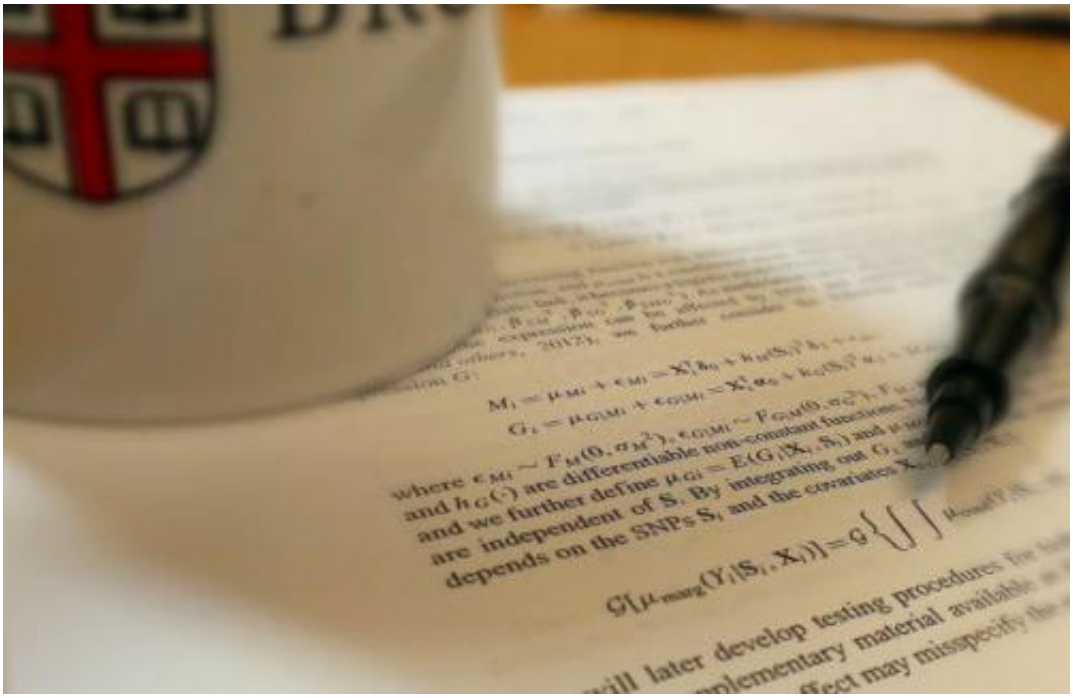
New epidemiology model combines multiple genomic data

April 8 2014



A new statistical model brings critical elements -- single-nucleotide differences in DNA and data on gene expression and methylation -- into the study of associations between genetics and disease. Credit: Brown University

The difference between merely throwing around buzzwords like "personalized medicine" and "big data" and delivering on their medical promise is in the details of developing methods for analyzing and interpreting genomic data. In a pair of new papers, Brown University epidemiologist Yen-Tsung Huang and colleagues show how integrating

different kinds of genomic data could improve studies of the association between genes and disease.

The kinds of data Huang integrates are single-nucleotide differences in DNA, called SNPs, data on gene expression, which is how the body puts genes into action, and methylation, a chemical alteration related to expression. All are potentially relevant to whether a person gets sick, but most analyses that connect genomics to disease account for only one. In papers now online in the journals *Biostatistics* and *Annals of Applied Statistics*, Huang describes the results of testing the model in analyses of asthma and brain cancer data.

"Our integrated approach outperforms single-platform approaches," Huang said. "Applied to real data sets, it works."

Improved performance

The statistical model Huang developed with Tyler VanderWeele and Xihong Lin of Harvard, co-authors on the Annals paper, isn't purely statistical. Its structure and assumptions are informed by the underlying biology. SNPs can be directly associated with disease, or that association can be mediated by whether genes, including the ones in which the SNPs reside, are expressed in healthy or sick patients.

The Annals paper describes the model witih SNPs and expression in detail and its application to data connecting the gene ORMDL3 to asthma. Using the model, the authors found 15 SNPs in the gene with significant associations with disease, compared to only five that have been apparent analyzing SNPs alone. The researchers also found that their "p-values," (a measure of an association's statistical significance) were substantially lower, and therefore stronger, when using the combined analyses their model allows, compared to traditional methods that track just one variable or attempt to mix multiple data sets.

They know the model isn't likely just churning up a lot of false positive SNPs because they also tested it against "null" data where it shouldn't find anything, and indeed it didn't.

Valid with different subjects

Huang further the extends the model, and again reports similar results in *Biostatistics* – new potentially relevant genes and lower p values – in the asthma data set as well as one involving the gene GRB10 and glioblastoma multiforme brain tumors. But this paper makes additional contributions. One of them is showing that the model can be useful even when SNP data and gene expression data come from different people, as long as the subjects are generally comparable. Another is that it integrates not only SNPs and expression, but also DNA methylation data, which is a chemical alteration of DNA associated with expression.

This is important because gene expression and DNA methylation can be tissue dependent. In the case of brain cancer, it's rarely plausible for an epidemiologist to retrieve brain tissue from the same subjects from whom they can more readily sample DNA.

In a new study Huang will conduct with Brown epidemiology colleague Dominique Michaud, he plans to apply the model to new sets of brain cancer data, including DNA from subjects with and without tumors as well as expression data from tissue of people who died, both of brain cancer or other causes.

There could be many other applications as well. The model's general structure of relationships between two variables (one which may mediate the other) and an outcome, he adds, allows it to be applied to similarly structured phenomena, not just to genomics and disease.

"I think our approach is representative of a new framework of data

integration," Huang said. "As long as you can lay out your biological question in terms of this kind of mediation model, then our approach can help you easily analyze your data."

More information: biostatistics.oxfordjournals.o … tics.kxu014.abstract

Provided by Brown University