Cancer genome statistics lesson

By Chris Cain, Senior Writer

The **Broad Institute of MIT and Harvard** has upgraded its original algorithm for cancer genome analysis to account for gene-specific differences in mutation rates.¹ Application of the improved algorithm could help identify previously overlooked mutations, and in one type of cancer it already allowed the researchers to narrow down the number of associated mutations.

Numerous large-scale cancer genome sequencing projects launched in the past decade are beginning to generate a wealth of genetic data on a growing number of tumor types. Interpretation of these data has required the development of new statistical approaches to help identify somatic mutations in tumors that are significantly associated with disease.

Several groups led by the Broad Institute have developed algorithms that account for factors such as genomewide background mutation rates, which vary from cancer to cancer, and may determine associations between genes and disease.

However, potential issues with the accuracy of these models, which may go unnoticed when applied to small data sets, can become amplified when analyzing large data sets.

"We knew there was a problem several years ago, but it didn't manifest that severely," Gad Getz, director of cancer genome computational analysis at the Broad Institute, told *SciBX*. "It became an acute problem as larger data sets were generated. When you have a small data set, for a mutation to be considered significant it must have a signal much higher than background. As sample sizes increase, you get the statistical power to see smaller variations compared with the background model, and if you have an inaccurate model, this becomes a real problem."

Thus, Getz and his team set out to build a more accurate algorithm for large-scale cancer genome analysis.

First, the team built a data set to inform the model that integrated whole-genome and whole-exome sequences from about 3,000 patient tumor samples, along with matched data from the patients' healthy tissue. About 92% of the data were collected at the Broad Institute.

The team first quantified the background mutation rate for 27 sequenced cancer types, then quantified the spectrum of mutations and found that they varied by cancer type in predictable ways. For example, melanomas were more prone to cytosine-to-thymidine mutations, which are caused by misrepair of UV-damaged base pairs.

Most importantly, the group used the sequencing information to determine regional differences in mutation rates across the genome,

enabling the analysis and quantification of gene-specific rates of mutation.

Interestingly, the levels of individual genes inversely correlated with their mutation rates, meaning genes with low expression were more frequently mutated. The average mutation rate for the 25% of genes with the lowest expression was roughly 3-fold higher than that for the 25% of genes with the highest expression. Although the link between gene levels and mutation rates had been previously shown, this analysis represents the most extensive quantification to date of the effects of transcription on mutation rate.

More unexpected was the finding that gene-specific mutation rates also varied based on the replication timing of a given gene during the cell cycle. The latest 10% of genes replicated had about a 2-fold higher mutation rate than the earliest 10% of genes replicated.

The authors used the new insights to update the Broad Institute's algorithm for cancer genome analysis, dubbed Mutational Significance (MutSig). The key improvement in the updated version, called MutSig Covariate (MutSigCV), is that it uses gene expression and replication timing—which co-vary with mutation rate—to estimate gene-specific mutation rates.

When applied to a recently published analysis of squamous cell lung cancer genome sequences conducted by members of Getz's team as part of **The Cancer Genome Atlas** (TCGA) project,² MutSigCV narrowed down the number of mutated genes significantly associated with the cancer from 450 to 11.

Results were published in *Nature*.

Writing in *Nature*, Getz and his team said inaccuracies in large-scale analysis likely have caused numerous false-positive results to show up in the literature. "The expectation has been that larger sample sizes will increase the power both to detect true cancer driver genes (sensitivity) and to distinguish them from the background of random mutations (specificity)." But "recent results seem to show the opposite phenomenon: with large sample sizes, the list of apparently significant cancer-associated genes grows rapidly and implausibly," according to the authors.

Indeed, the data in the paper show that cancer-associated genes such as olfactory genes, large genes and others, described as "highly suspicious" by the authors, had low expression and replicated late, suggesting they are prone to higher mutation rates. Thus, said the authors, mutations in these genes may not be truly associated with cancer more frequently than chance alone, contrary to predictions by algorithms that did not take gene-specific mutation rates into account.

"This is a major advance, and all groups need to consider these ideas and these tools—and they are already doing this," said Getz, who

"Previously the model was using a constant rate of mutation along the genome; now we have a variable rate. This tool could resurrect genes that would not be caught with a naïve model because significant changes in genes with mutation rates lower than average could be better detected."

—Gad Getz,
Broad Institute of MIT
and Harvard

noted that MutSigCV is now the routine algorithm used for analysis by TCGA.

Jun Wang, director of **BGI**, said similar efforts are under way at his institute.

“BGI has already accumulated a huge amount of cancer-omics data, and we are also considering optimizing the algorithm of detecting putative cancer driver genes. There is no doubt that this paper makes a good start for this kind of effort,” said Wang.

Collections catalog

Although advances in cancer genome analysis will undoubtedly improve the quality of the catalog of cancer-associated mutations, it may have little practical effect on the industry’s pursuit of cancer targets.

“The described algorithm will aid defining cancer-causing gene candidates and distinguishing them from passenger genes. Unfortunately, most cancer genome sequencing efforts remain more or less sophisticated counting exercises that contribute little to the identification of driver genes or genes important for understanding—or treating—cancer. Functional studies cannot be replaced with mathematical analyses,” said Christoph Lengauer, CSO of **Blueprint Medicines**.

Blueprint is developing selective kinase inhibitors targeting cancers driven by genomic alterations.

Markus Warmuth, president and CEO of **H3 Biomedicine Inc.**, said the new algorithm represents a major advance but cautioned there is a trade-off between being too stringent about statistical cutoffs and missing possible cancer-associated genes on the one hand and being too lax and generating catalogs of false positives on the other.

“To some degree, it depends on what you want to accomplish with the data. If you are at a major medical center, and you have 25 targeted therapies in your repertoire, you only want actionable mutations and you want to be as stringent as possible. However, in the research side or when looking at candidates for drug discovery, you may want to be a little less stringent to cast a wider net, and to follow up with functional studies to rule out false positives,” he said.

Getz emphasized that the new algorithm should not be viewed solely

as being more stringent than prior analysis methods. “We are not just ruling out genes with this method; it also has the potential to find more genes. Previously the model was using a constant rate of mutation along the genome; now we have a variable rate. This tool could resurrect genes that would not be caught with a naïve model because significant changes in genes with mutation rates lower than average could be better detected.”

Getz said his team will continue to improve upon MutSigCV as additional data are generated, and whole-genome sequencing information will be of particular importance. In this study, 2,957 whole-exome data sets were analyzed, compared with only 126 whole-genome data sets.

Yong Hou, director of the cancer research division of the BGI, agreed that increased access to whole-genome data sets is key. “The continuous optimizing of the algorithm for detecting putative cancer genes, accompanied by the accumulating of more and more cancer-omics data, especially on the whole-genome level, is still a major task for cancer genomics research,” he said.

The Broad Institute has filed for patents covering MutSigCV. The algorithm is freely available for not-for-profit use and will be available for licensing to for-profit organizations.

Cain, C. *SciBX* 6(27); doi:10.1038/scibx.2013.676

Published online July, 18, 2013

REFERENCES

1. Lawrence, M.S. *et al. Nature*; published online June 16, 2013; doi:10.1038/nature12213
Contact: Gad Getz, Broad Institute of MIT and Harvard, Cambridge, Mass.
e-mail: gadgetz@broadinstitute.org
Contact: Eric S. Lander, same affiliation as above
e-mail: lander@broadinstitute.org
2. The Cancer Genome Atlas Research Network. *Nature* **489**, 519–525 (2012)

COMPANIES AND INSTITUTIONS MENTIONED

BGI, Shenzhen, China

Blueprint Medicines, Cambridge, Mass.

Broad Institute of MIT and Harvard, Cambridge, Mass.

The Cancer Genome Atlas, Bethesda, Md.

H3 Biomedicine Inc., Cambridge, Mass.