Review of *Statistical significance for genomewide studies*

Grace Yi Chen

The paper this week introduces a measure of statistical significance called q value for controlling False Discovery Rate (FDR) when doing hypothesis testing simultaneously on thousands of features in genome-wide studies. With the development of high-throughput sequencing technologies, we need to perform statistical tests on thousands of features and identify as many significant as possible while maintaining low proportion of false positives. The q-value is thus introduced, and the authors give several motivating examples to demonstrate the benefit of using q-values in these cases. In terms of the calculation, the authors first plot the histogram of p-values and assume that the height of the flat portion gives a conservative estimate of the null p-value proportion. Then, we could formulate FDR in terms of the estimate of the proportion of false positive features among the significant portion. Q-value is the minimum FDR that have the feature as significant. The authors later compare the characteristics between p-value and q-value and share the software for q-value estimation.

Similar as the paper last week, this paper is well-written and easy for me to understand the construction of q-values to control FDR. This article has a good motivation to introduce q-values when doing multiple comparison in high-throughput sequencing experiments. As is mentioned in the motivating example, the analyses of high-throughput genetic data involve performing statistical tests on thousands of features in a genomewide study. We would expect that more than a few of the tested features are statistically significant. Controlling FWER would be strict to avoid false positive results and less powerful. However, we need to make some assumptions when estimating q-values. For example, the distribution of p-values that may not always be true if there are dependencies among the hypothesis tests.

Question:

1. Storey also introduces positive FDR from a Bayesian perspective. I wonder if the q-value here has a Bayesian interpretation.