New Cancer Somatic Variant Caller Promises Greater Accuracy

Mar 29, 2018 | John Gilmore

Premium

NEW YORK (GenomeWeb) – Researchers at the New York Genome Center (NYGC) have developed a somatic variant caller that detects single nucleotide variants and indels by simultaneously analyzing data from tumor and normal cells.

Last week, they published a description of the tool, called Lancet, in *Communications Biology*.

First author and NYGC senior bioinformatics scientist Giuseppe Narzisi said he originally developed a micro-assembly variant caller, Scalpel, while at Cold Spring Harbor Laboratory, which was designed to localize the assembly to small genomic regions.

At NYGC, Narzisi and his team saw that existing micro-assembly methods relied on separate assembly of tumor and matched normal data, which has limitations in regions with low supporting coverage, repeats, and large indels. In order to fix the issue, they further adapted the micro-assembly engine to examine tumor and normal data at the same time.

The team's Lancet algorithm applies a technique called a colored de Bruijn graph, which has been used for processes such as transcriptome assembly, according to senior author Michael Zody, senior director of computational biology at NYGC. Using a colored de Bruijn graph to jointly analyze reads from tumor and matched normal samples increases the accuracy of identifying mutations that are private to the tumor, especially indels, he said.

In an email, Narzisi explained that Lancet "allows researchers to visualize somatic mutations in a graph, providing additional support when confirming a variant of clinical importance."

He went on to highlight that visualization is relevant for indels that are longer than the read length, which are known to be challenging to detect and visualize using traditional alignment-based approaches.

In an email, Michael Schatz, a professor of computer science and biology at Johns Hopkins University, said he believes Lancet cleverly identifies somatic variants in tumor samples. He previously worked with Narzisi at Cold Spring Harbor to develop the Scalpel micro-assembly tool and noted that Narzisi's team showed substantially improved performance over other somatic mutation callers, especially for indel variants.

"Moving forward, I'm eager to apply [Lancet] to future cancer studies as well as to advance the algorithm for novel data types, especially long reads and linked reads," he said.

In their study, the researchers found that the Lancet algorithm scored somatic mutations more precisely than widely used somatic callers in cancer diagnostics, including MuTect, MuTect2, LoFreq, Strelka, and Strelka2.

For the analysis, the team made the best assembly across a designated region, and adding color-coded sections in the graph allowed them to decide which events are likely to be somatic versus germline. If the event was somatic, the team did not see any support for it in the germline sample.

Initially, Zody's team ran the algorithm on virtual tumors by introducing reads that support real germline SNVs and indels in one sample into another sample. By knowing the true somatic variants and changing their variant allele fractions (VAFs) in the sample, the team used the virtual tumors to test the different methods' ability to call somatic mutations at predefined VAFs. They found that Lancet outperformed all other somatic callers in the study on precision and recall curves, especially for indels.

Zody's team also ran Lancet on synthetic tumors originally generated for the DREAM Challenges (Dialogue for Reverse Engineering Assessments and Methods), a non-profit effort that engages questions regarding translational medicine and systems biology. The researchers artificially spiked real cancer mutations into reads from a normal individual and showed that Lancet also outperformed all other somatic callers for indel calling.

The NYGC researchers then analyzed a set of real data from a case of medulloblastoma. Unlike previous datasets, they compiled a curated list of somatic mutations to examine. They found that all callers favored sensitivity over specificity in the dataset, indicating they have been optimized for higher quality data, and false positive indels within short tandem repeats were highly discordant across callers in the medulloblastoma dataset. By comparison, Lancet reported only a small number of false positive indels without losing sensitivity.

Finally, the team looked at a trio of normal tissue, primary tumor, and colorectal cancer metastatic tissue to test the different methods' ability to identify somatic mutations between the lesions. Zody noted that most of the SNVs, but not indels, in the metastasis were also detected in the primary tumor, highlighting the issue of detecting somatic short tandem repeats and integrating indel calls across different methods.

Zody said that the study's main goal was to introduce the Lancet algorithm and demonstrate it works on different kinds of simulated datasets.

He explained that Lancet takes as an input the data from sequencing an entire genome or exome and then analyzes it in 600 base pair chunks. Because the analysis occurs in parallel, the researchers did not measure the amount of compute time required to analyze a single locus.

"Analyzing an entire tumor and normal genome pair at 80x/40x coverage takes about 3,000 compute hours, but this can parallelized to complete in a few days of real time on reasonably priced hardware," Zody said.

The methodology, he said, is especially good at identifying indels that are in the 10s to a few 100 bases long.

"What we see in cancer samples when we have other tools is that we're missing a lot of insertion and deletion variation in that range, some of which might be causal for cancer," he explained.

The researchers believe that Lancet is a step towards more comprehensive calls of somatic mutations in cancer genomes. According to Zody, researchers in the past have had difficulty finding mid-sized indels and therefore do not have a strong sense of their frequency or importance.

In clinical studies, Narzisi noted, it is standard practice to prioritize relevant variants for further validation using costly sequencing technology. Researchers therefore want to ensure that they focus their attention on mutations that actually exist in the sequence.

Zody acknowledged that Lancet struggles with blood cancers due to the difficulty of obtaining a normal control.

"In the case of blood-based cancers, it's a challenge of collecting a normal sample and making sure the normal doesn't have any tumor infiltrating it," Zody said. "Because it's a joint method, there might be some susceptibility to undercalling in the tumor if your normal tissue is skewed."

Zody also noted that compared to other types of algorithms, Lancet does run slower when sequencing the entire genome because of the time required to perform all the micro-assemblies. In order to solve the issue, the team only performed Lancet on a subset of the genome. If researchers have a list of candidate genes they want to guarantee detection of all complex variations for, they could perform the Lancet algorithm on those specific genes.

While Zody said that his team does not have a plan to commercialize the algorithm, they believe that it might have potential clinical applications. Lancet is currently free for academic and research use. Commercial entities that want to use the algorithm as part of a service can contact the NYGC for a licensing deal.

"As a general rule, we're happy to give a limited, no-cost evaluation license to commercial entities that want to use the software," he said. "Our primary goal is to have people use this tech, but [we] don't want people to go off and use this as part of a multi-million dollar package and we don't see any of the [financial] results."

One of the main goals of the team is to improve the Lancet algorithm's time performance. Zody explained that this will involve improving both the core algorithm and filtering for active regions or regions where it has a high probability of finding a variant.

In the future, the research center will also integrate Lancet into its cancer analysis pipelines, and Zody envisions the algorithm as an important element in the center's cancer research. "As part of our cancer pipeline, we think we'll see other studies in the future that are not necessarily Lancet-focused, but that they'll use [Lancet] as a tool for cancer research," he said.

According to Narzisi, the NYGC researchers will also collaborate with an affiliate member to use Lancet to search for low-level somatic mosaicism in isolated structural malformations by testing deep exome sequencing data generated at the NYGC.

Narzisi and his team will also extend the tool's capabilities in order to take advantage of recent long-range technologies, such as 10x Genomics linked reads. He believes that this will allow more accurate detection of somatic mutations in heterogeneous tumors. For example, Narzisi highlighted that the tool may precisely determine what fraction of cancer cells might harbor particular mutation, which is critical to understand tumor evolution.

Filed Under Informatics Cancer Sequencing colorectal cancer NYGC bioinformatics

We recommend

ICGC Study Highlights Challenges in Consistent Somatic Variant Calling, Tips for Improvement
GenomeWeb

Mt. Sinai to Open NGS Facility, Initially Equip with Ion Torrent Platforms
GenomeWeb

Team Unveils Reference-free Method for Simultaneously Detecting Multiple Somatic Mutation Types
GenomeWeb

Q&A: Toby Bloom on Establishing Informatics Infrastructure for the New York Genome Center
GenomeWeb

Einstein Researchers Publish Single-Cell Sequencing Prep Method, Launch Firm to Provide Service
GenomeWeb

Broad Institute's CRSP Lab Moves Tools, Best Practices from Research to Clinical Context
GenomeWeb

Powered by