Novel analyses improve identification of cancer-associated genes from microarray data

2 May 2014

Dartmouth Institute for Quantitative Biomedical Sciences (iQBS) researchers developed a new gene expression analysis approach for identifying cancer genes. The paper entitled, "How to get the most from microarray data: advice from reverse genomics," was published online March 21, 2014 in *BMC Genomics*. The study results challenge the current paradigm of microarray data analysis and suggest that the new method may improve identification of cancer-associated genes.

Typical microarray-based gene expression analyses compare gene expression in adjacent normal and cancerous tissues. In these analyses, genes with strong statistical differences in expression are identified. However, many genes are aberrantly expressed in tumors as a byproduct of tumorigenesis. These "passenger" genes are differentially expressed between normal and tumor tissues, but they are not "drivers" of tumorigenesis. Therefore, better analytical approaches that enrich the list of candidate genes with authentic cancer-associated "driver" genes are needed.

Lead authors of the study, Ivan P. Gorlov, Ph.D., Associate Professor of Community and Family Medicine and Christopher Amos, Ph.D., Professor of Community and Family Medicine and Director of the Center for Genomic Medicine described a new method to analyze microarray data. The research team demonstrated that ranking genes based on inter-tumor variation in gene expression outperforms traditional analytical approaches. The results were consistent across 4 major cancer types: breast, colorectal, lung, and prostate cancer.

The team used text-mining to identify genes known to be associated with breast, colorectal, lung, and prostate cancers. Then, they estimated enrichment factors by determining how frequently those known cancer-associated genes occurred among the top gene candidates identified by different analysis methods. The enrichment factor described how frequently cancer associated genes were identified compared to the frequency of identification that one could expect by pure chance. Across all four cancer types, the new method of selecting candidate genes based on inter-tumor variation in gene expression outperformed the other methods, including the standard method of comparing mean expression in adjacent normal and tumor tissues. Dr. Gorlov and colleagues also used this approach to identify novel cancer-associated genes.

The authors cite tumor heterogeneity as the most likely reason for the success of their variance-based approach. The method is based on the knowledge that different tumors can be driven by different subsets of cancer genes. By identifying genes with high variation in expression between tumors, the method preferentially identifies genes specifically associated with cancer. This same feature, tumor heterogeneity, may reduce the ability to identify critical gene expression changes when comparing mean gene expression in adjacent tumor and normal tissues, as tumors of the same type may have different sets of genes differentially expressed.

The results of the study challenge the model that comparing mean gene expression in adjacent normal and cancer tissues is the best approach to identifying cancer-associated genes. Indeed, the team identified high variation in adjacent "normal" tissue samples, which are typically used as control samples for comparison in analyses based on mean gene expression. The study suggests that methods based on variance may help get the most from existing and future global gene expression studies.

Provided by The Geisel School of Medicine at
Dartmouth