Drilling into big cancer-genome data

Vivien Marx

Paving roads through data mountains, consortia are developing workflows and tools for widespread use.

Cancer geneticist Matthew Meyerson, who is at the Dana-Farber Cancer Institute and the Broad Institute of MIT and Harvard, tracks the many ways tumors wreak chaos in orderly cells. He wants to squeeze into his schedule a dedicated time period in Gad Getz's lab at the Broad Institute to hone his computational skills for analyzing data about cancer genomes.

Such collaborations could become more common as scientists dive into data sets generated by large consortia including The Cancer Genome Atlas (TCGA) Research Network and the International Cancer Genome Consortium (ICGC).

To shape an experiment, Getz suggests that scientists first look at existing data. However, this shift in habits is not an easy sell, and doubts about tools and computational approaches abound. To make choosing among the options easier for the community, Getz and colleagues at other institutions are comparing and benchmarking software tools and making analysis pipelines more accessible. Separately, companies are expanding the ways to help customers work through big cancer-genome data.

Where do I find data?

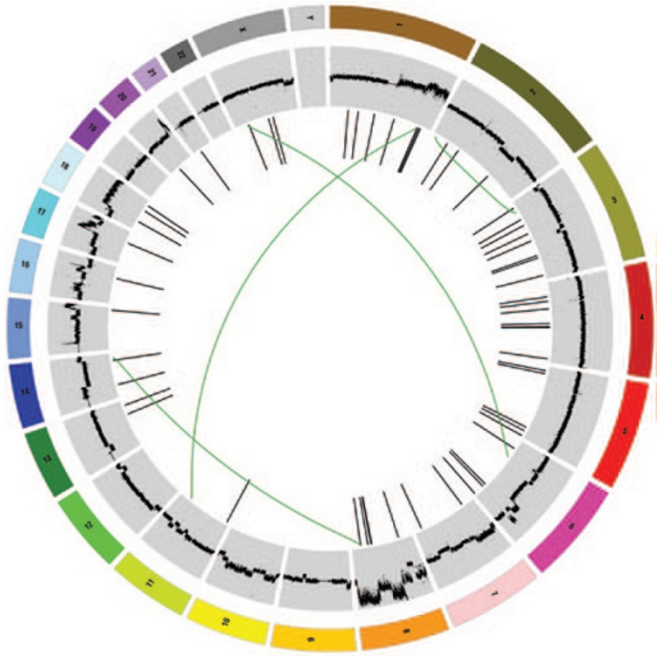
TCGA teams are profiling molecular differences between tumor cells and healthy cells in 500 patients and for more than 20 cancer types[1]. Since 2006, TCGA has explored these differences using a variety of platforms across more than 6,000 patient tumor-normal sample pairs, using single-nucleotide polymorphism, small RNA, transcriptome, exome and methylation data from sequencing and microarrays, says Kenna Shaw, TCGA program office director. For many samples, whole-genome sequence data are becoming available.

Those data, along with other sequencing results such as exome or mRNA sequence data, are held at the Cancer Genomics Hub at the University of California, Santa Cruz (UCSC), with controlled access for data that could allow individuals to be identified. Nonsequence data are kept at the TCGA data portal.

Recently, a researcher in Getz's group at the Broad downloaded genome sequences of patients' tumor and normal tissue from the Cancer Genomics Hub. "We downloaded something like 20 whole genomes, tumor-normal pairs, in 3 days," Getz says. "That's quite fast."

To create data packets that are easier to handle for analysis than the gigantic raw sequence files, the TCGA centers have created tiers. "Higher-level data—for example, the list of somatic mutations in exome data or copy-number changes along the genome, or expression levels of different genes—all of these are public data," Getz says. Those are much smaller in size than raw sequence files, he says, a difference that can help scientists shopping for more manageable files.

Since its launch in 2008, the ICGC has amassed around 250 terabytes of data from approximately 1,300 donors, in Lincoln Stein's rough estimation. He directs bioinformatics and computational biology at the Ontario Institute for Cancer Research (OICR), which is also the ICGC's data coordination center. ICGC scientists in Asia,



US National Institutes of Health/TCGA

TCGA is characterizing many tumor types. In this simplified Circos plot visualizing TCGA breast cancer data, scientists can integrate results and explore the genome data inter-relationships.

Somatic mutations are delivering "big surprises in terms of the types of genes to be found mutated in cancer," says Matthew Meyerson.

Europe and North America are characterizing over 24,000 tumor genomes from 50 tumor types, comparing tumor and normal tissue[2].

ICGC data are deposited in the European Genome Phenome Archive. Somatic variant data are openly accessible at the ICGC Data Portal but scientists must apply to access data such as raw sequence, germline mutations or clinical data.

In the past, each ICGC country housed its own data, but that strategy is changing. "The federated model, as we've discovered, has an Achilles' heel," Stein says. Network connectivity issues have on occasion made data inaccessible. All the interpreted data are now being copied into a centralized database administered by OICR. This transfer will be completed this autumn.

The year-long project is worth the effort because the new system scales well, Stein says. The database uses the distributed MongoDB architecture, which also offers high data availability, he says.

Which tools do I use?

A full toolbox is evidence of a vibrant developer community. Cancer genome analysis tools number "easily in the hundreds," says Stein, and every conference poster session brings more. "It's daunting for experts in the field as well."

"It's good to have many tools, but there is no systematic comparison of these tools," says Getz. Stein and his team find that many published tools have issues beyond a lack of documentation. "They don't install; they crash; they don't pass their own internal tests," Stein says. Although many tool builders test their algorithms for publication, they often do not finish the software engineering needed to stabilize the tool.

Addressing these hurdles, both TCGA and ICGC have begun benchmarking tools against a so-called gold-standard data set, which can take months. The OICR is wrapping up the benchmarking of nearly two dozen algorithms that detect structural genome rearrangement.

Stein believes that tool developers would save their colleagues time by distributing software preinstalled on a virtual machine, a so-called instance, on Amazon's Elastic Compute Cloud, Oracle's VirtualBox or VMware. Many scientists already load tools into the online and cloud-based genome data analysis platform Galaxy, which also has a software repository called Tool Shed.

Although not all tools are guaranteed to run, the more restricted environment of Galaxy's virtual machine offers a predictable version of the operating system with preinstalled libraries, says Stein. "People can star a tool that they like and dislike, so if it doesn't work, it will get low ratings, and we would probably not even bother with it in our benchmarking," he says.

One popular tool in use at OICR is the Broad's sequence-variant caller, GATK. It teases out the alterations between a person's tumor and normal tissue as well as variations from the human reference genome, says Quang Trinh, a computational biologist at OICR. Careful testing precedes the addition of any tool to the OICR production pipeline, an approach he hopes others can follow, too. "Each time you pick a tool, you have to run through the process of testing and validating, making sure what the false positive and false negative rate is, and so on," says Trinh.

Cancer genome analysis tools number "easily in the hundreds," says Lincoln Stein.

What about processing power?

Downloading and analyzing large data sets takes planning and plenty of computational horsepower. For researchers who do not regularly need continuous large-scale analysis, cloud computing can be an option.

After downloading a dataset, analysis at the Broad runs on-site in a high-performance computing environment. To expand their offerings and make data and tools more available to the community, Getz, UCSC's David Haussler and colleagues from other institutions are exploring cloud computing options, which must be made secure to process patient data. "These are things that are still in flux," Getz says. "But we're experimenting with building our compute pipeline on the cloud."

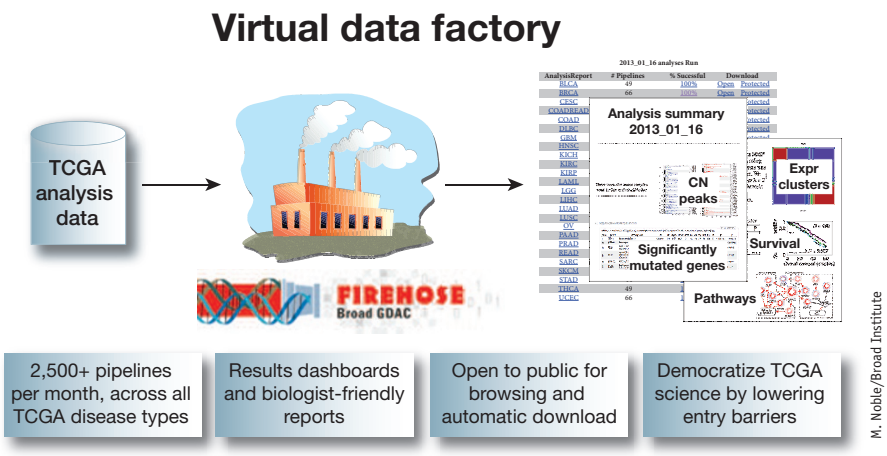
The cloud's best feature, he says, is elasticity. Researchers pay for the amount of compute time used, not for maintaining their own hardware. "You could say, 'OK, now I need 1,000 computers'," he says. "And then the next day you only need two." This solution works for both big genome centers and small labs, which can use clouds run by Amazon, Google, Microsoft, IBM or other providers.

Once the analysis is done, the virtual computers are released, and the data have to travel, which costs time and money. To address data transfer issues, Getz and his colleagues are exploring ways to keep data in the cloud. He says more help will come from increased access to the high-speed academic Internet backbone called Internet2, which includes 100-gigabit-per-second connections and is being set up by a consortium of universities, government agencies and companies.

Aren't pipelines for oil?

Genomics analysis pipelines cannot get oil from point A to point B, but they can transform data from A to Z. Every 2 weeks, the Broad's Genome Data Analysis Center (GDAC; http://gdac.broadinstitute.org/), with team members from the Broad, MD Anderson Cancer Center and Harvard Medical School, swoops up all the generated TCGA data, normalizes them and makes them available.

In a separate automated analysis pipeline series, these data sets are run through many



Every 2 weeks, the Broad's Genome Data Analysis Center (GDAC) takes the TCGA data, normalizes them and makes them available. Analysis pipelines that are run in a computational framework called Firehose take these data sets through many software tools.

software tools: for example, to detect significant copy-number alterations, correlate methylation status with clinical features or find significantly mutated genes, Getz says.

The pipelines run in a computational framework called Firehose, which also generates analysis reports.

Soon the Broad will open Firehose to all TCGA scientists and, eventually, the wider research community. "We want to make the system available so people can install their own tools and run more tools," Getz says. "The future aim is to generate something that looks like a publication automatically, with figures, supplementary information and figure legends," he says. The pipeline report still requires interpretation by scientists, but it jump-starts analysis.

One analysis challenge has been the Babel Problem, as Broad software engineer Michael Noble calls it. Scientists were not able to precisely refer to TCGA data slices, which reduced reproducibility. "They did not speak the same language," says Getz. To resolve this issue, Noble created "Version Stamp" to tag each data set and analysis run. Scientists can now identify the specific data they use for a particular analysis.

Firehose has a cousin called SeqWare developed by computational biologist Brian O'Connor during his postdoctoral fellowship at the University of California, Los Angeles. In 2011, he joined OICR, where he is senior software architect. SeqWare is a framework to package, archive and share sequence analysis workflows, O'Connor says. Built to be "location agnostic," it is not restricted to any one institution or type of computational infrastructure.

Whereas Firehose handles a substantial portion of analytical workflows for TCGA, SeqWare currently handles just the ICGC variant annotation pipeline, says O'Connor.

With SeqWare, data coming off the sequencer flow into a database that is monitored by a software-based 'decider' that triggers predetermined workflows for assembly, alignment and analysis. "This type of system has allowed us to automatically analyze thousands of samples with very little human interaction," he says. "Our plans are to release these workflows to the public, which would allow people to replicate our work at their own organization or on the Amazon cloud." There is also a portal for "nontechies to interact with the system and get analyzed data back."

Not all labs need platforms for large-scale automated analysis of terabases of sequence data. "However, that's changing," says O'Connor. As sequencing technologies evolve, individual labs increasingly produce data hills similar to the output of small genome centers from a few years ago.

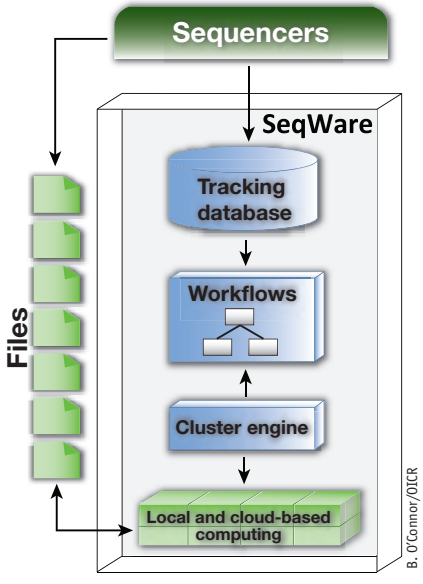
SeqWare-based workflows are among the many ICGC pipelines. As Stein explains, the consortium is currently addressing this multipipeline situation by benchmarking their pipelines in an exercise run by Ivo Gut of Spain's Centre Nacional d'Anàlisi Genòmica. At a recent meeting, the researchers reviewed the first results. "It's kind of interesting because nobody got exactly the same results," says Stein. The team plans to tally their findings into a series of best practices, which stand to help researchers use pipelines.

Can I buy the data analysis?

Beyond open-source tools, many commercial offerings exist. As the Broad widens the Firehose user base, some tools might be commercialized, says Getz, via a model that evolves tools by keeping them free for academics and nonprofits but requiring a fee from companies. "It's typically not that easy to get funding to support tools and make them commercial-level tools."

Taking SeqWare beyond academia, O'Connor launched a consulting company, Nimbus Informatics, providing an Amazon cloud–based version of SeqWare. He tailors workflows for clients: for example, helping Courtagen scale up their sequencing services and Life Technologies analyze the Iceman genome (http://icemangenome.net/), a mummy dating back to 3,300 BC.

Some companies focus on sequence data analysis for drug discovery or clinical uses. Cancer research right now is not unlike the phase when whole-genome sequencing "took off," says Thomas Knudsen, CEO of the bioinformatics firm CLC bio, which has customers in academia, biotech and pharma. First, the early adopters in large genome centers built their own tools, and then companies such as his offered theirs. Similarly, large-scale cancer research will



With the pipeline framework SeqWare, data flow off the sequencer into a database, where software triggers predetermined analysis workflows. SeqWare handles genome-variant annotation at the OICR.

soon broaden, but it is now more limited to groups with bioinformaticians for in-house software development.

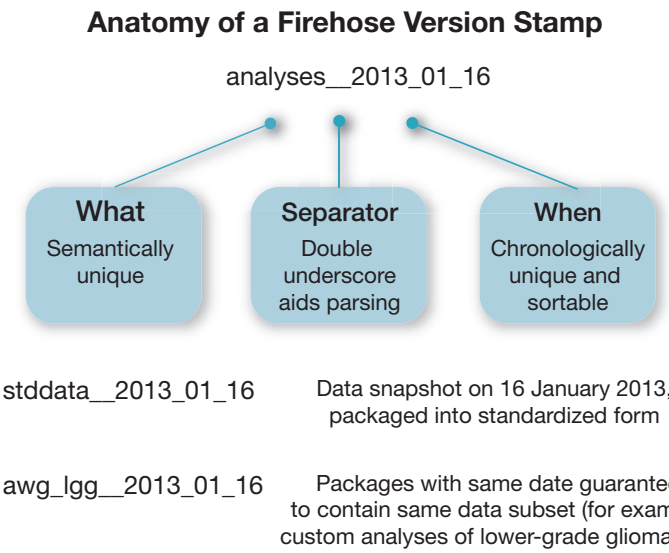
Knudsen's customer base has grown as more companies and clinical researchers have begun using second-generation sequencing. "Another trend is that we sell to more and more customers who replace their open-source pipelines and internally developed software with our solutions."

Jorge Conde, a cofounder of Knome, sees TCGA and similar projects as a source of growth for his firm's user base, which includes scientists seeking additional computational know-how. Customers can approach Knome to find genomic variants in data by using the company's platform, which integrates public data sources and analysis tools.

Early versions of academically produced software can start out "clunky and buggy," says bioinformatician Martin Ferguson, who consults for TCGA, setting up processes that ease data comparison across the hundreds of participating clinical sites.

Though they may have small beginnings, academic cancer genome analysis tools can develop significant business careers. One example is Compendia Bioscience, a 2006 University of Michigan spin-off that Life Technologies acquired last fall. Compendia's founders sought applications in drug development and clinical research. Its platform Oncomine includes cancer genome data, such as those from TCGA, and analysis tools, and it is free for academics but not for companies.

The momentum in cancer genomics and analysis stands to help cancer patients, says Elaine Mardis. "That's really at the end of the day why we are doing all of this."



The Broad Institute confronts the 'Babel Problem' that emerged when scientists used TCGA data but could not readily identify data sets. Version Stamp makes each automated analysis identifiable.

Firms are big users of TCGA data, slicing out what they need, often with commercial success, Ferguson says. Some of his clients are pharma companies. "They're using the data for anything they can: they're mining it for new targets, they're mining it for potential biomarkers that can be tested and turned into a companion diagnostic," he says.

Cancer genome projects are among the reasons Oracle built a platform for scientists to scale up genomic data analysis and include large public-domain data sets such as TCGA, and to view them across genotype and phenotype, says Jonathan Sheldon, global senior director of translational medicine in Oracle's health sciences business unit. "Frankly, bioinformaticians have to spend way too much time doing the mundane but necessary formatting and reformatting work to load these public data into systems ready for analysis—we are 'productizing' this step so they can focus on working with the disease scientists."

To this end, the company built an 'omics data model, which involves such tasks as defining data structures and how they relate to one another, and a platform that can analyze data from different sequencers and analysis pipelines, either locally or in a secured cloud-based computing environment or a combination of both, Sheldon says.

Recently, researchers at the Whitehead Institute for Biomedical Research used data posted in the 1000 Genomes Project and public genealogy information on the web to identify 50 individuals on the basis of short tandem repeats from sequence data[3]. This finding makes it "likely that the privacy landscape of genomic data per se is going to tighten up," Sheldon says.

Many companies offer to help researchers set up cloud-based genomics analysis. "It's perhaps a better economic solution than trying to put together your own compute farm, hiring an IT staff to maintain it and then hiring a bunch of programmers to build pipelines for you," says Elaine Mardis, who codirects The Genome Institute at Washington University School of Medicine. She also advises DNAnexus, a company offering these types of genome analysis services.

Beyond doubts, questions await

Researchers can use the available data and tools on their own hardware and the cloud to pursue their questions of interest. There are plenty of open questions to plumb because large genome centers do not have time for in-depth analysis, says Meyerson. Besides working to better understand the mix of normal and cancerous cells in a tumor, scientists seek to discern mutations that drive cancer progression. "The identification of drivers and passengers either computationally or experimentally remains a challenge," he says.

"We want to make the system available so people can install their own tools and run more tools," says Gad Getz.

The genes most commonly mutated in cancer are turning out to be ones that had been identified on either the gene or pathway level prior to the advent of second-generation sequencing, says Meyerson. But the data are also delivering unexpected results. He and TCGA colleagues discovered previously unreported loss-of-function mutations in the *HLA-A* gene in over 170 squamous cell lung cancers[4]. The team noted that this discovery speaks to cancer's ability to evade the immune system. Such somatic mutations are delivering "big surprises in terms of the types of genes to be found mutated in cancer," says Meyerson.

Although large-scale cancer genome projects such as those connected to TCGA and the ICGC are reaching the afternoons of their first iterations, the systematic characterization of cancer genomes is "at the very earliest moment of dawn," Meyerson says. "I think the cancer genome is more deeply disordered, especially in terms of rearrangements and all sort of unexpected structural events than we had ever anticipated."

He believes genome-based diagnosis will become common for cancer patients, and his commercial ventures reflect this view, including a licensed patent to LabCorp of America and the launch of Foundation Medicine, which offers sequencing-based cancer diagnosis.

Though individual scientists lacking computational expertise cannot yet take raw whole-genome sequence reads and find the important variants in their samples, they can "stitch together" open-source tools from the genome centers to analyze their own large data sets, says Mardis.

A little over 5 years ago, her team published the first whole-genome sequence comparison of tumor and normal tissue in an acute myeloid leukemia patient[5]. At the time, her project proposal about whole-genome analysis was met with "disbelief and derision," she says. Yet advances require pushing the technology and picking aggressive goals. "Doing things because it's early is often the only way to get them figured out."

Today's cancer patients often react to targeted therapies with dramatic improvements and then, "almost inevitably," relapse into therapy-resistant disease, she says. Scientists do not yet understand what fundamental changes in the genome explain such events, but the momentum in cancer genomics and analysis can address such conundrums, which stands to help cancer patients, she says. "That's really at the end of the day why we are doing all of this."

1. Chin, L., Hahn, W.C., Getz, G. & Meyerson, M. *Genes Dev.* **25**, 534–555 (2011).
2. The International Cancer Genome Consortium *et al. Nature* **464**, 993–998 (2010).
3. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E. & Erlich, Y. *Science* **339**, 321–324 (2013).
4. The Cancer Genome Atlas Research Network. *Nature* **489**, 519–525 (2012).
5. Ley, T.J. *et al. Nature* **456**, 66–72 (2008).

Vivien Marx is technology editor for *Nature* and *Nature Methods* (v.marx@us.nature.com).