Abstract

**Background:** Large-scale microarray integration studies present new challenges for statistical analysis. It has been shown that large-scale microarray integration studies are inherently composed of heterogeneous data with varying number of samples per gene. I study existing statistical methods used to identify genes which are over-or down-expressed only in a subset but not all samples in the disease group.

**Results:** I present a new statistical method borrowed from economics and modified to suit gene expression data in large-scale microarray integration studies with an aim of finding differentially expressed genes. I used both simulated and real data to test the detection power of the new method developed for outlier detection across multiple studies performed with different Affymetrix microarray generations and compared it to the existing methods. I demonstrate that the new method performed well in both simulated single study data and in large-scale integrated multi-study data.

**Conclusion:** By comparing and validating the results of the GTI algorithm, I demonstrate that it performed well in the case of simulated data and in a large integrated meta-analysis of multiple clinical datasets. We could uncover novel disease-driven differential gene expression patterns in glioblastomas, such as TYMS and CDKN2A.

Keywords: Gene Tissue outlier Index (GTI); Outlier bioinformatics algorithm; microarray data; differential gene expression.