PURESTORAGE®

intel

WHITE PAPER

Simplify and Accelerate Genomics Pipelines with Pure Storage and Intel

How Pure Storage FlashBlade//S and Intel Xeon processors can accelerate processing, enable scalability, support data security, and reduce costs in a single data platform.

Contents

Introduction

In the rapidly evolving fields of healthcare and life sciences, genomic analysis has become indispensable as a foundation for genetics and biomedical research. Genomics offers the key to unraveling genome sequences that can unlock new approaches for developing drugs and methods for treating debilitating conditions and congenital diseases.

Genomic analysis relies on high-throughput sequencing (HTS), a computationally intensive process that requires high-performing servers, responsive storage solutions, and low-latency networks to ensure optimal performance and speed. However, configuring, tuning, and maintaining that infrastructure can be a significant challenge for IT and storage administrators.

Genomic workloads are also characterized by massive and growing datasets, which test the reliability and scalability of the platforms on which they run. As these datasets expand, energy efficiency targets can become harder to reach, and both capital and operating costs can skyrocket from meeting growing storage needs; from growing space, power, and cooling costs; and from addressing the complexities of maintaining disparate compute and storage systems.

Traditionally, organizations have relied on parallel processing systems, such as WekaFS, IBM Storage Scale (GPFS), or Lustre, for running genomic workloads. These are popular options because they offer high performance and throughput for demanding genomics pipelines. But as testing and analysis by Intel and Pure Storage® demonstrate, organizations have another high-performance option available: a solution built with a Pure Storage FlashBlade® product.

By deploying high-performing compute servers along with a Pure Storage FlashBlade//S™ storage solution—all powered by Intel® Xeon® processors—organizations can realize high levels of performance while also addressing additional IT management challenges. For example, FlashBlade offers modular scalability for capacity and performance, a simplified experience for data and infrastructure management, and strong levels of security and reliability to reduce downtime and help protect sensitive data.

Strong levels of performance

Comparable to a traditional parallel processing file system

Simplified data and infrastructure management

Scales to petabytes of capacity and can replace four storage tiers (object storage, a tape system, burst buffers, and local solid-state drives [SSDs]) with one

Strong security and reliability

Always-on encryption, data protection, snapshots, disaster recovery features, and ransomware recovery to reduce downtime and help protect sensitive data

The Challenges of Traditional Infrastructure for Genomic Processing

The United States National Library of Medicine (NLM) estimates that, by 2025, genomics research will need between 2 and 40 exabytes of storage capacity just for the human genome[1]—a volume that reflects the rapid advancements and increasing scale of genomic sequencing technologies. This volume of continually growing data underscores the critical need for an advanced IT infrastructure to effectively manage and analyze such massive datasets.

At the same time, modernizing infrastructure for genomic workloads runs the risk of adding complexity. Configuring multiple servers and storage devices can be time consuming, as is balancing network loads for efficiency. Traditional solutions typically don't support upgrading compute and storage independently, which means they don't scale easily to accommodate changing needs and growing datasets.

To minimize these challenges, IT and storage administrators at health and life sciences organizations need infrastructure that can provide:

- Scalability to manage the increasing amounts of data needed for the genomics pipeline
- Sufficient processing power to support researchers who demand rapid time to results
- Data security and privacy to meet organizational needs and regulatory requirements
- Minimal management overhead for configuring and maintaining servers, storage, and networking
- Energy-efficient operations to keep costs down and to meet sustainability goals

The Pure Storage FlashBlade solution offers a compelling alternative to the traditional parallel processing system approach because it provides power optimizations and performance-efficient storage with fast access to both large and small files in extremely large datasets. FlashBlade is also designed with always-on manageability, scalability, and data-protection features built-in.

To determine if a FlashBlade system could replace a traditional parallel processing file system, Intel, in collaboration with Pure Storage, ran secondary genomics analysis tests using FlashBlade paired with the Intel® Server D50DNP Family.

Testing Description

For the testing, Intel and Pure Storage used the Genomics Analysis Toolkit (GATK), a set of genomics analytics tools published by the Broad Institute. In addition to GATK, the Broad Institute publishes best practices pipelines for running specific analyses. The GATK Best Practices pipeline takes in fragments of a single whole genome sequence (WGS) sample, aligns them to a reference genome, and identifies variants in the sample. The pipeline outputs both the aligned sample and the identified variants.[2]

The test engineers ran the GATK benchmark configuration while varying the cluster sizes and throughput, and they then measured genomes processed per node per day (G/N/D).

Testing Methodology

Intel used the publicly available NA12878 30X coverage WGS as the input dataset. A single input dataset was about 85GB, yielding an output of up to 480GB upon completion. The workflow consisted of 24 tasks: 6 multi-threaded tasks (broken into shards) and 18 single-threaded tasks. Intel used Cromwell as the workflow management system, configured with Slurm as the batch scheduler to submit throughput jobs.

The cluster setup consisted of one head node and 4 to 10 compute servers installed on an OpenHPC platform. See Figure 1 for the infrastructure configuration and Appendix A for workload details and full system configurations.

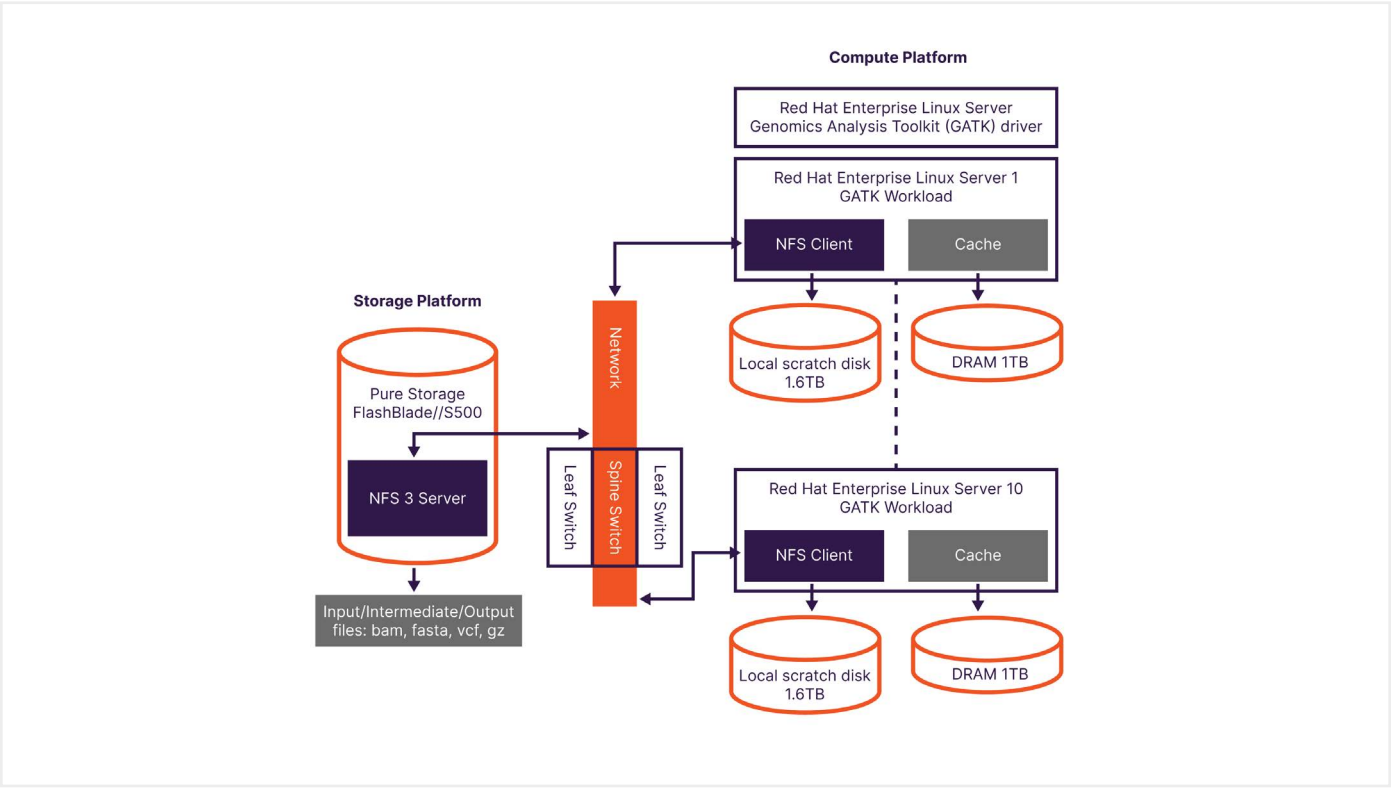


FIGURE 1 Infrastructure configuration for the Pure Storage FlashBlade//S solution and the compute servers, powered by Intel Xeon processors.

Testing Results

The solution delivered high-speed data throughput, as shown in Figure 2. This level of performance for data access is crucial for both high-performance computing (HPC) and AI workloads in genomics that need to quickly analyze vast amounts of genetic information to find patterns or anomalies. The FlashBlade solution also showed outstanding throughput scalability, ranging from 70 WGS samples to 350 WGS samples on a four-node and 10-node cluster setup. The storage utilization ranged from 34TB to 156TB for 70 WGS and 350 WGS samples, respectively.

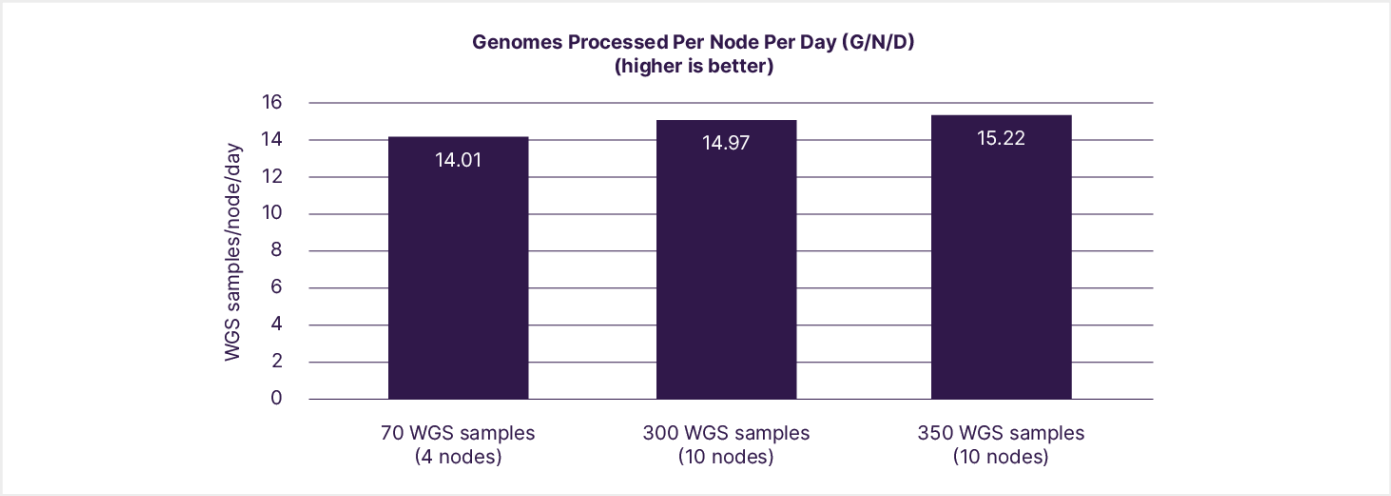


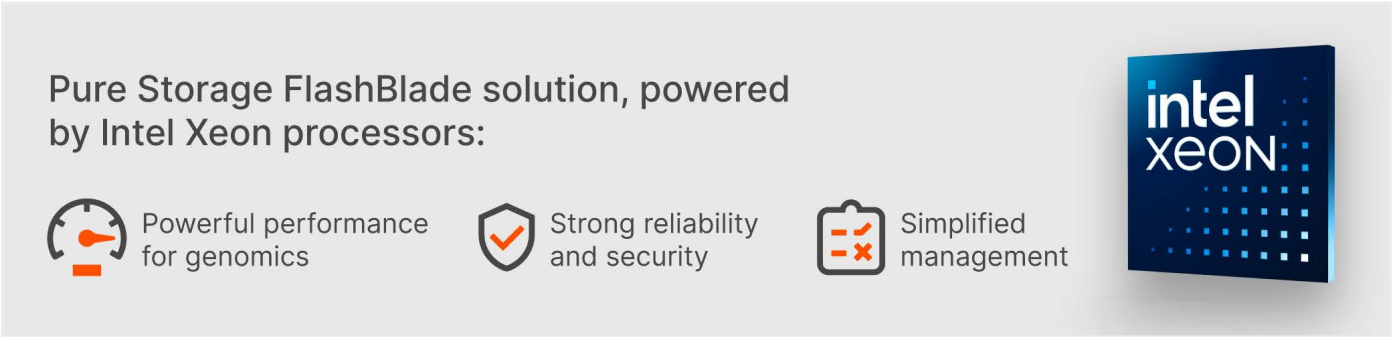
FIGURE 2 The FlashBlade solution, built with Intel Xeon processors on both the workload and storage platforms, demonstrated high throughput performance with 70, 300, and 350 WGS samples.

In addition, the compute servers were fully utilized during testing, as shown in Figure 3, while the FlashBlade system was not at full utilization. The extra headroom available on the FlashBlade//S solution demonstrates how the Pure Storage solution can help lower total cost of ownership (TCO) by scaling to meet ever-expanding datasets.

FIGURE 3 A snapshot of compute CPU utilization across all 10 compute systems shows consistently high CPU utilization.

These results can be compared to an earlier study performed by Intel on an OpenHPC parallel processing file system instead of a FlashBlade system, but powered by the same 5th Gen Intel Xeon processors and running the same GATK workload.[3] In that prior study, throughput reached 14.81 WGS on a four-node system running 70 WGS samples, demonstrating similar throughput performance (within 5.7 percent difference) compared to the FlashBlade system.

These results demonstrate the viability of performing secondary genomic analysis on a FlashBlade system, which can deliver similar levels of performance but with the added benefits of reduced complexity with added reliability and security.



These results are validated by real-world success stories, such as one from McMaster University. In this example, the researchers implemented FlashBlade to rapidly process genomic workloads, which are critical for research using DNA sequencing to quickly identify global microbial threats. Compared to the university's previous infrastructure, a less costly and highly scalable solution built with FlashBlade storage reduced the time to perform a slide scan analysis from between seven and 14 days to as little as one hour.[4]

The Performance and Efficiency Benefits of Intel Xeon Processors

The strong performance results measured in this study can be partly attributed to the Intel Xeon processors used in both the compute nodes and the FlashBlade systems. Although the current study did not examine generational performance improvements for Intel Xeon processors, the OpenHPC study referenced earlier did make this comparison. Those results are relevant because both that prior study and the current one used 5th Gen Intel Xeon processors in the compute nodes.

The results of that earlier testing, performed on a parallel distributed file system, demonstrated up to a 61 percent improvement in genomic analysis throughput with the 5th Gen Intel Xeon processors, compared to 4th Gen Intel Xeon Scalable processors.3 The newer generation processors also provide higher core counts, higher frequencies, higher memory bandwidth, and a high-capacity L3 cache, which combine to boost performance for GATK workloads.

In addition, software improvements from the Genomics Kernel Library (GKL) developed by Intel are integrated with the Broad Institute's GATK to make full use of Intel® Advanced Vector Extensions 512 (Intel® AVX-512). Intel AVX-512 is a set of instructions built to accelerate the performance of demanding workloads, including scientific simulations and analyses.

Beyond performance, 5th Gen Intel Xeon processors also provide power savings that can translate into a lower carbon footprint and that can help reduce power and cooling expenses. For example, the 61 percent performance gain provided by 5th Gen Intel Xeon processors in the previous study referenced above was achieved with a power consumption of only 0.87kWh per WGS sample, which is equivalent to 0.36kg of carbon dioxide emissions per WGS.[5]

FlashBlade Reduces Complexity and Boosts Availability, Security, and Scalability

The FlashBlade test configuration used in the current study provides several additional benefits while meeting or exceeding the performance levels of traditional parallel processing file system platforms such as Lustre. These include improved collaboration, strong regulatory compliance, simplified scalability, increased efficiency, and greater cost savings.

Collaboration by Design

Genomic research typically requires collaborative, real-time access to data. FlashBlade enables multiple researchers or AI systems to simultaneously access and process the same genomic datasets without bottlenecks, facilitating collaborative efforts in large-scale research projects. In addition, researchers can work from the same FlashBlade at the same time, both reading and writing data.

FlashBlade also facilitates rapid data retrieval for AI. With FlashBlade, AI algorithms can access and process genomic data efficiently, leading to quicker insights and decision-making, which is crucial for drug discovery, disease research, and other genomics workloads.

Data Security and Compliance

Genomic data is sensitive, requiring secure storage solutions that protect against breaches and data loss. FlashBlade offers robust always-on encryption, data protection, and compliance features, in addition to snapshots, disaster recovery features, and ransomware protection, to help ensure that genomic data remains accurate and uncorrupted throughout its lifecycle. When combined with Intel Xeon processors on both the compute and storage platforms, users gain data security starting at the silicon level for compute environments that require low latency and high throughput. For example, Intel Xeon processors help protect data in use by performing sensitive computations in hardware-based, attested Trusted Execution Environments (TEEs). Other Intel® technologies built into the processor help protect data from modification at the hardware level. These enhanced security features can help protect against malicious hypervisors, privileged bad actors, rogue guest operating systems, physical attacks, and side channel attacks.

Data privacy and integrity are also paramount concerns for highly regulated healthcare and biomedical organizations. By providing security and privacy features such as Object Lock, strong encryption, and retention, FlashBlade helps ensure compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA), the General Data Protection Regulation (GDPR), and others. As a result, FlashBlade is a reliable storage solution for organizations handling sensitive health and genomic data.

Scalability and Flexibility

FlashBlade scales to petabytes of capacity, while delivering fast access to billions of both large and small files. As AI applications in genetics research expand to include new use cases, such as machine learning (ML) models for personalized medicine or population genetics, FlashBlade helps ensure that infrastructure can scale to meet expanding computational and storage needs without costly and time-consuming upgrades.

In addition, FlashBlade can integrate with Amazon Simple Storage Service (Amazon S3) and the Equinix Metal platform, allowing for hybrid-cloud workflows. This integration makes it easier to move data between on-premises infrastructure and the cloud, optimizing storage costs and improving flexibility in genomics and AI projects.

FlashBlade//S arrays can simplify storage by replacing four storage tiers (object storage, a tape system, burst buffers, and local solid-state drives [SSDs]) with one tier. FlashBlade also minimizes the need for local storage or for overprovisioning cloud storage.

Cost Savings

Because it provides simplified management with efficient use of space and power, FlashBlade can provide a lower TCO compared to traditional on-premises and cloud storage options. A separate analysis performed by Illumina, Microsoft, and Pure Storage on Equinix Metal demonstrated that Pure Storage is 50 percent less expensive than other storage offerings available in Microsoft Azure.[6] These cost savings are particularly relevant to organizations in genomics and AI, where data storage needs can become expensive as research scales and data—typically kept indefinitely—grows to massive levels.

Spotlight on Technologies

FlashBlade//S is a high-performance scale-out system for file and object workloads that's designed to be easy to buy, set up, change, and improve. Powered by a high-performing Intel Xeon processor, FlashBlade//S delivers consistently fast and streamlined performance at every stage in genomic workflows.

FlashBlade//S also minimizes costs through efficient use of space and power that can help significantly reduce TCO for organizations. Its high-density design and scale-out architecture allow enterprises to increase storage capacity without proportional increases in space or energy consumption. And advanced data reduction technologies in FlashBlade reduce the physical storage required, further lowering both energy costs and the data center footprint.

Pure Storage Evergreen® subscription storage provides non-disruptive data-in-place blade upgrades, all-inclusive Purity software, and even flash storage upgrades so that you aren't burdened with hidden costs that might prevent necessary upgrades to support the business. This ensures you have the most up-to-date tools to help protect your genomics data and keep sequencing workflows running. And with a proven track record of non-disruptive upgrades, you can choose to update whenever you need without disrupting sequencing runs or disrupting time-sensitive, tertiary analysis.

Intel Xeon processors include built-in security and performance features that scale to fuel the base-calling algorithm and sequence-alignment tools used for secondary analysis in back-end systems.

Genomics Kernel Library (GKL), developed by Intel, accelerates commonly used, compute-intensive genomics kernels on Intel architecture. GKL includes the Intel AVX-512 optimizations, plus compression and decompression libraries, and it is distributed with GATK as open source software.

Pure Storage FlashBlade Brings Performance, Scalability, and Availability to Genomics Pipelines and Biomedical Research

As testing shows, FlashBlade can provide high levels of performance for genomics processing pipelines and biomedical research that rely on genomics data. And because FlashBlade offers scalability with high availability, efficiency, and security, it offers a compelling alternative to traditional complex parallel processing file system platforms for a variety of use cases, including:

- **Genomic sequencing:** FlashBlade can rapidly store and process vast amounts of sequencing data, facilitating faster genetic analysis.
- **Drug discovery and AI:** Pharmaceutical companies using AI to analyze results of genomic analysis for drug discovery can benefit from the speed and scalability of FlashBlade.
- **Population genomics:** Large-scale projects that analyze genetic data require scalable storage and real-time data processing, which FlashBlade provides.
- **AI-driven personalized medicine:** For AI models that tailor treatments based on a patient's genetic makeup, FlashBlade ensures quick access to large genomic datasets, enabling real-time decision-making.

Additional Resources

- Pure Storage genomics
- Intel Xeon processors and Intel Xeon Scalable processors

Appendix A

The following configurations were used in the testing referenced in this paper.

Server Configuration

Server	Intel Server D50DNP Family
Server CPU	Intel Xeon Platinum 8568Y+ processors 48 total cores/96 total threads, max turbo frequency 4GHz, processor base frequency 2.3GHz, cache 300MB
Intel® Ultra Path Interconnect (Intel® UPI) Speed	20 gigatransfers per second (GT/s)
Max Number of Intel UPI Links	4
Thermal Design Power (TDP)	350W
DRAM	1TB DDR5, 4,800 megatransfers per second (MT/s)
Network Interface Card (NIC)	100GB/s dual-port Intel® Ethernet Controller E810-CAM2

Storage Configuration

Storage	Pure Storage FlashBlade//S500
Chassis Capacity	964TB
File System	150.73TB used
Blade	10 × 43.17TB
Data Reduction	1.0:1
CPU	Intel Xeon Silver 4316 processor 20 total cores/40 total threads, max turbo frequency 3.4GHz, processor base frequency 2.3GHz, cache 30MB
Intel UPI Speed	10.4GT/s
Max Number of Intel UPI Links	2
TDP	150W
Flexible I/O (fio) Module	Intel Xeon D-1627 processor

Switch Configuration

Switch	Configuration
Cisco Leaf Switch	Cisco Nexus 9364D-GX2A switch, 64 ports, 1 leaf switch connected to 16 servers in the rack
Cisco Spine Switch	Cisco Nexus 9364D-GX2A switch, 64 ports, 1 spine switch with 32 uplinks from the leaf and 8 uplinks from storage XFM

Software Configuration

Software	Version
Operating System	Red Hat Enterprise Linux 9.4 (Plow), 5.14.0-427.18.1.el9_4.x86_64
GATK	4.4.0.0
Samtools	1.18
VerifyBamID	2.0.1
Burrow-Wheeler Aligner (BWA)	0.7.17
Picard	3.1.1
Java	OpenJDK 17
GKL	0.8.11
GNU Compiler Collection (GCC)	8.5.0-18.el8

Workload Description

Figure 4 shows the workflow used in testing.

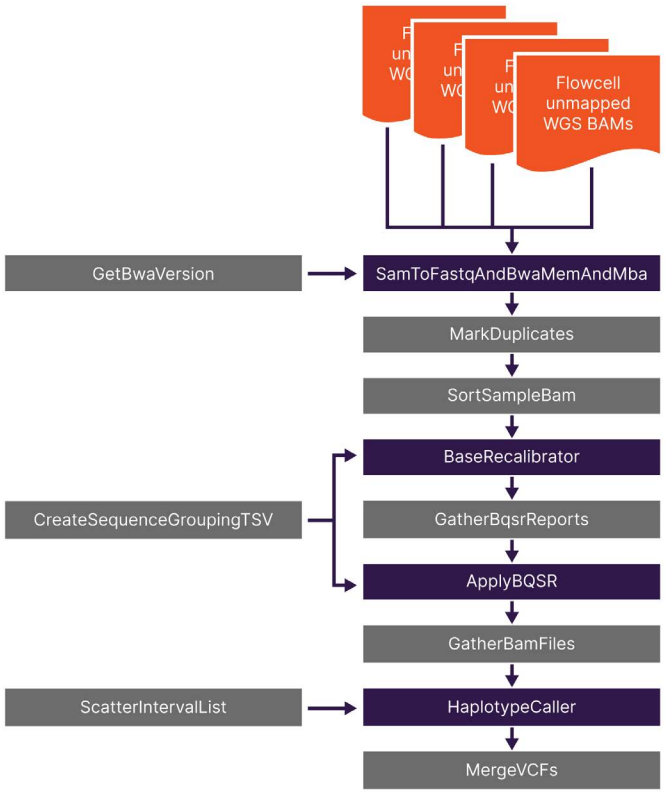


FIGURE 4 The GATK best practices pipeline for germline variant calling, which takes in fragments of a single WGS sample, aligns those fragments, and identifies variants in the sample.

1 United States National Library of Medicine. National Center for Biotechnology Information. "Big Data: Astronomical or Genomical?" July 2015.

2 Broad Institute. "Genome Analysis Toolkit (GATK) Best Practices." June 2024.

3 Intel. "Accelerating Genomics Analytics with 5th Gen Intel® Xeon® Scalable Processors While Keeping Compute Costs Low." May 2024.

4 Pure Storage. "McArthur Lab Fights Global Threats to Human Health." October 2023.

5 Calculating the kWh per WGS: Number of 30x whole genomes =350; total runtime (hours) = 58.57; average kilowatts across all 10 compute nodes and storage nodes = 5.185kW; total power (runtime x average kilowatts) = 303.69kWh; average power per whole genome = 303.69kWh/350 WGS = 0.868kWh/WGS. Calculating the CO2 per WGS: We used the US average of 0.410kg CO2eq/kWh as of July 2024 from source: https://app.electricitymaps.com/map?lang=en. Total CO2 per 10 nodes = 303.69kWh x 0.410kg CO2eq/kWh = 124.51kg CO2. Total CO2 per WGS = 124.51kg/350 WGS = 0.356kg CO2 per WGS.

6 Pure Storage. "Accelerating Secondary Genomic Analysis in a Hybrid Cloud." May 2023.

PS2722-01-en 02/25