intel.

Solution Snapshot

Intel® Select Solutions for Genomics Analytics

intel SELECT SOLUTIONS

The Challenge

There is currently a race to find a cure for diseases such as cancer and COVID-19.

Setting up and scaling a genomics analytics cluster poses many challenges that can create bottlenecks for genomics researchers. For example, experts predict that in a few years, the field of genomics may generate up to 40 exabytes of data per year.[1] Processing this amount of data, with the scalability required in a fast-changing field, is a substantial compute problem.

Today, Researchers need:

- Ability to run a broad range of applications with diverse system requirements
- Ways to address increasing datasets and complexity
- Faster time to results
- Reduced total cost of ownership (TCO)

Use Cases:

| Genomic Sequencing | Population Scaling | Cancer/Rare Disease Research | Agricultural Crop Research | Endangered Species Protection | COVID-19 Pandemic Research |

Genomics Analytics
Overview + Benefits

Intel® Select Solutions for Genomics Analytics add value to the alignment and variant calling step of genetic analyses. Intel has optimized performance in these steps by tuning Broad Institute's Genomics Analytics Toolkit (GATK) to take advantage of Intel® architecture.

Working together, Intel and the Broad Institute are pursuing three goals:

1. Develop an **optimized software stack** to analyze genomes faster and at greater scale.
2. Enable and qualify **turnkey configuration, setup, and deployment** of infrastructure to run genomics analysis.
3. **Take advantage of key technologies** to allow for ever-increasing datasets to be analyzed to deliver scientific and medical breakthroughs.
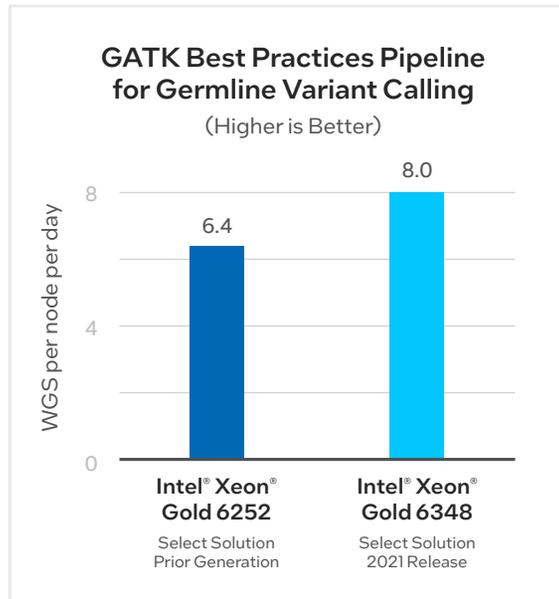
Intel® Select Solutions for Genomics Analytics are designed to scale from small (four-node minimum) to very large, clustered supercomputers. The customized systems can be quickly and dynamically configured to meet specific needs for organizations to scale as they grow their workloads. In addition, end-to-end optimized hardware and software solutions for genomic analysis.

- Optimized HPC cluster that can support multiple genomics pipelines as well as other HPC workloads
- Uses Intel® technology: 3rd Gen Intel® Xeon® processors and Intel® Solid State Drives (Intel® SSDs)
- WDL (workflow description language) scripts to allow users to replicate Genome Analysis Toolkit best practice pipelines and create their own pipelines quickly and easily

Proof Points

GATK Best Practices Pipeline for Germline Variant Calling
(Higher is Better)

WGS per node per day

	6.4	8.0

Intel® Xeon® Gold 6252
Select Solution Prior Generation

Intel® Xeon® Gold 6348
Select Solution 2021 Release

Pipeline for Germline Variant Calling benefits from the increased number of memory channels and support for 3200 MHz DRAM, along with higher core count (compared to 2nd Gen Intel® Xeon® Scalable processors).

Large-capacity Intel® SSD Data Center Family drives—either NAND-based or Intel® Optane™ SSDs—also contribute to the solution's high performance.

Why Intel for Genomic Analytics?

Intel® technologies integrated into Intel® Xeon® Scalable processors deliver further performance and reliability gains:

- **Intel® Advanced Vector Extensions 512 (Intel® AVX-512):** Boosts performance for the most demanding computational workloads, with up to double the number of FLOPS per clock cycle, compared to previous-generation Intel® processors.

- **Intel® Cluster Checker:** Inspects more than 100 characteristics related to cluster health. Intel® Cluster Checker examines the system at node and cluster levels, enabling all components to work together to deliver optimal performance.

- **Cluster Management Software Stack:** Provides a software stack required to deploy and manage Linux HPC clusters. The stack includes provisioning and development tools, resource management, I/O clients, and scientific libraries. Resource management tools such as Bright Cluster Manager, Warewulf, and xCAT support the software stack.

- **Intel® Cluster Checker Runtimes:** Supplies key software runtime elements required on each cluster to ensure optimal performance paths for applications. Intel® runtime performance libraries, such as Intel® Math Kernel Library and Intel® MPI Library, are available in the Intel® oneAPI toolkit. This toolkit helps deliver excellent performance that is optimized for clusters based on Intel® architecture. They can also simplify the development and deployment of data-centric workloads across CPUs, GPUs, FPGAs, and other accelerators.

Want More Information?

What are Intel® Select Solutions for HPC?

Intel® Select Solutions for HPC removes the guesswork of selecting the right set of high-performance computing products, hardware configurations, and software.

Resources

Intel® Select Solutions for HPC

Read more on the Intel® Select Solutions Genomics Analytics Solution Brief

Intel® Select Solutions for Genomics Analytics

3rd Generation Intel® Xeon® Scalable processors

Broad Institute of MIT and Harvard

Genome Analytics Toolkit