Solution Brief

Intel® Select Solutions | Version 2
High Performance Computing
April 2021

# Intel Select Solutions for Genomics Analytics

## Access performance, scale, and ease of deployment for genomics insight and discovery.

Advancements in genomics are opening new doors for understanding human diseases, and they are increasingly informing innovative precision treatment plans. Discoveries are dependent on processing, storing, and analyzing a growing amount of genomic sequencing data. In 2015, worldwide sequencing storage capacity approached a petabyte per year, and it continues to double every seven months.[1,2] At this rate, genomics sequencing will generate hundreds of petabytes per year in the next five years, and could require nearly a zettabyte of storage per year by 2025.[1,2]

The Broad Institute of MIT and Harvard (broadinstitute.org) is one of the world's largest producers of human genomic data, creating about 24 TB of new data per day. Currently, Broad Institute manages more than 50 PB of data.

Researchers require tools to analyze these enormous volumes of data in a timely manner to gain insights into disease and possible treatments. They need tools like the Genome Analysis Toolkit (GATK), a set of leading software methods created by the Broad Institute and trusted by the majority of genomics centers worldwide.

Broad Institute has released GATK 4.2 as its latest major version, under an open source license for all users, including for commercial purposes. An open source license makes GATK available to a wider audience of scientists and researchers and helps accelerate and advance genomics analytics worldwide.

> "Our goal is to reduce the challenges that researchers face to generate ever-more-meaningful insights from ever-larger sets of genomics data."
>
> — **Geraldine Van der Auwera**
> Director of Outreach and Communications, Broad Institute

**The Intel-Broad Center for Genomic Data Engineering brings together science and technology to optimize analytics and workflows.**
Photo credit: Erik Jacobs Photography, courtesy of Broad Institute

## Intel-Broad Center for Genomic Data Engineering

Intel and Broad Institute have collaborated on computing infrastructure and software optimization for years. In 2017, they launched a new effort—the Intel-Broad Center for Genomic Data Engineering is a five-year collaboration between the two organizations to simplify and accelerate genomics workflow execution using GATK, Burrow-Wheeler Aligner (BWA), Cromwell, Intel Genomics Kernel Library (Intel GKL), GenomicsDB, and other tools and techniques. Together, experts from Broad Institute and Intel build, optimize, and widely share tools and infrastructure to help scientists integrate and process genomic data. The result is a growing set of optimized best practices in hardware and software for genomics analytics on Intel architecture–based platforms that can be applied to research datasets stored in private data centers and that extend to private, public, and hybrid clouds.

With the massive growth of genomics data, the collaboration makes use of technology to enable genomics analytics at scale. It has already resulted in Intel Select Solutions for Genomics Analytics, a suite of optimized software, along with reference architectures for turnkey configuration, setup, and deployment to run genomics analysis that is qualified for GATK pipelines, Cromwell, and GenomicsDB.

## Intel Select Solutions for Genomics Analytics

The Intel-Broad Center for Genomic Data Engineering works to optimize GATK on Intel architecture and technologies and to define a reference architecture for genomics analytics.

The result is Intel Select Solutions for Genomics Analytics, developed by Intel and the Broad Institute and delivered by Intel solution providers. The solutions demonstrated a five-times overall performance improvement running GATK 4.0 compared to previous versions of the genomics software, and they reduce setup time for deploying an infrastructure to accelerate genomics workflows.[3] Performance gains include a 75 percent speedup for the BWA using Intel Solid State Drives (SSDs).[3] The validated performance and quality results have been certified by Broad Institute.

High-performance data analytics computing clusters and optimized workflows for genomics analytics are complicated hardware and software systems. Intel Select Solutions for Genomics Analytics are end-to-end optimized hardware and open source software configurations designed specifically to accelerate genomics analytics—both the deployment of systems and the software that runs on them—by providing verified stacks for setup and configuration of these complicated genomics pipelines.

Intel Select Solutions for Genomics Analytics are designed to scale from small to very large clustered supercomputers. The customized systems can quickly and dynamically be configured to meet specific needs. Organizations can scale as they grow their workloads. And Intel Select Solutions for Genomics Analytics include tools to discover, compose, and monitor resources with powerful, modern API-based software.

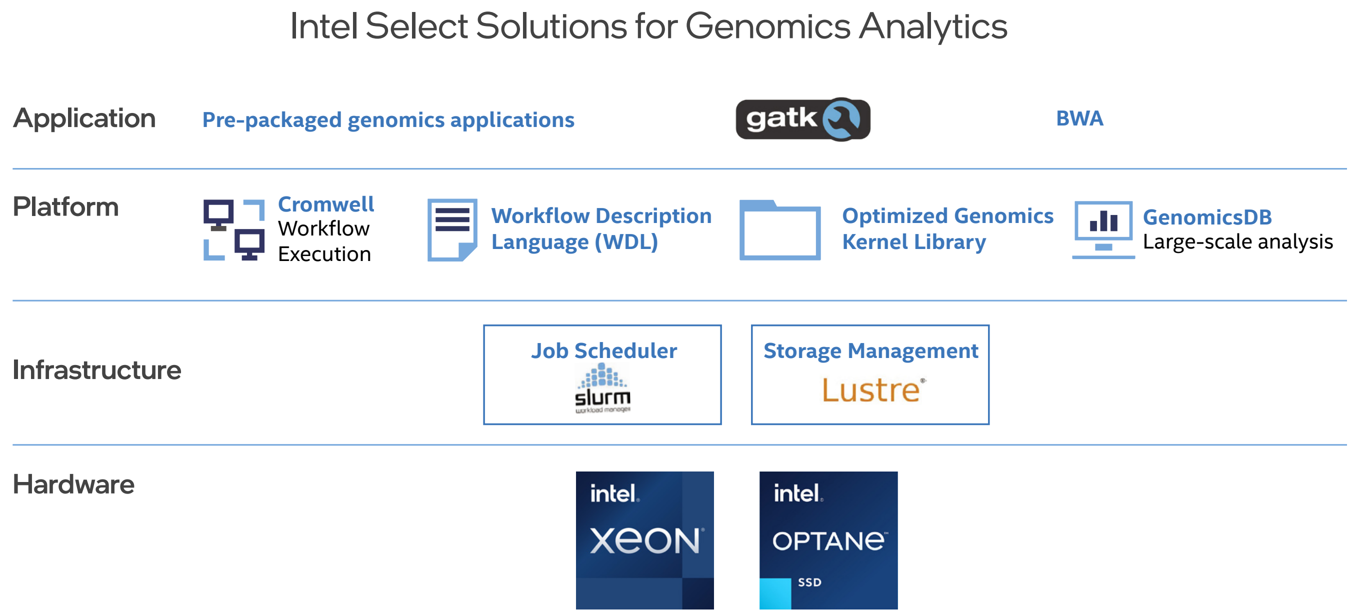
Intel Select Solutions for Genomics Analytics
Application
Pre-packaged genomics applications
gatk
BWA
Platform
Cromwell Workflow Execution
Workflow Description Language (WDL)
Optimized Genomics Kernel Library
GenomicsDB Large-scale analysis
Infrastructure
Job Scheduler
slurm
workload manager
Storage Management
Lustre
Hardware
intel XEON
intel OPTANE SSD

**Figure 1.** A high-level overview of the solution configuration

## What are Intel Select Solutions?

Intel Select Solutions are verified hardware and software stacks that are optimized for specific software workloads across compute, storage, and network. The solutions are developed from deep Intel experience with industry solution providers, in addition to extensive collaboration with the world's leading data center and service providers.

To qualify as an Intel Select Solution, a vendor must:

1. Follow the software and hardware stack requirements outlined by Intel
2. Replicate or exceed Intel's reference benchmark- performance threshold
3. Publish solution content to facilitate customer deployment

Solution providers can develop their own optimizations to add further value to their solutions.

## Intel Xeon Scalable processors

Intel Xeon Scalable processors:

- Offer high scalability for enterprise data centers
- Deliver performance gains for virtualized infrastructure compared to previous-generation processors
- Achieve exceptional resource utilization and agility
- Enable improved data and workload integrity and regulatory compliance for data center solutions

The family includes Intel Xeon Bronze processors, Intel Xeon Silver processors, Intel Xeon Gold processors, and Intel Xeon Platinum processors.

**Solution powered by:**

**Table 1.** The configuration for Intel Select Solutions for Genomics Analytics

| Ingredient | Intel Select Solutions for Genomics Analytics |
|---|---|
| 1 x Application Node | |
| Processor | 2 x Intel Xeon Gold 6252 processor (or higher), required |
| Memory | 12 x 16 GB DDR4 2,933 MHz 1DC (total capacity 192 GB or higher), required |
| Storage (boot) | 2 x 480 GB Intel SSD D3-S4510 or larger (mirrored OS), recommended |
| Data network | 1 x Intel Ethernet Connection X722 with Intel Ethernet Converged Network Connection X527-DA2/DA4 or Intel Ethernet Converged Network Adapter X710, 10 gigabit Ethernet (GbE) or higher, recommended |
| 4 x Compute Nodes | |
| Processor | 2 x Intel Xeon Gold 6252 processor (or higher), required |
| Memory | 12 x 32 GB DDR4 2,933 MHz 1DC (total capacity 384 GB or higher) per node, required |
| Storage (boot) | 2 x 480 GB Intel SSD D3-S4510 or larger (mirrored OS), recommended |
| Storage (capacity) | 1 x 1.6 TB Intel SSD DC P4610 (2.5-in PCIe 3.1 x4, 3D2, TLC) or larger, required |
| Data network | 1 x Intel Ethernet Connection X722 with Intel Ethernet Converged Network Connection X527-DA2/DA4 or Intel Ethernet Converged Network Adapter X710, 10 GbE or higher, recommended |
| Network Infrastructure | |
| Management Network | 1 x 10 Gbps 24x port switch or higher |
| Storage Infrastructure | |
| File System | Recommended, not required:<br>• Bandwidth—200 MB/s per client<br>• Capacity— 500 GB of capacity per genome to be processed and stored (120 TB per compute node allows for 30-day sample storage)<br>For systems larger than 4–8 nodes, a parallel file system (such as Lustre) is recommended. |
| Software | |
| **Required Software:**<br>• GATK, BWA, and GATK workflows optimized for Intel technologies<br>• Optimized Cromwell workflow<br>• Intel GKL with optimized routines for accelerating developer codes<br>• SJob scheduler (e.g., Slurm) for running clustered analytics jobs | **Optional Software:**<br>• Docker for running multiple jobs in isolated containers across a cluster<br>• Apache Spark for big data analytics processing<br>• Lustre, the open source parallel file system, for high-performance storage<br>• GenomicsDB, specializing in large-scale variant analysis |

## Software, firmware, and technology configuration

Intel Select Solutions for Genomics Analytics take advantage of the high-performance capabilities of Intel architecture, including Intel® Xeon® Scalable processors and Intel SSD Data Center Family drives. Table 1 shows hardware and software for the Intel Select Solutions for Genomics Analytics. To refer to a solution as an Intel Select Solution, a server vendor or data center solution provider must use these or better configurations. These solutions can be tailored with 2, 4, 16, 24, 36, or 48 of the specified compute devices and, when applicable, local and shared storage devices in order to meet the needs of individual environments.

## Technology selections for Intel Select Solutions for Genomics Analytics

In addition to the Intel hardware foundation used for Intel Select Solutions for Genomics, the following Intel technologies integrated in 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 floating point operations per second (FLOPS) per clock cycle, compared to previous-generation Intel processors.[4]
- **Intel Cluster Checker.** Inspects more than 100 characteristics related to cluster health. Intel Cluster Checker examines the system at both the node and cluster level, making sure all components work together to deliver optimal performance. It assesses firmware, kernel, storage, and network settings. It also conducts high-level tests of node and network performance using the Intel MPI Library benchmarks, STREAM, the High Performance LINPACK (HPL) benchmark, the High Performance Conjugate Gradients (HPCG) benchmark, and others. Intel Cluster Checker can be extended with custom tests, and its functionality can be embedded into other software.

> "For us, running GATK 4 on v1 of the Intel Select Solutions for Genomics Analytics delivered a 5x performance gain right away. We're working with Intel to make the GATK Best Practices pipelines run even faster, at even greater scale, and with easier deployment for genomic research worldwide."
>
> **— Geraldine Van der Auwera**
> Director of Outreach and Communications, Broad Institute

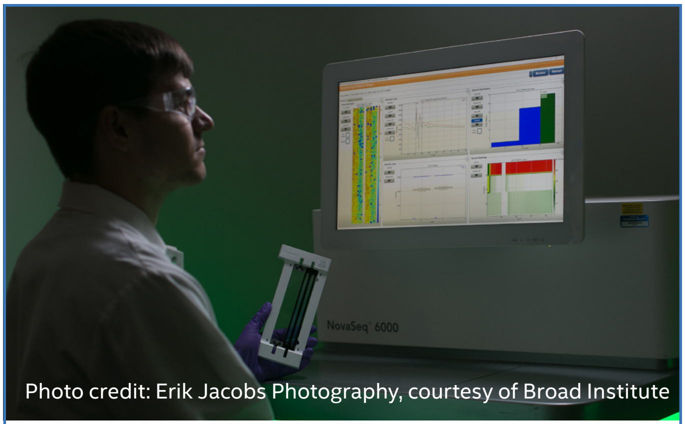

Photo credit: Erik Jacobs Photography, courtesy of Broad Institute

## Benefits of the Intel and Broad Institute collaboration

The work of Intel and Broad Institute offers many benefits to the genomics community and to the technologists and business managers that support it, including:

**Scientists who enjoy:**

- Support for optimized and efficient pipelines
- Optimized, turnkey solutions
- Prepackaged workflow description languages (WDLs) scripts
- Peer application support
- Low-touch IT support
- Access to more in-house genomics data
- Increased statistical power
- Open source software
- Flexible application architecture

**IT departments who need:**

- Ease of implementation
- Scalability
- Reduced setup time
- Open source software with no licensing costs
- Known reference architecture
- Vendor and solution support
- Optimal use of hardware versus workload (for example, prepackaged WDLs)

**Business owners who enjoy:**

- The ability to scale the solution to fit a budget
- Low price/watt
- Preconfigured solutions to reduce setup time and support costs
- Maximized value for in-house genomic data
- No license fees
- Open source application software
- Extendability to other applications

- **Intel Cluster Runtimes.** Supplies key software runtime elements that are required on each cluster to ensure optimal performance paths for applications. Intel runtime performance libraries, including Intel Math Kernel Library (Intel MKL) and Intel MPI Library, deliver excellent performance optimized for clusters based on Intel architecture.
- **Cluster Management Software Stack.** Provides a software stack required to deploy and manage Linux HPC clusters. The stack includes provisioning tools, resource management, I/O clients, development tools, and scientific libraries. Resource management tools such as Bright Cluster Manager, Warewulf, and xCAT support the software stack.

### Simplified code development with Intel Genomics Kernel Library (Intel GKL)

Intel GKL provides code used in genomics that is optimized for Intel architecture. Kernels include Intel AVX-512 implementations of Smith Waterman and PairHMM, two commonly used algorithms with GATK. The Intel GKL is distributed open source with—and called directly from—GATK. The library enables developers to focus on the function and operation of their code (instead of specific optimizations), while letting Intel GKL make use of the capabilities of Intel architecture.

### Scalability improvements with GenomicsDB

GenomicsDB is a unique variant store capable of supporting up to hundreds of thousands of genome variant data. It was first developed by Intel Labs and customized for Broad Institute's use cases.

GenomicsDB is packaged with GATK 4.2, and it helps significantly accelerate joint genotyping workflows. For example, without using GenomicsDB, Broad Institute took six weeks to generate a database from 2,300 whole genomes. With GenomicsDB, it was able to generate databases with five times more information in only two weeks.[5] That successfully enabled the Broad Institute–hosted Genomics Aggregation Database (gnomAD) project, which includes 15,000 whole genomes— one of the largest genomic data aggregations in the world.[5] In addition to being integrated into GATK 4.2, GenomicsDB is available open source through Omics Data Automation.

### Continuing development

There are large genomic databases around the world that can bring great benefits to worldwide research efforts. The ongoing work of the Intel-Broad Center for Genomic Data Engineering continues to develop Intel Select Solutions for Genomics Analytics to efficiently access those databases for analysis. In the future, incorporated technologies will provide the connectivity, performance, privacy, and security necessary for genomics in the cloud and shared environments.

### OEM partners—simplifying genomics analytics cluster deployment

The introduction of Intel Select Solutions for Genomics Analytics makes it easier to run genomics workloads. It also enables accelerated deployment of predictable clusters designed for genomics analytics. Thus, many integrators of high-performance systems have partnered with Intel and are offering design and deployment of solutions that will meet the needs of their customers in the genomics community.

## Access performance, scale, and ease of deployment for genomics analytics

The work of genomics science is critical to the understanding of disease and the creation of diagnostic tools and safe and effective therapies. Genomics data and analytics are quickly advancing as researchers use technology to build massive genomics data repositories and come to understand the power of that data. Broad Institute is one of the largest contributors of genomics data in the world, and its GATK software is the world's leading genome analysis tool for analytics and variant call research. The Intel-Broad Center for Genomic Data Engineering brings together science and technology to optimize genomics analytics codes and workflows and to define an optimized infrastructure—Intel Select Solutions for Genomics Analytics— to run those workloads. The results enable faster analysis and quicker times to deploy hardware solutions that are customized for genetics analysis. Several system integrators already offer services to install such systems that will continue to enable further discoveries through genetics.

For more on this and other Intel Select Solutions, visit: **intel.com/selectsolutions.**

## Learn more

The Intel-Broad Center for Genomic Data Engineering: **intel.com/broadinstitute**

"Big Data Genomics and Optimized Genomics Code": **intel.com/content/www/us/en/healthcare-it/solutions/genomicscode.html**

Intel Xeon Scalable processors: **intel.com/xeonscalable**

Deploy genomics workflows in the cloud: **oreilly.com/library/view/genomics-in-the/9781491975183/**

Where to buy: **intel.com/content/www/us/en/products/docs/select-solutions/where-to-buy**

Solution provided by:

[1] Stephens, Zachary D., et al. "Big Data: Astronomical or Genomical?" PLOS Biology. July 2015. doi.org/10.1371/journal.pbio.1002195.

[2] Robison, Reid J. "How Big Is the Human Genome?" Precision Medicine. January 2014. medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0.

[3] Intel. "Infrastructure for Deploying GATK Best Practices Pipeline." November 2016. intel.com/content/dam/www/public/us/en/documents/white-papers/deploying-gatk-best-practices-paper.pdf.

[4] Intel AVX-512 provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing Intel AVX-512 instructions may cause a) some parts to operate at less than the rated frequency, and b) some parts with Intel Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration. Learn more at intel.com/go/turbo.

[5] Geraldine Van der Auwera, Ph.D. Broad Institute. Bio-IT World. May 2017.

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Printed in USA 0421/JS/PRW/PDF ♻ Please Recycle 336731-007US