Next-Generation Sequencing Demands a Modern Data Experience

4 key data storage considerations to support genomics-driven precision medicine.

PURESTORAGE®

Table of Contents

Why You Should Care

> "By 2023, 40% of the top 25 healthcare and life science companies will have a genomics technology enterprise strategy and be actively leveraging genomics data in developing new products and therapies. By 2025, doctors will diagnose and treat 50% of their patients with aid from genomics, up from about 1.5% in 2013."[1]
>
> **GARTNER**

Genomic sequencing today is faster and much more affordable than ever before - from $300 million to sequence one human genome at the turn of the millennium to less than $600 today.[2] Moore's Law and the underlying technology of next-generation sequencing (NGS) has accelerated the pace of discovery, transforming medicine and research science. Genomics today is being used to enable early diagnoses and targeted treatments for cancer and rare diseases. At the same time, it is being deployed in public health enabling researchers to stay one step ahead of evolving infectious disease threats such as COVID-19. Scientists are using NGS to uncover the genetic basis of human health and disease while at the same time developing a better understanding of the plant and animal world around us.

NGS enables scientists to sequence an entire human genome in a single day. The rapid adoption of genomics in medicine and research is generating staggering volumes of data: One whole human genome sequence produces approximately 200 gigabytes of raw data, for example.[3] Each genomics analytics experiment can produce hundreds of gigabytes of data that must be stored and readily accessed for processing and analysis. As a result, genomics data storage needs are doubling every seven months.[4]

The requirement to manage huge amounts of data has launched genomics into the world of big data, making innovative storage technology critical.

In this eBook, we present a brief overview of the growth and applications of genomics analytics, its impact on technology, and four key considerations when implementing a storage technology infrastructure to support it:

1. Simplicity
2. Performance
3. Future-readiness
4. Security

[1] Healthcare and Life Science CIO's Genomics Series: Part 1 – Understanding the Business Value of Omics Data, Gartner Report, February 15, 2021.

[2] The Cost of Sequencing a Human Genome, National Human Genome Research Institute Fact Sheet, https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

[3] Is Cloud Computing the Answer to Genomics' Big Data Problem, Labiotech.eu, January 15,2020

[4] The Genomic Data Challenges of the Future, October 27, 2018, The Medical Futurist

Impact of Genomics Analytics on Technology

The dramatic growth of genomics data and its applications places new demands on informatics infrastructure to meet the needs for genomics workloads including:

- High-performance computing to rapidly assemble and analyze genomes
- Moving data seamlessly across multi-cloud and on-prem locations
- Securely sharing genomics data across researchers, EHRs, patients and other stakeholders
- Cost containment and optimization of storage requirement for archives

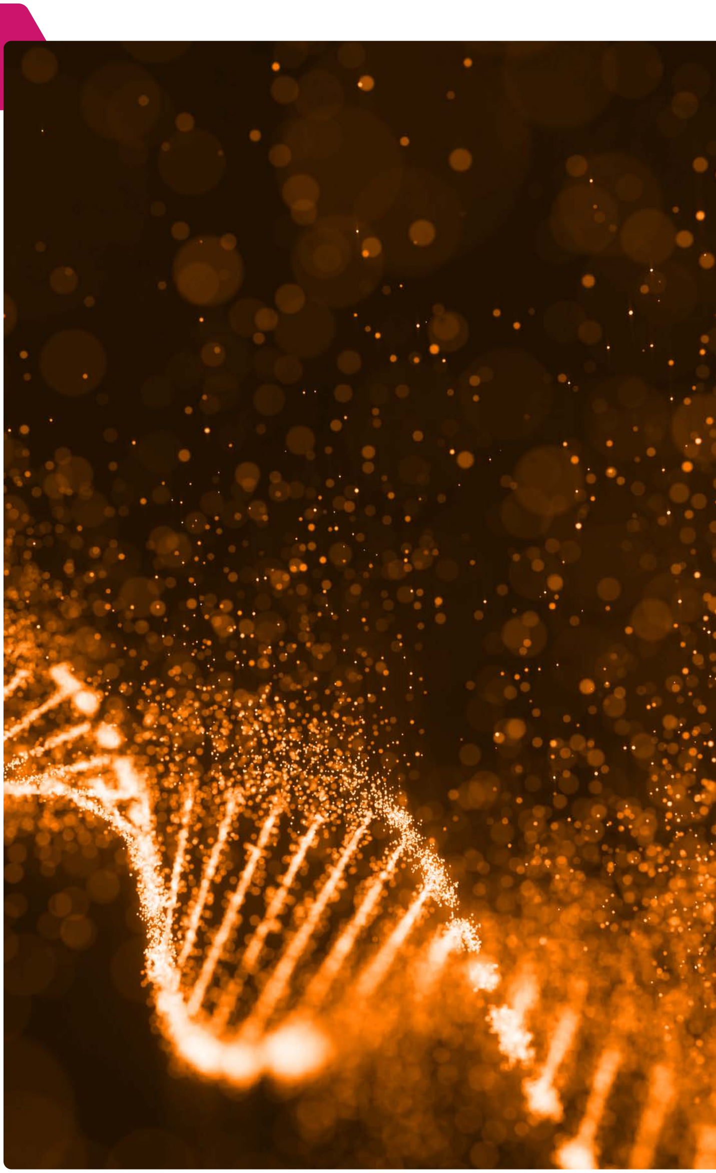
A typical genomics workflow consists of three basic stages: primary, secondary, and tertiary analysis and all require high performance computing capability.

The primary analysis stage involves converting raw instrument signal data into sequence data consisting of base calling, sequence reads, and quality scores. At this stage, sequencers generate millions of small- to medium-sized files which need to be continuously written to a storage location. Ideally, a storage system should be able to ingest such large data sets consisting of millions of small files without incurring any additional latency to the run time.

The secondary analysis stage assembles the reads and aligns them against a reference genome to then identify variants. This involves BAM file merging and sorting as well as variant calling read quantification. This stage is CPU and memory intensive involving sequential computing.

The tertiary analysis stage focuses on making sense of the observed data through annotation, curation, classification, interpretation, and ultimately clinical reporting. This stage requires a mix of sequential and random access and read intensive computing.

In a typical genomics infrastructure, the sequencing instrument sends raw data to interim storage, which can sometimes be unsuitable for advanced analytics. As a result, IT teams need to move that data to a high-performance computing (HPC) environment



for assembly and analytics, a time-consuming process that can jeopardize data integrity due to the multiple copy operations across disparate storage environments. In addition, legacy HPC systems are less suited to handle genomics workloads which have millions of small files, large files, and metadata, and can severely throttle performance, especially when data volumes are high and continuous.

The increased needs for rapid data access and analysis puts an enormous burden on IT to support high performance computing. Advances in technology must focus on removing the bottlenecks that cause I/O delays, heighten security risks, drive up costs, and increase complexity.

Healthcare and life sciences organizations are mobilizing to meet the needs of Genomics Big Data. According to a recent report from Gartner, by 2023, 40 percent of the top 25 healthcare and life science companies will have a genomics technology enterprise strategy and be actively leveraging genomics data in developing new products and therapies.[1]

PURE STORAGE IN ACTION

The Need for Speed in the Fight for Global Health

> “There’s no point in playing with traditional storage because it’s just not fast enough. With Pure Storage FlashBlade®, we can stay ahead of the curve as we fight global threats to human health.”

ANDREW G. MCARTHUR, PH.D.
MCMASTER UNIVERSITY

Objective

To combat superbugs and new coronaviruses, Andrew McArthur, Ph.D. and genomics professor and researcher at Canada’s McMaster University, needed to speed time to insights from the millions of genomic data points his lab processes every day.

Solution

McMaster’s McArthur Lab selected Pure Storage FlashBlade to support a new gene sequencing system, providing the power the researchers need – not only to accelerate live-saving drug discovery, but to monitor global threats to human health.

Results

- Speeds drug discovery by analyzing select data sets up to 24x faster
- Allows the team to monitor health threats more closely
- Scales to support additional research and clinical partnerships

[1] Healthcare and Life Science CIO’s Genomics Series: Part 3 – Prioritizing Omics Investments, Gartner report, February 15, 2021

Key Considerations for Storage Technology

The drive to enhance infrastructure to support the surge of genomics has especially impacted the need for advanced storage technology.

A study in PLOS Biology states that the research community:

> “needs to start designing and constructing data centers with fast, tiered storage systems to query and aggregate over large collections of genomes and ‘omics data.”[1]

Another study published in Giga Science stated that:

> “storing and analyzing the huge amounts of data generated by sequencing and other high-throughput technologies requires e-infrastructure providing high performance computing and large-scale storage resources.”[2]

As a key stakeholder in an organization that wants a viable -omics strategy, you should consider four key foundational infrastructure elements.

1. Simplicity

Supporting genomics research requires a technology infrastructure that supports both server message block (SMB) and native file system (NFS) protocols to simplify your workflow. The ease that comes with using a single storage environment that can speak both protocols can significantly reduce the need for IT resources as you proliferate the use of -omics in your organization. In addition, ongoing management should be easy and include automatic management, plus single pane of glass, no disruptive install and upgrade, and near-zero downtime.

2. High Performing

Assembling, mining, and interpreting sequences (secondary and tertiary analysis) requires a high-performance compute environment built on a storage architecture that delivers low-latency IOPS and high throughput. Secondary analysis in particular demands high performance storage to support metadata access and concurrency requirements. As the number of sequence reads grows, you need non-disruptive capacity scaling to prevent disruption of sequencing runs and scientific analysis. You need parallel processing technology that enables you to run through as many sequencing runs as quickly as possible to maximize your lab ROI and deliver on patient and lab outcomes.

[1] Big Data: Astronomical or Genomical? By Zachary D. Stephens, Sklar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael Schatz, Saurabh Sinha, Gene E. Robinson, PLOS Biology, July 7, 2015.

[2] Recommendations on e-infrastructure for next-generation sequencing, by Ola Spjuth, Erik Bongcam-Rudloff, Johan Dahlberg, Martin Dahlo, Aleksi Kallio, Luca Pireddu, Francesca Vezzi, and Eija Korpelainen, Giga Science, June 7, 2016.

3. Future-Ready

The applications and processes used in genomics are constantly evolving and IT infrastructures must evolve with them. This constant change drives the need to invest in a platform that's cloud-ready to support a hybrid cloud strategy. In addition, infrastructure should be AI-ready to enable you to leverage the reams of -omics data into predictive models. In addition, you want a system that's ready to scale to the petabytes of data as you need on the go with cloud-like flexible consumption models without the need for disruptive upgrades every few years.

4. Secure

Genomics data is a double-edge sword: It is powerful enough to help drive meaningful research, but it can make your system vulnerable to attacks because of the high value of the information it contains. You need to protect your data from attack every way possible, so your infrastructure needs security tools like always-on encryptions, HIPAA compliant systems, continuous backups, and ransomware protection. Disaster recovery and business continuity built into your infrastructure is critical to avoid disruptive downtime and loss of IP and critical patient data.

PURE STORAGE IN ACTION

Powering Artificial Intelligence in Medicine

> "Pure Storage's technologies offer speed, stability, and security. Speed - the computing speed of analyzing medical data has been greatly increased after introducing AIRI. Stability – AIRI supports massive data computing with great stability. Security – We've had no data security concerns with the system."

CHANG-KU FUO
DIRECTOR OF CENTER FOR ARTIFICIAL INTELLIGENCE IN MEDICINE CHANG GUNG MEMORIAL HOSPITAL

Objective

Chang Gung Memorial Hospital established the Center for Artificial Intelligence in Medicine to strengthen clinical applications of AI and deep learning and improve the quality of medical services and doctor-patient relationships.

Solution

The Center of AI in Medicine uses Pure Storage FlashBlade and AIRI to deliver the performance, security, and stability needed to power medical data analysis and medical research.

Results

- Improve the efficiency of medical research and support multiple projects simultaneously
- Accelerate extensive medical image analysis and genetic research
- Support secure and efficient system integration and computing

The Pure Solution

Pure Storage understands the infrastructure needs to support genomics research and has a range of solutions to help your teams meet their unique requirements for speed, scale, and purchasing agility. Pure Storage FlashBlade supercharges your genomics workflows by enabling:

1. **Simplicity:** Capture data directly from sequencers to conduct primary, secondary and tertiary analyses with a single, scale-out dynamic storage solution.
2. **Performance:** Speed up analyses with a high throughput platform that handles small and large files equally well. Purpose-built parallel processing and built-in compression enables up to 24x faster secondary analyses compared to traditional disk-based environments with proven ability to run multiple studies within hours.
3. **Scalability:** Scale to petabytes of capacity and fast access to hundreds of millions of files.
4. **Universality:** Native fast SMB/NFS multi-protocol support to support the entire genomics pipeline on the same shared storage platform.
5. **Operational Efficiency:** Maintain data integrity and regain time by eliminating data movement between storage environments and chances of copy errors.

PURE STORAGE IN ACTION

State-of-the-Art Agricultural Genomic Research

> "Pure Storage is purpose-built for flash technology while other solutions have been adapted for it. That's a major differentiator."

GONZALO VERA
HEAD OF SCIENTIFIC IT, CRAG

Objective

The Centre for Research in Agriculture (CRAG) needed a storage platform that could deliver the speed, scalability, and stability the organization required to support complex genomic research and large numbers of concurrent users.

Solution

By combining a FlashBlade data hub with disk-based archiving, Pure Storage broke through historic performance barriers to create a modern, agile platform for wide-scale testing. Users gained capabilities such as data snapshots while high load saturation limits accommodated expected growth in scientific projects.

Results

- 4x higher writing speeds (from 2.4 to 8 GiB/s)
- Increased reliability performance for intensive testing

6. **High Availability:** Maximize sequencer runtime with an always available and resilient storage platform.
7. **Data Protection:** Gain peace of mind with immutable snapshots and ransomware protection.
8. **Cloud-Readiness:** Combine the best of on-premises and the cloud, delivering cloud-like agility, flexibility, and consumption choices with the control of on-prem.

Pure Storage provides the infrastructure that enables faster time to science with storage so efficient, it's invisible. You realize significant ROI for a high-performance system with low costs thanks to a cloud-like service model for your genomics needs across your entire life sciences' enterprise – drug discovery models, imaging data, and a move to the hybrid cloud. Best of all for IT teams, Pure Storage solutions are easy to use, and simple to manage from a single pane of glass for both computing and storage needs.

Pure FlashBlade allows you to consolidate data into one storage environment that can take data directly from the sequencer and run high performance analytics. You eliminate the I/O steps between different storage environments enabling you to gain efficiency and simplicity.

The Bottom Line

As we progress into the future, next-generation sequencing (NGS)is already powering precision medicine and population genomics across the healthcare and life sciences industries.

For providers, genomics informs early screening and diagnostics, care support, and keeping experts aligned on large-scale cohorts through population genomics. For research organizations, NGS is advancing our understanding of molecular mechanisms, the genetic basis of diseases, and understanding the broader plant and animal world around us. For pharma and life sciences organizations, NGS is accelerating drug discovery and development in ways never imagined before.

NGS requires organizations to have a solid infrastructure foundation to receive data from the sequencer, shuttle it through assembly analyses, serve it to end users for advanced analyses, and finally archive it in a way that makes the most sense for the business. Only by building a powerful, modern infrastructure from the ground up can life sciences organizations fully realize the enormous benefits that genomics can provide to human health and beyond.

Pure Storage is committed to helping organizations get to a faster time to science.

For more information about our genomics, visit our website

purestorage.com

800.379.PURE

PURESTORAGE®