kang, G. R. Abecasis, C. J. Willer, S. Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. 50, 1335–1341 (2018).
11. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47(3):284-290. doi:10.1038/ng.3190
12. Mbatchou J, Barnard L, Backman JD, et al. A comparison of methods for correcting for population stratification in genome-wide association studies of rare variants. Hum Genet. 2019;138(7):749-764. doi:10.1007/s00439-019-02022-3
13. A. Verma, J.E. Huffman, A. Rodriguez, M. Conery, M. Liu, Y. Ho, Y. Kim, D. A. Heise, L. Guare, V. A. Panickan, H. Garcon, F. Linares, L. Costa, I. Goethert, R. Tipton, J. Honerlaw, L. Davies, S. Whitbourne, J. Cohen, D.C. Posner, R. Sangar, M. Murray, X. Wang, D.R. Dochtermann, P. Devineni, Y. Shi, T.N. Nandi, T.L. Assimes, C.A. Brunette, R.J. Carroll, R. Clifford, S. Duvall, J. Gelernter, A. Hung, S.K. Iyengar, J. Joseph, R. Kember, H. Kranzler, D. Levey, S. Luoh, V.C. Merritt, C. Overstreet, J.D. Deak, S.F.A. Grant, R. Polimanti, P. Roussos, Y.V. Sun, S. Venkatesh, G. Voloudakis, A. Justice, E. Begoli, R. Ramoni, G. Tourassi, S. Pyarajan, P.S. Tsao, C.J. O'Donnell, S. Muralidhar, J. Moser, J.P. Casas, A. G. Bick, W. Zhou, T. Cai, B. F. Voight, K. Cho, M.J. Gaziano, R.K. Madduri, S.M. Damrauer,

K.P. Liao. Diversity and Scale: Genetic Architecture of 2,068 Traits in the VA Million Veteran Program. medRxiv 2023.06.28.23291975; doi: https://doi.org/10.1101/2023.06.28.23291975

14. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., & Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics, 28(19), 2540-2542.
15. Kaasschieter, E. F. Preconditioned conjugate gradients for solving singular systems. J. Comput. Appl. Math. 24, 265–275 (1988).
16. Cook, Shane. Chapter 9 - Optimizing Your Application, In Applications of GPU Computing Series, CUDA Programming. 2013, Pages 305-440, ISBN 9780124159334, https://doi.org/10.1016/B978-0-12-415933-4.00009-0
17. Baji, T. Evolution of the GPU Device widely used in AI and Massive Parallel Processing, 2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM), Kobe, Japan, 2018, pp. 7-9, doi: 10.1109/EDTM.2018.8421507.
18. N. V. Sunitha, K. Raju and N. N. Chiplunkar, Performance improvement of CUDA applications by reducing CPU-GPU data transfer overhead, 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2017, pp. 211-215, doi: 10.1109/ICICCT.2017.7975190.
19. J. Honerlaw, Y. Ho, F Fontin, J. Gosian, M. Maripuri, M. Murray, R. Sangar, A, Galloway, A.J. Zimolzak, S.B. Whitbourne, J.P. Casas, R.B. Ramoni, D.R. Gagnon, T. Cai, K.P. Liao, J.M. Gaziano, S. Muralidhar, K. Cho. Framework of the Centralized Interactive Phenomics Resource (CIPHER) standard for electronic health data-based phenomics knowledgebase, Journal of the American Medical Informatics Association, Volume 30, Issue 5, May 2023, Pages 958–964, https://doi.org/10.1093/jamia/ocad030
20. H. Fang, Q. Hui, J. Lynch, J. Honerlaw, T. L. Assimes, J. Huang, M. Vujkovic, S. M. Damrauer, S. Pyarajan, J. M. Gaziano, S. L. DuVall, C. J. O'Donnell, K. Cho, K. Chang, P. W.F. Wilson, P. S. Tsao, R. Ramoni, J. Breeling, G. Huang, S. Muralidhar, J. Moser, S. B. Whitbourne, J. V. Brewer, J. Concato, S. Warren, D. P. Argyres, B. Stephens, M. T. Brophy, D. E. Humphries, N. Do, S. Shayan, X. T. Nguyen, E. Hauser, Y. Sun, H. Zhao, R. McArdle, L. Dellitalia, J. Harley, J. Whittle, J. Beckham, J. Wells, S. Gutierrez, G. Gibson, L. Kaminsky, G. Villareal, S. Kinlay, J. Xu, M. Hamner, K. Sue Haddock, S. Bhushan, P. Iruvanti, M. Godschalk, Z. Ballas, M. Buford, S. Mastorides, J. Klein, N. Ratcliffe, H. Florez, A. Swann, M. Murdoch, P. Sriram, S. S. Yeh, R. Washburn, D. Jhala, S. Aguayo, D. Cohen, S. Sharma, J. Callaghan, K. A. Oursler, M. Whooley, S. Ahuja, A. Gutierrez, R. Schifman, J. Greco, M. Rauchman, R. Servatius, M. Oehlert, A. Wallbom, R. Fernando, T. Morgan, T. Stapley, S. Sherman, G. Anderson, E. Sonel, E. Boyko, L. Meyer, S. Gupta, J. Fayad, A. Hung, J. Lichy, R. Hurley, B. Robey, R. Striker, H. Tang. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies, The American Journal of Human Genetics, Volume 105, Issue 4 (2019).

21. H. Hunter-Zinck, Y. Shi, M. Li, B. R. Gorman, S.-G. Ji, N. Sun, T. Webster, A. Liem, P. Hsieh, P. Devineni, P. Karnam, X. Gong, L. Radhakrishnan, J. Schmidt, T. L. Assimes, J. Huang, C. Pan, D. Humphries, M. Brophy, J. Moser, S. Muralidhar, G. D. Huang, R. Przygodzki, J. Concato, J. M. Gaziano, J. Gelernter, C. J. O'Donnell, E. R. Hauser, H. Zhao, T. J. O'Leary, VA Million Veteran Program, P. S. Tsao, S. Pyarajan, Genotyping Array Design and Data Quality Control in the Million Veteran Program. Am. J. Hum. Genet. 106, 535–548 (2020).
22. X.-M. T. Nguyen, S. B. Whitbourne, Y. Li, R. M. Quaden, R. J. Song, H.-N. A. Nguyen, K. Harrington, L. Djousse, J. V. V. Brewer, J. Deen, S. Muralidhar, R. B. Ramoni, K. Cho, J. P. Casas, P. S. Tsao, J. M. Gaziano, the VA Million Veteran Program, Data Resource Profile: Self-reported data in the Million Veteran Program: survey development and insights from the first 850736 participants. Int. J. Epidemiol. 52, e1–e17 (2023).
23. J. M. Gaziano, J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, P. Guarino, M. Aslan, D. Anderson, R. LaFleur, T. Hammond, K. Schaa, J. Moser, G. Huang, S. Muralidhar, R. Przygodzki, T. J. O'Leary, Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. 70, 214–223 (2016).
24. 1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68-74. doi: 10.1038/nature15393. PMID: 26432245; PMCID: PMC4750478.
25. Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, Board on Health Sciences Policy, Committee on Population, Health and Medicine Division, Division of Behavioral and Social Sciences and Education, National Academies of Sciences, Engineering, and Medicine, *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* (National Academies Press, Washington, D.C., 2023; https://www.nap.edu/catalog/26902).
26. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203-209. doi: 10.1038/s41586-018-0579-z. Epub 2018 Oct 10. PMID: 30305743; PMCID: PMC6786975.
27. James Reinders. 2007. Intel threading building blocks (First. ed.). O'Reilly & Associates, Inc., USA.
28. H. E. Daniels "Saddlepoint Approximations in Statistics," The Annals of Mathematical Statistics, Ann. Math. Statist. 25(4), 631-650, (December 1954).
29. Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80(1), 27–38. DOI: 10.1093/biomet/80.1.27
30. Nielsen, Frank (2016). "2. Introduction to MPI: The MessagePassing Interface". Introduction to HPC with MPI for Data Science. Springer. pp. 195–211. ISBN 978-3-319-21903-5.

31. https://github.com/clMathLibraries/clBLAS
32. Chandra R, Dagum L, Kohr D, Menon R, Maydan D, McDonald J. Parallel programming in OpenMP. Morgan kaufmann; 2001.
33. https://github.com/RBigData/tasktools
34. Konrad J. Karczewski, Rahul Gupta, Masahiro Kanai, Wenhan Lu, Kristin Tsuo, YingWang, Raymond K. Walters, Patrick Turley, Shawneequa Callier, Nikolas Baya, Duncan S. Palmer, Jacqueline I. Goldstein, Gopal Sarma, Matthew Solomonson, Nathan Cheng, Sam Bryant, Claire Churchhouse, Caroline M. Cusick, Timothy Poterba, JohnCompitello, Daniel King, Wei Zhou, Cotton Seed, Hilary K. Finucane, Mark J. Daly, Benjamin M. Neale, Elizabeth G. Atkinson, Alicia R. Martin Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. medRxiv 2024.03.13.24303864; doi: https://doi.org/10.1101/2024.03.13.24303864
35. https://pan.ukbb.broadinstitute.org/
36. Kolata, G. (2023, November 15). V.A. Recruits Millionth Veteran for its Genetic Research Database. The New York Times. https://www.nytimes.com/2023/11/15/health/million-veterans-database-va.html
37. Zhou, W., Bi, W., Zhao, Z. et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. Nat Genet 54, 1466–1469 (2022). https://doi.org/10.1038/s41588-022-01178-w

Acknowledgments

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. Last but not least, 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.

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 grant R01GM138597 (AV); National Institute Health grant T32 AA028259 (JDD); National Library of Medicine Grant 5R01LM010685 (RJC); National Human Genome Research Institute grant K99HG012222 (WZ); National Institute of Arthritis and Musculoskeletal and Skin Diseases grant P30AR072577 (KPL); National Institute of Diabetes and Digestive and Kidney Diseases

grant DK126194 (BFV); National Institute of Health grants NIR01AG067025, K08MH122911 (GV); National Institute of Health grants BX004189, R01AG065582, R01AG067025 (PR); Office of Research and Development, Veterans Health Administration award I01CX001849-01 (JG); Office of Research and Development, Veterans Health Administration awards BX004821, CX001737, BX005831 (YSV); Veterans Health Administration awards IK2-CX001780 (SMD).

Author contributions

Conceptualization: AAR, YK, TNN, KK, RB, JEH, MC, ML, KM, JH, DS, EB, GT, SM, KC, MJG, BFV, SD, KPL, WZ, AV, RKM; **Methodology**: AAR, YK, TNN, KK, RB, JEH, MC, ML, KM, JH, DS, BFV, SD, KPL, WZ, AV, RKM; **Investigation**: AAR, YK, TNN, KK, RB, JEH, MC, ML, KM, JH, DS, MC, SM, BFV, KC, MJG, SD, KPL, WZ, AV,RKM; **Visualization**: AAR, YK, TNN, KK, JEH, MC, SM, BFV, KC, MJG, SD, KPL, WZ, AV, RKM**; Funding acquisition:** EB, GT, PN, SM, KC, MJG, SD, KPL, AV, RKM; **Project administration:** AAR, JEH, MJG, SD, KPL, WZ, AV, RKM; **Supervision:** EB, GT, SM, PN, AV, MJG, SD, KPL, RKM; **Writing – original draft:** AAR, YK, TNN, KK, RB, JEH, AV, WZ, RKM; **Writing – review & editing:** AAR, YK, TNN, KK, RB, JEH, MC, ML, SM, BFV, KC, MJG, SD, KPL, WZ, AV, RKM.

Data and materials availability

The optimized SAIGE-GPU software can be accessed at the GitHub page https://exascale-genomics.github.io/SAIGE-GPU.

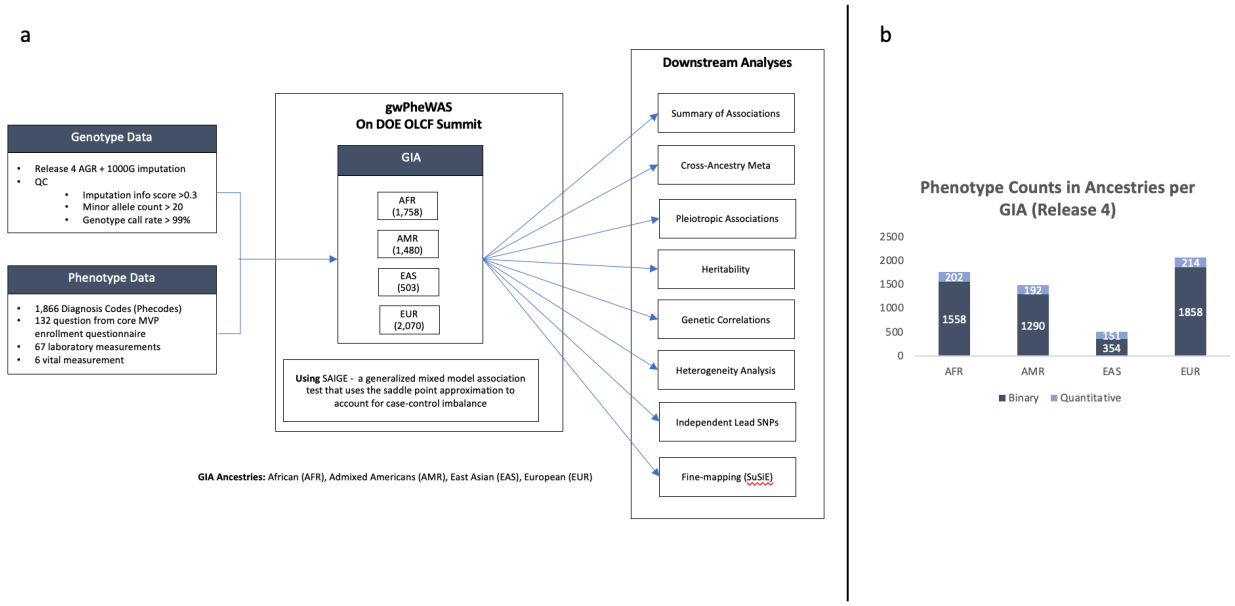

Fig. 1 Overview of genomic analysis in multiple population groups. a) Schematic representation illustrating the diverse set of GIA population groups. The analysis covers a deep catalog of traits extracted from electronic health records, clinical laboratory tests, vital signs, and survey responses. b) Chart categorizing traits into binary or quantitative types across different population groups. The height of each bar corresponds to the number of traits in each category, providing an overview of the trait composition for subsequent genomic analyses.

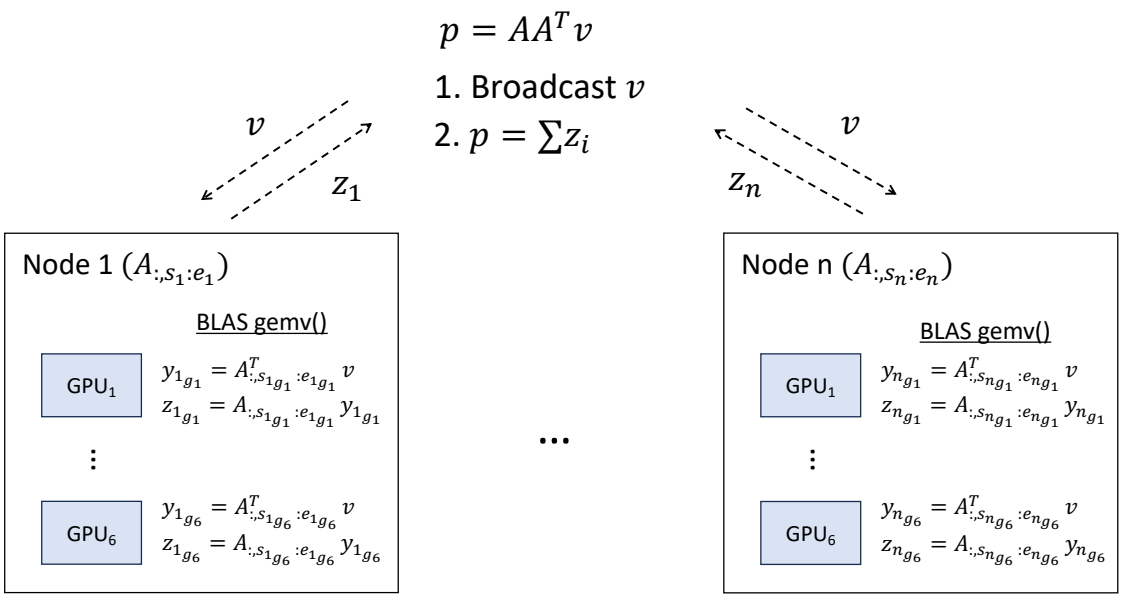

Fig. 2 Distributed BLAS gemv(), matrix-vector multiplication, using GPUs on the cluster. The columns of matrix A are distributed and preloaded on GPUs, with node i having columns with indices from s_i to e_i, and these columns are distributed on GPUs on that node. To compute $p = AA^T v$, we first broadcast v to GPUs, and each node computes a partial solution on GPUs. These partial solutions are aggregated to compute a solution p.

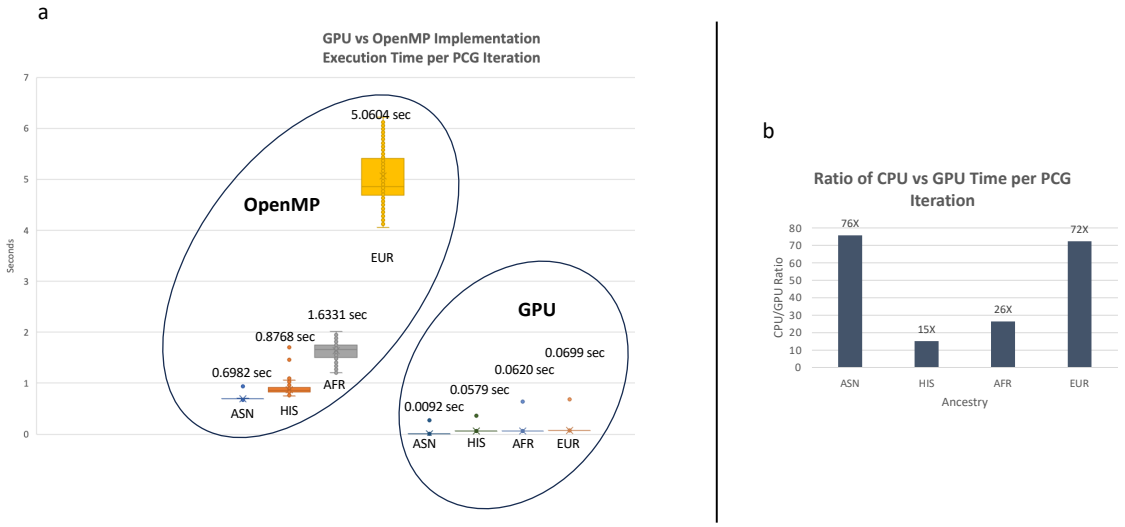

Fig. 3 Comparative Performance of GPU and CPU Implementations in SAIGE Step one - This figure compares the execution time for each iteration of matrix operations in SAIGE Step one for the European population group. a) Demonstration of the time required for a single PCG iteration on a GPU, showcasing the efficient parallelization within the GPU. b) Contrast with the OpenMP implementation on CPUs, emphasizing the significant speed improvement achieved with GPU acceleration. As the genotype matrix size increases, the advantage of using the GPU version becomes more pronounced, as highlighted by the diminishing execution time on the GPU compared to the CPU.

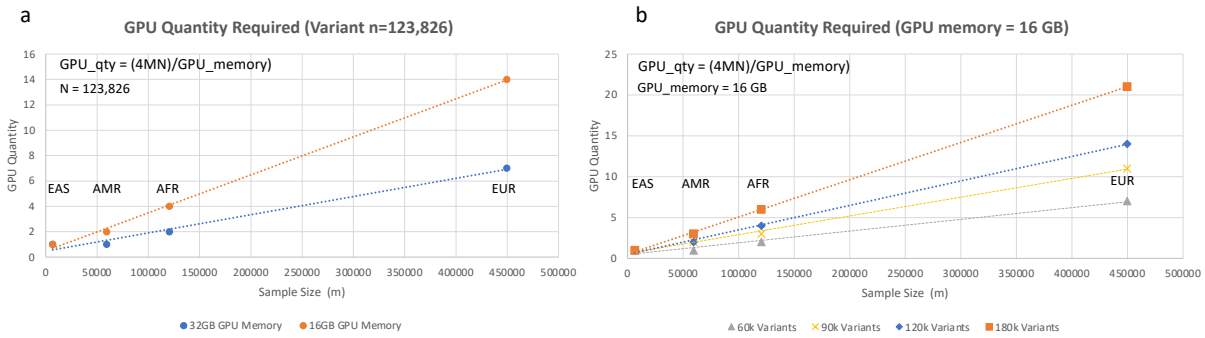

Fig. 4 GPU Node Requirements and Memory Impact – GPU node requirement highlight the linear relationship between genotype matrix size and the required number of nodes, offering insights into efficient GPU utilization. The GPU node requirement factored in the GPU memory, the byte size of a single precision floating-point number, and the conversion between bytes and gigabytes. A) Impact of changing the memory available in the GPU. B) Impact of changing number of genotype variants in the input matrix and fixing the GPU memory to 16 gigabytes per GPU, emphasizing considerations for diverse biobank cohorts and computational environments.

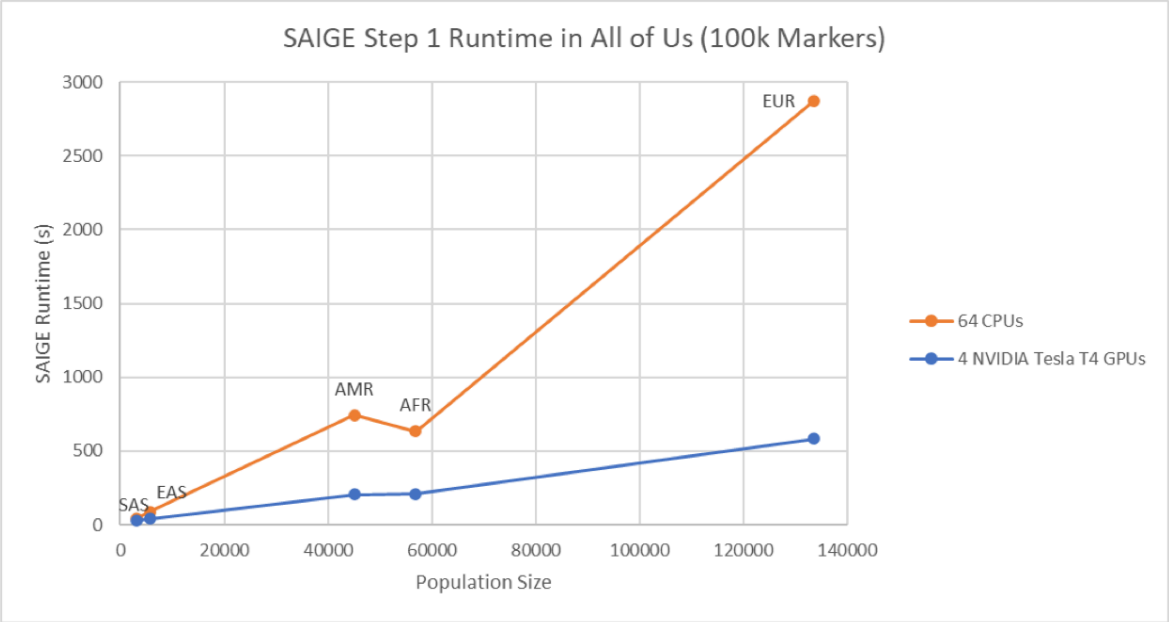

Fig. 5 SAIGE step one run time for All of Us data. The figure shows the time comparison of running SAIGE step one for the T2D phenotype on the Google Cloud Platform for the 5 population groups (EUR, AFR, AMR, EAS, SAS). The analysis was executed on 4 NVIDIA T4 GPUs for the SAIGE-GPU version and a 64-CPU VM for the SAIGE-CPU version.

Population Group	Participants (Release 4)
AFR	121,177
AMR	59,048
EAS	6,702
EUR	449,042

Table 1 Participant quantity in each grouping method per population group. Data was made available on OLCF Summit HPC to perform a GWAS analysis for all traits analysis and all population groups.

706

Population Group	Trait Quantity	Step One CPU hours for all traits (Projected)		Step One GPU hours for all traits (Production)
		Native SAIGE	SAIGE-OpenMP	SAIGE-GPU
AFR	1,760	322,768	78,266	1,336
AMR	1,482	72,284	60,295	411
EAS	505	27,162	12,253	116
EUR	2,072	1,371,960	330,372	12,420
Total	5,819	1,794,714	481,186	14,283

707

708 **Table 2** Projection times to complete GWAS for all traits (5,819) using SAIGE step one using
709 the different implementations of SAIGE: Native, OpenMP and GPU versions on CPU and GPU
710 environments.

711

Population Group	Subjects	Step one for Varicose Veins (hours)		
		Native SAIGE	SAIGE-OpenMP	SAIGE-GPU
AFR	121,725	5.10	1.06	0.38
AMR	51,124	1.50	0.97	0.28
EAS	8,003	0.97	0.58	0.23
EUR	458,307	25.75	4.10	1.50

Table 3 Execution times for SAIGE step one on Varicose Veins (ICD-9 code 454.1) using 3 versions of the SAIGE algorithm on the different OLCF infrastructures. CPU environment contained 32-core nodes, while the GPU nodes contain 42-cores and GPUs with 32 GB of RAM.

Step two for all traits (hours)						
Population Group	Trait Quantity	CPU Environment		GPU Environment		Fold-Change
		Hours for a Single Trait	Projection for all traits	Single Trait	Production run for all traits	
AFR	1,760	446	784,960	254	397,428	1.98
AMR	1,482	228	337,896	146	214,353	1.58
EAS	505	59	29,795	50	22,724	1.31
EUR	2,072	1,209	2,505,048	359	1,397,606	1.79
Total	5,819	1,942	3,657,699	809	2,032,111	1.80

Table 4 Execution time for SAIGE step two on Varicose Veins (PheCode 454.1) using 2 versions of the SAIGE algorithm on a CPU and GPU environment.

	All of Us		UK Biobank	
Category	European+	African American+	European+	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

Table 5 Cost and time execution comparison using All of Us and UK Biobank data on Google Cloud Platform for SAIGE-GPU version vs the native SAIGE version.