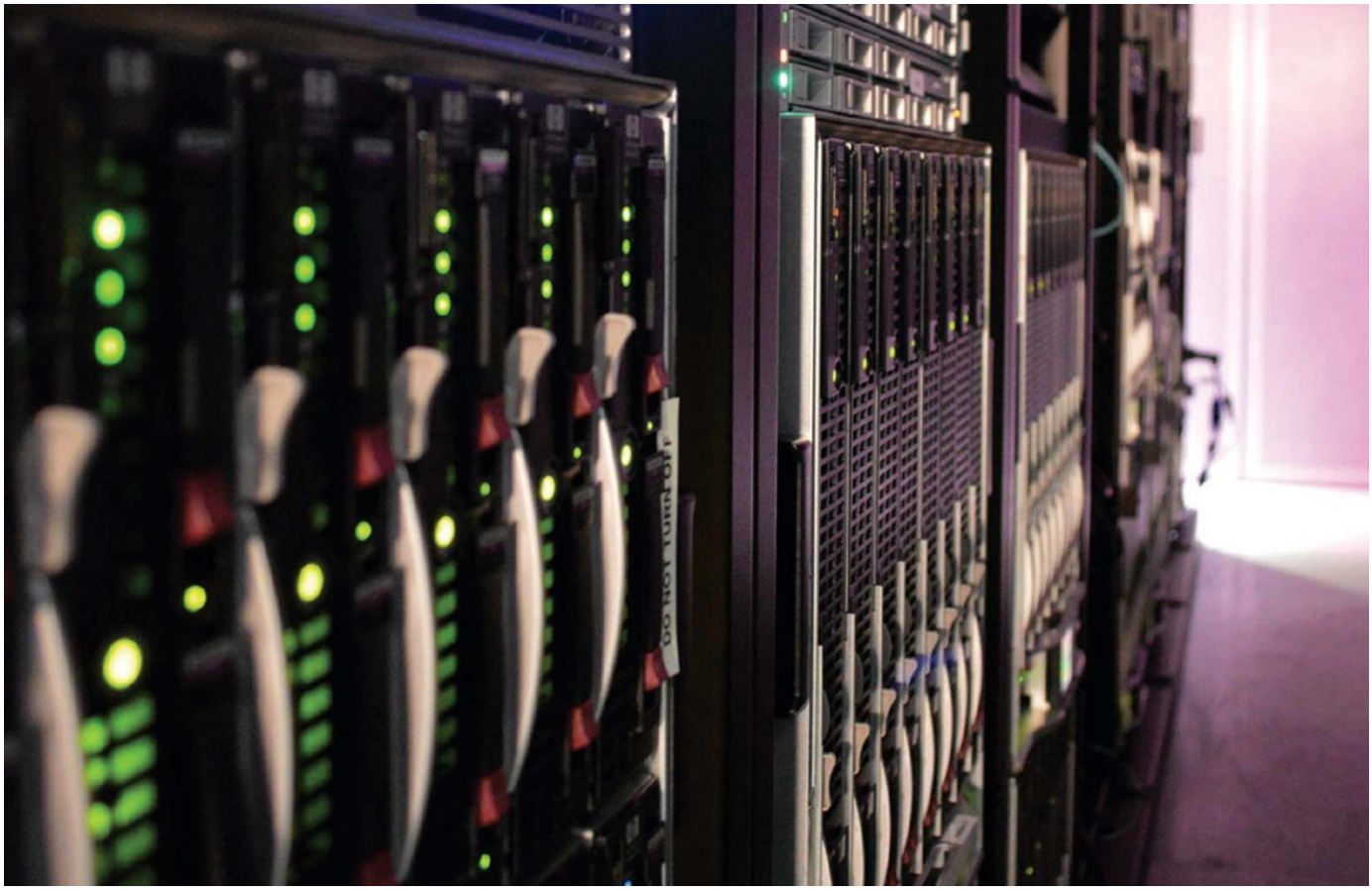
TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

As they grapple with increasingly large data sets, biologists and computer scientists uncork new bottlenecks.

EMBL-EBI



Extremely powerful computers are needed to help biologists to handle big-data traffic jams.

BY VIVIEN MARX

Biologists are joining the big-data club. With the advent of high-throughput genomics, life scientists are starting to grapple with massive data sets, encountering challenges with handling, processing and moving information that were once the domain of astronomers and high-energy physicists[1].

With every passing year, they turn more often to big data to probe everything from the regulation of genes and the evolution of genomes to why coastal algae bloom, what microbes dwell where in human body cavities and how the genetic make-up of different cancers influences how cancer patients fare[2]. The European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 petabytes (1 petabyte is 10^{15} bytes) of data and back-ups about genes, proteins and small molecules. Genomic data account for 2 petabytes of that, a number that more than doubles every year[3] (see 'Data explosion').

This data pile is just one-tenth the size of the data store at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. Every year, particle-collision events in CERN's Large Hadron Collider generate around 15 petabytes of data — the equivalent of about 4 million high-definition feature-length films. But the EBI and institutes like it face similar data-wrangling challenges to those at CERN, says Ewan Birney, associate director of the EBI. He and his colleagues now regularly meet with organizations such as CERN and the European Space Agency (ESA) in Paris to swap lessons about data storage, analysis and sharing.

All labs need to manipulate data to yield research answers. As prices drop for high-throughput instruments such as automated ▶

DNANEXUS

genome sequencers, small biology labs can become big-data generators. And even labs without such instruments can become big-data users by accessing terabytes (10^{12} bytes) of data from public repositories at the EBI or the US National Center for Biotechnology Information in Bethesda, Maryland. Each day last year, the EBI received about 9 million online requests to query its data, a 60% increase over 2011.

Biology data mining has challenges all of its own, says Birney. Biological data are much more heterogeneous than those in physics. They stem from a wide range of experiments that spit out many types of information, such as genetic sequences, interactions of proteins or findings in medical records. The complexity is daunting, says Lawrence Hunter, a computational biologist at the University of Colorado Denver. "Getting the most from the data requires interpreting them in light of all the relevant prior knowledge," he says.

That means scientists have to store large data sets, and analyse, compare and share them — not simple tasks. Even a single sequenced human genome is around 140 gigabytes in size. Comparing human genomes takes more than a personal computer and online file-sharing applications such as DropBox.

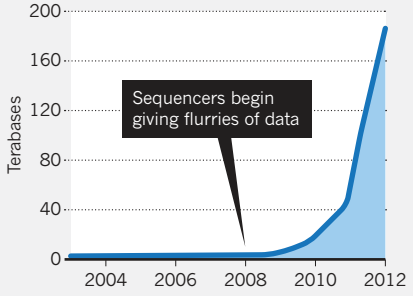
In an ongoing study, Arend Sidow, a computational biologist at Stanford University in California, and his team are looking at specific changes in the genome sequences of tumours from people with breast cancer. They wanted to compare their data with the thousands of other published breast-cancer genomes and look for similar patterns in the scores of different cancer types. But that is a tall order: downloading the data is time-consuming, and researchers must be sure that their computational infrastructure and software tools are up to the task. "If I could, I would routinely look at all sequenced cancer genomes," says Sidow. "With the current infrastructure, that's impossible."

In 2009, Sidow co-founded a company called DNAnexus in Mountain View, California, to help with large-scale genetic analyses. Numerous other commercial and academic efforts also address the infrastructure needs of big-data biology. With the new types of data traffic jam honking for attention, "we now have non-trivial engineering problems", says Birney,

SOURCE: EMBL–EBI

DATA EXPLOSION

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.



Andreas Sundquist says amounts of data are now larger than the tools used to analyse them.

LIFE OF THE DATA-RICH

Storing and interpreting big data takes both real and virtual bricks and mortar. On the EBI campus, for example, construction is under way to house the technical command centre of ELIXIR, a project to help scientists across Europe safeguard and share their data, and to support existing resources such as databases and computing facilities in individual countries. Whereas CERN has one supercollider producing data in one location, biological research generating high volumes of data is distributed across many labs — highlighting the need to share resources.

Much of the construction in big-data biology is virtual, focused on cloud computing — in which data and software are situated in huge, off-site centres that users can access on demand, so that they do not need to buy their own hardware and maintain it on site. Labs that do have their own hardware can supplement it with the cloud and use both as needed. They can create virtual spaces for data, software and results that anyone can access, or they can lock the spaces up behind a firewall so that only a select group of collaborators can get to them.

Working with the CSC — IT Center for Science in Espoo, Finland, a government-run high-performance computing centre, the EBI is developing Embassy Cloud, a cloud-computing component for ELIXIR that offers secure data-analysis environments and is currently in its pilot phase. External organizations can, for example, run data-driven experiments in the EBI's computational environment, close to the data they need. They can also download data to compare with their own.

The idea is to broaden access to computing power, says Birney. A researcher in the Czech Republic, for example, might have an idea about how to reprocess cancer data to help the hunt for cancer drugs. If he or she lacks the computational equipment to develop it, he or she might not even try. But access to a high-powered cloud allows "ideas to come from any place", says Birney.

Even at the EBI, many scientists access databases and software tools on the Web and through clouds. "People rarely work on straight hardware anymore," says Birney. One heavily used resource is the Ensembl Genome Browser, run jointly by the EBI and the Wellcome Trust Sanger Institute in Hinxton. Life scientists use it to search through, download and analyse genomes from armadillo to zebrafish. The main Ensembl site is based on hardware in the United Kingdom, but when users in the United States and Japan had difficulty accessing the data quickly, the EBI resolved the bottleneck by hosting mirror sites at three of the many remote data centres that are part of Amazon Web Services' Elastic Compute Cloud (EC2). Amazon's data centres are geographically closer to the users than the EBI base, giving researchers quicker access to the information they need.

More clouds are coming. Together with CERN and ESA, the EBI is building a cloud-based infrastructure called Helix Nebula — The Science Cloud. Also involved are information-technology companies such as Atos in Bezons, France; CGI in Montreal, Canada; SixSq in Geneva; and T-Systems in Frankfurt, Germany.

"If I could, I would routinely look at all sequenced cancer genomes. With the current infrastructure, that's impossible."

Cloud computing is particularly attractive in an era of reduced research funding, says Hunter, because cloud users do not need to finance or maintain hardware. In addition to academic cloud projects, scientists can choose from many commercial providers, such as Rackspace, headquartered in San Antonio, Texas, or VMware in Palo Alto, California, as well as larger companies including Amazon, headquartered in Seattle, Washington, IBM in Armonk, New York, or Microsoft in Redmond, Washington.

BIG-DATA PARKING

Clouds are a solution, but they also throw up fresh challenges. Ironically, their proliferation can cause a bottleneck if data end up parked on several clouds and thus still need to be moved to be shared. And using clouds means entrusting valuable data to a distant service provider who may be subject to power outages or other disruptions. "I use cloud services for many things, but always keep a local copy of scientifically important data and software," says Hunter. Scientists experiment with different constellations to

suit their needs and trust levels.

Most researchers tend to download remote data to local hardware for analysis. But this method is "backward", says Andreas Sundquist, chief technology officer of DNAnexus. "The data are so much larger than the tools, it makes no sense to be doing that." The alternative is to use the cloud for both data storage and computing. If the data are on a cloud, researchers can harness both the computing power and the tools that they need online, without the need to move data and software (see 'Head in the clouds'). "There's no reason to move data outside the cloud. You can do analysis right there," says Sundquist. Everything required is available "to the clever people with the clever ideas", regardless of their local computing resources, says Birney.

Various academic and commercial ventures are engineering ways to bring data and analysis tools together — and as they build, they have to address the continued data growth. Xing Xu, director of cloud computing at BGI (formerly the Beijing Genomics Institute) in Shenzen, China, knows that challenge well. BGI is one of the largest producers of genomic data in the world, with 157 genome sequencing instruments working around the clock on samples from people, plants, animals and microbes. Each day, it generates 6 terabytes of genomic data. Every instrument can decode one human genome per week, an effort that used to take months or years and many staff.

DATA HIGHWAY

Once a genome sequencer has cranked out its snippets of genomic information, or 'reads', they must be assembled into a continuous stretch of DNA using computing and software. Xu and his team try to automate as much of this process as possible to enable scientists to get to analyses quickly.

Next, either the reads or the analysis, or both, have to travel to scientists. Generally, researchers share biological data with their peers through public repositories, such as the EBI or ones run by the US National Center for Biotechnology Information in Bethesda, Maryland. Given the size of the data, this travel often means physically delivering hard drives — and risks data getting lost, stolen or damaged. Instead, BGI wants to use either its own clouds or others of the customer's choosing for electronic delivery. But that presents a problem, because big-data travel often means big traffic jams.

Currently, BGI can transfer about 1 terabyte per day to its customers. "If you transfer one genome at a time, it's OK," says Xu. "If you sequence 50, it's not so practical for us to transfer that through the Internet. That takes about 20 days."

BGI is exploring a variety of technologies to accelerate electronic data transfer, among them *fasp*, software developed by Aspera in Emeryville, California, which helps to deliver data for film-production studios and the oil and gas industry as well as the life sciences. In an experiment last year, BGI tested a *fasp*-enabled data transfer between China and the University of California, San Diego (UCSD). It took 30 seconds to move a 24-gigabyte file. "That's really fast," says Xu.

Data transfer with *fasp* is hundreds of times quicker than methods using the normal Internet protocol, says software engineer Michelle Munson, chief executive and co-founder of Aspera. However, all transfer protocols share challenges associated with transferring large, unstructured data sets.

The test transfer between BGI and UCSD was encouraging because Internet connections between China and the United States are "riddled with challenges" such as variations in signal strength that interrupt data transfer, says Munson. The protocol has to handle such road bumps and ensure speedy transfer, data integrity and privacy. Data transfer often slows when the passage is bumpy, but with *fasp* it does not. Transfers can fail when a file is partially sent; with ordinary Internet connections, this relaunches the entire transfer. By contrast, *fasp* restarts where the previous transfer stopped. Data that are already on their way do not get resent, but continue on their travels.

"There's no reason to move data outside the cloud. You can do analysis right there."

Xu says that he liked the experiment with *fasp*, but the software does not solve the data-transfer problem. "The main problem is not technical, it is economical," he says. BGI would need to maintain a large Internet connection bandwidth for data transfer, which would be prohibitively expensive, especially given that Xu and his team do not send out big data in a continuous flow. "If we only transfer periodically, it doesn't make any economic sense for us to have this infrastructure, especially if the user wants that for free," he says.

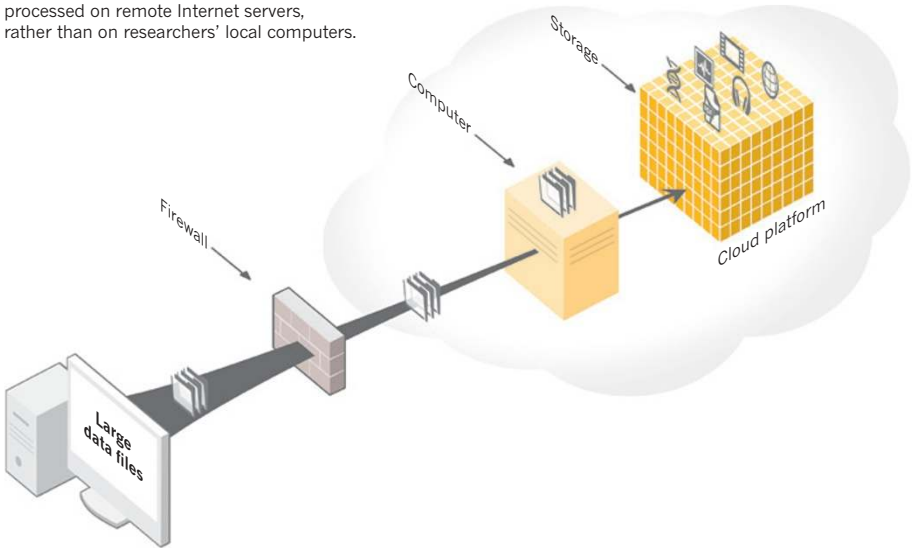
Data-sharing among many collaborators also remains a challenge. When BGI uses *fasp* to share data with customers or collaborators, it must have a software licence, which allows customers to download or upload the data for free. But customers who want to share data with each other using this transfer protocol will need their own software licences. Putting the data on the cloud and not moving them would bypass this problem; teams would go to the large data sets, rather than the other way around. Xu and his team are exploring this approach, alongside the use of Globus Online, a free Web-based file-transfer service from the Computation Institute at the University of Chicago and the Argonne National Laboratory in Illinois. In April, the Computation Institute team launched a genome-sequencing-analysis service called Globus Genomics on the Amazon cloud.

Munson says that Aspera has set up a pay-as-you-go system on the Amazon cloud to address the issue of data-sharing. Later this year, the company will begin selling an updated version of its software that can be embedded on the desktop of any kind of computer and will let users browse large data sets much like a file-sharing application. Files can be dragged and dropped from one location to another, even if those locations are commercial or academic clouds.

The cost of producing, acquiring and disseminating data is decreasing, says James Taylor, a computational biologist at Emory University in Atlanta, Georgia, who thinks that "everyone should have access to the skills and tools" needed to make sense of all the information. Taylor is a co-founder of an academic platform called Galaxy, which lets scientists analyse their data and share software tools and workflows for free. Through

HEAD IN THE CLOUDS

In cloud computing, large data sets are processed on remote Internet servers, rather than on researchers' local computers.



SOURCE: ASPERA

Web-based access to computing facilities at Pennsylvania State University (PSU) in University Park, scientists can download Galaxy's platform of tools to their local hardware, or use it on the Galaxy cloud. They can then plug in their own data, perform analyses and save the steps in them, or try out workflows set up by their colleagues.

Spearheaded by Taylor and Anton Nekrutenko, a molecular biologist at PSU, the Galaxy project draws on a community of around 100 software developers. One feature is Tool Shed, a virtual area with more than 2,700 software tools that users can upload, try out and rate. Xu says that he likes the collection and its ratings, because without them, scientists must always check if a software tool actually runs before they can use it.

KNOWLEDGE IS POWER

Galaxy is a good fit for scientists with some computing know-how, says Alla Lapidus, a computational biologist in the algorithmic biology lab at St Petersburg Academic University of the Russian Academy of Sciences, which is led by Pavel Pevzner, a computer scientist at UCSD. But, she says, the platform might not be the best choice for less tech-savvy researchers. When Lapidus wanted to disseminate the software tools that she developed, she chose to put them on DNAnexus's newly launched second-generation commercial cloud-based analysis platform.

That platform is also designed to cater to non-specialist users, says Sundquist. It is possible for a computer scientist to build his or her own biological data-analysis suite with software tools on the Amazon cloud, but DNAnexus uses its own engineering to help researchers without the necessary computer skills to get to the analysis steps.

Catering for non-specialists is important when developing tools, as well as platforms. The Biomedical Information Science and Technology Initiative (BISTI) run by the US National Institutes of Health (NIH) in Bethesda, Maryland, supports development of new computational tools and the maintenance of existing ones. "We want a deployable tool," says Vivien Bonazzi, programme director in computational biology and bioinformatics at the National Human Genome Research Institute, who is involved with BISTI. Scientists who are not heavy-duty informatics types need to be able to set up these tools and use them successfully, she says. And it must be possible to scale up tools and update them as data volume grows.

Bonazzi says that although many life scientists have significant computational skills, others do not understand computer lingo enough to know that in the tech world, Python is not a snake and Perl is not a gem (they are programming languages). But even if biologists can't develop or adapt the software, says Bonazzi, they have a place in big-data science. Apart from anything else, they can offer valuable feedback to their computationally fluent colleagues because of different needs and approaches to the science, she says.

Increasingly, big genomic data sets are being used in biotechnology companies, drug firms and medical centres, which also have specific needs. Robert Mulroy, president of Merrimack Pharmaceuticals in Cambridge, Massachusetts, says that his teams handle mountains of data that hide drug candidates. "Our view is that biology functions through systems dynamics," he says.

Merrimack researchers focus on interrogating molecular signalling networks in the healthy body and in tumours, hoping to find new ways to corner cancer cells. They generate and use large amounts of information from the genome and other factors that drive a cell to become cancerous, says Mulroy. The company stores its data and conducts analysis on its own computing infrastructure, rather than a cloud, to keep the data private and protected.

Drug developers have been hesitant about cloud computing. But, says Sundquist, that fear is subsiding in some quarters: some companies that have previously avoided clouds because of security problems are now exploring them. To assuage these users' concerns, Sundquist has engineered the DNAnexus cloud to be compliant with US and European regulatory guidelines. Its security features include encryption for biomedical information, and logs to allow users to address potential queries from auditors such as regulatory agencies, all of which is important in drug development.

CHALLENGES AND OPPORTUNITIES

Harnessing powerful computers and numerous tools for data analysis is crucial in drug discovery and other areas of big-data biology. But that is only part of the problem. Data and tools need to be more than close — they must talk to one another. Lapidus says that results produced by one tool are not always in a format that can be used by the next tool in a workflow. And if software tools are not easily installed, computer specialists will have to intervene on behalf of those biologists without computer skills.

ASPERA

Various data-transfer protocols handle problems in different ways, says Michelle Munson.

STANFORD UNIV./DNANEXUS

Arend Sidow wants to move data mountains without feeling pinched by infrastructure.

Even computationally savvy researchers can get tangled up when wrestling with software and big data. "Many of us are getting so busy analysing huge data sets that we don't have time to do much else," says Steven Salzberg, a computational biologist at Johns Hopkins University in Baltimore, Maryland. "We have to spend some of our time figuring out ways to make the analysis faster, rather than just using the tools we have."

Yet other big-data pressures come from the need to engineer tools for stability and longevity. Too many software tools crash too often. "Everyone in the field runs into similar problems," says Hunter. In addition, research teams may not be able to acquire the resources they need, he says, especially in countries such as the United States, where an academic does not gain as much recognition for software engineering as for publishing a paper. With its dedicated focus on data and software infrastructure designed to serve scientists, the EBI offers an "interesting contrast to the US model", says Hunter.

US funding agencies are not entirely ignoring software engineering, however. In addition to BISTI, the NIH is developing Big Data to Knowledge (BD2K), an initiative focused on managing large data sets in biomedicine, with elements such as data handling and standards, informatics training and software sharing. And as the cloud emerges as a popular place to do research, the agency is also reviewing data-use policies. An approved study usually lays out specific data uses, which may not include placing genomic data on a cloud, says Bonazzi. When a person consents to have his or her data used in one way, researchers cannot suddenly change that use, she says. In a big-data age that uses the cloud in addition to local hardware, new technologies in encryption and secure

transmission will need to address such privacy concerns.

Big data takes large numbers of people. BGI employs more than 600 engineers and software developers to manage its information-technology infrastructure, handle data and develop software tools and workflows. Scores of informaticians look for biologically relevant messages in the data, usually tailored to requests from researchers and commercial customers, says Xu. And apart from its stream of research collaborations, BGI offers a sequencing and analysis service to customers. Early last year, the institute expanded its offerings with a cloud-based genome-analysis platform called EasyGenomics.

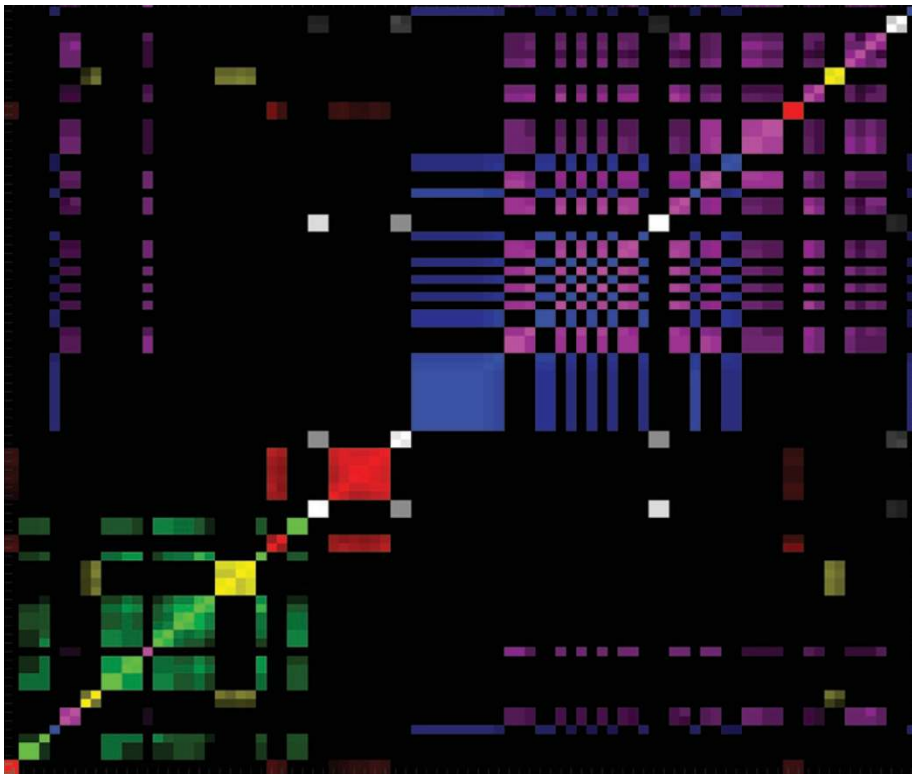
In late 2012, it also bought the faltering US company Complete Genomics (CG), which offered human genome sequencing and analysis for customers in academia or drug discovery. Although the sale dashed hopes for earnings among CG's investors, it doesn't seem to have dimmed their view of the prospects for sequencing and analysis services. "It is now just a matter of time before sequencing data are used with regularity in clinical practice," says one investor, who did not wish to be identified. But the sale shows how difficult it can be to transition ideas into a competitive marketplace, the investor says.

When tackling data mountains, BGI uses not only its own data-analysis tools, but also some developed in the academic community. To ramp up analysis speed and capacity as data sets grow, BGI assembled a cloud-based series of analysis steps into a workflow called Gaea, which uses the Hadoop open-source software framework. Hadoop was written by volunteer developers from companies and universities, and can be deployed on various types of computing infrastructure. BGI programmers built on this framework to instruct software tools to perform large-scale data analysis across many computers at the same time.

"The cultural baggage of biology that privileges data generation over all other forms of science is holding us back."

If 50 genomes are to be analysed and the results compared, hundreds of computational steps are involved. The steps can run either sequentially or in parallel; with Gaea, they run in parallel across hundreds of cloud-based computers, reducing analysis time rather like many people working on a single large puzzle at once. The data are on the BGI cloud, as are the tools. "If you perform analysis in a non-parallel way, you will maybe need two weeks to fully process those data," says Xu. Gaea takes around 15 hours for the same number of data.

To leverage Hadoop's muscle, Xu and his team needed to rewrite software tools. But the investment is worth it because the Hadoop framework allows analysis to continue as the data mountains grow, he says.



MERRIMACK PHARMACEUTICALS

A simplified array of breast-cancer subtypes, produced by researchers at Merrimack Pharmaceuticals, who use their own computational infrastructure to hunt for new cancer drugs.

They are still ironing out some issues with Gaea, comparing its performance on the cloud with its performance on local infrastructure. Once testing is complete, BGI plans to mount Gaea on a cloud such as Amazon for use by the wider scientific community.

Other groups are also trying to speed up analysis to cater to scientists who want to use big data. For example, Bina Technologies in Redwood City, California, a spin-out from Stanford University and the University of California, Berkeley, has developed high-performance computing components for its genome-analysis services. Customers can buy the hardware, called the Bina Box, with software, or use Bina's analysis platform on the cloud.

FROM VIVO TO SILICO

Data mountains and analysis are altering the way science progresses, and breeding biologists who get neither their feet nor their hands wet. "I am one of a small original group who made the first leap from the wet world to the *in silico* world to do biology," says Marcie McClure, a computational biologist at Montana State University in Bozeman. "I never looked back."

During her graduate training, McClure analysed a class of viruses known as retroviruses in fish, doing the work of a "wet-worlder", as she calls it. Since then, she and her team have discovered 11 fish retroviruses without touching water in lake or lab, by analysing genomes computationally and in ways that others had not. She has also developed software tools to find such viruses in the genomes of other species, including humans. Her work generates terabytes of data, which she shares with other researchers.

Given that big-data analysis in biology is incredibly difficult, Hunter says, open science is becoming increasingly important. As he explains, researchers need to make their data available to the scientific community in a useful form, for others to mine. New science can emerge from the analysis of existing data sets: McClure generates some of her findings from other people's data. But not everyone recognizes that kind of biology as an equal. "The cultural baggage of biology that privileges data generation over all other forms of science is holding us back," says Hunter.

A number of McClure's graduate students are microbial ecologists, and she teaches them how to rethink their findings in the face of so many new data. "Before taking my class, none of these students would have imagined that they could produce new, meaningful knowledge, and new hypotheses, from existing data, not their own," she says. Big data in biology add to the possibilities for scientists, she says, because data sit "under-analysed in databases all over the world". ■

Vivien Marx *is technology editor at* Nature *and* Nature Methods.

1. Mattmann, C. *Nature* **493,** 473–475 (2013).
2. Greene, C. S. & Troyanskaya, O. G. *PLoS Comput. Biol.* **8,** e1002816 (2012).
3. EMBL–European Bioinformatics Institute *EMBL-EBI Annual Scientific Report 2012* (EMBL–EBI, 2013).