Cloud Deployment Quick Guide Series

Best Practices and Instance Recommendations

Getting Started with Genomics Analytics on Google Cloud Platform

"Take advantage of unique GATK optimizations and Intel® Xeon® Scalable processors' support for Intel® AVX acceleration to speed time to insight."

Overview

Increasingly, the scientific community is looking to harness cloud-based computing for genomics analytics. This is especially true for projects with unpredictable compute utilization and when datasets are already located in the public cloud. The Genome Analysis Toolkit (GATK) provides industry-standard tools for identifying single nucleotide polymorphisms (SNPs) and indels in germline DNA and RNA-seq data.

This guide provides high-level recommendations for using the Google Cloud Platform (GCP) and GATK, specifically for a Germline Variant Calling pipeline, a form of secondary analysis.

Workload Considerations

Choosing the right cloud infrastructure is important when performing whole genome variant calling in the cloud. This workload is commonly deployed on a high performance computing (HPC) infrastructure. Keep in mind the following considerations when selecting instance and storage types:

- **Datasets:** Whole Genome Sequences (WGS) are hundreds of GBs in size and Whole Exome Sequences (WES) are tens of GBs in size.
- **Pipelines:** Genomic datasets are processed using workflows, often called "pipelines," consisting of multiple heterogeneous tools. The Broad Institute publishes their GATK Best Practices Pipelines in Dockstore here, and provides recommendations for running them on GCP here.
- **Workload characteristics:**
  - High I/O at the beginning and end of the pipeline.
  - Computations are highly parallel (low MPI usage).
  - Computations are compute-intensive for a fraction of the heterogeneous workflow.
  - Throughput (WGS/node/day) scales with CPU core count; single sample processing time is dependent on CPU clock speed.

Deployment Architecture Recommendations

The Broad Institute provides workflow scripts written in the Workflow Definition Language (WDL), for each of their GATK Best Practices Pipelines, along with open source access to GATK and Cromwell. Whether the infrastructure backend is a local batch scheduler or the GCP compute engine, workflows are submitted via Cromwell (see Figure 1). This provides flexibility for users to utilize either local compute nodes or GCP based on their preference.

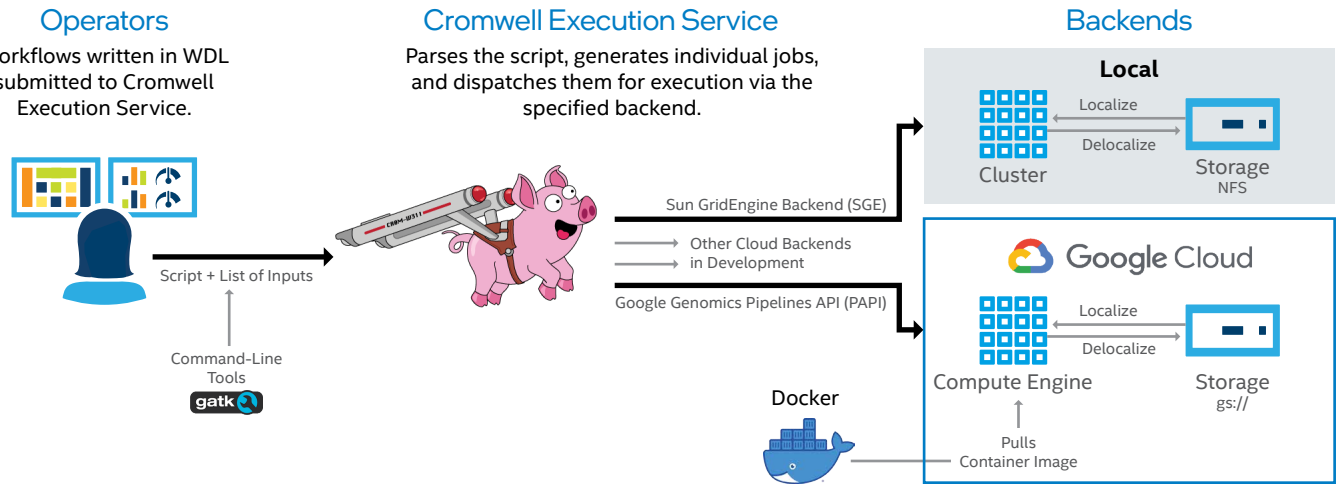

Figure 1. Defining the workflow
Source: Broad Institute; https://f1000research.com/slides/6-1381

Cloud Instance Recommendations

Intel and the Broad Institute collaborated to create the Genomics Kernel Library (GKL), a collection of Intel-optimized libraries used throughout the genomics workflow. GKL includes compression and decompression libraries, as well as Intel® Advanced Vector Extensions 512 implementations of common genomics tools. GKL is distributed open source with the GATK and enables faster runtimes and more samples processed per day. Performance also benefits from the use of fast local storage, including Intel® 3D NAND NVMe SSDs.

Within GCP, the instances using Intel AVX-512 are N1, N2, and C2 (see Table 1).

Table 1. Machine Type Comparison

Machine types	Memory (per vCPU)	vCPUs	Processors
General-purpose (N2)	0.5–8 GB	2–80	• 2nd gen Intel® Xeon® Scalable processors
General-purpose (N1)	0.95–6.5 GB	1–96	• 6th gen Intel® Core™ processor • 5th gen Intel® Core™ processor • 4th gen Intel® Core™ processor
Compute-optimized (C2)	4 GB	4–60	• 2nd gen Intel Xeon Scalable processors

Source: https://cloud.google.com/compute/docs/machine-types

Getting Started with GATK Best Practice Pipelines

Users have two options for Best Practice Pipeline workflow automation on GCP:

- **Terra:** A data platform built on GCP by the Broad Institute and Verily. Includes a Cromwell server that runs workflows on GCP via the Google Genomics Pipelines API. The Broad Institute publishes Terra Workspaces for each Best Practices Pipeline that include example data, official workflows, and examples for running them. Instance types can be customized in the workflow configurations.
- **Google Genomics Pipelines API (PAPI):** Creates a VM, runs containerized workflows, and destroys the VM. It also automates the distribution of pipelines. Instance types can be specified when launching VMs, similar to any other GCP workload.

An introduction to both options is available in the ebook, Genomics in the Cloud: Using Docker, GATK, and WDL in Terra.

For details on how to use the GATK, visit the **Broad Institute site.**