Managing the Modern Genomics Data Flood

By John Russell, Contributing Editor, Bio•IT World
Managing the Modern Genomics Data Flood

What was once just the “promise” of genomics to revolutionize biomedical science is rapidly becoming reality as declining DNA sequencing costs — spurred by next generation technologies — turn sequencing into a mainstay of modern research. A single Illumina HiSeq 2000 next generation sequencer (NGS), for example, now generates roughly 600 gigabases per eight-day run, a startling jump from 200 gigabases roughly a year ago. Of course, even more powerful sequencing technologies are far along in development.

For researchers, the result is an abundance of riches buried in an avalanche of data. Moreover, the rush of sequence data is prompting a surge of functional genomics research (proteomics, pathway analysis, etc.) creating its own data avalanche. Managing this perfect storm of multi-discipline data is now one of the central challenges in life sciences — from de novo sequencing of new genomes and genome-wide association (GWA) studies searching for disease-related genes to deep sequencing of cancer tumors and other omics-related research.

A recent Nature Methods report crystallizes the problem: “Researchers at the National Center for Genome Resources (NCGR) … published a very complete genomic analysis of identical twins in which only one in each of three pairs had multiple sclerosis. Reads for whole-genome shotgun sequencing numbered in the billions; reads for transcriptomes and methylomes numbered in the tens of millions. The database had billions of rows,” recalls Neil Miller, deputy director of software engineering at NCGR. “It was frightening even to me.”

The datasets, quite simply, are huge, and the computational demands are daunting. BGI, the largest sequencing center in the world, has gone so far as to launch a new peer-reviewed publication and associated database entitled GigaScience, “to meet the needs of a new generation of biological and biomedical research as it enters the era of ‘Big Data.’” One observer estimates that sequencing 100,000 people, not a large number in the scheme of things, will produce at least 30 petabytes of data. Complete Genomics estimates individual files from high-resolution sequencers will eventually exceed 1 terabyte (TB).

TECHNICAL COMPUTING EXPERTISE

SGI, long a leader in technical computing, understands the challenges associated with managing and storing “Big Data” and has developed a portfolio of powerful, flexible storage systems. SGI customers such as the Technical University of Denmark (CBS, DTU), NNF Center for Protein Research (CPR), Institute of Cancer Research, BBSRC Genome Analysis Center (TGAC), Imperial College London, BioGemma, Novo Nordisk
Managing the Modern Genomics Data Flood

A/S and Merck have all adopted SGI solutions to accelerate their research in life sciences.

“SGI has always led efforts to advance the state of the art for the creation of Big Data. One consequence is that the SGI storage ecosystem has been constantly optimized to store and manage the massive amounts of data our compute platforms can generate. Nowhere is this pairing of leading edge compute and storage technology more important than in genomics research,” says Floyd Christofferson, storage marketing director at SGI.

To a considerable extent, sequencing centers — especially those also offering various analysis services — have the most challenging data management issues in genomics. Here are few of their major IT storage pain points:

- **Storage scalability.** Modern sequencers already produce up to 180TB per instrument per year, and that’s expected to grow 100-fold as capacities increase and more annotation data is captured. Scalability is a critical attribute for any genomics data storage system.

- **I/O bandwidth.** Just keeping up with real-time instrument output is a fundamental challenge, and a wide variety of applications and users require access to the data to turn it into useful knowledge. High I/O storage bandwidth is essential for these tasks.

- **Data integrity.** From the initial sequence data to assembled genomes and the results of computational analyses, ensuring data integrity is critical. Some data must be kept for years to meet regulatory requirements, to support ongoing drug R&D and to guard intellectual property.

- **Tiered storage.** Raw sequence data is heavily accessed initially for data cleaning and assembly, but much less often once this is done. Assembled genome and variant data are accessed regularly by researchers working on projects, sometimes years later. Managing this dynamic prioritization of data access and storage is an ongoing challenge.

Bottlenecks in any of these areas can stymie research. For example, it’s not unusual for initial *de novo* assemblies to bog down or take so long as to be impractical to attempt in part because of intractable memory I/O issues. Many sequencing clients, particularly academic researchers, don’t have sufficient IT capacity to store their data, and sequencing centers must offer these clients both storage and ready access to it. Cost is also an important consideration. Besides the initial investment in data management hardware and software, there are ongoing IT support, power consumption and valuable datacenter space costs to account for.
Managing the Modern Genomics Data Flood

THE SGI ARCFINITI PLATFORM

Mindful of the many thorny issues that arise with large volume data management, SGI developed ArcFiniti, a disk-based active archive solution particularly well suited for genomics and life sciences where so much data is unstructured. Utilizing Intel® Xeon® E5 processor family, ArcFiniti is an integrated hardware and software platform aimed specifically at coping with the growth of unstructured file-based data. For many organizations, the scarcity of such solutions is adversely affecting efficiency and cost, as expensive active storage is increasingly filled with inactive — but still valuable — data such as raw sequence data.

“ArcFiniti brings together, in an integrated package, the best of SGI’s expertise in both the hardware and software for managing extremely high volumes of data. In this way, SGI storage and data management archives are now available as factory-integrated, easy-to-install solutions," says Christofferson.

As a disk-based solution tuned for this kind of workflow, ArcFiniti offers performance, accessibility and long-term data integrity advantages over tape. Leveraging patented SGI technology to significantly reduce power consumption and ensure data integrity, ArcFiniti is available in five different configurations, ranging from 156TB to 1.4PB of usable storage in a single rack before compression. This density results in significant infrastructure savings over conventional archive systems, while also enabling users easy, immediate network access to archived data.

SGI ArcFiniti Options: Expand as Your Needs Grow

ArcFiniti comes in five factory-configured models to suit any archive environment. Lower capacity models can be upgraded to higher capacity, maxing out at just over 1.4PB of usable archive in a single rack.

<table>
<thead>
<tr>
<th></th>
<th>A156</th>
<th>A234</th>
<th>A468</th>
<th>A936</th>
<th>A1404</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Usable Capacity</td>
<td>178TB</td>
<td>256TB</td>
<td>490TB</td>
<td>970TB</td>
<td>1.44PB</td>
</tr>
<tr>
<td>Primary Disk Cache</td>
<td>22TB *</td>
<td>22TB *</td>
<td>22TB *</td>
<td>34TB *</td>
<td>46TB *</td>
</tr>
<tr>
<td>Power Managed Archive Tier</td>
<td>156TB</td>
<td>234TB</td>
<td>468TB</td>
<td>936TB</td>
<td>1.4PB</td>
</tr>
<tr>
<td>GigE Ports</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>10GigE Ports</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

* Standard configuration. Optional 12TB of additional cache capacity available.
Managing the Modern Genomics Data Flood

Other important ArcFiniti benefits include:

- **Proactively protect data.** ArcFiniti uses patented DiskAerobics® software from SGI to continually monitor the health of the disk sub-systems and the integrity of all data stored there. Whether data is parked in ArcFiniti for a few weeks or many years, it is always available and protected. By proactively monitoring disk health, problems are detected before they can cause data loss, providing online data protection while virtually eliminating time-consuming RAID rebuilds.

- **High performance I/O.** High-performance NFS and CIFS access enables easy consolidation of archives across multiple application and compute environments. Multiple Gigabit Ethernet or 10Gigabit Ethernet ports in each of the five different ArcFiniti models speed high-volume throughput, maintaining maximum flexibility.

- **Automated virtualization.** Automated, policy-based tier virtualization is used to minimize storage costs. Each ArcFiniti model contains a primary cache of high-speed disks, as well as an archive tier for long-term retention. These two tiers are virtualized, such that all files are visible as on-line and available at all times. ArcFiniti manages the migration and data integrity checks to ensure that data is stored in the right tier for maximum protection and cost savings without the need for user intervention.

While tape is a good option for some types of data storage, for long-term preservation and immediate access an affordable disk-based solution offers the greatest flexibility. ArcFiniti provides a low-cost option to tape without the risks. Unlike conventional arrays that consume energy to enable online access, ArcFiniti’s archive tier is able to power down while still providing online visibility to all files in the archive. This decreases the cost of maintaining an archive, while also increasing the lifespan of the drive systems.

**CASE IN POINT: COMPLETE GENOMICS**

Founded in 2005, Complete Genomics is a pioneer in offering affordable sequencing and genome analysis services to researchers. The company developed a third generation sequencing platform and began offering commercial services in 2010. Its customer base is broad and includes academia and the biopharmaceutical industry. Recently, Complete Genomics chose and deployed an SGI ArcFiniti solution that could handle current demands and grow as needed.

In essence, Complete Genomics encounters virtually every data management issue associated with large-scale genomics research.
Complete Genomics’ standard service provides whole human genome sequencing, with assembly, variant detection and annotation across all variation types including SNPs, indels, copy number variations (CNVs), structural variations (SVs) and mobile element insertions (MEIs). Minimum average coverage is either >40x or >80x reads, depending on the level selected. Deliverables include sequence variants, functional annotations and data summary reports, as well as the full set of supporting data for these results (reads, scores and mappings).

To meet this goal, Complete Genomics set out to offer affordable outsourced sequencing services to researchers requiring higher throughput and a completely new set of data management and analytics capabilities. At the same time, the IT staff at Complete Genomics had to balance the need for a large and expandable storage and archiving solution — with the potential to grow from terabytes to tens and even hundreds of petabytes as new sequencers come into use — with the need to minimize total cost of ownership in order to remain competitive.

The scale of the data problem faced by Complete Genomics is typical of genomics environments everywhere: Extreme amounts of data being produced in parallel by multiple sequencers. A single genome report, for example, weighs in at about 30GB, not including the over 500GB of reads mapping data that goes along with it. Although a single sequence run takes about 11 days to run, Complete Genomics has 30 sequencers that can run...
Managing the Modern Genomics Data Flood

16-18 genomes in parallel. This data feeds a post processing pipeline across a cluster of servers.

The result is that the Complete Genomics sequencing pipeline generates about 30TB of new data each day in a relentless flood. Not only does the storage infrastructure need to be sized to cost-effectively handle this volume, but it needs to do so in a way that absolutely ensures data integrity. An 11-day cycle is not something anyone can afford to repeat.

Another key issue with this type of workflow, which is seen in many big data environments, is the fact that less than 10% of the data is generally needed to be read again. The problem is, even though only a small amount of the data ever needs to be accessed again, when it is needed, it is absolutely critical to get it quickly.

This presents a challenge: How to justify the cost of storing inactive data on high-cost active disk arrays. Transactional disk arrays generate heat, burn power and have a high acquisition cost. And yet, when the data is needed, it is there. What was needed for this workflow was an active archive, something with a performance profile of transactional disk arrays, but with much lower acquisition and operational costs.

For several years, Complete Genomics had relied upon a commercial cloud provider for its data storage needs. The cost and the fact that it was physically impossible to effectively move this volume of content into and out of the cloud in a timely fashion caused them to look for another solution. They needed to find highly available, highly protected and highly reliable storage, with absolute security against data loss or corruption. And they needed to find a way to do this in an active archive.

Complete Genomics considered tape, disk and hybrid solutions combining the two technologies. In the end, the company opted away from tape. While tape is valid in many workflows, the slow access time and lack of inherent data protection meant that cost and usability did not fit their requirements.

The SGI ArcFiniti active archive system rose to the top as the ideal solution that combines the speed, data protection and cost profiles needed for managing this type of workflow.

ACHIEVE LOWER TCO

Relative to tape storage options, the greater density per square foot of SGI ArcFiniti reduces the required space footprint and electrical requirements to maintain the system. Complete Genomics expects TCO savings of about 40% over three years.
Managing the Modern Genomics Data Flood

“After comparing conventional disk arrays, tape libraries and cloud-based offerings, this solution provided us with the lowest cost, highest performance and best ease of use. ArcFiniti was the best solution for our needs,” says Bruce Martin, senior vice president of development at Complete Genomics.

ArcFiniti deployment is straightforward. The client fills out a form that identifies the type of networking environment being used at the site. Typically this is NFS or CIFS. “Basically all IP addresses are listed and preconfigured into the system along with any specific business rules a client may wish. The system is shipped to the location, installed, plugged into their network and almost immediately accessible,” says SGI Marketing Director Christofferson. “Migrating data into ArcFiniti is also straightforward and can be done at rates exceeding 30 TB per day.”

Clearly, turning genomics into personalized medicine, improving crops and gaining a further understanding of basic knowledge are as much an exercise in high performance computing as they are a product of new experimental technologies. The blizzard of data pouring from these new instruments must be managed and stored effectively before it can be analyzed. What’s more, the plummeting costs of NGS will soon push genomics into mainstream medicine, and the medical world is taking note as indicated by the excerpt from an American College of Physicians (ACP) article, “Can the Electronic Record Contain an Entire Genome?”

 “…Research labs are currently grappling with this issue of data overload. For example, the informatics core at Washington University’s NIH-funded sequencing center had about five petabytes of data storage capacity that was reported to be 80% to 90% full in the spring of 2010; it is currently undergoing expansion to approximately double its size, thanks in part to a $14 million 2009 Recovery Act grant. The core is so power-intense that it has its own electrical substation. To provide some idea of scale, according to Wikipedia, Google processes about 24 petabytes of data a day and AT&T moves around about 19 petabytes of data through its networks daily in the U.S.”

The new computational demands spawned by NGS are staggering. SGI understands these challenges. The company’s proven large-scale data management solutions are deployed in many industries, including life sciences, and the high performance ArcFiniti data archiving solution is particularly well suited for the genomics data management challenge. Scalability, ease of use, reliability and low cost of ownership are just a few of the benefits ArcFiniti offers. For more information, visit sgi.com/products/storage/archive/arcfiniti/.

---


ii. ACP Internist, Feb. 2011, “Can the Electronic Medical Record contain an entire genome”

---

SGI, the SGI logo, ArcFiniti, UV and DiskAerobics are registered trademarks or trademarks of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries. Intel, the Intel logo, Intel Xeon, and Xeon inside are trademarks of Intel Corporation in the U.S. and other countries. Other names and brands may be claimed as the property of others.