Will Computers Crash Genomics?

New technologies are making sequencing DNA easier and cheaper than ever, but the ability to analyze and store all that data is lagging

Lincoln Stein is worried. For decades, computers have improved at rates that have boggled the mind. But Stein, a bioinformaticist at the Ontario Institute for Cancer Research (OICR) in Toronto, Canada, works in a field that is moving even faster: genomics.

The cost of sequencing DNA has taken a nosedive in the decade since the human genome was published—and it is now dropping by 50% every 5 months. The amount of sequence available to researchers has consequently skyrocketed, setting off warnings about a "data tsunami." A single DNA sequencer can now generate in a day what it took 10 years to collect for the Human Genome Project. Computers are central to archiving and analyzing this information, notes Stein, but their processing power isn't increasing fast enough, and their costs are decreasing too slowly, to keep up with the deluge. The torrent of DNA data and the need to analyze it "will swamp our storage systems and crush our computer clusters," Stein predicted last year in the journal *Genome Biology*.

Human Genome 10th Anniversary 2001 • 2011

Funding agencies have neglected bioinformatics needs, Stein and others argue. "Traditionally, the U.K. and the U.S. have not invested in analysis; instead, the focus has been investing in data generation," says computational biologist Chris Ponting of the University of Oxford in the United Kingdom. "That's got to change."

Within a few years, Ponting predicts, analysis, not sequencing, will be the main expense hurdle to many genome projects. And that's assuming there's someone who can do it; bioinformaticists are in short supply everywhere. "I worry there won't be enough people around to do the analysis," says Ponting.

Recent reviews, editorials, and scientists' blogs have echoed these concerns (see Perspective on p. 728). They stress the need for new software and infrastructures to deal with computational and storage issues.

In the meantime, bioinformaticists are trying new approaches to handle the data onslaught. Some are heading for the clouds—cloud computing, that is, a pay-as-you-go service, accessible from one's own desktop, that provides rented time on a large cluster of machines that work together in parallel as fast as, or faster than, a single powerful computer. "Surviving the data deluge means computing in parallel," says Michael Schatz, a bioinformaticist at Cold Spring Harbor Laboratory (CSHL) in New York.

Dizzy with data

The balance between sequence generation and the ability to handle the data began to shift after 2005. Until then, and even today, most DNA sequencing occurred in large centers, well equipped with the computer personnel and infrastructure to support the analysis of a genome's data. DNA sequences churned out by these centers were deposited and stored in centralized public databases, such as those run by the European Bioinformatics Institute (EBI) in Hinxton, U.K., and the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Researchers elsewhere could then download the data for study. By 2007, NCBI had 150 billion bases of genetic information stored in its GenBank database.

Then several companies in quick succession introduced "next-generation" machines, faster sequencers that spit out data more cheaply. But the technologies behind these machines generate such short stretches of sequence—typically just 50

CREDIT: ALVARO ARTEAGA/ALVAREJO.COM

to 120 bases—that far more sequencing is required to assemble those fragments into a cohesive genome, which in turn greatly ups the computer memory and processing required. It once was enough to sequence a genome 10 times over to put together an accurate genome; now it takes 40 or more passes. In addition, the next-generation machines produce their sequence data at incredible rates that devour computer memory and storage. "We all had a moment of panic when we saw the projections for next-generation sequencing," recalls Schatz.

Those projections are already being realized. A massive study of genetic variation, the 1000 Genomes Project, generated more DNA sequence data in its first 6 months than GenBank had accumulated in its entire 21-year existence. And ambitious projects like ENCODE, which aims to characterize every DNA sequence in the human genome that has a function, offer jaw-dropping data challenges. Among other efforts, the project has investigated dozens of cell lines to identify every DNA sequence to which 40 transcription factors bind, yielding a complex matrix of data that needs to be not only stored but also represented in a way that makes sense to researchers. "We're moving very rapidly from not having enough data to going, 'Oh, where do we start?' " says EBI bioinformaticist Ewan Birney.

Moreover, as so-called third generation machines—which promise even cheaper, faster production of DNA sequences (*Science*, 5 March 2010, p. 1190)—become available, more, and smaller, labs will start genome projects of their own. As a result, the amount and kinds of DNA-related data available will grow even faster, and the sheer volume could overwhelm some databases and software programs, says Katherine Pollard, a biostatistician at the Gladstone Institutes of the University of California (UC), San Francisco. Take Genome Browser, a popular UC Santa Cruz Web site. The site's programs can compare 50 vertebrate genomes by aligning their sequences and looking for conserved or nonconserved regions, which reveal clues about the evolutionary history of the human genome. But the software, like most available genome analyzers, "won't scale to thousands of genomes," says Pollard.

The spread of sequencing technology to smaller labs could also increase the disconnect between data generation and analysis. "The new technology is thought of [as] being democratizing, but the analytical capacity is still focused in the hands of a few," warns Ponting. Although large centers may be stretching their computing, and their laborpower, to new limits, they basically still have the means to interpret what they find. But small labs, many of which underestimate computational needs when budgeting time and resources for a sequencing project, could be in over their heads, he warns.

Clouds on the horizon

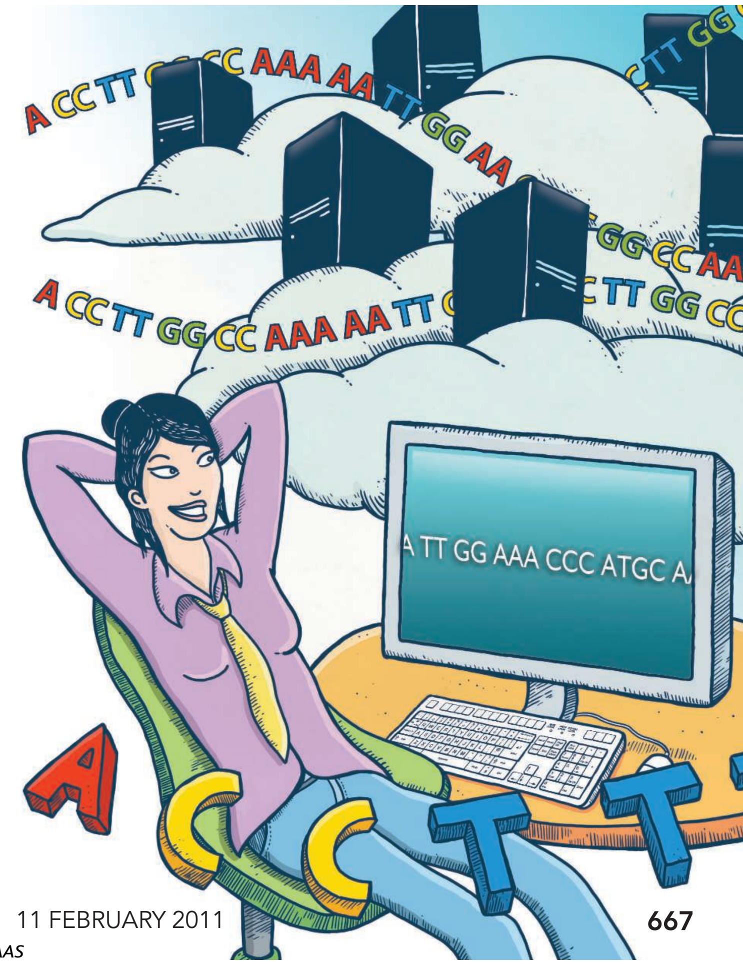
James Taylor, a bioinformaticist at Emory University in Atlanta, saw some of the demands for data analysis coming. In 2005, he and Anton Nekrutenko of Pennsylvania State University (Penn State), University Park, pulled together various computer genomics tools and databases under one easy-to-use framework. The goal was "to make collaborations between experimental and computational researchers easier and more efficient," Taylor explains. They created Galaxy, a software package that can be downloaded to a personal computer or accessed on Penn State's computers via any Internet-connected machine. Galaxy allows any investigator to do basic genome analyses without in-house computer clusters or bioinformaticists. The public portal for Galaxy works well, but, as a shared resource, it can get bogged down, says Taylor. So last year, he and his colleagues tried a cloud-computing approach to Galaxy.

Cloud computing can mean various things, including simply renting off-site computing memory to store data, running one's own software on another facility's computers, or exploiting software programs developed and hosted by others. Amazon Web Services and Microsoft are among the heavyweights running cloud-computing facilities, and there are not-for-profit ones as well, such as the Open Cloud Consortium.

For Taylor's team, entering the cloud meant developing a version of Galaxy that would tap into rented off-site computing power. They set up a "virtual computer" that could run the Galaxy software on remote hardware using data uploaded temporarily into the cloud's off-site computers. To test their strategy, they worked with Penn State colleague Kateryna Makova, who wanted to look at how the genomes of mitochondria vary from cell to cell in an individual. That involved sequencing the mitochondrial genomes from the blood and cheek swabs of three mother-child pairs, generating in one study some 1.8 gigabases of DNA sequence, about 1/10 of the amount of information generated for the first human genome.

Analyzing these data on the Penn State computers would have been a long and costly process. But when they uploaded their data to the cloud system, the processing took just an hour and cost $20, Taylor reported in May 2010 at the Biology of Genomes meeting in Cold Spring Harbor, New York. "This is a particularly cost-effective solution when you need a lot of computing power on an occasional basis," he says. With the help of the cloud, he has access to many computers but doesn't have the overhead costs of maintaining a powerful computer network in-house. "We're going to encourage more people to move to the cloud," he adds.

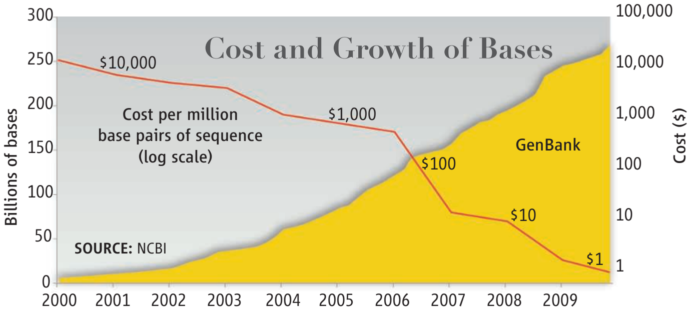
CSHL's Schatz and Ben Langmead, a computer scientist at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, are already there and are helping to make that shift possible for others. In 2009, the pair published one of the



CREDIT: ALVARO ARTEAGA/ALVAREJO.COM

first results from marrying cloud computing and genomics. They wanted to identify common sites of DNA variation known as single-nucleotide polymorphisms (SNPs), but to do so they needed to hunt through short sequences of human DNA totaling an amount equivalent to 38 copies of the human genome. With the help of a cloud-based cluster of 320 computers, they identified 3.7 million SNPs in less than 4 hours and for less than $100. "We estimate it would have taken a single computer several hundred hours for the analysis," says Schatz.

At the Biology of Genomes meeting, Langmead and Schatz unveiled two new cloud-computing initiatives. Langmead described a computer program called Myrna that determines the differential expression of genes from RNA sequence data and is designed for the parallel processing performed by cloud-computing facilities. Schatz introduced another program, Contrail, that can assemble genomes from data that next-generation sequencing machines generate and deposit into a cloud.



SOURCE: NCBI

So much for so little. The decline in sequencing costs (red line) has led to a surge in stored DNA data.

Low cost and speed aren't the only advantages of the cloud approach, says Langmead. "The cloud user never has to replace hard drives, renew service contracts, worry about electricity usage and cooling, deal with flooding or other natural disasters, et cetera," he points out. For small labs that lack their own powerful computer clusters, "cloud computing may represent the democratization of computation," says Schatz.

But cloud computing is "not mature," cautions Vivien Bonazzi, program director for computational biology and bioinformatics at NHGRI. Putting data into a cloud cluster by way of the Internet can take many hours, even days, so cloud providers and their customers often resort to the "sneaker net": overnight shipment of data-laden hard drives. And with the exception of Galaxy, Myrna, and a few other computer tools, not much genomics software is configured for the massively parallel processing approach taken by cloud computers. "It is currently too difficult to develop cloud software that's truly easy to use," says Langmead.

Also, cloud computing works best if an analysis can be divided into many separate tasks handled by multiple processors. But the connections among the cloud's processors can be fairly slow, so computations requiring processors to talk to each other can get bogged down, says Langmead. Some researchers worry that the burgeoning cloud-computing industry won't agree on standards that will allow for connections between clouds, such that data stored on one cloud can be accessible to another. "Cloud computing is hot and sexy," says Bonazzi. "But it's not the answer to everything."

Storage issues

Cloud computing offers a possible solution to other problems facing the bioinformatics community: data storage and transfer. Because storage costs are dropping much more slowly than the costs of generating sequence data, "there will come a point when we will have to spend an exponential amount on data storage," says Birney.

That has created pressure to let go of the field's long-standing tendency to archive all raw sequence data. Because the raw material from next-generation machines is in the form of high-resolution images, it soaks up huge amounts of computer storage. So scientists are considering discarding the original image files once they produce the preliminarily processed sequence data, which is more easily kept. Eventually, it may be more economical to save no raw data and just resequence a DNA sample if necessary. But for now, as to what should be kept, "there's a lot of thrashing still to happen," says Bonazzi.

Putting the data in an off-site facility could relieve some of the pressure, says OICR's Stein. The economies of scale available to large cloud-providing companies can produce significant cost savings, meaning it might be cheaper to rent transient storage space from the cloud in some cases. Storage costs at the Amazon Web Server top off at 14 cents a gigabyte per month, according to Amazon's Deepak Singh. "In comparison, it commonly costs 50 cents to $1 per gigabyte for high-end storage on a local system," Schatz says. For NCBI, however, it's still more cost-effective to keep GenBank and its other databases in-house, says Don Preuss of NCBI.

Putting data in a cloud may help in other ways as well. Right now, anyone wanting to analyze a genome has to download it from a public archive such as GenBank—and as these data sets get larger, such transfers become slower. Moreover, downloaded copies of these data sets, some now out of date, have proliferated around the world, each one taking up storage space that eats into bioinformatics budgets. In his vision, says Stein, "you have one copy of the data located in this common cloud that everyone uses" and it won't be necessary to download or upload the data between computers for processing.

Encouraged by the genomics community, NCBI has put a copy of the data from the pilot project of the 1000 Genomes effort into off-site storage run by a cloud-computing provider. And U.S. East Coast users of Ensemble, the EBI sequence database, are automatically funneled into a cloud environment as part of a test of the strategy.

One worry about this approach is the security of the data. Data involving the health of human subjects, which is being linked more and more to genome information, requires extra precautions that make some researchers hesitant about clouds. However, at least one cloud-computing company already has clients whose human data are covered by the strict health information protection laws of the United States, so there are indications that this concern can be allayed.

All these issues came to the fore last year, when NHGRI hosted several meetings on cloud computing and on informatics and analysis, says Bonazzi. Also, at a retreat last summer, the case was made for more bioinformatics training and education. "One thing that is clear is that as computation becomes more and more necessary throughout biomedical research, the way these [infrastructure] resources are funded will have to change to be more efficient," says Taylor. For now, NHGRI has no programs in place to address these needs. "But they are on our radar," says Bonazzi.

Like Stein, she worries about swamped storage systems and overwhelmed computer clusters. But Bonazzi remains sanguine. "Do I think these problems will be solved?" she says. "I'm optimistic." And even Stein is trying to think positively. "I'm very good at predicting disasters that never happen," he says. There's always sunlight above the clouds.

–ELIZABETH PENNISI

Published by AAAS