# 'Purity' of tumor samples may significantly bias genomic analyses

December 4 2015

A new study by UC San Francisco scientists shows that the proportion of normal cells, especially immune cells, intermixed with cancerous cells in a given tissue sample may significantly skew the results of genetic analyses and other tests performed both by researchers and by physicians selecting precision therapies.

It has long been known that tumors may contain a variety of healthy cells as well as cancerous cells, and it is believed that this heterogeneity underlies resistance to various cancer therapies. But this problem has not been thoroughly investigated, said the researchers, who show that factoring precise measures of tumor "purity" into common analytical techniques may clarify some basic principles of cancer biology as well as open new therapeutic avenues for patients. (In research parlance, pure tumors are those that are entirely, or mostly, composed of cancer cells.)

In one medically relevant example from the new work, reported in the Nov. 4, 2015 issue of *Nature Communications*, the team found that measures used to predict the effectiveness of checkpoint-inhibitor drugs, the most widely used form of cancer immunotherapy, are accurate only when the extent of infiltration of immune cells into the tumor was explicitly quantified. When this aspect of tumor purity was not accounted for, estimates of the likely success of immunotherapy were either too high or too low.

"Tumor purity is a big problem when you're dealing with fresh tissue from real patients rather than with cell lines, and there has been no

systematic analysis of this issue," said first author Dvir Aran, PhD, a postdoctoral associate in the laboratory of Atul Butte, MD, PhD, director of UCSF's Institute for Computational Health Sciences (ICHS). "In the case of immunotherapy, it's an expensive treatment and it can have side effects," Aran said, "so it's important to know which patients are most likely to benefit. If we pay more attention to the immune cells that are actually in tumors we may have more success."

For their study, the research group—which also included ICHS member Marina Sirota, PhD, assistant professor of pediatrics at UCSF—made use of a massive dataset known as The Cancer Genome Atlas (TCGA), a joint initiative of the National Cancer Institute and the National Human Genome Research Institute. The TCGA dataset is derived from samples of tumors and normal tissue from 11,000 patients, and represents 33 types of cancer.

Using this resource, the team used four different methods to measure tumor purity in more than 10,000 TCGA samples representing 21 cancer types, and examined how purity might affect the reliability of three of the most common genomic methods used in cancer research: correlation, clustering, and differential analysis.

Correlational techniques reveal so-called co-expression networks—genes that tend to be expressed together most frequently in tumor samples—with the aim of identifying molecular pathways that drive malignancy and metastasis. In a type of bladder cancer known as bladder carcinoma, for example, two genes called JAK3 and CSF1R tend to be jointly expressed at high levels, which suggests that they somehow act together to drive the cancer.

But the UCSF team found that the tandem expression of JAK3 and CSF1R in bladder carcinoma varied widely if tumor purity was taken into account—in the purest samples there was little correlation between

the expression of the two genes, calling their potential joint role in a cancer-driving pathway into question.

Similar potentially disruptive effects of tumor purity were seen in clustering, which groups cancers into subtypes based on molecular markers, with the hope of arriving at more precise treatments, and in differential analysis, which compares gene expression in tumors and normal tissue in order to uncover genetic flaws distinctive to cancer.

In their analyses of samples of lung cancer, kidney cancer, and thyroid cancer, the group found that, if tumor purity were not taken into account, differential analysis could yield misleading results on the relative expression of proteins called CTLA-4 and CD86, both important targets in cancer immunotherapy.

A high mutational burden in a given tumor, essentially a higher number of cancer-driving genes carrying mutations, is often thought to be a genetic signature that predicts a positive response to immunotherapy, but again the UCSF team found that this measure is highly correlated with the purity of a given sample: "purer" tumor samples that consisted mostly of cancerous cells had a lower overall mutational burden.

Since inflammatory responses induced by immune cells are known to increase mutation rates in cancer cells, the team proposes that tumors carrying more mutations may respond better to immunotherapy simply because more immune cells are already present in the tumor, and can be activated against it.

"Mutational burden is a useful measure, because it identifies genes and pathways that may lead tumors to respond to conventional targeted drugs," said Sirota. "But if it is the greater infiltration of immune cells in a tumor that makes it more sensitive to immunotherapy, we should try to measure that directly as well." To that end, Aran said that he hopes that

simple and inexpensive tests of [tumor](#) purity might soon be built into genomic studies in both the laboratory and the clinic.

"Cancer isn't just one big blob," said Butte. "Instead, tumors are a complex microenvironment containing a number of cell types—normal and cancerous—that all act upon one another. If we hope to advance our understanding of cancer and to devise new treatments we need to truly understand how tumors are made up, and to take that makeup into account when we do genomic studies."

The scientists said that their findings underscore the power of computational approaches to health in general, and the importance of large, open-access datasets like TCGA in particular.

"Datasets are the ultimate commodity," Butte said. "Unlike oil or water, which can only be used once, data can continually generate new insights." And, added Sirota, "though each research team may be looking at the same dataset, they can ask completely different questions."

Provided by University of California, San Francisco