Minding Your P's and Q's When Studying Genomes

John Storey, Department of Biostatistics

John Storey

In scientific research, you might think it would be impossible to have too much data. New biomedical research technology, however, appears to be giving us more than we can handle. Biologists can now gather literally hundreds of thousands of data points in a single experiment. For example, it is increasingly common to study the activity of the approximately 35,000 different genes in the human genome. Although every cell in an organism contains the same DNA, and thus the same genes, not all of these genes are "expressed," or active. Biologists use a technology called DNA microarrays to find out which genes are active in a biological sample. They can also compare microarrays—for instance, from normal tissue and diseased tissue (e.g., cancer tumor tissue)—in order to determine which genes are expressed differently in the two types.

Humans are actually very bad at looking at a list of numbers—even if just a few dozen—and drawing accurate conclusions. So when analyzing data, scientists often use "statistical hypothesis testing" in order to decide whether an interesting phenomenon is real or not. Traditionally, something called a p-value is calculated, which measures the probability that the data are showing an interesting phenomenon *simply by chance*. In testing for different expression in healthy tissue and diseased tissue, a single gene's p-value gives the probability the gene appears that differentially expressed by chance alone. P-values have been around for a long time, and statisticians have developed many ways in which p-values can be applied. But looking at 35,000 p-values at once is a new problem, and an overwhelming one at that. The traditional p-value approach actually no longer makes sense in such situations.

When searching among 35,000 genes for those that behave one way in healthy tissue, but differently in diseased tissue, then, how does one decide which are really different and which are different by chance alone? Biostatisticians like Dr. John Storey, a new Assistant Professor with a joint appointment in the Department of Biostatistics (SPHCM) and Genome Sciences (School of Medicine), are taking a new look at statistical hypothesis testing in order to tackle this problem.

The method that Dr. Storey and his colleagues have developed is based on the concept of the false discovery rate. At the end of a study, biologists usually pick a set of genomic features that they deem statistically significant for showing a real phenomenon, such as different gene activity in healthy and diseased tissue. The false discovery rate is the proportion of false positives among these statistically significant genomic features. "False positives" are the things that we want to avoid—the features appearing to be significant by chance alone; "true positives" are the real features, which we hope make up most of our list of things we call significant. For example, in a DNA microarray experiment, if 100 genes are said to be expressed differently at a false discovery rate of 5%, then 5 of these are expected to be false positives and 95 to be true positives. Dr. Storey has developed a new measure, called the q-value, that can be used to decide which genomic features are significant in a study. The q-value is similar to the p-value in that each genomic feature can be assigned a q-value. However, it is defined in terms of the false discovery rate. As an example, if all genes with q-values less than or equal to 5% are called significant in a DNA microarray study, then it is expected that 5% of these are false positives – that is, one gets a list of genes with a false discovery rate of 5%.

In addition to DNA microarrays, many other kinds of genomics studies also involve testing thousands of genomic features, where many are expected to be truly significant. Studies involving large-scale genotyping and the analysis of genome sequences are two examples. A general objective in genomics is to identify as many real features as possible without incurring too many false positives. Dr. Storey's q-value method allows genomics researchers to flexibly and carefully do this. •

Further Reading: Storey JD and Tibshirani R. (2003) Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100:9440-9445