

**An adjusted test statistic based on shrunken sample variance for identifying differentially expressed genes in microarray experiments**

Akihiro Hirakawa<sup>1</sup>, Yasunori Sato<sup>2</sup>, Chikuma Hamada<sup>3</sup>, Isao Yoshimura<sup>3</sup>

<sup>1</sup>Office of New Drug I, Pharmaceuticals and Medical Devices Agency, Japan

<sup>2</sup>Department of Biostatistics, Harvard School of Public Health, USA

<sup>3</sup>Faculty of Engineering, Tokyo University of Science, Japan

DNA microarrays are a powerful technology for monitoring the expression levels of thousands of genes simultaneously. They provide the basis for a variety of applications, including tumor classification, molecular pathway modelling, and functional genomics. One of the main objectives in the analysis of microarray experiments is the identification of genes that are differentially expressed under two experimental conditions, where researchers expect to determine the smallest possible set of genes that can accurately predict prognostic outcome in clinical practice. To identify true differentially expressed genes, choosing an appropriate test statistic for comparing gene expression levels between two groups is essential. The traditional test statistic is a  $t$ -statistic or Mann-Whitney U-statistic, but they can not reduce both false positive and negative identifications. Therefore, a test statistic which can simultaneously reduce possibilities of false positives and false negatives is required. Principal source of false identifications is the underestimation and overestimation of variance of gene expression levels. For instance, the Welch  $t$ -statistic leaves both the underestimation and overestimation of variance uncontrolled, resulting in an increased risk of identifications of both false positives and false negatives. The  $t$ -type score proposed by Pan et al. (2003) with a correction term added to the denominator of the Welch  $t$ -statistic can reduce false positives by suppressing underestimation of variance, but it leaves overestimation uncontrolled. To suppress overestimation, we propose to use a variance stabilized  $t$ -type score which is obtained by replacing the standard error in the denominator of the  $t$ -type score with other estimator based on shrunken sample variances of James-Stein type. The shrunken sample variances, which utilize information across genes under the James-Stein shrinkage concept, might suppress overestimation of variance. The variance stabilized  $t$ -type score, therefore, is expected to reduce both false positives and negatives. We conducted a simulation study to compare the performances of the Welch  $t$ -statistic,  $t$ -type score, and variance stabilized  $t$ -type score. As the result, the variance stabilized  $t$ -type score was proved to be better than or at least as good as the  $t$ -type score. In particular, the variance stabilized  $t$ -type score outperformed the  $t$ -type score when the sample size was smaller than 5 in each group or the proportion of differentially expressed genes was smaller than 5%. The application of Significance Analysis of Microarray (SAM) (Tusher et al., 2001) to colorectal cancer data (Provenzani et al., 2006) based on the variance stabilized  $t$ -type score suggested that the underestimation and overestimation of variance were simultaneously controlled.