THEORETICAL FALSE DISCOVERY RATE AND SAMPLE SIZE FOR MICROARRAYS

J. Song, P.C. Thomson

University of Sydney and CRC for Innovative Dairy Products, Camden, NSW, Australia

Email: *jiesong@camden.usyd.edu.au*

Recent attempts to account for multiple testing in the analysis of large-scale microarray data have focused on controlling the false discovery rate (FDR). In the microarray context, FDR is the proportion of genes which are declared to be differentially expressed (DE) but which in fact are not. Sample size (the number of slide replicates) is one of the factors that influence FDR, and it is the only one can be controlled by experimenters. FDR can be reduced with more sample size. In our study, at first we calculate FDR mathematically based on the following model: $y_{ij} = \begin{cases} g_i + \varepsilon_{ij} & \text{for DE genes} \\ \varepsilon_{ij} & \text{for non-DE genes} \end{cases}$, where y_{ij} is the observed gene expression value for i^{th} gene on the j^{th} slide, g_i is the i^{th} differentially expressed gene effect, and ε_{ij} is the noise associated with the i^{th} gene on the j^{th} slide. The distribution of the *t* test statistics will follow a mixture of central *t* (non-DE) and noncentral *t* (DE) distributions, from which the FDR is evaluated. We describe an approach to explicitly connect the sample size with the theoretical FDR. Based on this approach, we recommend the minimum sample size for microarray experimental design is six slides. We have also compared the theoretical FDR with the FDR estimated by simulation, and as expected the results are the same, apart from sampling error.