Cancer genomes: discerning drivers from passengers

Vivien Marx

Tumors impart hints at what drives their progression. Parsing those signals takes old and new approaches.

In tumors, scientists can detect any number of point mutations and larger genomic alterations such as insertions, deletions, inversions and translocations, all of which make these diseased tissues dissimilar from healthy ones. Some mutations—driver mutations—lead a cancer to grow, spread and, often, take a patient's life. Passenger mutations tend to not contribute to cancer growth.

The ability to discern between the two types of mutations can lead to a deeper understanding of cancer biology and empower the development of cancer therapeutics. But the complexity of cancer genomes does not make it easy for researchers to tell drivers and passengers apart. As second-generation sequencing matures, new tools and approaches are helping scientists discover what drives a given cancer.

Who could be driving?

There are around 100 genes that are known cancer drivers. When researchers look across the sequences of many tumor samples, they will find 'mountains', which are mutations occurring in many tumors. One such highly mutated driver is *TP53*, the gene encoding tumor protein p53; Kirsten rat sarcoma viral oncogene homolog (*KRAS*) is another¹.

But mutation levels of many genes in cancer cells can also be 'hills' with frequencies that are not much greater than those of noncancer cells^{1,2}. And these genes might well be drivers, too. Frequency is not a clear metric for distinguishing drivers because mutation frequency varies widely between healthy and diseased cells and across tissue types. "We know it varies across the genome," says Kenneth Kinzler, a cancer genome researcher at Johns Hopkins University's Sidney Kimmel Comprehensive Cancer Center. This variation also means that methods that rely solely

Around 100 genes are known cancer drivers—genes that are highly mutated across many tumor samples. But there may be more.

on mutation frequency to discern drivers from passengers "may be problematic," he says.

Large-scale cancer genome sequencing projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have created well-endowed gene catalogs and portals for the research community to use that render visible this variation in mutation frequency (Box 1). As these projects reach the end of their first chapters, there are various ways to leverage these catalogs, hunt for signals of drivers in the data, and develop new methods and approaches.

Is there a rule for picking driver signals?

As a small lab, Kinzler says he, Bert Vogelstein and their colleagues do not sequence thousands of tumors at a time,

nor do they have a large group of biostatisticians at their disposal. They have decided to focus on the genes that are "unequivocally, clearly driver genes," Kinzler says.

They apply what they call the ratiometric rule, which is about mutation patterns as opposed to mutation frequencies. The rule distinguishes between oncogenes, which need to be hyperactive to cause cancer, and tumor suppressor genes, which cause cancer when they stop working. For an oncogene, 20% of the recorded mutations in the gene must occur at the same position and cause a single switched amino acid in the protein that the gene encodes. For a tumor suppressor gene, more than 20% of the mutations in the gene must be clearly inactivating¹.

Kinzler sees advantages to this ratiometric approach over other methods, which have "pretty significant false discovery

BOX 1 SOME CANCER GENOMICS DATA PORTALS

cBioPortal for Cancer Genomics (<http://www.cbioportal.org/public-portal/>): this portal is developed and maintained by the Computational Biology Center at Memorial Sloan-Kettering Cancer Center. It offers curated data sets from over 50 published studies, including from the TCGA.

The Cancer Genome Atlas (TCGA) Data Portal (<https://cghub.ucsc.edu/>): data sets include genomic as well as clinical information and analysis related to cancer genomes. TCGA sequence data are held at the University of California at Santa Cruz Cancer Genomics Hub.

TumorPortal (<http://www.tumorportal.org/>): developed and maintained by the Broad Institute of Harvard and MIT, this site hosts a 'pan-cancer' data set from 21 tumor types. It provides visualization of computationally processed gene information. It shows which genes are mutated in many tumor samples and types.

International Cancer Genome Consortium (ICGC) Data Portal (<http://dcc.icgc.org/>): a portal for raw data from ICGC and TCGA projects. Data can be filtered to visually present information such as top-mutated genes.

rates"—perhaps even as high 10%, he says—which can skew a list of driver genes. "It's a question of what you want your list to look like."

What types of signal analysis tools exist?

A tumor is not born as a diseased tissue with many mutations; rather, these mutations accumulate over time. Driver mutations lend a tumor a tiny growth advantage, which may be as low as an estimated 0.4% increase in the difference between cell birth and cell death rates³. Brain tumor cells divide around once every three days in glioblastoma, which is a type of brain tumor, and once every four days in colon cancers. This small growth advantage compounds with time, which enables a tumor mass to form.

One common strategy to identify cancer drivers is to look for these signals of positive selection in the pattern of somatic mutations in genes across tumors, says Nuria Lopez-Bigas, a computational biologist at Pompeu Fabra University in Barcelona, Spain.

Some tools look at the rate of cancer mutations above background rates or clustering patterns of mutations, and others rank mutations according to functional impact. As members of the ICGC have pointed out, new methods are needed to, for example, better predict the effect of sets of mutations on protein and cellular function and to identify mutations that influence resistance or sensitivity to certain therapies⁴.

Lopez-Bigas and her team developed

IntOGen-mutations, a web-based computational platform to look at cancer mutations, genes and pathways across tumor types. It applies a number of software tools, and its database is continuously updated with publicly available information from tumor sequencing projects⁵. The platform lets scientists analyze newly sequenced tumors so that they can better interpret mutations in them, she says.

She and her group believe that combining multiple complementary signals of positive selection is the best way to obtain a more "comprehensive and reliable" list of driver genes⁶. Lopez-Bigas and her team devel-

oped algorithms, such as OncodriveFM and OncodriveCLUST, that can identify cancer drivers from a list of somatic tumors. These algorithms are built into IntOGen, which bases all computation on stored lists of somatic mutations. With this architecture, the platform becomes "scalable to the analysis of hundreds of thousands of tumors," she says.

She and her team are now working on a new platform called The Cancer Genome Interpreter. The idea is to include drug information so that the software will help interpret which mutations in a tumor are important and whether options for targeted therapies exist that are directed toward those mutations.

Some tumor signal analysis tools have a longer history than others. MutSig is an algorithm developed by Gad Getz at the Broad Institute of Harvard and MIT and his colleagues in 2007 to help analyze the first whole-exome sequencing experiment performed in the Vogelstein lab⁷. These were the days before high-throughput methods: the exome was sequenced by synthesizing primer pairs for each exon, followed by PCR and then Sanger sequencing.

MutSig has continued to evolve in its ability to process different types of positive selection signals from cancer genomes. This evolution, as Getz and Broad Institute computational biologist Michael Lawrence explain in a joint comment, began as the team looked at tumor types with high background mutation frequencies. And that was when the limitations of their original

Luis Diaz, Nickolas Papadopoulos, Kenneth Kinzler, Shubin Zhou and Bert Vogelstein (and Victor Velculescu, not shown) collaborate on cancer genes that are clear drivers.

Broad Institute researchers Michael Lawrence (left) and Gad Getz (right) believe the catalog of cancer variants is far from complete.

approach “came into focus,” they say.

To improve MutSig they tackled a TCGA data set of lung squamous cell tumors, realizing that the local background rate in the genome is correlated to many co-varying genomic quantities. When they engineered the algorithm, they found that it was able to remove many false positives from a long list of mutated lung cancer genes (Box 2).

Are there noncoding driver mutations?

Many approaches mainly analyze regions in the genome that encode proteins and that can be mutated to give overactive or defective forms. But recent studies indicate that noncoding regions of the genome, which can be responsible for regulating gene activity, might also harbor cancer drivers. Noncoding drivers could potentially outnumber coding ones, say Lawrence and Getz. But for now, the community is “completely blind to them” because whole-exome sequencing has been the workhorse to date.

The focus on coding regions has been a practical one. As a way to hold costs down, most cancer genome sequencing projects have focused on exome sequencing, says Lopez-Bigas. “Now with the focus and economics shifting to whole-genome sequencing, we’re all under the gun to get our act together beyond the splice sites,” say Lawrence and Getz.

A team of scientists at the Broad Institute, Dana-Farber Cancer Institute, Harvard Medical School and MD Anderson Cancer Center describe two highly recurrent mutations in melanoma that lie outside of protein-coding regions⁸.

Specifically, they found two somatic mutations in a regulatory region, the promoter of the telomerase reverse transcrip-

tase gene (*TERT*). They note that in addition to coding sequences, recurrent somatic mutations in regulatory genomic regions “may represent important driver events in cancer.”

As researchers begin to look at these noncoding regions for possible driver mutations, bioinformaticians will develop tools to find and help characterize the role of these mutations. Lopez-Bigas and her team plan on leverag-

ing computational approaches to identify drivers in coding regions as part of the hunt for signals of positive selection in noncoding regions. It might lead to an additional analysis module in IntOGen-mutations. It is “going to be difficult,” she says of the development process, but this important endeavor will help put together a comprehensive catalog of driver mutations.

Studying noncoding regions will help researchers to understand cancer, Kinzler says. The *TERT* promoter finding is an example of a noncoding driver mutation that had been previously missed, he says, and there are likely to be others in this category. But he does not believe these findings will present a “completely different understanding” of cancer drivers such as new pathways implicated as the origins of cancer.

BOX 2 EVOLUTION OF A DRIVER-HUNTER

Born in 2007, the algorithm MutSig has become an increasingly seasoned mutation signal processing tool.

From MutSig to MutSigCV. In the beginning, the tool looked at one signal: the abundance of nonsilent mutations in relation to background mutation patterns. Over time, the Broad team changed the way the algorithm calculates gene-specific background rates and integrated co-varying factors including gene density, chromatin structure and replication timing, which is the distance to the nearest site where DNA replication initiates, all of which reduces the false positive rate among cancer genes.

MutSigCL. Clustered mutations may implicate a gene as a driver. The scientists added to the algorithm the ability to calculate the probability that the positional clustering of mutations was due to chance.

MutSigFN. This algorithm estimates the functional impact of each mutation. The team added the capability to estimate how likely it is that a missense mutation might be deleterious to the protein encoded by the mutated gene. Evolutionary conservation is key for this signal detection: if a mutation occurs in a location conserved among multiple branches of the evolutionary tree, the algorithm knows to consider it significant.

The developers, heeding the fact that clustering and conservation may be correlated, calculate a joint *P* value for CL + FN. They also take into account all three signals—CV, CL and FN—to deliver a *P* value for a given gene, leading to the three-signal analysis that the team applies to cancer genomes.

Source: Broad Institute

MuSiC-SMG / MutSigCV

Identifies genes mutated more frequently than background mutation rate

OncodriveFM

Identifies genes with a bias towards high functional mutations (FM bias)

OncodriveCLUST

Identifies genes with a significant regional clustering of mutations

ActiveDriver

Identifies genes significantly enriched in mutations affecting phosphorylation-associated sites

N. Lopez-Bigas/Pompeu Fabra University

Discerning drivers from passengers is enabled by combining tools to identify signals of positive selection in the pattern of mutations⁶. Here: MuSiC/MultiSigCV from Washington University and the Broad Institute, OncodriveFM and OncodriveCLUST from Pompeu Fabra University, and ActiveDriver from the University of Toronto.

Where is the complete cancer catalog?

A comprehensive cancer gene variant catalog is desirable to better understand cancer biology. It can show the spectrum and frequency of mutations across tumors to help identify drivers and also serve as a reference for doctors and drug discovery scientists selecting a treatment regimen based on particular molecular disruptions in a patient's tumor.

The TCGA is analyzing 500 samples from each of over 20 types of cancer, and the ICGC is looking at 50 different tumor types. Although large data sets have been generated thus far in these and other projects, the catalog is incomplete.

"It is very clear that, although many new cancer genes have been identified, we are missing many more," says Lopez-Bigas. Teams are identifying cancer driver genes that are known and others that have not been previously discovered. In many cases, the positive selection signals for these mutations are low, which in turn "directly implies that when we sequence more tumors, we will find more genes that act as drivers."

In their analysis of over 4,700 cancers across 21 tumor types, Lawrence and his

colleagues found known cancer genes and almost three dozen undiscovered ones that contribute to a host of cellular processes including cell death and cell growth². The team notes that "major gaps" remain in the knowledge about genes mutated at frequencies between 2% and 20%.

Pompeu Fabra University

Nuria Lopez-Bigas is working on a new platform called The Cancer Genome Interpreter.

based on the fact that many new genes were detected as we increased the sample size," says Getz. The team calculates that to obtain a catalog of somatic point mutations at both high and intermediate frequency requires analysis of approximately 2,000 tumors for each of 50 tumor types, which means molecular characterization of 100,000 tumors.

Feasibility of such an enterprise may be within reach because second-generation sequencing has made molecular characterizations less cost prohibitive. Given cancer's toll, the authors note, such a project should be "a biomedical imperative."

Having such a catalog will be a "landmark" in terms of the understanding of human cancer, says Kinzler. At the same time, the research community must discuss how to reach that goal. For some tumor subtypes, such as breast cancer or prostate cancer with distinct pathological or molecular characteristics, 2,000 tumors is a "significant fraction" of all of the cases in the United States in one year, he says.

One might indeed choose to sequence the genome of every cancer patient, Kinzler says, as there is "no doubt" about the value of additional cancer genome sequencing. Looking at only the highly mutated genes will not detect every cancer variant. It might not even find all the clinically useful genes, he says, but it will find a "good part of the ones that you can do something with."

Scientists want, one day, to be able to deliver mutation information about every patient to clinicians. "Ideally, it will tell you how they will respond to therapy or how you should treat them," Kinzler says.

Sequencing many thousands of patients to find low-level mutations could lead to findings that are difficult to convert to clinically useful knowledge. "Right now there are not a lot of actionable mutations," he says. But the number of such mutations that can guide treatment regimens will grow. Targeted therapies have been available for only a few decades and are "quite remarkable" compared with classic chemotherapies that have been in place for over half a century, he says.

Although resistance to both old and new compounds tends to develop, "we understand that resistance in much more molecular detail," Kinzler says, which will help to further cancer treatment. Tumor behavior is dictated by mutations as well as a tumor's developmental state and the cell type from which it arises. To forecast tumor behavior and what might drive it, scientists and clinicians need to take into consideration that "it's the sum of a lot of different processes," he says.

Knowledge from an *n* of 1

Another approach to hunting for drivers in cancer genomes is to look at small numbers of 'exceptional' patients whose genomes might provide important clues. Singular

anecdotes—that is, cases when the number of subjects, n , equals 1—are not typically seen as authoritative indicators for clinicians. But perhaps they can deliver cancer genome mutation signals.

Not all tested compounds receive approval by the US Food and Drug Administration. In some cases, a small number of patients respond well, whereas others do not. Last year, the US National Cancer Institute began an “Exceptional Responders Initiative” to find this small percentage of patients—perhaps between 1% and 10%—who responded to the tested drug. Perhaps, organizers thought, molecular signatures in the genomes of these patients can help explain their favorable response. These clues can further the search for influential cancer mutations in other patients.

David Solit and his colleagues at Memorial Sloan-Kettering Cancer Center and the University of California at San Francisco, along with a colleague from Foundation Medicine, which is a company focused on genome-based diagnostics and treatment, explored the background

of one exceptional patient in a clinical trial⁹.

The patient, who suffered from metastatic bladder cancer, was treated with the Novartis drug Afinitor (everolimus), which is prescribed for patients to prevent rejection of tumor organs but is also used in cancer. In this patient, and unlike in other patients in the trial, the drug led to a remission of the cancer that has now lasted almost three years. The drug inhibits the mTOR pathway. Sequencing revealed a mutation in the tuberous sclerosis complex 1 gene (*TSC1*), a regulator of mTOR pathway activation. These results suggest that mTORC1-directed therapies “may be most effective in cancer patients” whose tumors harbor *TSC1* somatic mutations, the scientists note in their paper.

Driver mutations perhaps do not always lead to this kind of clinical response, Kinzler says, but he suspects they might. This focus on a small number of cases with an exceptional clinical response is an efficient approach to finding ‘actionable’ driver mutations, he says. It is also practical because, from the start of a study, tumor samples are

available that are associated with a certain type of clinical response.

Scientists can then launch their analysis from this positive clinical response—a phenotype—and work their way back to these patients’ genome to hunt for reasons that explain these different, positive responses. Finding driver mutations in cancer is a challenge and will stay important. In some cases, finding these mutations can be exceptionally good news.

1. Vogelstein, B. *et al. Science* **339**, 1546–1558 (2013).
2. Lawrence, M.S. *et al. Nature* **505**, 495–501 (2014).
3. Bozic, I. *et al. Proc. Natl. Acad. Sci. USA* **107**, 18545–18550 (2010).
4. Gonzalez-Perez, A. *et al. Nat. Methods* **10**, 723–729 (2013).
5. Gonzalez-Perez, A. *et al. Nat. Methods* **10**, 1081–1082 (2013).
6. Tamborero, D. *et al. Sci. Rep.* **3**, 2650 (2013).
7. Getz, G. *et al. Science* **317**, 1500 (2007).
8. Huang, F.W. *et al. Science* **339**, 957–959 (2013).
9. Iyer, G. *et al. Science* **338**, 221 (2012).

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