Validity and reliability of preprocessing and differential expression combinations for Affymetrix GeneChip microarrays

黃冠華教授
國立交通大學統計學研究所

Abstract

Background

Microarray technology has been widely used for several years and a large number of computational analysis tools have been developed. We focus on the most popular platform, Affymetrix GeneChip arrays. To evaluate which combinations of preprocessing and differential expression method perform well, we considered 4 popular preprocessing methods (MAS 5.0, RMA, dChip and PDNN) and 5 popular differential expression methods (fold-change, two sample t-test, SAM, EBarrays and limma). We used three spike-in datasets to assess the validity, and ROC curves were used for the evaluation. To evaluate the reliability, we used another dataset from the MAQC project, which was generated using samples hybridized to Affymetrix platform at two different test sites. Overlap rates between two test sites were compared.

Results

Validity was more sensitive to preprocessing method, whereas reliability was more sensitive to differential expression method. Considering both validity and reliability, six combinations were recommended when a small percentage of the genes were spiked-in to be differentially expressed, RMA+fold-change, RMA+SAM, RMA+limma, PDNN+fold-change, PDNN+SAM, and PDNN+limma. Three combinations were recommended when the percentage of the spiked-in genes were large, dChip(PM-only)+fold-change, dChip(PM-only)+SAM, and dChip(PM-only)+limma.

Conclusions

Our results provide general guidelines for selecting preprocessing and differential expression methods in analyzing Affymetrix GeneChip array data.