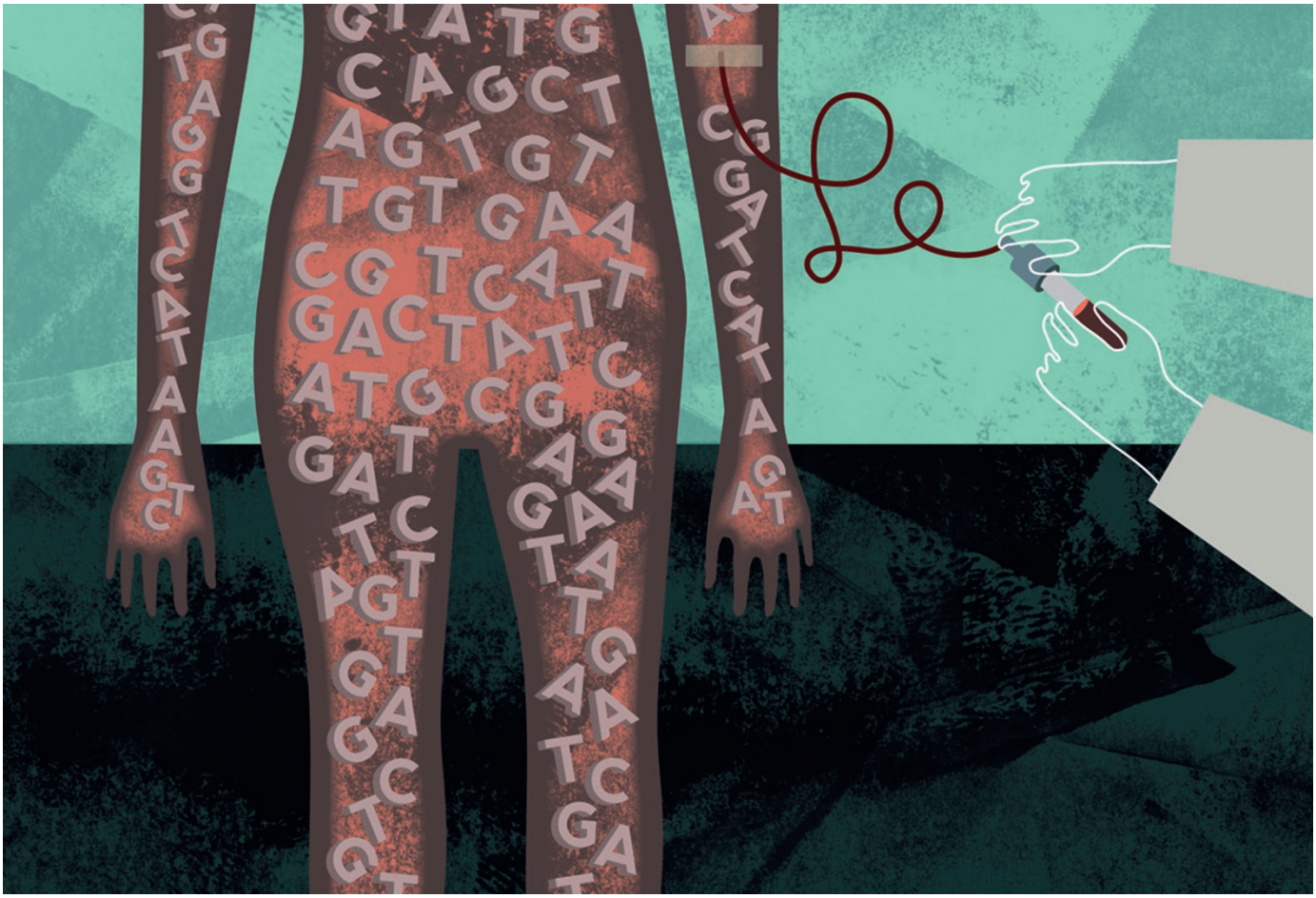

ELIN SVENSSON

GENETICS

Big hopes for big data

Technology is allowing researchers to generate vast amounts of information about tumours. The next step is to use this genomic data to transform patient care.

BY JILL U. ADAMS

Adrian Lee has dedicated his career to studying breast cancer, which is to say he is actually tackling many different diseases at once. "No two breast cancers are the same," says Lee, a pharmacologist and chemical biologist at the University of Pittsburgh in Pennsylvania. "Cancer is way more complex than we know."

Lee is using genomic technology to fully describe cancers of the breast and apply that knowledge to guide treatment decisions for individual patients. "We can now analyse multiple variables from a single specimen, such as changes in DNA, changes in RNA and changes in methylation," he says. "Genome-wide scans allow for better systems biology and allow us to learn what's gone wrong in a particular tumour."

Sequencing tumours is faster, cheaper and easier than ever. With many researchers collecting sequence data and uploading these to public databases such as the The Cancer Genome Atlas (TCGA), opportunities to describe the many different cancers that arise in breast tissue are upon us. "The challenge used to be generating the data," says Nicholas Navin, a geneticist at The University of Texas MD Anderson Cancer Center in Houston. "Those issues have been resolved. Now the challenge is data processing and data analysing — interpreting the mutations and communicating those to oncologists."

At the University of Pittsburgh, researchers are working to link the molecular signatures of people with breast cancer to a host of clinical data, including demographic information associated with risk such as age, ethnicity and body weight. They are mining electronic health records for clinical correlates, treatment interactions and outcomes. "We've got a big haystack and we're trying to find the needle," says Lee. "But we're also trying to incriminate the needle, by linking it to lots of things." Collecting all that data from patients' electronic records adds up, Lee says. It takes infrastructure — Pittsburgh has already accumulated 5 petabytes, or 5 million gigabytes, which is enough data to overload around 40,000 new iPhone 6 devices.

Making the connection between the reams of data coming out of sequencing laboratories and the individual women fighting breast cancer takes big-time computing power. Big data needs researchers who are comfortable with statistical noise and those who are old hands at the iterative process required to create flexible computer programs.

FROM DATA TO KNOWLEDGE

Big-data researchers take a large data set and look for patterns. The idea is to identify

mutations that can be targeted with drug treatment. It is the essence of personalized medicine: screen a patient's tumour for a set of biomarkers to choose the best treatment to fight the cancer. Big-data researchers believe that analysing the data of the thousands of tumours that have come before will reveal patterns that can improve screening and diagnosis, and inform treatment.

Lee and his colleagues have illustrated how big-data science led to a rethink of breast cancer[1]. They used two public databases — TCGA and METABRIC (Molecular Taxonomy of Breast Cancer International Consortium), which contain data on the entire set of genes, RNA transcripts and proteins of thousands of breast-cancer tumours — to parse out potential differences in the molecular signatures of breast tumours in younger compared with older women. Women who are diagnosed before the age of 40 tend to have worse disease: they are more likely to have later-stage cancers, poorer prognoses and worse survival outcomes than older women.

The team analysed tumour data from women under 45 years old, who were probably premenopausal, and women over 55 years old, who were probably postmenopausal. "We looked at everything you can look at," Lee says, including mutations in the genome, mutations in RNA, tumour gene expression, variations in the number of copies of certain genes and levels of DNA methylation. They found that tumours in premenopausal women follow a different playbook, especially in terms of gene expression.

As researchers find rarer and rarer mutations, the question of significance becomes more and more daunting, Lee says. He has just finished looking at a spreadsheet of 2,000 mutations. "One of them is the ER mutation," he says, referring to a mutation in the oestrogen receptor — a common mutation in breast cancers. "But how do I sift through the others? That's the fundamental problem."

One way to do it is to analyse the cellular pathways that the mutations affect. That means using algorithms developed to integrate all the collected molecular information and categorizing it into the common growth or cell-cycle pathways. Researchers can use this sorted information to describe tumours in terms of affected pathways rather than simply affected molecules. In one such effort, bioinformatician Josh Stuart of the University of California Santa Cruz developed a computational method that integrates a variety of genomic data sets with known cell-signalling pathways. "We know how gene circuits work in normal cells. Now we're asking, what got broken in this tumour cell?" Stuart says. "It's surprisingly successful."

Lee's group used the computational analysis PARADIGM in their study[1]. The approach proved particularly revealing for oestrogen-receptor-positive breast cancers in premenopausal women. The method demonstrated that although the individual molecules that showed abnormalities varied, they often occurred within a particular set of pathways that signal for integrins — proteins involved in the formation of tumour-associated blood vessels.

The evident importance of integrins in the tumours of premenopausal women with oestrogen-receptor-positive breast cancers suggests that these molecules could be a therapeutic target. "There are integrin inhibitors out there," Lee says, and some of them have been tested in clinical trials.

FROM KNOWLEDGE TO APPLICATION

As big-data researchers churn through large tumour databases looking for patterns of mutations, they are adding new categories of breast cancer. In 2012, two consortiums published papers on their data-driven approaches to breast-cancer genomics. The TCGA Network, made up of dozens of research institutions in the United States and Europe, came up with four overarching groupings of breast tumours based on genetic and epigenetic abnormalities[2]. They found that only three genes (*TP53*, *PIK3CA* and *GATA3*) were mutated in more than 10% of the samples, demonstrating that rare mutations are now an important part of breast-cancer typing. The METABRIC group, a consortium of UK and Canadian institutions, integrated genetic data — copy-number and gene-expression changes — with long-term clinical outcomes into 10 families of tumour types. Combined with clinical data, both these new groupings have the potential to allow oncologists to make better prognoses and treatment decisions[3].

> ***"We know how gene circuits work in normal cells. Now we're asking, what got broken in a tumour cell."***

"We're still refining our approach," says Oscar Rueda, a biostatistician at Cancer Research UK's Cambridge Research Institute, which is part of the METABRIC effort. They are now fully sequencing the 2,000 samples used in the research. Rueda says that the hope is to identify driver mutations, which have a role in the initiation of cancer. "There are a hundred different mechanisms by which cells go bad," he says.

Big-data approaches may eventually reveal cellular pathways that had previously been overlooked. Avi Ma'ayan of the Icahn School of Medicine at Mount Sinai is working on a pathway database to create a resource for future potential targets. His effort comes under the umbrella of the National Institutes of Health Library of Integrated Network-based Cellular Signatures (LINCS), which uses data generated at institutions such as the Broad Institute of Massachusetts Institute of Technology. High-throughput labs at the Broad Institute test a host of drugs — both experimental ones as well as those with regulatory approval — on ten different cell lines to study how the drugs interact with cellular activity.

"You get a signature of what happens to cells," Ma'ayan says. "And signatures can be queried for new uses of drugs." If clinical researchers want to turn off a particular cellular pathway in cancer, they could use Ma'ayan's database to search for drugs that have that action.

CLINICAL TRANSLATION

The next step is to apply the newly gained knowledge of actionable mutations to patient care. Research hospitals collect data on patients for their own care and to add to the knowledge base. At MD Anderson Cancer Center, for instance, people with a new cancer diagnosis are screened for a selection of cancer genes. "It's not the whole genome, but a panel of 200 genes with actionable mutations," Navin says. As research knowledge grows, so does the panel. In the past year, the original 200 genes have already expanded to 300, he says.

Navin's speciality is single-cell sequencing, which allows his lab to study tumour cells that are circulating in the blood. One might only collect 10 or 20 cells in a sample. "Previous analytic methods couldn't process such a small number of cells," he says. The single-cell approach opens the possibility that patients could be monitored over the course of treatment with a noninvasive test, such as a blood sample. Oncologists could then check if the tumour cells are responding to therapy or if resistance is emerging.

Big data intersects with the clinic in the form of I-SPY 2, a clinical trial of experimental breast-cancer drugs. "We're collecting real time data on patients," says Laura van't Veer, a molecular oncologist at the University of California San Francisco.

Patients are enrolled at diagnosis and, based on their tumour signature, placed into one of eight pre-defined types. The women are then treated with standard treatment and an experimental targeted drug, while van't Veer and her colleagues monitor which tumours respond to which targeted therapies. The goal is to evaluate biomarkers that improve response to targeted therapies. "With standard chemotherapy, we see 30–35% complete remission," says van't Veer. "Among our 8 subtypes, we sometimes get up to 50–60% remission."

Plenty of challenges lie ahead. A single tumour can host a baffling diversity of mutations, which change over time. Still, Ma'ayan remains the optimist. "The more money and effort we can throw at the problem, the more snapshots we can get. With better resolution, we can improve our understanding of the whole process," he says. "It's not infinite. Although it can feel like it." ■

Jill U. Adams *is a freelance science writer in Albany, New York.*

1. Liao, S. *et al. Breast Can. Res.* **17,** 104 (2015).
2. TCGA Network. *Nature* **490,** 61–70 (2012).
3. Curtis, C. *et al. Nature* **486,** 346–352 (2012).