White Paper
Intel Genomics Solutions
Intel® Xeon® Processor E5 v3 Family

ACCELERATING TIME-TO-INSIGHT (TTI)

For Whole Genome Analysis Using Intel® Server Building Blocks

Seven Bridges genomics

"Seven Bridges' optimized pipeline implementations are a huge asset for our national projects. When we help countries analyze hundreds of thousands of genomes, every performance improvement counts."
— Deniz Kural, CEO, Seven Bridges Genomics

Abstract

Based on the Intel® Xeon® processor E5 v3 family, the Intel Genomics Cluster solution can significantly improve performance for genome analysis versus previous-generation hardware platforms. As described in this paper, performance tests conducted by Intel and Seven Bridges Genomics demonstrate up to 41 percent faster performance for whole genome analysis versus a typical solution based on previous-generation processors.[1]

With these performance gains, researchers and clinicians can potentially obtain results for a whole genome analysis almost five hours sooner. They can also use this 4-node system as a building block for constructing large, local clusters to support multiple high-volume, next-generation sequencers.

Opportunities and Challenges in Whole Genome Variant Calling

The transition toward next-generation, high-throughput genome sequencers is creating new opportunities for researchers and clinicians. Population-wide genome studies and profile-based clinical diagnostics are becoming more common and more cost-effective. At the same time, such high-volume and time-sensitive usage models put more pressure on bioinformatics pipelines to deliver meaningful results faster and more efficiently.

The Intel Genomics Cluster solution is designed to help organizations meet the demand for fast, high-volume genome analysis for next-generation sequencers. This four-node cluster building block is powered by the Intel Xeon processor E5 v3 family and Intel® Solid-State Drives (Intel® SSDs). It provides multiple, high-bandwidth connectivity options to enable high-speed data input and efficient performance when multiple building blocks are combined in large cluster configurations.

Intel worked closely with Seven Bridges Genomics' bioinformaticians to design the optimal genomics cluster building block for direct attachment to high-throughput, next-generation sequencers. Though most use cases will involve variant calling against a known genome, more complex analyses can be performed with this system. A single 4-node building block is powerful enough to perform a full transcriptome. As demands grow, additional building blocks can easily be added to a rack to support multiple next-generation sequencers operating simultaneously.

An Optimized Building Block for Genomics Clusters

The Intel Xeon processor E5 v3 family provides significant new performance capabilities versus previous generation processors. It provides more cores, threads, and cache and supports faster memory (DDR4 versus DDR3). This processor family includes Intel® Advanced Vector Extensions 2.0 (Intel® AVX2), which doubles the floating point operations that can be performed per second (Flops) versus first-generation Intel AVX. Intel AVX2 also doubles the width of vector integer instructions.

These increased processor capabilities are ideal for the complex computations of genome analysis. A single, 4-node Genomics Cluster based on the Intel® Xeon® processor E5-2695 family provides 112 processor cores and 224 threads. It also provides an optimal balance of high-speed memory and Intel SSD storage, which enables the processor cores to operate at or near peak capacity during complex calculations. Cooling and air flow are also optimized to support continuous cluster operation in demanding research and clinical environments.

Table of Contents

Verifying Performance for Whole Genome Analysis

To help customers quantify the potential benefits of the Intel Genomics Cluster solution, Intel and Seven Bridges Genomics ran a series of performance tests using the Seven Bridges Genomics software platform. Performance for a whole genome pipeline running on one node of the test cluster was compared with the performance of the same software platform running on a comparable server node based on the previous generation Intel® Xeon® processor E5 v2 family. Configuration details are shown in Table 1.

Seven Bridges Genomics provides a full software platform for analysis of NGS data. To test the Intel Xeon processor E5 family, Seven Bridges Genomics optimized a very common use case: whole genome variant calling, in which single nucleotide polymorphisms (SNPs) and insertions/deletions are identified by aligning read files with a known reference genome. Such use cases involve processing data from both single and paired-end reads. The complete pipeline is shown at the end of this paper.

The subset of the pipeline used for the performance tests includes four distinct computational phases (Figure 1):

- **Phase A:** Alignment, deduplication, and sorting of the raw data reads
- **Phase B:** Local realignment around Indels
- **Phase C:** Base quality score recalibration
- **Phase D:** Variant calling and variant quality score recalibration

As shown in Table 1, BWA* was used for the computations in phase A, GATK2.39Lite* was used for phases B through D, and the Rabix* executor was used to manage the workflow. This pipeline can accept FASTQ files as an input. The output is a single variant calling format (VCF) file per sample. This pipeline is configured to Broad Institute's best practices and is broadly applicable to clinical environments.

The results of the performance tests are shown in Table 2. The Intel Genomic Cluster solution based on the Intel Xeon processor E5-2695 v3 family completed a whole genome pipeline in just 429 minutes versus 726 minutes for the solution powered by the prior-generation Intel Xeon processor E5 v2 family.

Based on these results, researchers and clinicians can potentially complete a whole genome analysis almost five hours sooner using the newer system. They can also run more analyses in less time to support the demands of additional users and multiple next-generation sequencers.

The following sections provide more detailed information about the workloads and system behavior in each phase of the analysis.

Notes on software optimization

- The performance tests described in this paper were performed on the local cluster using Java* SE Runtime Environment 7, 1.7.0_76-b13, as the underlying Java virtual machine. This edition of Java is optimized for Intel AVX2. To achieve comparable performance results, the same or a later version of Java Platform should be used.
- Since performing these tests, Intel and the Broad Institute have performed comparable optimizations on a number of pipeline tools, including the tools used in these performance tests. Researchers and clinicians teams will potentially see even higher performance gains than documented in this paper when using these optimized pipeline tools. For more information, visit http://www.intel.com/content/www/us/en/healthcare-it/genomicscode.html?wapkw=genomics

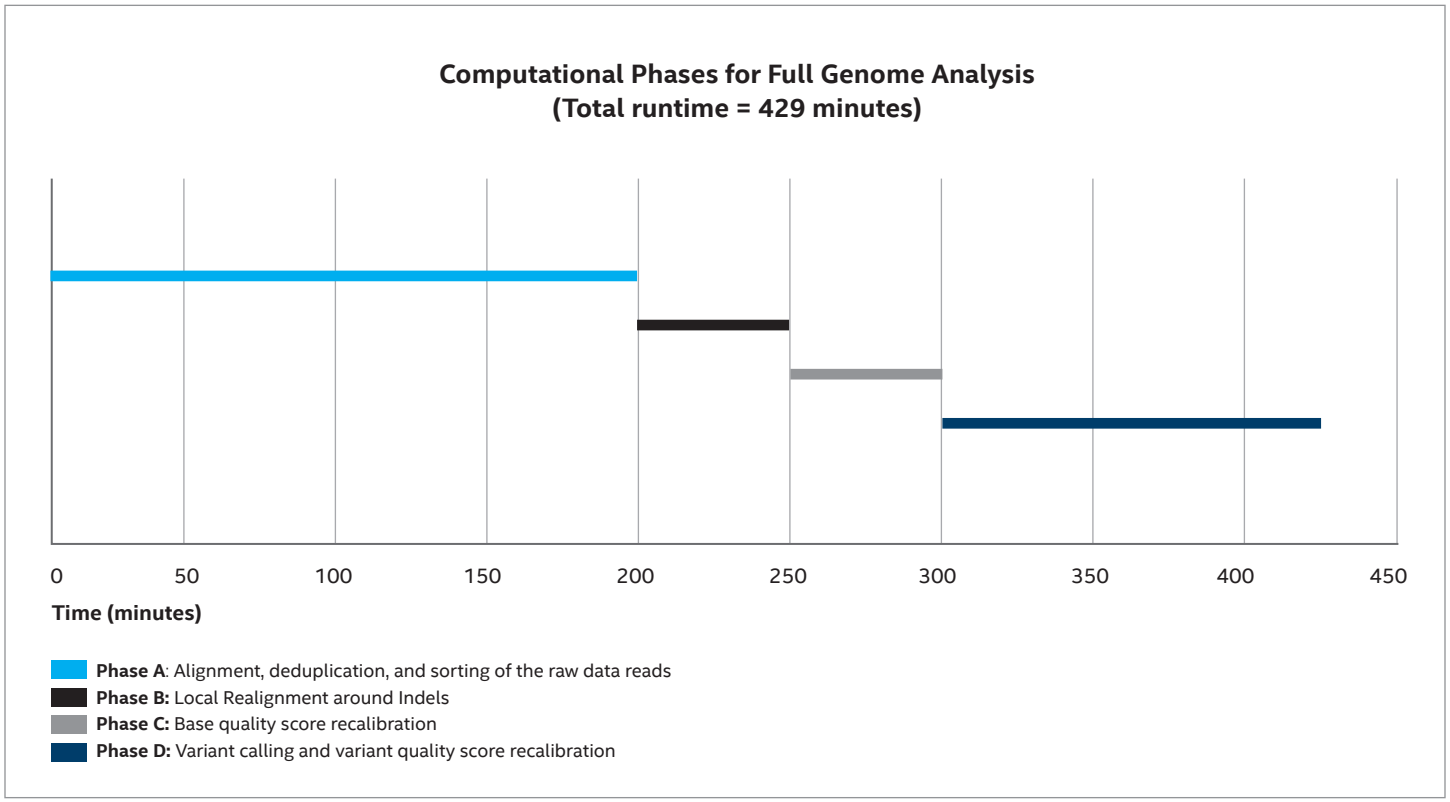

Figure 1. Computational phases for a whole genome analysis, plus their approximate runtimes using one node of the Intel Genomics Cluster building block based on the Intel® Xeon® processor E5 v3 family (see Table 2 for more precise runtimes).

Table 1. Test Configurations

	Baseline Cluster (1-node)	**Test Cluster** (1-node Intel Genomics Cluster)
Hardware (per node)		
	2 x Intel® Xeon® processor E5-2680 v2	**2 x Intel® Xeon® processor E5-2695 v3**
Memory	60 GB	256 GB DDR4 2133
Networking	10 GbE	10 GbE
Storage	2 x 320 SSD	Intel® SSD DC S3500 Series (300 GB) Intel® SSD DC S3700 Series (800 GB)
Software		
Operating System	Red Hat Enterprise Linux* 6.6.9	
Java Virtual Machine (JVM)	Java Platform SE 76	
Pipeline Applications	BWA* 0.7.11, GATK2.39Lite*, Rabix* scheduler v0.22	

Table 2. Performance Results for Whole Genome Analysis

Phase	**SW**	**Tool**	**Baseline Cluster** Intel® Xeon® processor E5-2680 v2 (runtime in minutes)	**PCSD Genomics Cluster** Intel® Xeon® processor E5-2695 v3 (runtime in minutes)	**Improvement**
A	BWA	Samblaster, Sambamba	421	216	49%
B	GATK 2.39 Lite	RealigerTargetCreator IndelRealigner	32 27	25 19	22% 30%
C		BaseRecalibator PrintReads	19 87	17 73	11% 16%
D		UnifedGenotyper VariantRecalibrator	99 39	66 13	33% 67%
		Total Time	**724**	**429**	**41%**

Phase A: Alignment, deduplication, and sorting of reads

During phase A, the raw data is aligned, deduplicated, and sorted for efficient comparison with the reference genome. In this case, alignment was performed using BWA 0.7.11, against the GRCh b37 human genome, and the workload was parallelized to execute simultaneously on 56 threads. The produced SAM file was fed directly to Samblaster (via unix pipe) for deduplication. The data was then piped to the Sambamba[2] View command for SAM to BAM conversion, and to Sambamba Sort for coordinate sorting, also using 56 threads.

The workload during this phase is compute-intensive, and benefits from the large number of cores and threads provided by the Intel Xeon processor E5 v3 family. It also benefits from Intel® Streaming SIMD Extensions (Intel® SSE), which enable a 3:1 ratio of integer and packed double-precision floating point calculations.

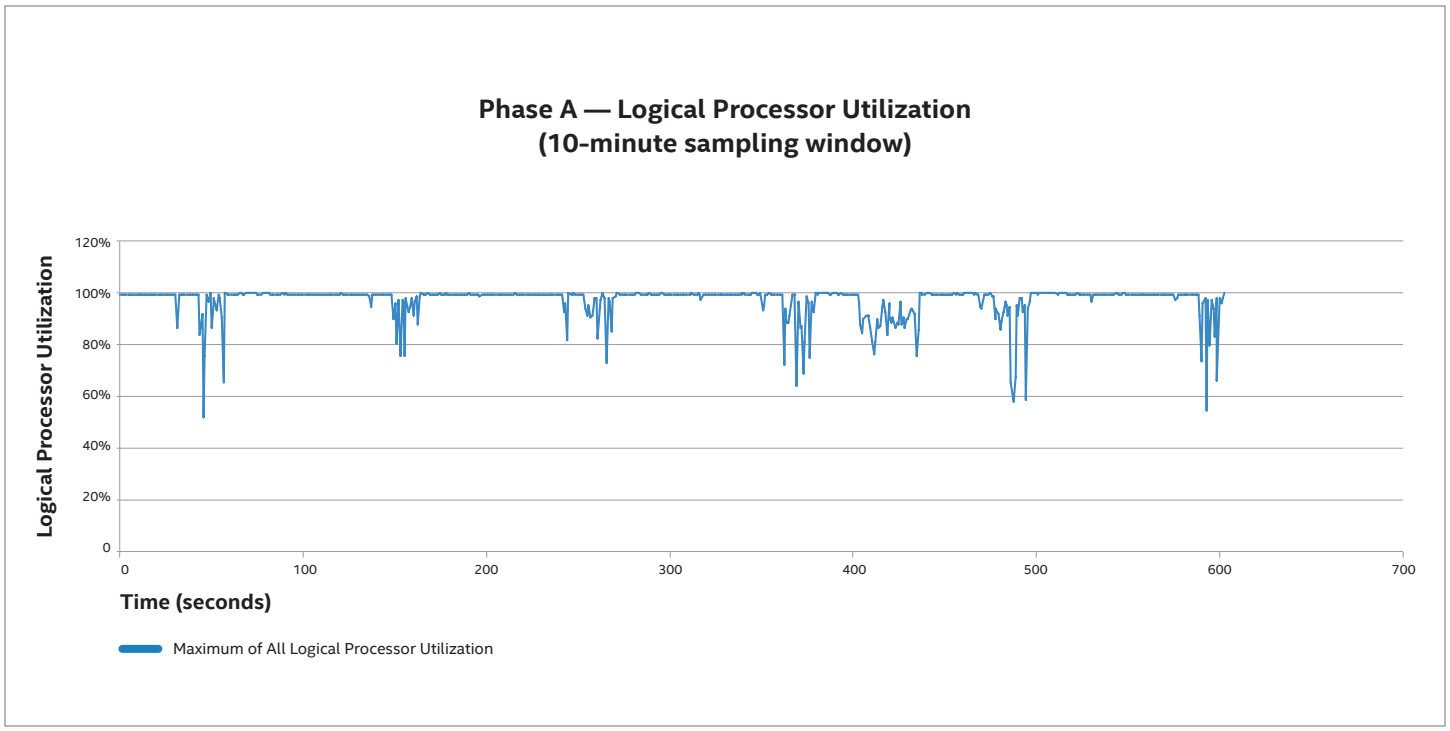

Figure 2. Logical processor utilization was at or near 100 percent throughout phase A, showing that the optimized workload and balanced hardware platform enabled exceptionally efficient use of available processing resources.

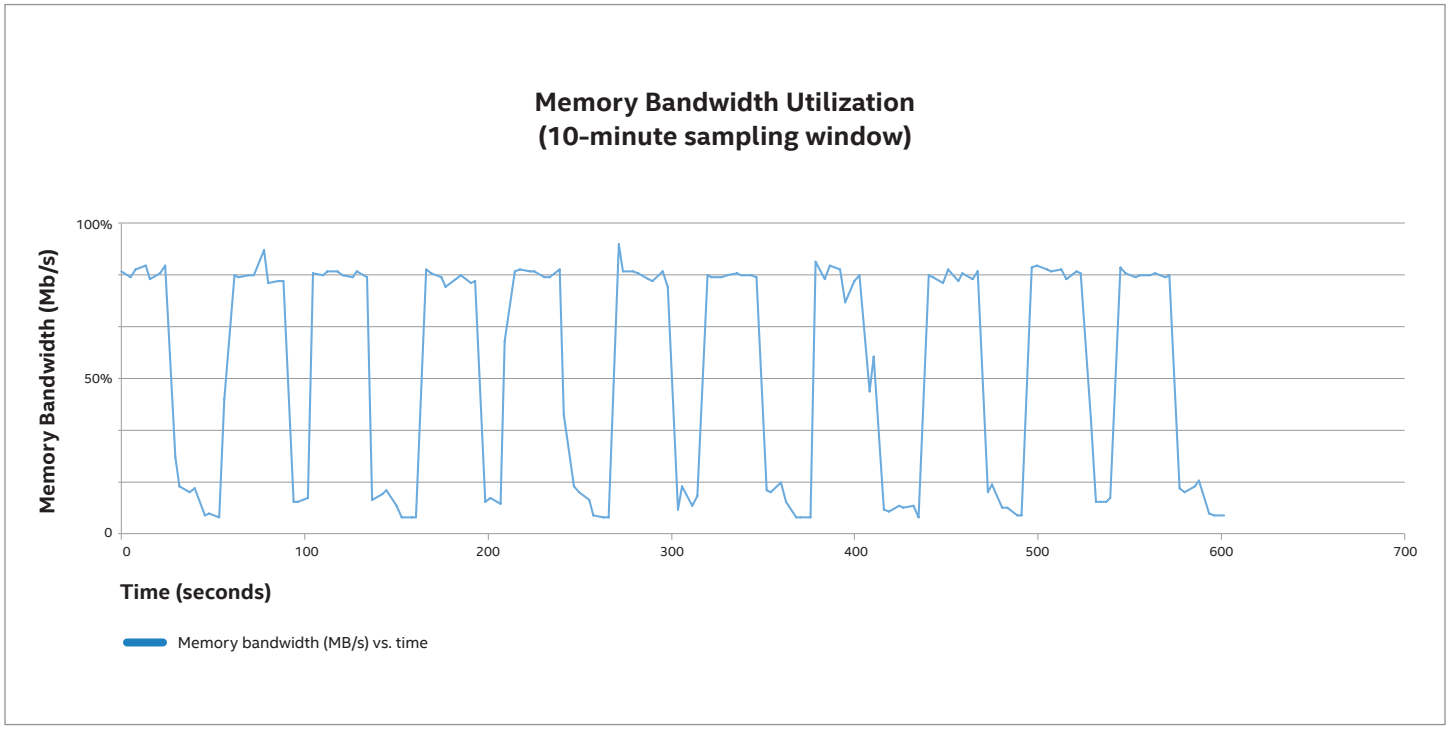

Figure 3. Memory bandwidth utilization peaked at 70-80 percent of maximum during Phase B, demonstrating efficient use of available resources. As shown by the relatively flat peaks, the tests also demonstrated good NUMA locality, another sign of efficient memory utilization.

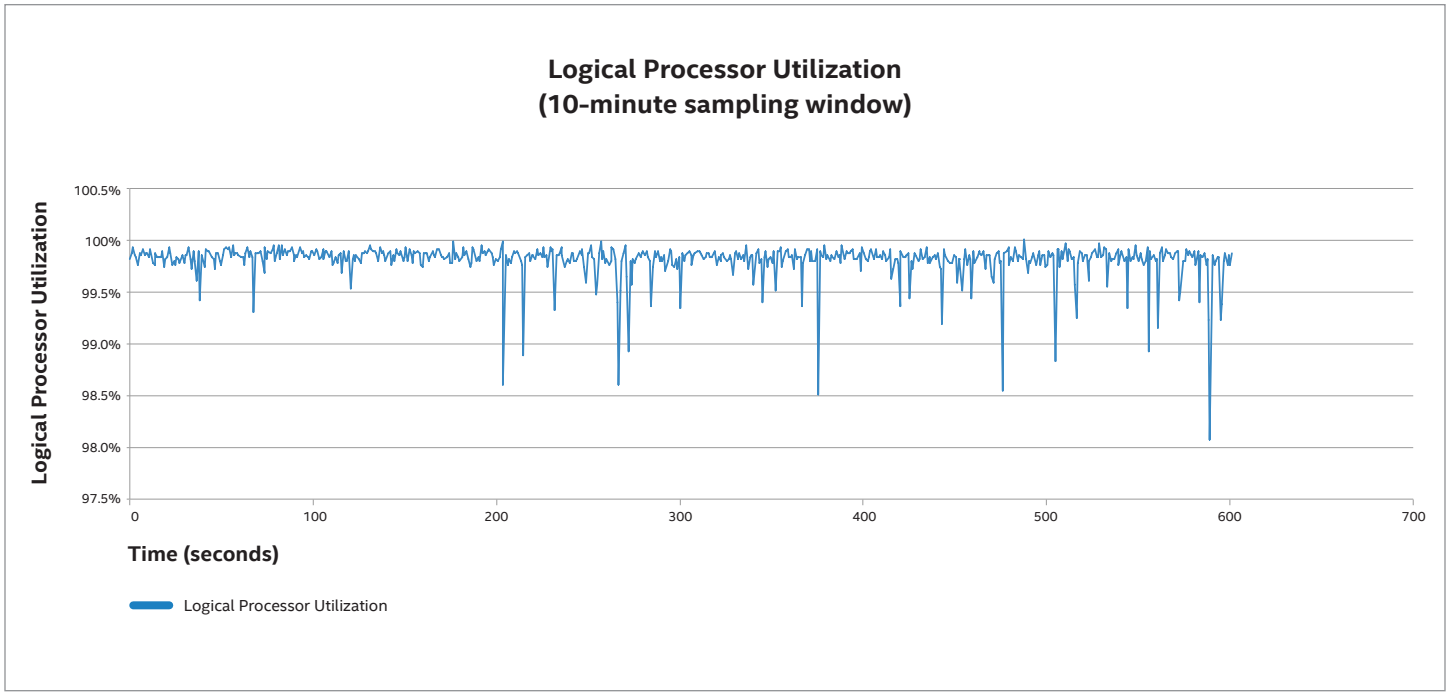

Figure 4. Logical processor utilization was consistently high during phase-C processing, showing that the optimized workload and balanced hardware platform enabled efficient use of available processing resources.

As shown in Figure 2, logical processor utilization during phase A testing was at or near 100 percent throughout a representative, 10-minute sampling window, with occasional and expected drops due to periods of high cache miss rates and peak memory references. This high utilization rate shows that the optimized software and balanced hardware platform enable exceptionally efficient use of available processing resources to deliver sustainable high performance.

Phase B: Local Realignment around Indels

During Phase B, the data is realigned around Indels to provide higher accuracy at the most critical data points. Two tools from GATK 2.39Lite were used: RealignerTargetCreator and the IndelRealigner.

To simplify and accelerate the computations, the reference genome was divided into 100 chunks. Larger chromosomes were divided into multiple chunks, with the divisions occurring at the centers of regions that had at least 500 consecutive N bases in the reference genome. Each chunk was treated as a separate process, which the Rabix scheduler assigns to an available processor thread.

The workload during this phase contains a mix of single and double precision calculations, which benefit from Intel AVX2. However, throughput during this stage is primarily dependent on memory performance.

As shown in Figure 3, memory performance was quite good, with sustained throughput of roughly 70-80 percent of maximum. The steady performance at peak utilization indicates good NUMA locality, which verifies that the system is managing data efficiently for reduced latency.

Phase C: Base quality score recalibration

During phase C, the base quality score is recalibrated to determine the reliability of the base reads. Two tools in GATK 2.39Lite were used: BaseRecalibrator to calculate the table, and PrintReads to apply the recalibration. BaseRecalibrator was run on 56 threads and was restricted to chromosome 20 as a sub-sampling strategy. PrintReads was run using 100 threads, using the same parallelization strategy that was used in Phase B.

The workload in this phase takes full advantage of Java Platform SE 76, which, as discussed earlier, is highly optimized for Intel AVX2. As shown in Figure 4, processor utilization was consistently high, demonstrating efficient use of available compute resources. NUMA locality was also high, with 70-80 percent of memory references targeting local memory.

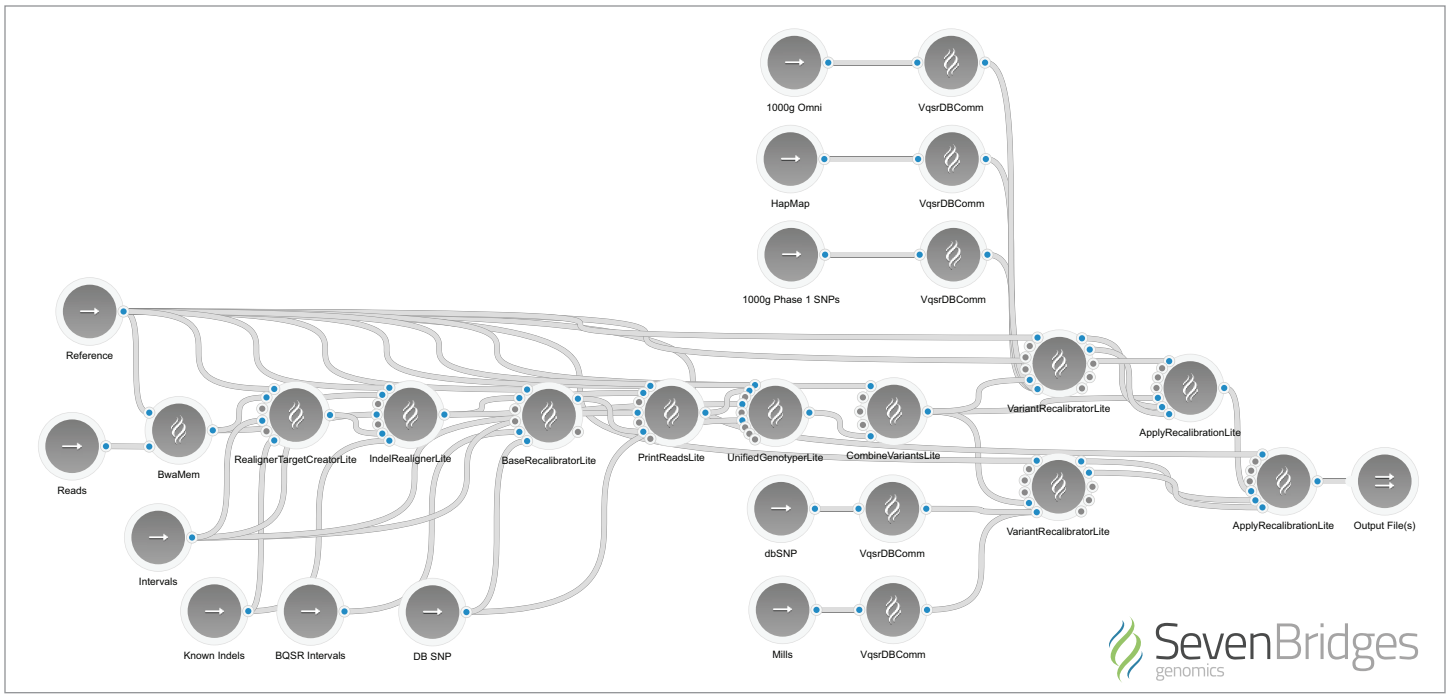

Figure 5. Seven Bridges Genomics provides a complete software platform for simplifying and accelerating a whole genome pipeline.

Phase D: Variant calling and variant quality score recalibration

In Phase D, the genome variants are called and a quality score is recalibrated to quantify the reliability of the results. Variant calling was performed using the UnifedGenotyper in GATK 2.39Lite. The workload was distributed across 100 threads, using the same parallelization strategy that was used in Phase B and C. Variant quality score recalibration was performed in two steps. The software settings were optimized for SNPs and Indels, and each step was run across all 56 available cores.

The computations in this phase depend primarily on double precision floating point arithmetic, which make good use of Intel AVX2. As shown in Table 2, the performance gains were particularly high for this phase.

Summary

The Intel Genomics Cluster solution provides a scalable, high-performance building block for conducting whole genome analysis using next-generation sequencers. Based on the Intel Xeon processor E5 v3 family, this system can perform a whole genome analysis up to 41 percent faster than a comparable system powered by the previous-generation Intel Xeon processor E5 v2 family. With these performance gains, a whole genome analysis can potentially be completed almost five hours sooner. Additional systems can be combined within the same data center rack to support multiple next-generation sequencers.

Learn More

Seven Bridges Genomics: https://www.sbgenomics.com/

Intel Xeon processor E5 v3 product family: www.intel.com/xeone5

Optimized Genomic Code: www.intel.com/healthcare/optimizecode

[1] Source: Tests conducted by Intel and Seven Bridges Genomics. Baseline system: Amazon EC2 c3.8xlarge with 2 x Intel® Xeon® processor E5-2680 v2, 60 GB memory, 10 GbE networking, 2 x 320 GB SSD. Test cluster: 1-node of local Intel Genomics Cluster, each node with 2 x Intel® Xeon® processor E5-2695 v3, 256 GB memory (8 x 32 GB DDR4 2133), 10 GbE networking, 300 GB Intel® SSD DC S3500 Series drive and 800 GB Intel® SSD DC S3700 Series drive. Software for both the baseline and test system included: Red Hat Enterprise Linux* 6.6.9, Java Platform*, Standard Edition 8, BWA* 0.7.11, GATK2.39Lite*, Rabix* scheduler v0.22. The baseline system completed a whole genome analysis in 724 minutes. The test system completed the same analysis in 429 minutes, reducing the time to results by 4 hours and 55 minutes, for a 41 percent improvement.

[2] Tarasov, Artem, et al. "Sambamba: fast processing of NGS alignment formats." Bioinformatics (2015): btv098.

0515/ABK/HBD/PDF 332451-001US