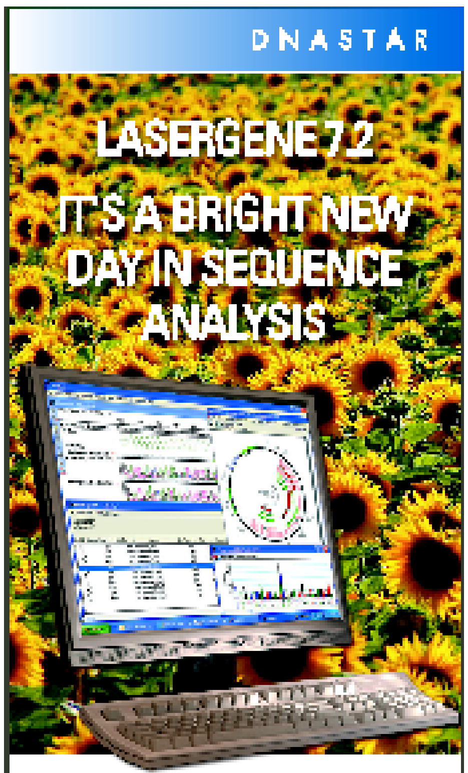
Managing Data from Next-Gen Sequencing

Tremendous Volume Can Be Handled by Tools Available Now; Complexity Is Another Matter

William Van Etten, Ph.D.

Using novel sequencing chemistries, microfluidic systems, and reaction-detection methods, next-generation sequencing vendors offer 100- to 1,000-fold increased throughput and 100- to 1,000-fold decreased cost as compared to conventional Sanger DNA sequencing.



Where high-throughput sequencing was previously limited to top sequencing centers, these new instruments are bringing large-scale sequencing as a research tool into institutional core facilities, small research groups, and the labs of individual principal investigators.

This research tool is not limited to the de novo sequencing of whole genomes. Rather, the nature of next-generation sequencing data, with many more and generally shorter reads at lower cost, make it applicable to many forms of resequencing experiments (e.g., genotyping, comparative genomics, and phylogenetic studies).

As more and more groups perform next-generation sequencing operations, two data-management problems have been revealed—data volume and data complexity. The data-volume conundrum is new to the recent converts, issues with data complexity are new to all.

Data Volume

Each next-generation sequencer is unique in terms of the volume and nature of the data it generates over time and is greatly affected by how the instrument is used. Generally, purchasers of $0.5–2 million instruments intend to operate it at near full capacity, generating anywhere from 600 GB (gigabytes) to 6 TB (terabytes) of data per run over a period of one to three days per run.

One terabyte of data is not large by today's standards. An external terabyte disk can be purchased for less than a few hundred dollars at any office supply store. Accumulating one terabyte per day, maintaining it for rapid online access, and archiving it for permanent storage is a familiar problem to the sequencing center and maybe even the core facility, but this is a novel problem for the small research group and individual principal investigator.

William Van Etten, Ph.D., is a founding partner and director of services at BioTeam, a consulting practice focused on delivering technology solutions to life science researchers. Web: www.bioteam.net. E-mail: bill@bioteam.net.

I have observed data being copied daily onto $300 external 1 TB USB drives and stacked high upon shelves. At the opposite extreme, I have also seen the implementation of six-figure, highly scalable, cluster file systems that are automatically replicated to a remote, off-site mirror for disaster recovery and automatically archived to tape using robotic tape libraries for permanent storage. Which is right? What do I recommend?

Still by far, the least expensive and most reliable method of storing massive volumes of DNA sequence data is within a DNA molecule, and I'm only being somewhat flippant in reminding the reader of this fact. The correct solution, though, is site-and use specific.

If you were to compare the cost of a six-figure storage system to a few thousand dollars in biological reagents and a day's instrument time, repeating an experiment is both a plausible and prudent means of recovering lost sequence data.

Researchers are resistant to discarding any data, however, they are accustomed to repeating experiments when necessary. An instrument's one terabyte per day of data consists of 90% primary, binary image data that doesn't compress well and 10% secondary, text sequence and meta-data (quality scores and alignment annotations) that compresses well. The primary data is needed in at least two cases.

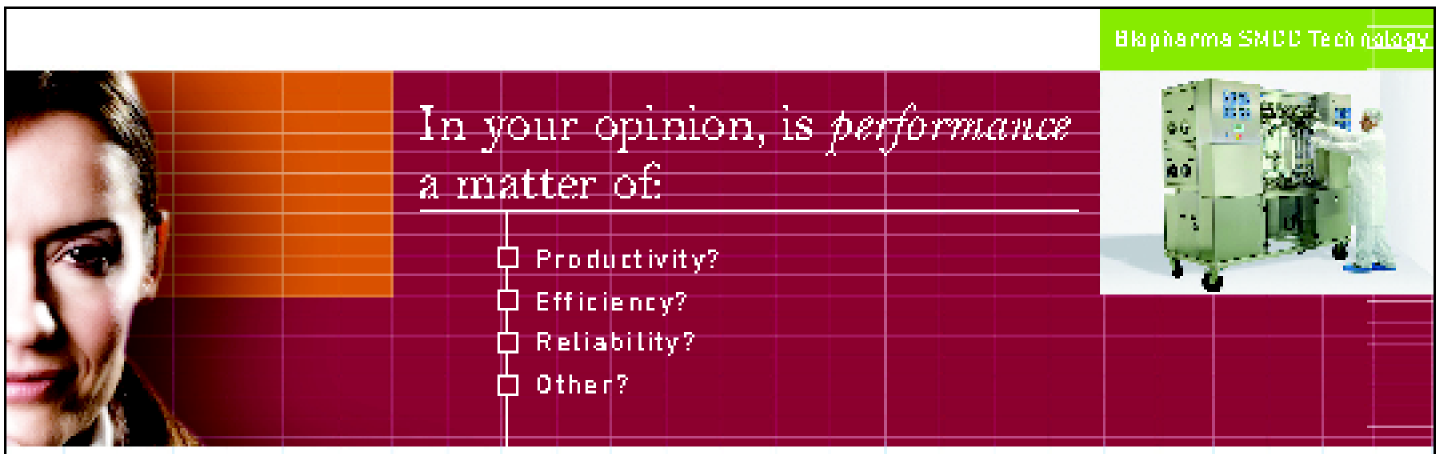
In the event the analysis algorithms change and improve, the user might want to reprocess primary data using updated software. Or, if an error is found in how the analysis was performed, the user might want to reprocess primary data with the same software but different parameters.

These instruments, their corresponding software, and their use are so new that events requiring the reprocessing of primary data aren't entirely unlikely. If you're willing to repeat experiments should adverse events occur, you can reduce the need for storage by a factor of ten.

For the small lab, a reasonable approach might be to maintain one month of primary data on a server with 10–20 TB of direct attached redundant array of inexpensive disks (RAID) storage. This is usually sufficient to provide for the accumulation of primary data and permit its reprocessing over a month's time.

RAID is important in the event that an individual disk fails (and they will), and direct attached is important because moving a terabyte of data over an Ethernet network takes hours to days. As a run's data becomes a month old, the primary data is discarded and the secondary data (~100 GB) is archived to tape (~400 GB) or external 1 TB USB drives stacked considerably lower on shelves.

See Data Management on page 44



Data Management

Continued from page 42

I hate mentioning specific brands or technologies since they become outdated so quickly, however, at this time an attractive and highly affordable solution for the small research group or core facility is Sun Microsystems' (www.sunmicrosystems.com) Sun Fire X4500 Server, otherwise known as Thumper.

Thumper, a quad-core file server in a 4 U rack-mount form factor, contains 48 disks managed by Sun's ZFS (Zeta file system). It comes in 12, 24, 36, and 48 TB sizes and unlike other common file systems, it is clever and feature rich.

ZFS is a local, not a clustered, file system, meaning that volume sizes and redundancy are limited to a single device. Late last year, Sun acquired the Lustre distributed file system, and it is anticipated that features from Lustre will be incorporated into new offerings later this year.

For high-end users, Isilon (www.isilon.com) offers a clustered file system that scales both file systems and file servers up to 1.6 petabytes (PB; 1,600 TB). The ease in which storage and servers are scaled is startling. Twice, I have encountered Isilon storage in use for next-generation sequencing, and in both cases the users were pleased with their purchase.

Data Complexity

In the old-days of conventional Sanger DNA sequencing, the data from many instrument runs generally contributed to a single common experiment, where dozens of resulting files were uniquely named with a user-defined convention that identified the experiment that the run belonged to.

With next-generation sequencing, a run consists of thousands of files created within a common directory structure. This directory has a unique user-defined name, however, this name is generally associated with the operation of the instrument, not a particular experiment. With increased throughput, a single run is more likely to include data from many distinct and often unrelated experiments.

The directory structure identifies the physical layout of the microfluidics system (e.g., lanes, cells, and tiles) and bears no discernable relationship to experiments. And as research goes, some portion of today's run might have data related to some portion of yesterday's run, and so on. Herein lies the data complexity problem.

Each next-generation sequencing instrument vendor has attempted to address this problem with varying degrees of success. Generally, vendors provide a graphical user interface (GUI) to apply experiment-specific annotations that map to regions of the microfluidics system and another GUI to review results and reports that incorporate these annotations. Many next-generation sequencing users feel that these solutions have fallen short of their needs.

In many of the next-generation sequencing deployments that BioTeam (www.bioteam.net) has participated in, we have implemented a data-management solution that we call wikiLIMS. This system leverages open-source media wiki software (the same software behind Wikipedia) to address several data-management issues including automating the movement of data from one or more next-generation sequencers to a central file server, automatically creating a wiki entry for that run in a user-defined format and layout, providing a familiar graphical wiki user interface to the free-form annotation of data within the run, offering historical tracking and version control of annotation edits provided by media wiki, and providing HTML graphical elements to launch the reprocessing of runs on external computer servers.

Each next-generation sequencing instrument has its niche (e.g., longer reads, greater accuracy, and greater data density) and for these reasons, many researchers have instruments from more than one vendor. These labs have used wikiLIMS as a means of providing a single, coherent data-management interface to sequencing data from multiple vendors including other non-sequencing laboratory instruments.

Widespread adoption of next-generation sequencing practices has created data-management challenges. The data-volume problem, however, is being resolved by ever cheaper and increasingly innovative hardware solutions. The greater problem, as users of next-generation sequencing will attest, is managing the complexity of data. GEN