

Stratified False Discovery Rates and q -values for Gene Expression Microarrays

Angelo J. Canty and Shaheena Bashir

Department of Mathematics and Statistics, McMaster University, Canada

One use of gene expression microarrays is to allow scientists to look for genes which are differentially expressed between two conditions. Due to the size of the microarray this leads to a large multiple testing problem which is often handled using a False Discovery Rate (FDR) or q -value approach (Storey, 2002). In this approach the rejection region is chosen such that the estimated FDR, or maximum q -value for genes in the rejection region, is bounded by some user specified limit. This approach, however, treats all target genes as being equally likely to be true positives a priori which may not be a reasonable assumption in some experiments. In work on rodent models for Type 1 Diabetes, we often examine the effect of changing the genetic derivation of one or more well-defined intervals in the genome. We can then compare these *congenic* animals with the parental strain to which they are identical by descent everywhere except in these genetic intervals. The intervals are usually regions known or suspected to be implicated in susceptibility or resistance to Type 1 Diabetes. In looking for the genes that cause this effect, we often compare gene expression between the congenic and parental strains. We clearly expect that there will be a higher proportion of differentially expressed genes in the congenic interval than the rest of the genome. By stratifying the microarray based on membership of this interval we can control the FDR separately in each of the intervals. If there really is a higher proportion of true positives in the congenic interval, this method can give us better estimates of the FDR for the congenic interval and/or increased power to find those effects. We can also calculate stratified q -values based on stratified FDR. Similar to the work of Sun et al (2006) we examine two approaches to apply stratified FDR in the context of congenic experiments. One method retains a fixed rejection region to define significant effects. The rejection region is found by bounding the overall FDR or maximum q -value. We can then calculate the stratified FDR and q -values separately in each of our strata based on this overall rejection region. The second approach defines separate rejection regions for each stratum by bounding the stratified FDR or stratified q -values in each stratum. In this presentation, we will describe these two methods, present a simulation study and applications to a number of congenic rodent experiments from our study of Type 1 Diabetes. We will also compare the stratified approaches to the usual unstratified methods.

References

- [1] Storey JD. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**: 479–498.
- [2] Sun L, Craiu RV, Patterson AD and Bull SB (2006) Stratified False Discovery Control for Large-Scale Hypothesis Testing with Application to Genome-Wide Association Studies. *Genetic Epidemiology* **30**: 519–530