Next-generation sequencing: adjusting to data overload

Monya Baker

To keep pace with accelerating sequencing machines, genomics researchers clean house and move toward the cloud.

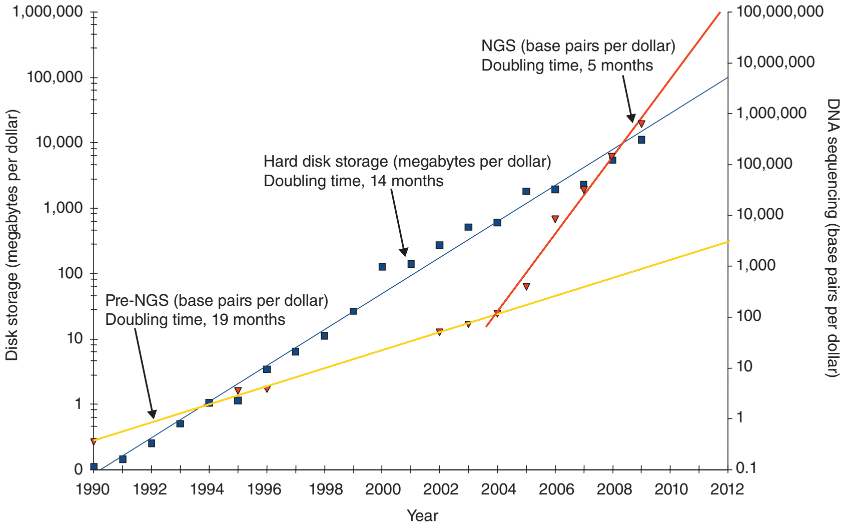
In 2005, Michael Schatz, then at The Institute for Genomic Research in Maryland, USA, felt justifiably proud of a computer program he wrote to visualize genomic data. 'Hawkeye' allowed researchers to survey a genome to weed out errors and confirm biological findings. Researchers could zoom from an overview of all the chromosomes down to the individually sequenced DNA 'reads' from which the genome had been assembled. Then Applied Biosystems (ABI) introduced its next-generation sequencing machine, which was quickly followed by instruments from Illumina and Roche 454. Throughput per machine increased 500,000-fold, says Schatz; papers from 2007 and 2008 show the number of reads per genome increasing by ~100-fold.

Hawkeye got stuck. "It fell into this trap where it would try to load all this information into memory before visualizing it," Schatz says. The program is still used for small genomes, he says, "but for large data sets it's not really feasible."

Hawkeye is far from an isolated example. With sequencer data yield increasing faster than computers can keep up, next-generation sequencing has forced researchers to rethink more than their software. Everything from storage to processing power to data output is being retrofitted or redesigned to meet the demands of ever-faster sequencing machines.

According to recent calculations from the Ontario Institute for Cancer Research, the advent of next-generation sequencing has literally changed the shape of the cost curves[1]. Since the advent of the ABI SOLiD, the cost of sequencing a base has fallen by half about every five months. The cost of storing each byte of data is dropping, too, but more slowly, halving roughly every 14 months.

Michael Schatz, who writes software for the cloud, believes the genomics community needs to explore parallel computing more aggressively.



Because of next-generation sequencing (NGS), the cost of sequencing a base is dropping faster than the cost of storing a byte. Scales are logarithmic and not corrected for inflation or costs of personnel, overheads and depreciation. Image is reprinted from ref. 1.

But such statistics only begin to explain the problem, says Vivien Bonazzi, program director for informatics at the US National Human Genome Resource Institute (NHGRI), part of the National Institutes of Health. When a sequence includes information about quality and alignments, the number of bytes needed to store a data for a single base expands quickly, she says. "We talk about megabases, but if you don't look at the issues with megabytes, then you're going to be out of sync."

One difficulty comes from scientists' reluctance to delete data files. In most analyses, data are converted from raw image files into series of called bases. These sequences or reads are then subjected to an assembly or alignment process that determines how they all line up with each other and, usually, a reference genome. Finally, these assembled reads are used in an analysis that determines, for example, what variants are present or what genes are expressed. Throughout the process, many intermediate files are created, copied and shared between researchers.

Vivien Bonazzi at the NHGRI is helping molecular biologists coordinate bases and bytes.

For most human genomes, a complete analysis identifies about 4 million variations from a reference genome, which could probably be represented in tens of megabytes of processed data, says David Craig of Translational Genomics Research Institute (TGen). "We're using terabytes to describe this because we're not throwing anything away," he says. "It's ridiculous. We really have to think about what data we are going to keep, what data we are going to throw out, and what we are going to back up."

Improvements in and experience with next-generation sequencing have already convinced some researchers to keep less data. When the 1000 Genomes Project launched in 2008 with plans to sequence genomes from people all over the world, participating researchers had to submit raw and processed intensity files, plus the string of sequence reads or 'base calls' derived from them as well as other related files. The data required worked out to around 50 bytes per base, estimates David Dooling, head of informatics at The Genome Center at Washington University in St. Louis. This May, he says, researchers participating in the project decided that the original requirements were overkill and agreed to submit only base calls along with estimates of their quality, thus slashing storage requirements to close to one byte per base. (They also decided to boost the number of people whose genomes would be sequenced to 2,500.)

Analysis paralysis

But even if researchers start throwing out some files, they are still faced with copious amounts of data to process (**Box 1**). New software resources are being developed to deal with these kinds of data. A collaboration of developers at Penn State, Emory University and other institutions has created Galaxy, which integrates multiple genomics software tools, kits them out with similar interfaces and links tools to data warehouses. The goal is for developers to write software that biologists can use and for biologists to easily share results of analyses. But as valuable as this effort is, it does not solve the universal problem among genomics researchers: the capacity of their sequencers is fast outstripping the capacity of their computers.

Researchers at the National Center for Genome Resources (NCGR) in Santa Fe, New Mexico, USA recently published a very complete genomic analysis of identical twins in which only one in each of three pairs had multiple sclerosis[2]. 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."

Physicists routinely deal with petabytes of data for tasks such as calculating conditions at the beginning of the universe or modeling climate change, but such computational approaches often do not work for genomics analysis, says John McPherson of the Ontario Institute for Cancer Research. Most physics number crunching relies on internal variables and calculations. In contrast, crunching through sequencing data requires computers to compare reads with an assembled or reference genome. This requires on-site storage, something that experts running supercomputers shun. "The clusters we've dealt with in the past usually wipe the disk clean every night," says McPherson. "A lot of the supercomputers that are available aren't designed for our kind of data."

Cloudy future

Even at big, well-resourced sequencing centers researchers are investigating an emerging technology known as cloud computing as a solution to the onslaught of data. Cloud computing allows scientists to rent both storage and processing power virtually by accessing servers as they are needed. The technology is even more appealing to institutes without a vast computer infrastructure, says Bonazzi. An increasing number of grant applications coming into NHGRI

David Craig at the Translational Genomic Research Institute worries the thousand-dollar genome could be a thousand dollars of reagents and ten thousand of computer storage.

David Dooling at The Genome Center at Washington University in St. Louis has created a software pipeline to compare genomes of cancerous and normal cells.

include plans to make use of the cloud, she says.

Genome databases from Ensembl, GenBank and preliminary data from the 1000 Genomes Project are already accessible via clouds. The earliest cloud offering, Elastic Cloud Computing from Amazon, was introduced the same year ABI commercialized next-generation sequencing. Since then, Amazon has been joined by other commercial providers such as Microsoft Azure as well as by academic projects, such as the industry-university collaborations Open Cloud Consortium and Cloud Computing University.

A few academics like Schatz, currently at Cold Spring Harbor Laboratory, have turned their attention to writing software that can analyze genomic data in the cloud. In less than four hours, Crossbow, a program cowritten by Schatz and Ben Langmead, can take billions of reads, align them to a reference genome and evaluate the newly assembled genome for single-nucleotide variations, or SNPs, the simplest type of human genetic variation.

Programs that work in the cloud must be written so that they can fragment tasks and perform them in parallel, explains Schatz. “Each computer does a little bit of work, and then there’s a round of aggregation, and then each computer does a little bit more work; it’s an iterative approach where work gets distributed, aggregated and redistributed.”

That is not how many of the existing programs run. By some accounts, NCGR provided the first cloud for genomics studies. Back in 2007, it introduced a program called Alpheus, which allowed researchers without vast computational resources to upload their data to NCGR and analyze datasets over a web-based interface. Even though Alpheus was already written to take advantage of parallel processing, rewriting it to run in a typical third-party cloud cluster is expected to take several solid weeks of effort, says Miller.

Other programs may not be salvageable at all, says McPherson. Even though many of the programs genomics researchers use are open source and so available for tinkering, these programs are often so idiosyncratic and poorly documented that the only person who could adapt them is the person who wrote them, and often those individuals have moved on to other projects.

Commercial players like GenomeQuest and DNAnexus hope that these issues will bring them business. DNAnexus formally rolled out its first cloud offering this year, a product that cofounder Andreas Sundquist recently described as “a genome browser crossed with Google Maps”(ref. 3). GenomeQuest offers services that align reads on its own cloud server, allowing genome-wide analysis and comparisons of thousands of genomes. “It’s an application that runs in a web browser that manages all the enormous volume data coming from a next-generation sequencing machine,” says Richard Resnick, vice president of software and services. Though GenomeQuest does have proprietary software, it also accommodates other software packages and has incorporated several assemblers, including the de novo assembler Newbler, into its platform.

Cloud hopping

But even when software works in the cloud, moving data between the cloud and researchers’ own infrastructure remains a challenge. “You can put gobs of stuff up there,” says Bonazzi, “but the problem is bandwidth.” Sometimes the quickest way for researchers to move large amounts of data to the Amazon cloud is to physically mail a hard drive to Amazon. For Crossbow, Schatz electronically transferred 100 gigabytes of compressed short-read data from the European Bioinformatics Institute to Amazon’s cloud in northern Virginia, USA. Even though he used 40 computers to transfer data simultaneously and has an internet connection about 1,000 times faster than a standard home connection, the transfer still took an hour and fifteen minutes.

BOX 1 SEEING RESULTS

Just as important as storing and analyzing genomics data is displaying data to make it more comprehensible, a task known as visualization. Too often, though, visualization means having a computer put data on a screen without any regard for the human being trying to understand it, says Vivien Bonazzi, program director for informatics at the NHGRI. “Sometimes it looks like a Picasso gone wrong.”

A variety of software programs have been written to visualize genomics data[5], but not all deal well with the volume produced by next-generation sequencing. One of the most popular includes the Integrated Genomics Viewer from the Broad Institute, which was designed from the start to work with large, integrated datasets and is regularly updated with new functionalities and genomes. Additions of rice, opossum and zebrafish genomes were announced in May 2010. Another oft-used program is EagleView from Gabor Marth’s laboratory at Boston College, which was released in 2008. It offers an easy way to see the quality of a genome assembly and can display base qualities, machine-specific signals, genome feature annotations and other types of information.

One challenge of creating good visualization tools is that not all informaticians think about end users, but another reason is more profound: current representations of genomes are not ideal. A reference genome is generally shown as a continuous segment with a variety of annotations, but a continuous segment has trouble accommodating many types of variants, such as large rearrangements or insertions. Repetitive sequences are also a problem. “Imagine a string of Ts,” says TGen’s Craig. “You know one is deleted, but you don’t know which one. How do you describe that? If you describe something in terms of the base that comes before or after, what about inversion?” To make matters more complicated, different communities have different needs depending on whether they are interested in SNPs, insertions and deletions, larger structural variation, epigenetic modifications or gene expression.

Bonazzi is one of several researchers who think the one-dimensional string of letters could be better represented as something multidimensional, such as a graph that can accommodate several variables. “We need to think about how we visualize the data differently now that the type of data has changed,” she says.

Neil Miller at the National Center for Genome Resources believes storing data on the cloud will require a "mental shift" that researchers will have to get comfortable with.

The problem gets worse once a dataset is put to work: every analysis of a dataset generates yet more data, so research groups who work on data in the cloud could get stuck keeping their data there unless they expand their own storage capacity, says TGen's Craig. And researchers are not only concerned about their abilities to access their own data; they worry about unauthorized access: researchers need to be convinced that they can ethically and legally fulfill their obligations to protect the privacy of human subjects if data are housed by a third party.

One solution researchers are exploring is university-owned clouds. These can reduce security concerns and data-transfer costs, but they also require the university to purchase excess storage and processing capacity, and so eliminate one of the main benefits of cloud computing.

Researchers like Washington University's Dooling think the solution will be hodgepodge, a mixture of public and private clouds and other forms of distributed computing. Ideally, he says, informaticians will write middleware that allows research groups to submit data-processing jobs along with a few specifications, and computing jobs would be automatically routed to the most appropriate place.

Mix-and-match assembly

In addition to finding the right mix of computational infrastructure, researchers are looking to find the right combination of data types. With the advent of next-generation sequencers, reads became shorter and more numerous. Not only are reads from these machines now getting longer, but an onslaught of third-generation sequencers are hitting the market and promising to produce even longer reads.

Longer reads are easier to assemble correctly. They will be particularly useful for assembling genomes for unsequenced organisms as well as for tracking unrepresented human variation and detangling genomic abnormalities that crop up in cancer cells, but they will not solve the data overload problem brought on by next-generation sequencing. "Third-generation sequencing makes everything even more challenging," says Lincoln Stein, director of the Informatics and Biocomputing Platform at the Ontario Institute for Cancer Research. "Longer reads improve data quality and make sequence assembly easier but don't solve the fundamental problem that the data [are] growing too fast for conventional solutions."

In fact, researchers are likely to require more bytes per base as they begin to work out the potential and limitations of the technology, says Dooling. "Third generation is in its infancy," he says, "so people are going to revert back to wanting to keep all the data."

Every new sequencing platform relies on a different technology and stores its sequence data in a different file type, a situation that has contributed to a plethora of programs to align reads to a reference genome or even assemble a genome based on overlap. Right now, Illumina, ABI and Roche 454 instruments are used to produce most of the sequencing data, but other platforms are poised to gain in popularity. Pacific Biosciences just announced that it can detect methylation simultaneously with base identity[4]; Helicos and Complete Genomics continue to attract fans, and Ion Torrent plans to roll out its platform later this year.

Because the scientific community has not yet settled on a standard format, multiple sequencing platforms mean resources are spent converting raw files and making these files compatible. "Multiplying the number of platforms will make this an exponentially larger headache" as more sequencing platforms emerge, says Bonazzi. "You'll need 25 converters, and it will drive everyone to distraction."

But as more platforms come on board, researchers will also be able to combine them to find the best intersection of gaining information and minimizing costs. "If you have a new organism to sequence, you won't do it all on PacBio [Pacific Biosciences machines], you'll layer them," says Dooling. "You'll use those reads to inform a huge bolus of reads from the SOLiD or Illumina [machines]." Just as combining computational strategies will be essential to dealing with sequencing data, layering the advantages of second- and third-generation platforms will be part of the innovation that takes genomics to the next level.

1. Stein, L.D. *Genome Biol.* **11**, 207 (2010).
2. Baranzini, S.E. *et al. Nature* **464**, 1351–1356 (2010).
3. Davies, K. *Bio-IT World* D4 (May–June 2010).
4. Flusberg, B.A. *et al. Nat. Methods* **7**, 461–465 (2010).
5. Nielsen, C.B., Cantor, M., Dubchak, I., Gordon, D. & Wang, T. *Nat. Methods* **7**, S5–S15 (2010).

Monya Baker is technology editor for *Nature* and *Nature Methods* (m.baker@us.nature.com).

SUPPLIERS GUIDE: COMPANIES OFFERING NEXT-GENERATION SEQUENCING SERVICES OR RELATED PRODUCTS

Company	Web address
454 Sequencing (a Roche company)	http://www.genome-sequencing.com/
Accelrys	http://accelrys.com/
Amazon Elastic Compute Cloud	http://aws.amazon.com/ec2/
Applied Biosystems	http://www.appliedbiosystems.com/
BC Platforms	http://www.biocomputing.fi/
CLC Bio	http://www.clcbio.com/
CycleComputing	http://www.cyclecomputing.com/
DataDirect Networks	http://www.ddn.com/
DNAnexus	http://www.DNAnexus.com/
Eureka Genomics	http://www.eurekagenomics.com/
GATC Biotech	http://www.gatc-biotech.com/
Gene Codes	http://www.genecodes.com/
Genedata duplicate of Sage Bio	http://www.genedata.com/
GenomeQuest	http://www.genomequest.com/
Geospiza	http://www.geospiza.com/
Helicos	http://www.helicosbio.com/
Hexagrid	http://www.hexagrid.com/
Illumina	http://www.illumina.com/
JMP (part of SAS)	http://www.jmp.com/software/genomics/
Laragen	http://www.laragen.com/
Microsoft Cloud Services	http://www.microsoft.com/cloud/
MITRE Corporation	http://www.mitre.org/
Nucleics	http://www.nucleics.com/
Oxford Nanopore Technologies	http://www.nanoporetech.com/
Pacific Biosciences	http://www.pacificbiosciences.com/
Partek	http://www.partek.com/
PREMIER Biosoft	http://www.premierbiosoft.com/
Sage Bionetworks	http://www.sagebase.org/
SAS/Genetics	http://www.sas.com/industry/pharma/genetics/
SeqWright	http://www.seqwright.com/
SoftGenetics	http://www.softgenetics.com/
Striking Development	http://www.paracel.com/