Genomic Variant Store

Product Sheet

BROAD INSTITUTE

HIGHLIGHTS

Large-scale joint calling

The GVS scales up to 400,000 genomes* and 100,000 exomes with better performance than existing solutions.

Powered by GATK and WDL

The GVS combines GATK and WDL to create a practical solution for the average researcher.

Advancing frontiers with All of Us (AoU)

The GVS will scale to joint call 1 million human genomes and is the backend data solution for cohort building in the AoU Researcher Workbench.

PRODUCT OVERVIEW

Current joint calling solutions, such as those recommended in the GATK Best Practices, can scale to ~2,000 human genome samples. Larger cohorts increase the sensitivity and specificity of each variant call, but as cohort sizes increase, so do the computational and financial costs of performing joint calling. As a result, variant discovery with large, powerful cohorts ranges from impractical to impossible for most researchers.

Developed by the Data Sciences Platform at the road Institute of MIT and Harvard, the Genomic Variant Store (GVS) combines the power of BigQuery and GATK Best Practices for variant discovery to support joint calling on more than 400,000 human genomes*. In addition to improving the scale, cost and reliability of joint calling, the GVS will also enable researchers to create subsets of large joint callsets and search within callsets based on the properties of variants observed in the data.

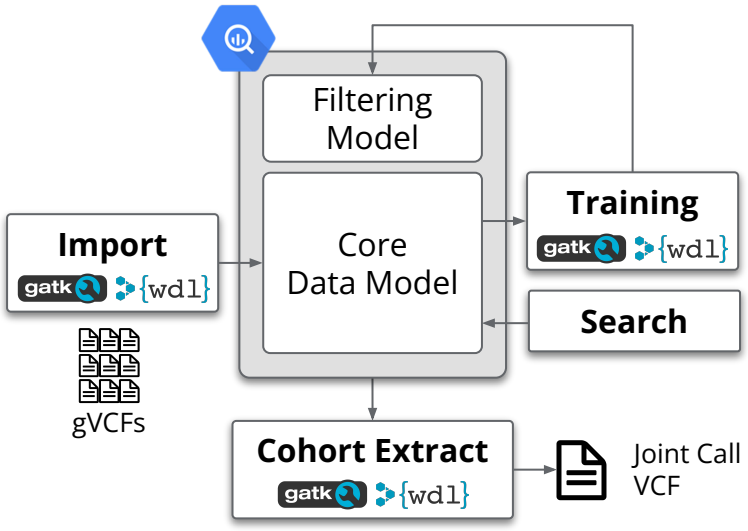

Figure 1. Overview of the GVS. Briefly, gVCF files are imported into BigQuery, which encompasses the core data model and filtering model. Alternate alleles are identified and the Variant Extract Train Score (VETS) filtering model is trained. The VETS model is applied and callsets are extracted to VCF files as output. GATK and WDL tools are used where indicated.

Creating human whole genome callsets is faster, more reliable, and **10x cheaper** than before

Joint calling that scales to **400,000** human genomes and **100,000** human exomes*

Variant filtering driven by machine learning with high **precision** and **sensitivity**

WHAT'S NEXT?

We plan to continue developing the GVS in a number of ways, including:

- Continue scaling towards 1 million genomes!
- Enabling cohort building and subsetting within callsets

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.

*For the beta release, we are prioritizing testing callsets of up to 25,000 human genomes and 100,000 human exomes or blended genome exomes. If you don't have a callset in mind or would like to create one larger than this, we'd still love to hear about your interest in the GVS to inform our development roadmap.