Using Big Data to Better Understand Cancerous Mutations

Artificial intelligence and machine learning can help identify genetic changes that lead to cancer—and which treatments work best.

July 12, 2022 By Greg Glasgow and University of Colorado Cancer Center

---

Artificial intelligence and machine learning are among the latest tools being used by cancer researchers to aid in detection and treatment of the disease.

One of the scientists working in this new frontier of cancer research is University of Colorado Cancer Center member Ryan Layer, PhD, who recently published a study detailing his research that uses big data to find cancerous mutations in cells.

“Identifying the genetic changes that cause healthy cells to become malignant can help doctors select therapies that specifically target the tumor,” says Layer, an assistant professor of computer science at CU Boulder. “For example, about 25% of breast cancers are HER2-positive, meaning the cells in this type of tumor have mutations that cause them to produce more of a protein called HER2 that helps them grow. Treatments that specifically target HER2 have dramatically increased survival rates for this type of breast cancer.”

Scientists can evaluate cell DNA to identify mutations, Layer says, but the challenge is that the human genome is massive, and mutations are a normal part of evolution.

“The human genome is long enough to fill a 1.2 million-page book, and any two people can have about 3 million genetic differences,” he says. “Finding one cancer-driving mutation in a tumor is like finding a needle in a stack of needles.”

Scanning the Data

The ideal method of determining what type of cancer mutation a patient has is to compare two samples from the same patient, one from the tumor and one from healthy tissue. Such tests can be complicated and costly, however, so Layer hit upon another idea — using massive public DNA databases to look for common cell mutations that tend to be benign, so that researchers can identify rarer mutations that have the potential to be cancerous.

“There was a project called the Genome Aggregation Database, or gnomAD, out of the Broad Institute, where they put together a bunch of different studies that were going on within the Broad

into the single largest genetic database that anybody has ever even thought about," Layer says. "It was 65,000 individuals at first, and now it's around half a million individuals. At the time I was at the University of Utah doing research in the undiagnosed rare disease clinic, and the usefulness of that database was just beyond belief."

Even if he was able to sequence a child with cancer and her parents, Layer says, there often were so many genetic mutations that it was difficult to determine which one was causing the disease. Using gnomAD, he could look to see how often a certain variant occurred in a larger population, greatly reducing the number of therapeutic targets.

Verifying Variants

Inspired by that experience, Layer began looking at other ways to use big data to identify potentially cancerous mutations. Knowing that detection of complex DNA mutations called structural variants (SV) frequently can result in false negatives, he and his colleagues developed a process that focuses on verification instead of detection. This method searches through raw data from thousands of DNA samples for any evidence supporting a specific structural variant.

"We scanned the SVs identified in prior cancer studies and found that thousands of SVs previously associated with cancers also appear in normal healthy samples," Layer says. "This indicates that these variants are more likely to be benign, inherited sequences rather than disease-causing ones."

The team also found that its method performed just as well as the traditional strategy that requires both tumor and healthy samples, opening the door to reducing the cost and increasing the accessibility of high-quality cancer mutation analysis.

"With all the data that exists for cancer, we were able to show that this method is really powerful for identifying not necessarily the driving mutation in cancer, but what variants are unique to the tumor, versus the rest of your body," he says. "That way, tumor treatment can become super-personalized. We can say, 'If you have this mutation, use this drug; if you don't have this mutation, don't use that drug.'"

Sharing the Research

Layer's lab has now deployed a website where doctors can enter information on structural variants found in a patient tumor to see how common — and potentially dangerous — they are. He is also looking to build a larger cancer-focused dataset to help better understand how and where tumors are formed.

"Our work so far has been to take a structural variant and look to see how frequent it is in a healthy population," he says. "But what if we make indexes that allow you to search our populations? Let's say you take a sample of a tumor in a lung and you find structural variants — now you can search those against prostate cancer and breast cancer and all the other cancers, and it might help you identify, 'What is the origin of the tumor?' 'Has it metastasized, or did it originate in the lung?' We can search the tumor databases to try to find other matched tumors for

more personalized medicine-inspired treatments.”

This article was originally published June 29, 2022, by the University of Colorado Cancer Center. It is republished with permission.
