DNA Sequencing Caught in Deluge of Data

Kathy Kmonicek for The New York Times

W. Richard McCombie, a professor of human genetics at the Cold Spring Harbor Laboratory, examining some cells.

By ANDREW POLLACK
Published: November 30, 2011

BGI, based in China, is the world's largest genomics research institute, with 167 DNA sequencers producing the equivalent of 2,000 human genomes a day.

BGI churns out so much data that it often cannot transmit its results to clients or collaborators over the Internet or other communications lines because that would take weeks. Instead, it sends computer disks containing the data, via FedEx.

"It sounds like an analog solution in a digital age," conceded Sifei He, the head of cloud computing for BGI, formerly known as the Beijing Genomics Institute. But for now, he said, there is no better way.

The field of genomics is caught in a data deluge. DNA sequencing is becoming faster and cheaper at a pace far outstripping Moore's law, which describes the rate at which computing gets faster and cheaper.

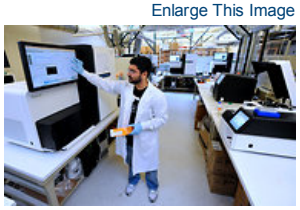
The result is that the ability to determine DNA sequences is starting to outrun the ability of researchers to store,

RECOMMEND
TWITTER
LINKEDIN
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Related

Times Topics: DNA | Biotechnology

Enlarge This Image



Kathy Kmonicek for The New York Times

Nabil Azmay, a technician at the Cold Spring Harbor Laboratory, checks a HiSeq 2000 that reads 300 billion bases per run.

Enlarge This Image

Kathy Kmonicek for The New York Times

MOST E-MAILED | MOST VIEWED

1. WELL
How Exercise Benefits the Brain

2. OP-ED COLUMNIST
My Man Newt

3. PERSONAL HEALTH
It Could Be Old Age, or It Could Be Low B12

4. HOLIDAY GIFT GUIDE
The 10 Best Books of 2011

5. STATE OF THE ART
A Thermostat That's Clever, Not Clunky

6. Judy Lewis, Secret Daughter of Hollywood, Dies at 76

7. OP-ED COLUMNIST

R. Richard McCombie and Michael Schatz at Cold Spring data center.

BGI

Freerk van Dijk, of the Netherlands Genome Project

transmit and especially to analyze the data.

"Data handling is now the bottleneck," said David Haussler, director of the center for biomolecular science and engineering at the University of California, Santa Cruz. "It costs more to analyze a genome than to sequence a genome."

That could delay the day when DNA sequencing is routinely used in medicine. In only a year or two, the cost of determining a person's complete DNA blueprint is expected to fall below $1,000. But that long-awaited threshold excludes the cost of making sense of that data, which is becoming a bigger part of the total cost as sequencing costs themselves decline.

"The real cost in the sequencing is more than just running the sequencing machine," said Mark Gerstein, professor of biomedical informatics at Yale. "And now that is becoming more apparent."

But the data challenges are also creating opportunities. There is demand for people trained in bioinformatics, the convergence of biology and computing. Numerous bioinformatics companies, like SoftGenetics, DNAStar, DNAnexus and NextBio, have sprung up to offer software and services to help analyze the data. EMC, a maker of data storage equipment, has found life sciences a fertile market for products that handle large amounts of information. BGI is starting a journal, GigaScience, to publish data-heavy life science papers.

"We believe the field of bioinformatics for genetic analysis will be one of the biggest areas of disruptive innovation in life science tools over the next few years," Isaac Ro, an analyst at Goldman Sachs, wrote in a recent report.

Sequencing involves determining the order of the bases, the chemical units represented by the letters A, C, G and T, in a stretch of DNA. The cost has plummeted, particularly in the last four years, as new techniques have been introduced.

The cost of sequencing a human genome — all three billion bases of DNA in a set of human chromosomes — plunged to $10,500 last July from $8.9 million in July 2007, according to the National Human Genome Research Institute.

That is a decline by a factor of more than 800 over four years. By contrast, computing costs would have dropped by perhaps a factor of four in that time span.

The lower cost, along with increasing speed, has led to a huge increase in how much sequencing data is being produced. World capacity is now 13 quadrillion DNA bases a year, an amount that would fill a stack of DVDs two miles high, according to Michael Schatz, assistant professor of quantitative biology at the Cold Spring Harbor Laboratory on Long Island.

There will probably be 30,000 human genomes sequenced by the end of this year, up from a handful a few years ago, according to the journal Nature. And that number will rise to millions in a few years.

In a few cases, human genomes are being sequenced to help diagnose mysterious rare diseases and treat patients. But most are being sequenced as part of studies. The federally financed Cancer Genome Atlas, for instance, is sequencing the genomes of thousands of tumors and of healthy tissue from the same people, looking for genetic causes of cancer.

One near victim of the data explosion has been a federal online archive of raw sequencing data. The amount stored has more than tripled just since the beginning of the year, reaching 300 trillion DNA bases and taking up nearly 700 trillion bytes of computer


A Banker Speaks, With Regret


8. 12 Things You Didn't Know Facebook Could Do


9. OP-ED COLUMNIST
The Arab Awakening and Israel


10. DNA Sequencing Caught in Deluge of Data

Go to Complete List » Show My Recommendations

memory.

Straining under the load and facing budget constraints, federal officials talked earlier this year about shutting the archive, to the dismay of researchers. It will remain open, but certain big sequencing projects will now have to pay to store their data there.

If the problem is tough for human genomes, it is far worse for the field known as metagenomics. This involves sequencing the DNA found in a particular environment, like a sample of soil or the human gut. The idea is to take a census of what microbial species are present.

E. Virginia Armbrust, who studies ocean-dwelling microscopic organisms at the University of Washington, said her lab generated 60 billion bases — as much as 20 human genomes — from just two surface water samples. It took weeks to do the sequencing, but nearly two years to then analyze the data, she said.

“There is more data that is infiltrating lots of different fields that weren’t particularly ready for that,” Professor Armbrust said. “It’s all a little overwhelming.”

The Human Microbiome Project, which is sequencing the microbial populations in the human digestive tract, has generated about a million times as much sequence data as a single human genome, said C. Titus Brown, a bioinformatics specialist at Michigan State University.

“It’s not at all clear what you do with that data,” he said. “Doing a comprehensive analysis of it is essentially impossible at the moment.”

Other scientific fields, like particle physics and astronomy, handle huge amounts of data. In those fields, however, much of the data is generated by a few huge accelerators or observatories, said Eugene Kolker, chief data officer at Seattle Children’s Hospital.

“In the life sciences, anyone can produce so much data, and it’s happening in thousands of different labs throughout the world,” he said.

Moreover, DNA is just part of the story. To truly understand biology, researchers are gathering data on the RNA, proteins and chemicals in cells. That data can be even more voluminous than data on genes. And those different types of data have to be integrated.

“We have these giant piles of data and no way to connect them” said H. Steven Wiley, a biologist at the Pacific Northwest National Laboratory. He added, “I’m sitting in front of a pile of data that we’ve been trying to analyze for the last year and a half.”

Still, many say the situation will be manageable. Jay Flatley, chief executive of Illumina, the leading supplier of sequencing machines, said he did not think information handling was a bottleneck or that it was causing people to hold off on buying new sequencers.

Researchers are increasingly turning to cloud computing so they do not have to buy so many of their own computers and disk drives.

Google might help as well.

“Google has enough capacity to do all of genomics in a day,” said Dr. Schatz of Cold Spring Harbor, who is trying to apply Google’s techniques to genomics data. Prodded by Senator Charles E. Schumer, Democrat of New York, Google is exploring cooperation with Cold Spring Harbor.

Google’s venture capital arm recently invested in DNAnexus, a bioinformatics company. DNAnexus and Google plan to host their own copy of the federal sequence archive that had once looked as if it might be closed.

The amount of data stored for a human genome will drop sharply. Sequencers produce huge amounts of raw data that then has to be analyzed and processed by software to produce the result.

With the field still young, many researchers store all the raw data, so it can be re-analyzed if better software is developed in the future.

In uncertain times, "scientists cling to their data," said David J. Dooling, assistant director of the genome institute at Washington University in St. Louis.

But there is now so much raw data that it is becoming not feasible to re-analyze it. So researchers will increasingly store just the final results. In the case of human genomes, they might store even less — only the difference between a particular genome and some reference genome.

Professor Brown of Michigan State said: "We are going to have to come up with really clever ways to throw away data so we can see new stuff."

A version of this article appeared in print on December 1, 2011, on page B1 of the New York edition with the headline: A Genome Deluge.