SCIENTISTS DECONSTRUCT CANCER COMPLEXITY THROUGH GENOME AND TRANSCRIPTOME ANALYSIS

At Cold Spring Harbor Laboratory, scientists used SMRT® Sequencing to decode one of the most challenging cancer genomes ever encountered. Along the way, they built a portfolio of open-access analysis tools that will help researchers everywhere make structural variation discoveries with long-read sequencing data.

When Mike Schatz realized a few years ago that his PacBio® System had reached the throughput needed to process human genomes, he decided to give it a real challenge: the incredibly complicated, massively rearranged SK-BR-3 breast cancer cell line. The genome consists of 80 chromosomes, and that's just the tip of the complexity iceberg.

"We were really interested in sequencing a human genome that would be maximally impactful and that was aligned with our research interest in cancer genomes, where it's been well documented that structural variations play a major role," says Schatz, now an associate research professor of computer science at Johns Hopkins University and an adjunct associate professor of quantitative biology at Cold Spring Harbor Laboratory, where the analysis took place. He notes that despite its importance, structural variation has not been thoroughly studied because short-read sequencers cannot reliably identify these large genomic elements. "One of the really special properties about the PacBio Sequencer is in addition to being able to call SNPs or small variants, we also get to look for large variants such as structural variation," he says.

But as Schatz and his collaborators at Cold Spring Harbor Laboratory and the Ontario Institute for Cancer Research delved into this work, they realized that existing variant callers were tailored to short-read data. To make the most of the large amount of long-read information they were generating, the team wrote a suite of new analysis tools optimized for SMRT Sequencing data. "The tools catering to short-read data just aren't made to capture the awesome information that we can now take advantage of," says Maria Nattestad, a graduate student in Schatz's lab who wrote several of the new algorithms. "Building our own tools was really the only way to go here."

Those tools, which are especially important for understanding structural variation, are now being publicly released to fuel further SMRT Sequencing studies of human genomes. Also coming out soon is the team's detailed analysis of the SK-BR-3 genome and transcriptome, which includes a high-quality assembly as well as a new understanding of gene fusions, the evolutionary history of this cell line, and more.

'Genome Gymnastics'

SK-BR-3 wasn't chosen at random. The Her2-amplified breast cancer cell line was originally isolated in 1970 and has been used extensively in preclinical research for Her2-targeting therapeutics, such as Herceptin. Cited in hundreds of papers, the cell line is dramatically rearranged, possibly because it came from a lung metastasis rather than from a primary tumor.

The cell line was also well known at Cold Spring Harbor, where scientists had used it for other projects and were already familiar with its peculiarities. The widespread rearrangements and gene fusions appealed to Schatz, who believes that such complication may be characteristic of some cancers. "That's really been the driver behind why we're so interested to study it," he says, noting that these extremely rearranged cancer genomes tend to be associated with the poorest patient outcomes.

> "We're at the point where, if you're trying to assemble a genome *de novo*, be it for cancer or model organisms, PacBio has really emerged as the best technology."

De novo sequencing and assembly were the first steps in making sense of the SK-BR-3 genome. With 72-fold SMRT Sequencing coverage, "we got an outstanding assembly of this genome even though it's so complicated," Schatz says, citing a contig N50 size of 2.5 Mb compared to a state-of-the-art short-read assembly with a contig N50 of just 3 kb. "That's nearly a thousand-fold more contiguous going from short-read to long-read assemblies, and it's through that improved assembly that the majority of structural variants were detected."

Using custom-built analysis tools including variant callers Sniffles, by Schatz lab member Fritz Sedlazeck, and Assemblytics, by Nattestad, the scientists found more than 10,000 structural variants in the SK-BR-3 genome ranging in size from 50 bases to millions of base pairs long. "That's an important result, because other studies of cancer genomes with short-read technology only capture a small fraction of that," Schatz says.

Assemblytics, an assembly-based variant caller, detected insertions and deletions of hundreds of novel *Alu* elements throughout the genome, among other structural variants. "That's a good confirmation that we're seeing biologically relevant results," Nattestad says.

Another major discovery involved meticulously characterizing the complicated process that led to

the cell line's Her2 oncogene amplification. The background of the amplification had never been determined, but a new tool called SplitThreader unfolded the region's history. "We see that there's been this fascinating set of processes where that gene region has been fused into other chromosomes, and from there have been a whole series of further amplifications and inversions," Schatz says. "We call that genome gymnastics. It was a very sophisticated process that resulted in this current state."

SplitThreader works by "threading through the genome and finding the likely variants that are connected to each other" to identify the most likely evolutionary path, says Nattestad, who developed the tool. "It shows how the different amplifications and variants have come together to amplify the Her2 oncogene." By reconstructing the history of regions like this, she adds, it may eventually be possible to find entirely new classes of cancer variations.

Two Variants, One Fusion

The team also used the Iso-Seq™ method to analyze the full transcriptome of SK-BR-3, finding as much complexity at the RNA level as they saw in the DNA. "In the Iso-Seq analysis, we see many tens of thousands of novel isoforms," Schatz says. "That's a really strong testament to the long reads, which fully capture an isoform in one sequence — unlike short reads, where you have to infer isoform structure." Her2, for instance, proved to be one of the most complicated genes at the transcriptome level, with dozens of novel isoforms detected. "That was all resolved through PacBio Iso-Seq and had never been documented before," he adds.

Gene fusions were a major focus of the team's transcriptome analysis. In addition to validating many fusions previously reported in this cell line, they also found several novel ones, including some from the translocation between chromosome 8 and the Her2 oncogene on chromosome 17. Perhaps more importantly, the researchers discovered the genetic mechanism responsible for some gene fusions that had always been a mystery: scientists could see them in RNA, but no direct link between the genes had ever been found in the DNA.

"We were able to figure out why this happened, and it's because there's more than one fusion in the genome necessary for bringing these two genes right next to each other," Nattestad says. "In four cases, we found fusions that take place through two variants. You go from one gene into an intermediate sequence through one variant, span along there for a little while, split off again through a second variant, and then you hit the next gene."

The process perfectly explains these gene fusions, but was not previously discovered because scientists didn't have access to full-length isoforms of the gene product nor the high-quality genome sequence made possible by the long reads. "If you only look at one variant at a time, you don't capture much of the complexity or much of the biological story of what this cancer is undergoing," Nattestad adds. Using the Iso-Seq analysis, on the other hand, "shows us the sheer complexity of this genome."

Mike Schatz, Ph.D., is a computational biologist and an expert at large-scale computational examination of DNA sequencing data.

Maria Nattestad is a bioinformatics graduate student in the Schatz lab.

Moving Forward

Conquering SK-BR-3 has Schatz and his team eager to tackle additional cancer genomes, both from cell lines and from patient samples. "We broke into a lot of new territory through this project, especially at the level of algorithmics where we invented entirely new structural variation callers and new systems for analyzing them," Schatz says. "Now we're really excited to apply this to additional samples and see if the complicated oncogene amplifications that have taken place in SK-BR-3 are also driving mechanisms in other cancers."

"We'll continue to focus on samples that have complicated structural variation because that's where a lot of our strength lies," Nattestad says, noting that BRCA1-mutated cancers tend to fit those criteria.

Future projects will rely on SMRT Sequencing; indeed, the Schatz lab is anticipating installation of the higher-throughput Sequel™ System to meet demand. "We're at the point where, if you're trying to assemble a genome *de novo*, be it for cancer or model organisms, PacBio has really emerged as the best technology," Schatz says.

PN: CS115-033116