INTERNATIONAL JOURNAL OF
SYSTEMS AND SYNTHETIC BIOLOGY

1(2) December 2010, pp. 199-226

Evaluating Preprocessing and Differential Expression Combinations for Affymetrix GeneChip Microarrays via Spike-in, RT-PCR and Cross-laboratory Datasets

Ya-Li Wang and Guan-Hua Huang

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

ABSTRACT

Microarray technology for gene expression 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. Despite the rich research on selecting the optimal method of preprocessing and/or detecting differential expression, this paper is unique in the following aspects. First, we have explored suitable combination of preprocessing and differential expression methods. Second, we have evaluated both accuracy and inter-laboratory consistency on a variety of benchmark datasets with distinct characteristics. Third, we have compared stochastic-model-based and physical-mode-based preprocessing algorithms and gene-specific and empirical-Bayes' differential expression detection. We consider popular preprocessing methods: MAS 5.0, PLIER, RMA, dChip and PDNN, and differential expression methods: fold-change, two sample t-test, SAM, limma and EBarrays. Two spike-in datasets and a "real-world-sample" microarray dataset accompanying RT-PCR measurements are used to assess accuracy, and ROC curves are used for the evaluation. To evaluate inter-laboratory consistency, we use a dataset from the MAQC project, which contains arrays generated at two different laboratories using replicated samples. Inter-laboratory overlap rates of differentially expressed gene lists are compared. Our results show that accuracy is more sensitive to preprocessing methods, whereas inter-laboratory consistency is more sensitive to differential expression methods. We conclude that the signal intensity levels are the main factor that explains different performances between methods. We also recommend performing loess normalization at the probe set level.

Keywords: accuracy; inter-laboratory consistency; overlap rate; ROC curve

INTRODUCTION

Microarray for gene expression is a device designed to simultaneously measure the expression levels of many thousands of genes in a particular tissue or cell type. It is widely used in many areas of biomedical research, especially Affymetrix GeneChip platform. Millions of probes with length of 25 nucleotides are designed on an Affymetrix array. Two categories of probes are designed: "perfect match (PM)" probe perfectly matches its target sequence and "mismatch

*Corresponding author: *E-mail: ghuang@stat.nctu.edu.tw*