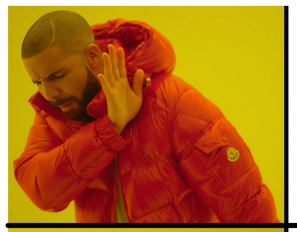
Genotype Matrix Multiplications Go Brrr

Sparse Linear Algebra Accelerates Genotype Representation Graph Computation at Biobank Scale

Yifan Li*, Qingyao Sun*, Drew DeHaas, Max Xiaohang Zhao, Adam R. Boyko, Shaila A. Musharoff, April Xinzhu Wei, Giulia Guidi

Correspondence: gg434@cornell.edu

Background



Whole-genome sequencing creates massive datasets: All of Us (AoU) CDRv8 srWGS♯ has 415K individuals and 870M variants, taking about **2.94 TiB in `.vcf.gz`**.



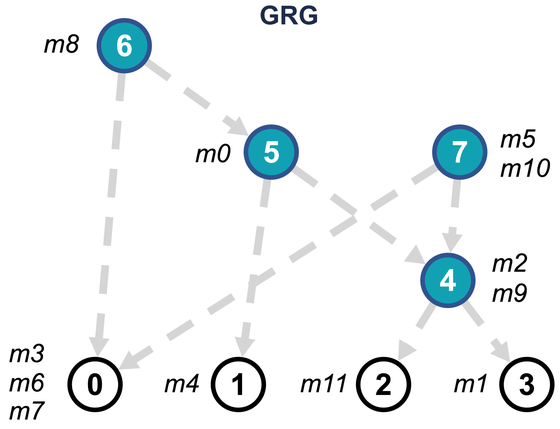
Genotype Representation Graph (GRG) [1] **losslessly** compresses shared haplotype structure: the same dataset takes **106 GiB in `.grg`**, with direct genotype matmul support.

Genotype Matrix G

Mutation (m4, m9) × Haplotype

0	0	0	1	0	1	1	1	1	0	1	0	…
1	0	0	0	1	0	0	0	1	0	0	0	…
1	0	1	0	0	1	0	0	1	1	1	1	…
1	1	1	0	0	1	0	0	1	1	1	0	…

GRG

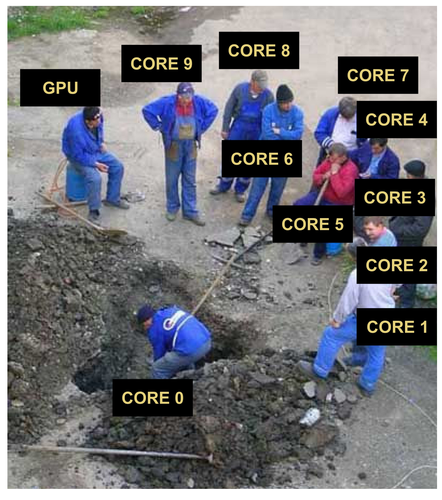


The same phased genotype data in tabular format and in GRG.

Motivation

GRG greatly reduces both data movement and computation for biobanks, but GRGL matmul is implemented using **single-threaded graph traversal** within each chromosome.

How to leverage CPU multi-threading or GPU acceleration? Is there a **hardware-agnostic** approach that works on all accelerators?



Acknowledgement

We thank All of Us participants for their contributions and the NIH All of Us Research Program for providing the data. This material is based on work supported by the National Science Foundation under Award No. IIS-2435801. The broader study used computational resources at NERSC and on NCSA's DeltaAI, the latter through the ACCESS program.

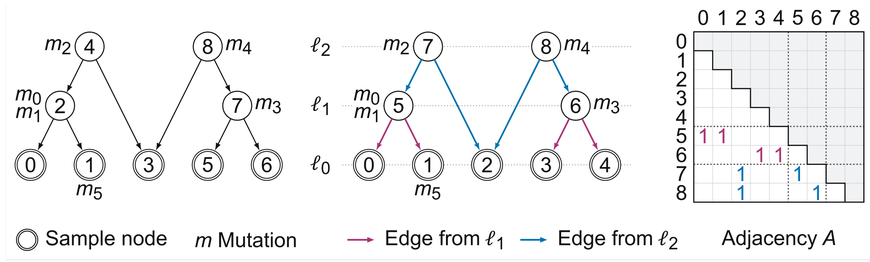
Reference

[1] DeHaas, D., Pan, Z. & Wei, X. ***Enabling efficient analysis of biobank-scale data with genotype representation graphs****. Nat Comput Sci 5, 112–124 (2025).*

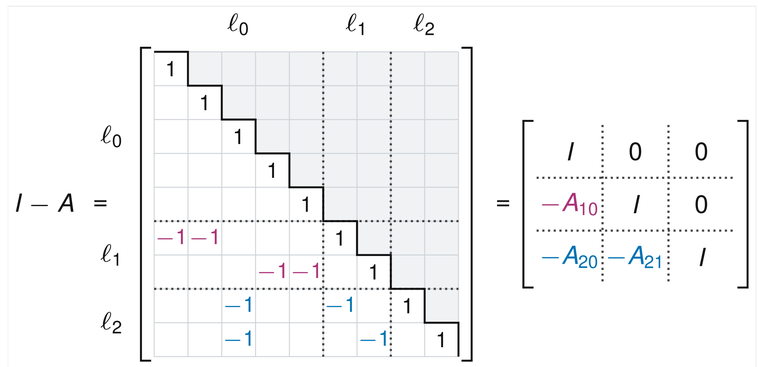
[2] DeHaas, D., Adonizio, C., Pan, Z. & Wei, X. ***General, orders-of-magnitude faster whole-genome analysis with genotype representation graphs****. bioRxiv (2026).*

Mikado

Under a topological ordering, every DAG has a **triangular** adjacency matrix A. We make it **block triangular** by relabeling the nodes in the level-set order.



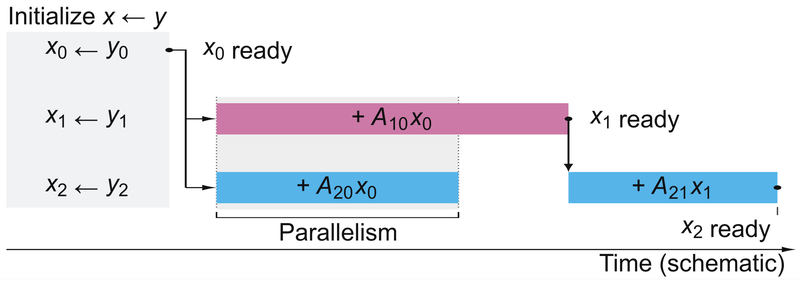
To compute Gv (resp. $G^\intercal w$), we ask "what haplotypes can each mutation reach?" (resp. "what mutations can reach each haplotype?"). Thus, the genotype matrix G can be obtained from the **reachability matrix**: $T = (I - A)^{-1}$.



Denote the maps between nodes and haplotypes (resp. mutations) by S (resp. M), and genotype matmul becomes a **sparse triangular solve (SpTRSV)** in block form.

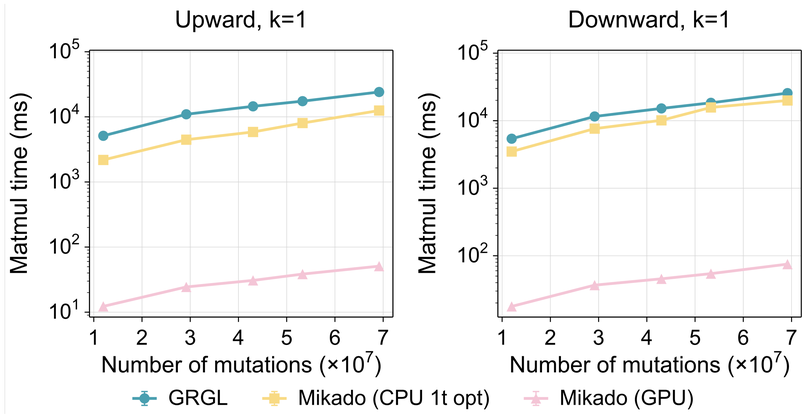
$$Gv = S\times\left(T^\intercal\times(M^\intercal v)\right), \qquad G^\intercal w = M\times\left(T\times(S^\intercal w)\right)$$

We perform block forward/backward substitution by pipelining vendor-optimized **sparse matrix-vector multiplication (SpMV)** kernels.



Results

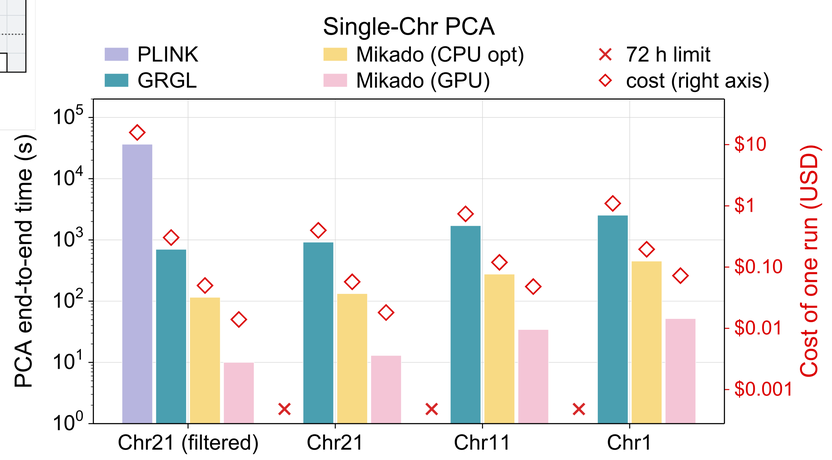
Mikado (GPU) on A100 80GB achieves geomean speedup **452×** ($G^\intercal w$; left) and **326×** (Gv; right) across chromosomes 1, 6, 11, 16, and 21 over single-threaded GRGL [1].†



Application: PCA

For PCA on Chr 21 biallelic SNPs (top 10 components), **PLINK2, GRGL, and Mikado (GPU) took 10.1 h, 700.8 s, and 10.0 s, resp.**, with speedup

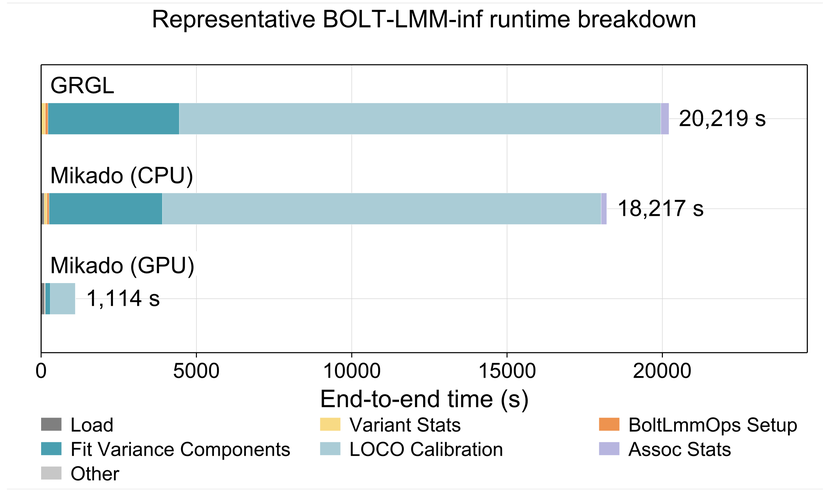
$$\frac{\text{PLINK2}}{\text{Mikado (GPU)}} = \frac{\text{PLINK2}}{\text{GRGL}} \times \frac{\text{GRGL}}{\text{Mikado (GPU)}}$$
$$= 52.0 \times 70.2 \approx 3{,}650.$$



No LD pruning; [2] motivates full-variant LOCO PCA.

Application: mixed-model association

We use BOLT-LMM-inf to run GWAS on standing height after filtering the GRGs to 221K individuals (EUR) and 45M variants (MAF $\geq 5\times10^{-5}$, HWE p $\geq 10^{-12}$).‡ Mikado (GPU) on A100 80 GB is **18.7×** faster than GRGL at **1.6×** the hourly cost, resulting in **11.7×** cost-effectiveness.



Conclusion

Our reformulation of GRG traversal enables the use of vendor-optimized SpMV kernels across Intel/AMD CPUs and NVIDIA/AMD GPUs.

Now, what would *you* build with Mikado?

bioRχiv THE PREPRINT SERVER FOR BIOLOGY

♯ We used the srWGS statistical phasing callset from the All of Us Research Program Controlled Tier Dataset CDRv8, available to authorized users on the Researcher Workbench.
† All GRG-based timings use prebuilt `.grg` or `.grg_spmv` inputs.
‡ BOLT-LMM-inf inputs occupy 25 GiB in vbyte-encoded `.grg` and 47 GiB in `.grg_spmv` (CSR with fixed-width integer indices). The analysis used 208K individuals with height measurements.