Business Brief

Healthcare and Life Sciences

Open Source Software Optimizations for CPU-Based Genomic Analysis

Intel has been working with partners for 10+ years to optimize genomics workloads for the advanced performance features of Intel® Xeon® processors. Today, research and clinical laboratories have choice among solutions that tailor cutting-edge results to their specific needs on commercial off the shelf (COTS) servers.

Technology refinements large and small have reduced by orders of magnitude the time and cost requirements to accurately analyze genomes and detect genetic variants. From personalized medicine to public health programs that encompass whole populations, the ability to readily analyze genetic information at scale provides unprecedented research and clinical opportunities.

Intel's role in the genomics ecosystem

The pioneering collaboration between Intel and the Broad Institute, stretching back more than a decade, has helped define the field of genomic analysis. A central achievement of this co-engineering effort has been to produce dramatic performance improvements for Broad's open source Genome Analysis Toolkit (GATK) on Intel® Xeon® processors. Ongoing collaboration between Intel and Broad continues to improve GATK results on successive generations of Intel processors and to advance the state of computing for genomics more broadly.

In particular, Intel worked together with the Broad to accelerate performance of some key components of GATK. Those optimizations are part of the open source GATK, and they continue to be maintained by Intel. Lenovo's Genomics Optimization and Scalability Tool (GOAST) builds on the GATK with substitutions for tools and components that improve throughput for certain workloads. In parallel with those efforts, Sentieon produces commercial genomics analysis software that is compatible with any sequencing hardware, optimized for very high speed on Intel Xeon processors.

This range of solutions demonstrates the range of choice available to the research and clinical communities for addressing specific genome analysis requirements with Intel CPUs. One way of characterizing the commonalities and differences among them is by considering the degree to which each has been engineered to emphasize and balance criteria of scalability, throughput and speed, including with optimizations and other enablement for Intel CPUs.

Design Criteria for Genomics Analysis Solutions

Scalability to large cohorts relates to a solution's efficiency and suitability for research such as worldwide population genotyping studies. One potential tradeoff in other usages is that software components engineered to handle massive data may be inefficient at smaller scale.

Throughput for high-volume analysis is concerned with how many samples can be handled in a set period of time, such as number of studies per month. The importance of throughput is emphasized in settings such as batch analysis by a pharmaceutical research team or in an oncology lab analyzing multiple samples from a single tumor or from multiple patients.

Speed for fast turnaround of results is the ability to process a single sample in the shortest time possible. Scenarios that favor speed over throughput include providing near-immediate newborn genetic screening for rare diseases or predicting adverse drug reactions from a single patient genome in an acute-care facility.

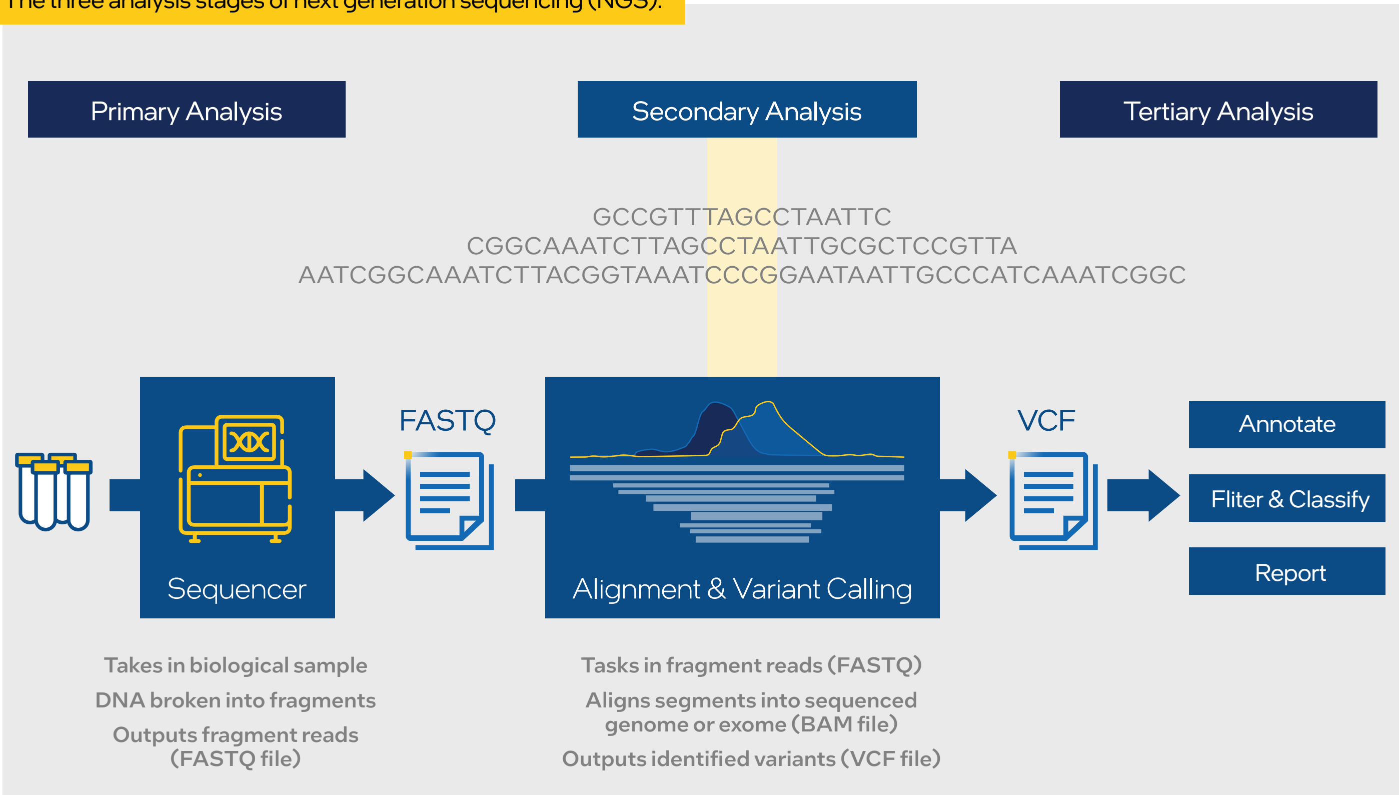
Hardware platforms and the genomics pipeline

The stages in the analysis pipeline for genetic data can be roughly characterized as primary, secondary and tertiary analysis, with each successive stage building on the one before it. Primary analysis involves the initial processing and quality control of raw data generated by the sequencer. Secondary analysis aligns sequence reads to a reference genome, assembles them into longer contiguous sequences and identifies genetic variants. Tertiary analysis involves interpretation and functional annotation of the variants, for integration with other data and processes.

As the most computationally intensive of these stages, secondary analysis is the primary optimization focus for genomics analysis on Intel architecture. Those optimizations enable differential emphasis among the priorities of scalability, throughput and speed by various solutions, for choices that can help deliver advantages in the time and cost requirements for various usage models. In addition, the Intel CPU roadmap delivers ongoing advances across its balanced hardware platforms that benefit secondary analysis:

- **High-performance cores**. A cadence of increasing per-core performance, hardware acceleration and software optimizations using new processor instructions increase data throughput.
- **Memory subsystem innovation**. Faster memory with higher bandwidth, plus larger caches, help accelerate memory-intensive processes such as Burrows-Wheeler Aligner algorithms.
- **Increasingly robust I/O**. Datasets that commonly reach hundreds of gigabytes in size benefit from enhanced PCIe resources for fast access to local and distributed storage, as well as improved CPU-to-CPU interconnects.

The three analysis stages of next generation sequencing (NGS).


Primary Analysis
Secondary Analysis
Tertiary Analysis
GCCGTTTAGCCTAATTC
CGGCAAATCTTAGCCTAATTGCGCTCCGTTA
AATCGGCAAATCTTACGGTAAATCCCGGAATAATTGCCCATCAAATCGGC
FASTQ
VCF
Sequencer
Alignment & Variant Calling
Annotate
Fliter & Classify
Report
Takes in biological sample
DNA broken into fragments
Outputs fragment reads (FASTQ file)
Tasks in fragment reads (FASTQ)
Aligns segments into sequenced genome or exome (BAM file)
Outputs identified variants (VCF file)

Genomics tools optimized for Intel architecture

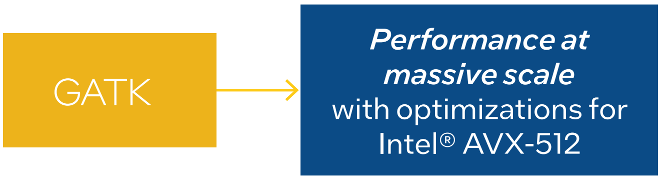
Popular toolsets for genomics secondary analysis take advantage of the ubiquity of Intel CPUs, both on-premises and in public cloud infrastructure. Optimizations for Intel architecture help drive high performance and low cost per genome. Performance also enables high accuracy to discern the difference between errors made in the sequencing process versus true genetic variants. Repeated processing to achieve sequencing depth for increased accuracy and reliability magnifies the impact of performance improvements from hardware advances and software optimization.

The capabilities of optimized secondary analysis tools for genomics, including the Broad GATK and solutions from Lenovo and Sentieon, are discussed in the remainder of this section.

Broad Institute's GATK

The open source GATK is a set of command-line tools to analyze sequencing data, primarily for variant discovery. The tools can be used individually or chained together to create workflows for different use cases. GATK is built for high performance at massive scale, and Intel has provided many optimized code components to address performance bottlenecks and take advantage of the hardware features and capabilities of Intel processors.

Key GATK optimizations implement Intel® Advanced Vector Extensions 512 (Intel® AVX-512) technology to minimize latency and loop overhead. Intel AVX-512 instructions increase vectorization in GATK code, condensing and fusing operations into fewer steps to improve computing throughput. Enablement of GATK software with Intel AVX-512 helps shorten time to completion for data processing tasks, including transformation algorithms.



GATK

Performance at massive scale with optimizations for Intel® AVX-512

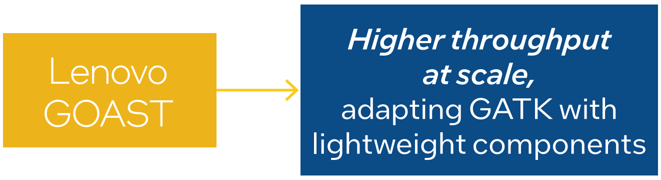
The architecture of GATK consists of mostly Java applications and a couple of C++ applications, designed to run sequentially. At a high level, the two main categories of tasks for secondary analysis are alignment and variant calling. Alignment uses tools to align a genome sequence outputted from the sequencer to a standard reference genome. Variant calling is the process of identifying substitutions, additions and deletions (variants) in the test sample, compared to the reference. Intel and Broad performance engineers have worked together to optimize both alignment and variant calling processes.

One of the C++ tools used in the alignment stage is the Burrows-Wheeler Aligner: Maximal Exact Match (BWA-MEM2), which finds the longest substring of the sequence under analysis that matches the reference genome. Intel has worked with Heng Li, the author of BWA-MEM2, to make it run faster on Intel CPUs. The GATK variant-calling phase includes the Hidden Markov Model (currently PairHMM), which finds sequence alignments.

Collaboration between Intel and the Broad optimized both BWA-MEM2 and PairHMM for higher GATK performance using Intel AVX-512. That work is ongoing, and Intel is helping optimize software updates. These optimizations to accelerate commonly used, compute-intensive genomics kernels on Intel Xeon processors are collected in the Genomics Kernel Library (GKL, also written in C++), which ships as open source along with GATK. Intel and the Broad also jointly created GenomicsDB, a datastore technology for genomic variants and likelihoods.

Lenovo GOAST

The Lenovo GOAST Bioinformatics Solution adapts GATK for throughput advantages when used at more modest scale. The solution has a free license, distributed as part of a solution that also includes a tuned architecture based on Lenovo hardware. Developed in collaboration with Intel, Lenovo GOAST enables lighter weight operation by replacing Cromwell from GATK with the Snakemake framework as workflow manager and implements Conda to manage software installations.



Lenovo GOAST

Higher throughput at scale, adapting GATK with lightweight components

In addition to those changes, GOAST inherits the GKL optimizations for BWA-MEM2 and PairHMM based on Intel AVX 512 that are described above. While GOAST is limited in terms of the jobs it can handle compared to GATK, it can deliver on the order of 3x throughput advantages.

Sentieon's DNAscope and DNAseq

Sentieon's commercial genomics solutions draw on GATK but are written entirely in assembly language and C++ to optimize processing speed, for faster time to completion and improved cost efficiency. DNAseq is designed to deliver the same exact results as GATK but faster. DNAscope includes additional enhancements to variant-calling sensitivity that improve accuracy.

Sentieon DNAscope

Higher speed and accuracy; written entirely in Assembler and C++

Sentieon solutions are sequencer-agnostic, maintaining compatibility with equipment that outputs either short reads or long reads. The ability to process both short and long sequences simplifies the collective use of disparate data from multiple platform sources. This factor is increasingly valuable as labs follow industry trends of looking at larger and larger populations in search of the next generation of insights.

Conclusion

More than a decade of Intel software enablement for CPU-based genomics analysis has contributed to the creation of a strong open source and commercial ecosystem. The industry-standard, open source Broad GATK incorporates the GDK developed and maintained by Intel to optimize the software with Intel AVX-512 technology. The Intel Xeon processor roadmap continually builds on GATK performance with balanced enhancements to execution, memory and I/O resources, both on-premises and on popular public cloud infrastructure. The broader genomics solution ecosystem, including innovations from Lenovo and Sentieon, provides added flexibility across usages.

Learn more

Intel Healthcare and Life Sciences Solutions
Broad Institute: Genome Analysis Toolkit
Broad Institute: GenomicsDB
Lenovo GOAST Bioinformatics Solution
Sentieon DNAscope pipeline

No product or component can be absolutely secure.

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

Your costs and results may vary.

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

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a nonexclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

0924/GR/MESH/PDF 356890-001US