Cancer: hunting rare somatic mutations

Vivien Marx

Emerging ways to lower the error rate when hunting low-frequency mutations.

Rare mutations can be a tumor's secret weapon: a targeted drug might kill most cells in a tumor, but a subclonal population of cells can manage to stay unscathed and then grow out.

High-throughput sequencing has let labs discover genetic alterations in many cancers. But hunting rare mutations such as single-nucleotide changes, small insertions and deletions and gene fusions that might be in 0.1% to 1% of cells in a tumor is hard given that these frequencies are below the sensitivity of current assays. "Due to sequencing errors, variants present in fewer than 5% of reads are typically disregarded," says Michael Schmitt, oncology fellow at the University of Washington, who splits his time between the lab and patient treatment.

There are many sources of assay-induced errors—from library prep to sequencing itself—for the various regions throughout the genome, says Jason Bielas, a researcher at the Fred Hutchinson Cancer Center. He says a finding might appear to be a gene fusion, for example, but during library prep "there is a lot of jumping that can occur during the PCR reactions, from one piece of DNA to another piece of DNA."

Experimental and computational assays for measuring low-frequency events in cancer are maturing and gaining traction as researchers seek to characterize the heterogeneity of tumors, says Bielas. Highly sensitive assays are needed to discern assay-induced errors from "true variants" in the cell population, he says. He, Schmitt and others are developing methods to help give labs that discerning capability.

Rare mutations come in multiple guises. A specific genetic alteration might be found in only a few specific cancer types, or it might be found in a fraction of the cells in one person's tumor.

Brand X

Now that many cancer genomes have been sequenced, researchers can explore how best to hunt and find low-frequency mutations. It's a challenging hunt.

Rare among cancers

"When we looked at lung cancer we never found it, when we look at breast cancer it's not there," says Axel Hillmer, a cancer researcher at the A*STAR Genome Institute of Singapore. He, together with David Virshup at Duke-National University of Singapore and colleagues at universities in the United States, Germany and Switzerland, investigated whole genomes of patients and found a specific type of genetic rearrangement that renders *TP53* inactive: a translocation with breakpoints in intron 1 of the *TP53* gene[1]. The change is unique to osteosarcoma, a cancer that arises in the bone. *TP53*, a well-known tumor suppressor gene mutated in many cancer types, encodes a protein that helps maintain genome stability.

The team first analyzed four osteosarcoma patients' genomes, then expanded their analysis to nearly 300 patients and over 1,000 genomes from people with other cancer types. They also drew on samples from the Bone Tumor Reference Center, a resource at the University of Basel in Switzerland. Hillmer now wants to explore why this genetic rearrangement occurs only in the osteoblast lineage because "there must be something in these cells, which makes this part easier to break."

It seems that a related genetic rearrangement, an inversion with one breakpoint in intron 1 of *TP53*, can also occur in the germline. Hillmer and his colleagues found this change in the germline of people with a rare condition called Li-Fraumeni syndrome (LFS), who suffer a heightened cancer risk, including for osteosarcoma. The researchers have not yet determined the frequency of the *TP53* inversion in LFS, but when germline mutations lead to *TP53*'s loss of function, almost all tissue types are prone to develop cancer.

These rearrangements, both somatic and germline, are easily missed. Two decades ago, *TP53* rearrangements were hunted by Southern blotting, and now researchers apply whole-genome sequencing. In the genomes of people with osteosarcoma, the team noticed something unusual: the rearrangements' breakpoints occurred in the same small genomic region, suggesting a biological, disease-related function. It was similar to a recurrent gene fusion in chronic myeloid leukemia.

In their hunt for this rare mutation, the scientists took a multi-tiered approach, first performing paired-end tag sequencing on samples from four patients with osteosarcoma (DNA-PET) and then sequencing the *TP53* regions. DNA-PET leverages the unique sequences at the 5′ and 3′ DNA ends of long DNA fragments, says Hillmer. The DNA fragment is sequenced first from one end and then from the other, which helps in the hunt for insertions, deletions and rearrangements once reads are aligned to the human reference genome. But DNA-PET requires much DNA. And although the prices have dropped for whole-sequencing,

Corinna Kielwein

Axel Hillmer and his colleagues noticed something unusual in the genomes of people with osteosarcoma. It was a specific type of genetic rearrangement that is easily missed.

it remains pricier than targeted sequencing, which is why Hillmer and his colleagues designed a custom capture of the entire *TP53* gene region, including both coding and noncoding regions. "This assay is cost effective, when it is used for many samples," he says.

The team did targeted paired-end sequencing of the *TP53* locus, and this pulled down DNA fragments containing the points where the *TP53* intron 1 region is fused to the partner chromosomal regions of the rearrangement. And it was this targeted sequencing approach, says Hillmer, that helped them find the 445-kilobase inversion with breakpoints in intron 1 of *TP53* in the genomes of the family with LFS.

They also developed a 'break-apart' fluorescence *in situ* hybridization (FISH) test with probes that surround the *TP53* gene. With the test, if 10% of cells showed separated hybridization signals, cells were considered to have a breakpoint in this region. To improve resolution over that obtainable with FISH, they used microarrays, such as Affymetrix' CytoScan, to localize breakpoints and determine their frequencies.

Methods that identify break and fusion points in structural rearrangements are "still not very well established," says Hillmer. Each lab has its own metric, and groups have "their own sniff and feel" approach to these data. They might use *P* values calculated in various ways, and there is no community-wide way to define a generic *P* value cutoff.

It's computationally challenging to identify that a particular piece of DNA has moved from one chromosome to another, says Hillmer. Researchers are more likely to identify the rearrangement point when many reads around a breakpoint are at a good distance from one another, he says. But in other cases alignment to the reference is challenging and breakpoints can be missed.

Computational biologist Jinghui Zhang and her team at St. Jude Children's Research Hospital have developed Clipping Reveals Structure, or CREST, one of several software tools that help scientists hunt breakpoints, also in rare somatic mutations[2]. A sequence read might span a breakpoint and align to the reference genome on either side of that point, but there is an unaligned portion that does not map, a 'soft' portion. These soft-clipped reads are assembled into a longer contiguous sequence and then aligned against the reference genome to find the second breakpoint of the structural variant.

Zhang and her lab tested CREST on whole-genome sequence from children with leukemia and on a human melanoma cell line. The tool's main talent, she says, is its ability to detect breakpoints at base-pair resolution in an aligned genome. "CREST actually does not have any prior knowledge regarding the breakpoints so it has the same power for detecting common or rare rearrangements," she says.

CREST is best used when scientists have germline and tumor sequence from the same person, which helps them filter out common variants. Zhang and her team are happy with CREST's strengths, but highly repetitive DNA sequence is challenging for the software, as are very long insertions at the chromosomal breakpoints.

Rare within a tumor

A number of tool developers address the specific challenges of finding rare somatic mutations that might be in only 1% or less of cells in a single tumor. Barcoding is one helpful approach. For example, the Safe-Sequencing approach developed in the lab of Bert Vogelstein at Johns Hopkins University School of Medicine adds single-stranded barcodes that are 12–24 nucleotides long to DNA snippets[3]. When barcoded DNA is amplified and sequenced, the reads with errors are readily spotted.

Jesse Salk

While ice climbing in the Canadian Rockies, Michael Schmitt co-developed the idea for Duplex-seq, which uses the information on both DNA strands.

Duplex Sequencing, or Duplex-seq, developed in the lab of Lawrence Loeb at the University of Washington, also uses barcodes[4]. Schmitt and his colleague Jesse Salk hashed out the idea on an ice-climbing trip in the Canadian Rockies and co-developed the method in the lab. As Schmitt explains, an altered nucleotide is scored as real only if both DNA strands contain this change. The method's error rate is under one false mutation per billion nucleotides, which means it can detect a mutant allele in a single cell, he says. Not every question will need such high accuracy, he says. Other methods work for "intermediate" frequencies of 0.1% to 1%. The lab is now using the method to study intratumor heterogeneity.

In Duplex-seq, the two DNA strands are tagged with a random double-stranded nucleotide sequence. Next, sequencing adaptors are added. The team has modified the method with an enrichment strategy: there are several rounds of hybridization with biotinylated oligonucleotides that are ligated onto the DNA[5]. Targeted capture approaches with biotinylated probes are common in labs, says Schmitt, yet one limitation is that standard capture protocols do not scale well to targets that might be 20 kilobases long, he says. But the two successive rounds of capture in Duplex-seq resolve that issue, he says, such that he and his colleagues obtain more than 95% of reads that map to the targeted genes.

The researchers applied their method to hunt for rare mutations in the *ABL1* gene and found they could identify rare mutations associated with resistance to a drug for chronic myeloid leukemia. As Schmitt explains, labs are gaining a sense of the patient-to-patient mutation diversity in cancer, and it makes them eager to study the landscape of mutational diversity within individual patients. The lack of tools for these quests motivated him to work on Duplex-seq, which he hopes to use for single-cell genomic analysis. With this goal in mind, the Loeb lab has been optimizing their method by, for example, improving enzymatic steps and ligation efficiencies. Now the scientists are "actively investigating ways to make our method easier for others to take on," says Schmitt. Those potential opportunities might include providing the adaptors as part of a kit or making the assay available as a service.

A number of companies have approached the lab to help. One of these companies is Integrated DNA Technologies, which manufactures reagents such as oligonucleotides. Caifu Chen, who directs IDT's research and development projects, thinks Duplex-seq can help scientists discern low-frequency variants in sequence data with greater confidence. Assays must work consistently and reproducibly in the hands of many, says Chen, which led him to discuss adaptor synthesis and ligation with the Loeb lab.

The Duplex Sequencing adaptor needs to stay in the duplex form, but single-stranded adaptors might also be formed, says Chen, when the enzymatic extension process is incomplete or if DNA melts when making the adaptor. Ligating a single-stranded adaptor to a target DNA molecule will result in the upper and lower DNA strands carrying different molecular tags, a shift that would reduce the power of the method, he says. Adaptors must be individually made in a clean room, and care is needed to avoid contamination. Little things such as adaptors matter, says Chen, who likens adaptors in Duplex-seq to the mirrors in a car that must perform well.

Another method for finding low-frequency mutations is CypherSeq, developed in the Bielas lab[6]. Bielas and his team are using it to discover early indicators of ovarian cancer: mutations in the genomes of cells from a Pap smear. Circulating DNA and tumor cells are other potential applications.

When looking for rare mutations, scientists might be trying to discern one mutation in a group of one million molecules, says Bielas, which calls for having at least one million molecules on hand. If purification reduces the number of molecules to one thousand, then the hunt will be about finding one molecule in one thousand and "that's your new resolution," he says. A method can be sensitive, but the number of recovered molecules sets other limits, which can be challenging when handling precious patient samples and a limited amount of DNA.

With CypherSeq, two biotinylated, target-specific primers are needed, one for each DNA strand. DNA is ligated into circular bacterial vectors that contain the targeted region of interest, and each strand is barcoded with a short sequence of nucleotides and then amplified with the rolling-circle amplification reaction (RCA).

The result of RCA is a concatenated biotinylated strand with many copies of the template. Any errors that were introduced during amplification are along this concatenated strand and can be eliminated computationally, says Bielas. They will not occur in the majority of reads from one barcode. Given that the concatenate contains a biotin, scientists can pull out the strand and leave behind the template. At a later date, the researcher could probe the template from this same patient again. "You can keep on going back for different sites," he says, or return to the same site, for example, to obtain deeper coverage there. "The barcodes preserve the information in time," he says.

Fred Hutchinson Cancer Research Center

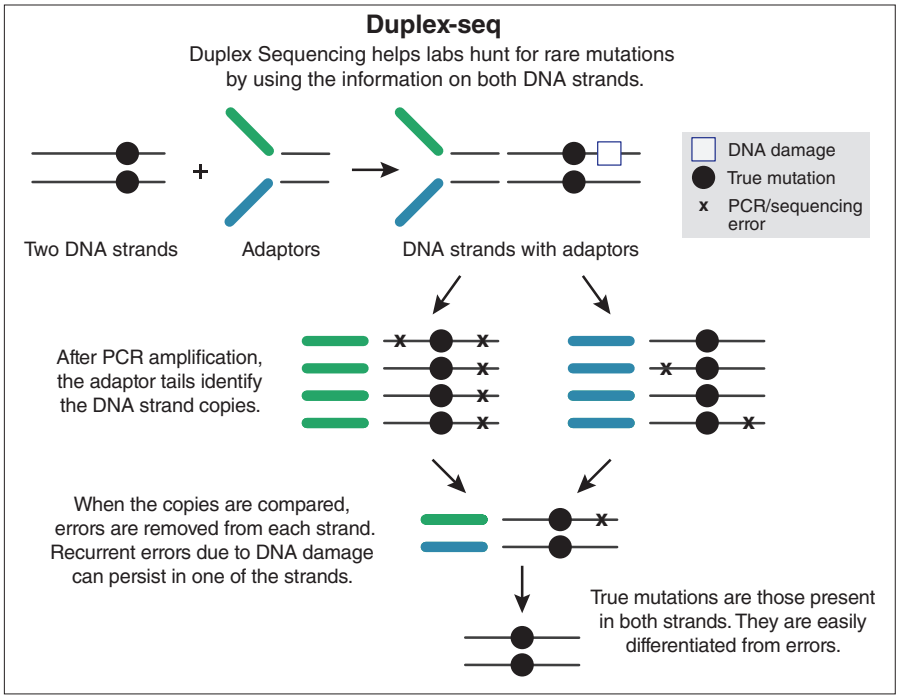
To discern errors and artifacts from true variants, Jason Bielas recommends experiments with spike-in controls.

Schmitt says that his Duplex-seq method has higher accuracy than other approaches given that they are based on sequencing duplicates of single-stranded DNA. Common types of DNA damage get in the way of accurately identifying mutations from single-stranded DNA, says Schmitt. One such example is an abasic site, a spot with an absent base where one of four bases might have been. When the DNA is copied by PCR or RCA, says Schmitt, the polymerase inserts an 'A' opposite the abasic site, causing the abasic site to be read as 'T' and generating a false mutation, not a true one. When sequencing both strands, he says, one strand might have an abasic site, but it is unlikely that the paired strand from the same single molecule is damaged at that very same position.

Believing is seeing

To rule out sequencing error, researchers rely on high coverage, says Zhang. And CREST currently requires such high-coverage data. CypherSeq helps researchers obtain high-quality data, so they do not need high coverage for whole-genome sequencing. "With new applications like CypherSeq, we shall be able to explore the use of low-coverage data as sequencing error is not a major concern," says Zhang.

Hillmer has used CREST and has begun exploring Duplex-seq and Safe-Seq for a lung cancer project he and his colleagues are launching. Barcoding allows a PCR product to be traced back to the original DNA and thus controls for error. Hillmer likes that Duplex-seq lets him interrogate many parts of the genome. "I like the methodology but it doesn't work so well in our hands," he says. His colleagues have used Safe-Seq. The teams are exchanging notes



M. Schmitt, U of Washington / Erin Dewalt, Nature Publishing Group

as they pick methods for their new project. Hillmer also plans to explore CypherSeq.

Bielas has a research colleague who believes he has found a rare genetic fusion in cancer. "And I don't believe it," says Bielas. He is skeptical because his colleague does not know his assay's detection limit. "You're unsure if what you're detecting is an error, an artifact, or a true variant," says Bielas. To avoid this risk he recommends experiments where the correct results are known. By using spike-in controls with known mutants across a range of frequencies or dilutions, he says, researchers can determine their assay's precision, accuracy and specificity.

Labs also need to consider their algorithm's eyesight. In the Bielas lab, a spike-in experiment generated negative results even though a mutation should have been detected on both DNA strands at the tested frequency. They investigated and found that the algorithm was "seeing" a true, spiked-in variant, determining it was an artifact and eliminating it. "It was a mistake in the algorithm itself," says Bielas. Algorithm validation is crucial especially when hunting rare variants, he says.

Scientists will want to troubleshoot whichever assay or analysis pipeline they choose to make sure the variants are called accurately. A spike-in experimental validation "allows you to test every aspect of the assay," says Bielas.

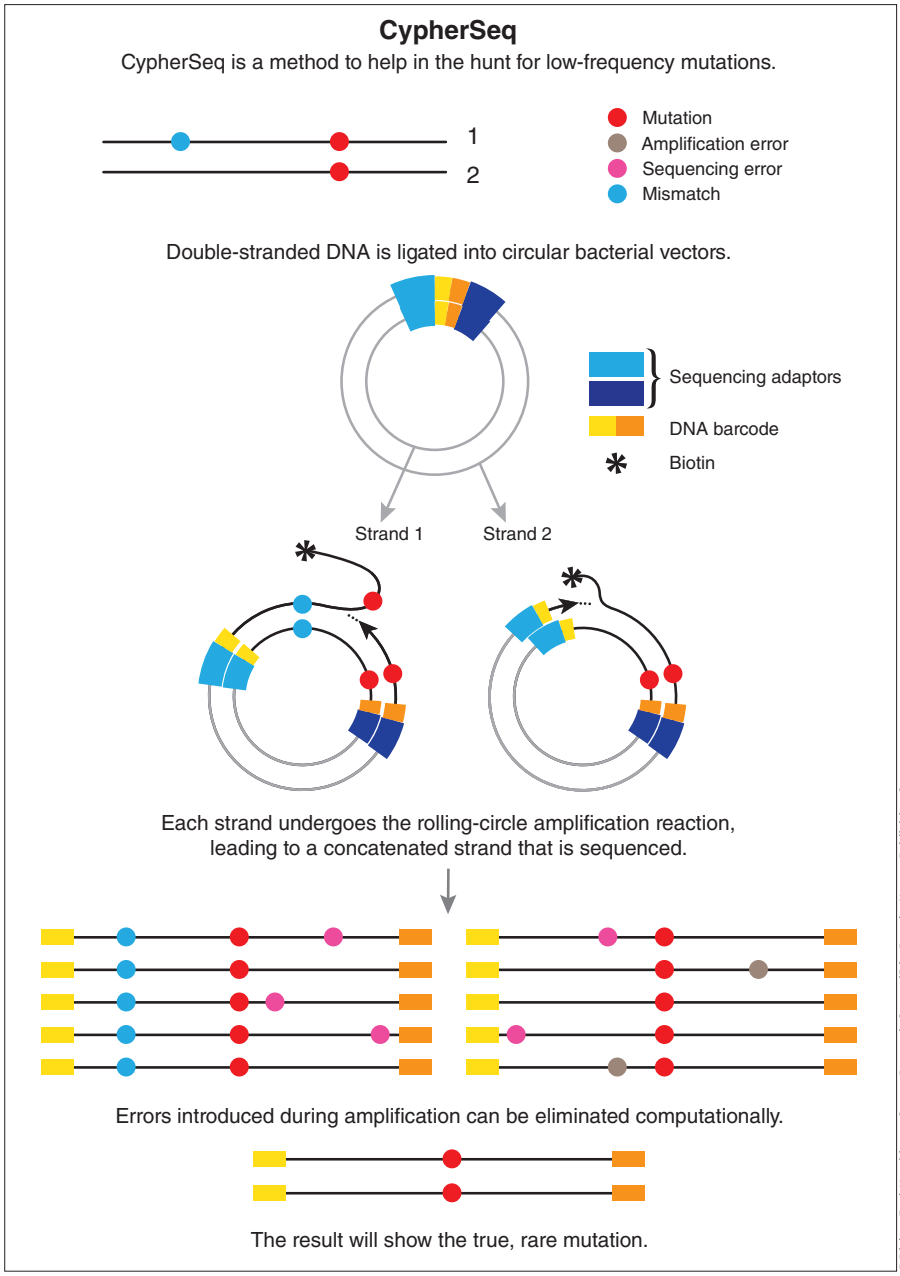
Why the hunt matters

A rare somatic mutation is not automatically a cancer driver or super-driver, says Zhang. Nor is it irrelevant for a tumor's behavior. When labs find a rare somatic mutation they will want to consider existing knowledge and collect multiple lines of evidence to see whether and how it might affect cancer progression.

A variant's biological context matters. "For example, if we suspect that a rare kinase gene fusion may play an important role in a specific cancer, we need to evaluate the protein domains being affected by the gene fusion with the possibility of auto-activation of kinase," says Zhang. Scientists will want to check whether the gene expression signature of the same tumor indicates kinase activation based on the expression signal of downstream targets for that kinase. "Overall, it is much harder to determine the 'driver' status for rare variants," she says.

In 2012, Matthew Meyerson at Dana-Farber Cancer Institute and colleagues at Harvard Medical School, the Broad Institute of MIT and Harvard, Massachusetts General Hospital, Brigham and Women's Hospital and research centers in Mexico and Canada published a large-scale breast cancer mutation analysis of samples from 103 patients[7]. Among the findings was a gene fusion unique to triple-negative breast cancer. This gene fusion between *MAGI3* (membrane associated guanylate kinase, WW and PDZ domain containing 3) and *AKT3* (v-akt murine thymoma viral oncogene homologue 3), which are on different arms of the same chromosome, was a rare kind of structural rearrangement. It was also a possible insight into cancer biology and, in the age of drugs that target cancers with specific molecular traits, a treatment hint.

In 2015, researchers at Weill Cornell Medical College and colleagues indicated that after performing FISH and RT-PCR, they were not able to find this gene fusion in their tumor samples[8]. In response, Meyerson and his colleagues reassessed their data using an array that covers exons, noncoding regions and intronic regions involved in gene fusions[9]. They found four gene fusions in frozen tissue but could not find the fusion in samples that had been formalin fixed and embedded in paraffin. They screened additional samples



J.Bielas, Fred Hutchinson Cancer Research Center/Erin Dewalt, Nature Publishing Group

and realized that their finding had actually been a subclonal population of tumor cells with a frequency much lower than their original estimate.

Schmitt sees this reassessment of the gene fusion event in breast cancer as "an excellent example of the self-correcting nature of science." He praises the care the authors took in following up on the raised concern and their use of an independent method to screen 366 additional paired samples. Zhang sees in this incident the importance of replication studies to validate the frequency of a mutation or a fusion event in cancer. Errors can creep in at many stages in the experimental process as well as during computational analysis. As Meyerson explains, "I actually think that finding sources of error is one of the most important things when trying to find truth in science." An aspect that will propel the hunt for rare genetic alterations is, in his view, "statistical power that comes from having more and more samples."

There is now a huge opportunity to discover the 5%, 2%, 1% and 0.5% frequency mutations with an accurate approach, says Meyerson. Large-scale sequencing efforts such as The Cancer Genome Atlas and the International Cancer Genome Consortium have enabled cancer genome discovery. But, he says, "paradoxically," the funding for these sorts of studies "has pretty much dried up" right as sequencing prices have dropped and it has become feasible to dig deeply into cancer genomes.

Structural rearrangements are especially tough to find. "I think that we're missing most of the structural alterations of the genome," says Meyerson. Scientists need more genomic sequence. But with short reads, labs will not be able to find all of the rearrangements.

It is statistically challenging to know what the background rate of genetic variation in cancer is, which is important for functionally discerning the role of mutations both rare and common, says Meyerson. Frequent genetic alterations in cancer might have been selected for, or there just might not be much counterselection.

Existing state-of-the-art technology is good at finding point mutations in the exome and other nonrepetitive sequence. It remains challenging to find structural variations throughout the genome. What will truly help for the discovery of rare alterations is the statistical power that comes from studying more samples, says Meyerson. Technology advances now let labs begin to gain an understanding of structural alterations inside and outside of coding regions, he says. It's an experimental challenge and "absolutely a software question," he says. A result can turn out to be an error, he says, and the flip side is also possible: a result that researchers think is an error can actually be real. Both incidents can affect patient treatment and progress in science.

1. Ribi, S. *et al. Oncotarget* **6**, 7727–7740 (2015).
2. Wang, J. *et al. Nat. Methods* **8**, 652–654 (2011).
3. Kinde, I. *et al. Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
4. Schmitt, M.W. *et al. Proc. Natl. Acad. Sci. USA* **109**, 14508–14513 (2012).
5. Schmitt, M.W. *et al. Nat. Methods* **212**, 423–425 (2015).
6. Gregory, M.T. *et al. Nucleic Acids Res.* **44**, e22 (2016).
7. Banerji, S. *et al. Nature* **486**, 405–409 (2012).
8. Mosquera, J.-M. *et al. Nature* **520**, E11–E12 (2015).
9. Pugh, T.J., Banerji, S. & Meyerson, M. *Nature* **520**, E12–E14 (2015).