Scaling population genomics to a million: the Genomic Variant Store as a cloud-native variant data management platform

Aaron Hatcher, Miguel Covarrubias, Nalini Ganapati, Sofia Labrecque, Lee Lichtenstein
SLABREQ@BROADINSTITUTE.ORG

HIGHLIGHTS

Biobank-scale, population-representative calling – 535,000+ genomes, >1.3B variant sites jointly called for *All of Us* CDR v9, the largest open WGS callset in the world. < 4 cents/sample per GVCF.
Sparse storage that scales linearly – By decoupling storage from analysis format, GVS scales linearly with sample count instead of superlinearly. Callsets are extracted on demand into analysis formats.
Validated accuracy – GATK VETS matches or exceeds legacy VQSR accuracy on SNP/Indels. VETS scales past 535K with no downsampling required.

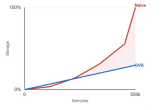
WHAT IS JOINT CALLING AND WHY IS IT HARD?

Researchers are generating petabytes of genomic data to understand human biology, identify the causes of diseases, and develop new treatments. Common analyses in genomics, such as association studies, require accurate SNP/Indel variant calls on large cohorts. Joint calling improves variant quality by analyzing genomes together across a cohort and consolidating storage. The joint-called variants can then be exported in formats for downstream analysis.

When we started building the Genomic Variant Store (GVS), performing joint calling on "just" 15,000 whole genomes was costly and required manual effort from engineers. Now, we have joint-called **>535,000 whole genomes**, and expect to be able to reach millions of samples. The challenges we faced in smaller-scale joint calling persisted in large-scale joint calling. Before we dive into the solution, let's break down the challenges:

THE CHALLENGES: VARIANT STORAGE AND ANALYSIS AT BIOBANK SCALE

~1B variants × >500k participants = 5×10^{14} cells, each with confidence scores & metadata



Explosive Storage Needs
Naive storage of the full genotype matrix grows quadratically with sample count.



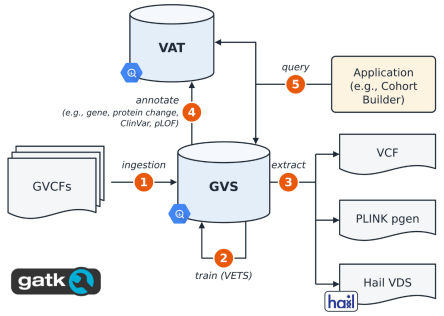
The N+k Problem
When k new samples arrive, re-running joint calling on all N+k samples wastes prior compute.

THE SOLUTION: DECOUPLE STORAGE FORMAT FROM ANALYSIS FORMAT

Traditional approaches to joint calling used VCF — a text file — as both storage and analysis format. VCF scales superlinearly and couples storage and analysis data formats. GVS decouples them by using an optimized internal storage model with analysis-ready formats that are extractable on demand.

INTRODUCING THE GENOMIC VARIANT STORE (GVS)

The Genomic Variant Store is a cloud-native solution for biobank-scale joint calling, built on Google BigQuery (a serverless columnar warehouse). GVS ingests gVCF files, applies quality filtering, and extracts joint-called SNP/Indel callsets in multiple analysis formats.



(1) SNP/Indel GVCFs are ingested into a BigQuery database with a specialized schema for variants at scale.
(2) A filtering model (VETS) is trained to increase precision
(3) Samples and variants can be extracted into an analysis-ready format for consumption by researchers. GVS supports filtering, by both genomic regions and samples, to reduce costs.
(4) The Variant Annotation Table (VAT) can be rendered. This table includes standard functional annotations, such as gene, protein change, and variant classification. It also includes frequency information from the instance of GVS and gnomAD, separately.
(5) Applications (e.g., graphic user interface (GUI)-based) can query the GVS table for variant information. This table can be joined with the VAT for more powerful queries. Operations and troubleshooting staff can use the GVS tables to quickly query the variants to locate errors (not shown).

Incremental Ingestion & VETS Quality Filtering

New batches of samples are added without reprocessing prior data, solving the N+k problem. After ingestion, GATK VETS (Variant Extract-Train-Score) updates the variant filter set — an isolation-forest detection model using allele-specific annotations. VETS is VQSR's (Variant Quality Score Recalibration) faster, more scalable successor.

(CONT.) INTRODUCING GVS

Multi-Format Extraction

GVS extracts to analysis-ready formats on demand: VCF, Hail VDS, and PLINK pgen — by genomic region and/or specified samples. Adopting a new format requires only a new extraction job, not callset reprocessing. This is the property that makes GVS a platform rather than a pipeline.

Fast, Auditable Answers at Biobank Scale

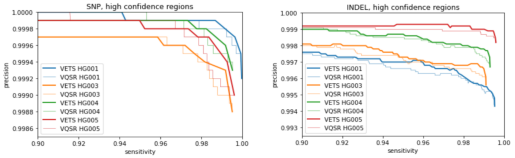
GVS's queryable structure enables rapid answers to ad-hoc operational questions — why a variant is being filtered, participant counts with specific variants, or the origin of a data error. For example, when a program investigator reported that a common variant was not in a release, a SQL query spanning 414,000 samples and 1.5 trillion alternate genotypes returned the answer in one second at a cost of USD $0.01.

Example Queries Over 414,830 *All of Us* CDR v8 Samples x 1.3B Variants

Query	Time	Joined with VAT
All participants with chr11:6868417-C-A (a known cilantro-taste variant)	6 sec	No
All participants with a stop-gained variant in BRCA1	56 sec	Yes
All participants with a ClinVar P/LP variant in any of the fourteen MODY genes with GQ>20 and total AD>20	67 sec	Yes

GVS+GATK VARIANT EXTRACT-TRAIN-SCORE (VETS) TOOLCHAIN

Previous variant calling workflows used GATK VQSR (Variant Quality Score Recalibration) which is over 10 years old and did not scale well to either small (<10 samples) callsets or large (>100,000). We are now using a new machine learning model to train our pipeline on what is a variant and what is likely noise, allowing us to achieve better sensitivity and precision on both Indels and SNPs. Benchmarking with VETS used less memory and ran 13X faster than VQSR, and real world usage has shown it to easily scale to analyses of at least ~535k WGS with no need for downsampling.



Precision and sensitivity of a GVS joint callset filtered w GATK VETS toolchain vs. w GATK VQSR. The data was 3,000 human whole genomes from AnVIL.

TRY IT OUT

We're looking for beta testers for the GVS on Terra and we'd love to hear from you! If you're interested, let us know by filling out our form at broad.io/variantstore to tell us about your work.

DOWNSTREAM DATA PRODUCTS ENABLED BY GVS

Because GVS is a single, authoritative, incrementally-updated source of truth, multiple data products and services derive their inputs from it without reprocessing.

Cohort Builder

The *All of Us* Researcher Workbench Cohort Builder enables researchers to filter participants by variant presence, ancestry, and clinical phenotype. Cohort Builder uses cached results from GVS and the *All of Us* instance of the VAT functional variant annotations to answer queries like: "Give me all participant IDs with a cancer diagnosis and the following variants..."

***All of Us* + AnVIL Imputation Service**

An imputation service with a reference panel of 515,579 whole genomes (414,830 *All of Us* CDR v8 + 100,749 CCDG/AnVIL), phased with Beagle 5.2 across 665M high-quality sites — with 49% non-European ancestry. Available at allofus-anvil-imputation.broadinstitute.org (FedRAMP/NIST 800-53 Rev 5 Moderate). Learn more at the lightning talk: Wednesday, October 21, 1:30 - 1:33pm EDT or poster: Friday, October 23, 2:30 - 4:30pm EDT

SUMMARY

The Genomic Variant Store (GVS) is a cloud-native platform built on Google BigQuery that decouples variant storage from analysis, making biobank-scale joint calling practical for the first time.

- **Sparse storage that scales linearly** — storing alt alleles and ref blocks only, instead of the full genotype matrix, avoids the superlinear growth that makes VCF-based storage unworkable at biobank scale.
- **GATK VETS delivers validated, scalable filtering** — matching or exceeding legacy VQSR precision and sensitivity, while scaling past 535,000 genomes with no downsampling, at a scale VQSR could not reach.
- **Production-proven** at 535,000+ genomes and 1.3 billion variant sites for *All of Us* CDR v9, the largest open whole-genome callset in the world.

ACKNOWLEDGEMENTS

Thank you to everyone who has developed the GVS including: Bec Asch, Eric Banks, Genevieve Brandt, Laura Gauthier, Susan Jacobs, Daniel King, Samuel Lee, Kevin Lydon, Benjamin Neale, Konrad Karczewski, Kristian Cibulskis, Andrea Haessly, Akum Shergill, Jose Soto, David Roazen, Megan Shand, Morgan Aster, Tim Poterba, Louis Bergelson, Jonn Smith, and more.
Work was made possible by NIH awards: OT2OD035404, OT2OD03821, OT2OD002750, and U24HG010262