BRENDAN MONROE

BIOINFORMATICS

# Big data versus the big C

*The torrents of data flowing out of cancer research and treatment are yielding fresh insight into the disease.*

BY NEIL SAVAGE

In 2013, geneticist Stephen Elledge answered a question that had puzzled cancer researchers for nearly 100 years. In 1914, German biologist Theodor Boveri suggested that the abnormal number of chromosomes — called aneuploidy — seen in cancers might drive the growth of tumours. For most of the next century, researchers made little progress on the matter. They knew that cancers often have extra or missing chromosomes or pieces of chromosomes, but they did not know whether this was important or simply a by-product of tumour growth — and they had no way of finding out.

"People had ignored it for a long time, primarily because it's really hard to understand," says Elledge, of Brigham and Women's Hospital in Boston, Massachusetts. "What we didn't know before is that it's actually driving cancer."

Elledge found that where aneuploidy had resulted in missing tumour-suppressor genes, or extra copies of the oncogenes that promote cancer, tumours grow more aggressively (T. Davoli *et al. Cell* **155,** 948–962; 2013). His insight — that aneuploidy is not merely an odd feature of tumours, but an engine of their growth — came from mining voluminous amounts of cellular data. And, says Elledge, it shows how the ability of computers to sift through ever-growing troves of information can help us to deepen our understanding of cancer and open the door to discoveries.

Modern cancer care has the potential to generate huge amounts of data. When a patient is diagnosed, the tumour's genome might be sequenced to see if it is likely to respond to a particular drug. The sequencing might be repeated as treatment progresses to detect changes. The patient might have his or her normal tissue sequenced as well, a practice that is likely to grow as costs come down. The doctor will record the patient's test results and medical history, including dietary and smoking habits, in an electronic health record. The patient may also have computed tomography (CT) and magnetic resonance imaging (MRI) scans to determine the stage of the disease. Multiply all that by the nearly 1.7 million people diagnosed with cancer in 2013 in the United States alone and it becomes clear that oncology is going to generate even more data than it does now. Computers can mine the data for patterns that may advance the understanding of cancer biology and suggest targets for therapy.

Elledge's discovery was the result of a computational method that he and his colleagues developed called the Tumor Suppressor and Oncogene Explorer. They used it to mine large data sets, including the Cancer Genome Atlas, maintained by the US National Cancer Institute, based in Bethesda, Maryland, and the Catalogue of Somatic Mutations in Cancer, run by the Wellcome Trust Sanger Institute in Hinxton, UK. The databases contained roughly 1.2 million mutations from 8,207 tissue samples of more than 20 types of tumour.

The researchers selected a set of parameters that helped to identify the genes they were looking for, such as the mutation rate or the ratio of benign mutations to those that cause a gene to stop functioning. They then applied statistical classification methods to differentiate between suppressor genes and oncogenes. About 70 suppressor genes and 50 oncogenes were already known for these tumour types, but Elledge and his colleagues increased that to about 320 and 200, respectively (although

➲ NATURE.COM
Find out more about software design for bioinformatics:
go.nature.com/auca5y

that number could fall, because some genes could turn out to be false positives). They also identified pathways in the growth process that might make good drug targets.

Making this sort of finding requires large data sets. "Any individual cancer cell's a mess, but if you look at enough tumours, you get a pattern," Elledge says. "The only way you can figure this out is if you look at them globally."

## EASY TO USE

Analysing the genomes of 8,200 tumours is just a start. Researchers are "trying to figure out how we can bring together and analyse, over the next few years, a million genomes", says Robert Grossman, who directs the Initiative in Data Intensive Science at the University of Chicago in Illinois. This is an immense undertaking; the combined cancer genome and normal genome from a single patient constitutes about 1 terabyte ($10^{12}$ bytes) of data, so a million genomes would generate an exabyte ($10^{18}$ bytes). Storing and analysing this much data could cost US$100 million a year, Grossman says.

To make it easier to access whatever subset of data researchers need, Grossman and his colleagues have developed Bionimbus, a cloud-based, open-source platform for sharing and analysing genomic data from the Cancer Genome Atlas.

The results can be powerful. Megan McNerney, a pathologist at the University of Chicago, used Bionimbus to track down a gene involved in acute myeloid leukaemia (AML). Scientists already knew that some patients with the disease had lost part of chromosome 7, but could narrow down the gene involved only to 15–20 candidates. McNerney selected 23 patients from the database and used the computer to compare their RNA sequences to see if something might be missing. She discovered that one copy of the gene *CUX1*, which normally encodes a tumour-suppressor protein, had been deleted in these patients (M. E. McNerney *et al. Blood* **121,** 975–983; 2012). Testing in fruit flies and mice showed that removal of one copy of the gene led to an overgrowth of certain blood cells and, eventually, to leukaemia. Her discovery may not have produced a cure for AML, but it has increased the understanding of a disease for which the median survival time has been stuck at less than a year for four decades, and it might also lead to more-accurate prognoses.

McNerney says that even her small-scale project has shown the benefits of mining data. "It's transforming cancer biology enormously," she says. "Big data has made leaps that we couldn't make otherwise."

Genomics — and data from other '-omics, such as proteomics and epigenomics — are not the only sources of data being sifted. The American Society of Clinical Oncology (ASCO) in Alexandria, Virginia, is developing a platform called CancerLinQ, which trawls through patients' electronic health records. These records increasingly include genomic data, as well as diagnoses and notes on treatment, and measures of how well patients are responding to therapy. The system has gathered records from 177,000 people with breast cancer for a pilot project. Developers hope that the system will be fully operational by the summer of 2015, with other solid tumours to follow.

Clifford Hudis, a breast-cancer specialist at the Memorial Sloan Kettering Cancer Center in New York and president of ASCO, says that CancerLinQ could make discoveries missed by clinical trials. As approved drugs are deployed more widely, the system could gather data on side effects, drug interactions and outcomes in different patient populations. It might also notice, for instance, if doctors stray from US Food and Drug Administration guidelines for drug dosage, based on their assessment of how the dose affects their patients. "If there are 100 cases in a row of doctors independently disregarding the guideline, it helps to teach the computer that the guideline's wrong," Hudis says. The computer might discover, for instance, that doctors get better results when they adjust the dosage according to the patient's age.

> ***"There are some fundamental challenges being caused by our ability to capture so much data."***

Discoveries can also be made from combining genomics and standard medical-imaging records. "High-performance computing and big data are enabling us to look across modalities," says David Foran, a pathologist and head of informatics at the Rutgers Cancer Institute of New Jersey in New Brunswick. The centre produces high-resolution digital images of tissue samples and compares them between patients, looking for patterns that might aid prognosis. It expects to generate 40,000–100,000 images.

Researchers might see genetic clues indicating that some patients will respond to a particular drug therapy, for instance, and then look at their CT and MRI scans to see whether changes in the cancer match up with the genetic prediction. Or they might find a correlation between mutations, therapy choice and smoking history. "The computer program can simultaneously look at the patterns in all of them," Foran says.

Comparing so much data greatly expands doctors' expertise, Foran adds. "When you go to see a physician, especially an oncologist, you're relying on his past experience. What we're doing now is training the computer to look at large cohorts of thousands and hundreds of thousands." It is as if the doctor were making treatment decisions based on personal experience of hundreds of thousands of patients.

Gene sequences and electronic health records are new sources of data, but there is a lot of historical information available, too. Johns Hopkins Hospital in Baltimore, Maryland, for instance, has paper-based pathology reports that date back to its opening in 1889. Before it switched to computer records in 1984, the hospital generated more than half-a-million records. Every US state has years or decades of historical cancer records, as do other countries. Denmark, for instance, has cancer records going back to 1943. And Public Health England last year launched a database of all cancers currently being diagnosed across the country, including 11 million records going back 30 years. Adding all that history into the mix widens the field of possible clues that computers can search through.

## HARD TO ANALYSE

But it is the new technologies that are creating an information boom. "We can collect data faster than we can physically do anything with them," says Manish Parashar, a computer scientist and head of the Rutgers Discovery Informatics Institute in Piscataway, New Jersey, who collaborates with Foran to find ways of handling the information. "There are some fundamental challenges being caused by our ability to capture so much data," he says.

A major problem with data sets at the terabyte-and-beyond level is figuring out how to manipulate all the data. A single high-resolution medical image can take up tens of gigabytes, and a researcher might want the computer to compare tens of thousands of such images. Breaking down just one image in the Rutgers project into sets of pixels that the computer can identify takes about 15 minutes, and moving that much information from where it is stored to where it can be processed is difficult. "Already we have people walking around with disk drives because you can't effectively use the network," Parashar says.

Informatics researchers are developing algorithms to split data into smaller packets for parallel processing on separate processors, and to compress files without omitting any relevant information. And they are relying on advances in computer science to speed up processing and communications in general.

Foran emphasizes that the understanding and treatment of cancer has undergone a dramatic shift as oncology has moved from one-size-fits-all attacks on tumours towards personalized medicine. But cancers are complex diseases controlled by many genes and other factors. "It's not as if you're going to solve cancer," he says. But big data can provide new, better-targeted ways of grappling with the disease. "You're going to come up with probably a whole new set of blueprints for how to treat patients." ■

**Neil Savage** *is a freelance science and technology writer based in Lowell, Massachusetts.*