Gene 512 (2013) 82–88

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Methods Paper

# A novel gene expression index (GEI) with software support for comparing microarray gene signatures

Haseeb Ahmad Khan *

*Analytical and Molecular Bioscience Research Group, Department of Biochemistry, College of Science, King Saud University, Riyadh, Saudi Arabia*

ARTICLE INFO

*Article history:*
Accepted 29 September 2012
Available online 8 October 2012

*Keywords:*
Microarray
Gene signatures
Statistical comparisons
Algorithm
Software
Gene expression index

ABSTRACT

This study was aimed to examine the validity of commonly used statistical tests for comparison of expression data from simulated and real gene signatures as well as pathway-characterized gene sets. A novel algorithm based on 10 sub-gradations (5 for up- and 5 for down-regulation) of fold-changes has been designed and testified using an Excel add-in software support. Our findings showed the limitations of conventional statistics for comparing the microarray gene expression data. However, the newly introduced Gene Expression Index (GEI) appeared to be more robust and straightforward for two-group comparison of normalized data. The software automation simplifies the task and the results are displayed in a comprehensive format including a color-coded bar showing the intensity of cumulative gene expression.



## 1. Introduction

Microarray technique requires the post-experimental organization and analysis of vast amount of data. One of the foremost challenges associated with microarray experiment is collecting, managing, and analyzing the emerging data. Recently, scientists have shown interest in evaluating the quality of microarray experiments to ensure the robustness and authenticity of molecular profiling and its clinical applications. Liu et al. (2012) have compared multiple microarray platforms and observed that commercial arrays are more consistent than "in-house" arrays and one-dye platforms are more consistent than two-dye platforms. Li et al. (2012) have presented a method for the determination of functional differences or similarities in microarray data generated from multiple array platforms across all the functions that are presented on these platforms. Fan et al. (2011) have suggested that it is possible to successfully apply multiple-class prediction models across different commercial microarray platforms, offering a number of important benefits such as accelerating the possible translation of biomarkers identified with microarrays to clinically-validated assays. Russ and Futschik (2010) have constructed a consolidated list of platform-independent tissue-specific genes using a set of complementary measures for reliable data interpretation and obtaining biologically more meaningful results.

Although gene clustering is an important tool for the identification of like-groups in a microarray, this methodology cannot be used for two-group comparisons. Another prime goal of microarray analysis is to identify a subset of genes that are differentially expressed between the control and treated samples; this information actually constitutes a gene signature. However, the simplest albeit clinically more valuable microarray experiment is to study changes in gene expression levels between a reference sample (control or untreated) and a diseased or treated sample (Chamberland et al., 2009) or between two populations (Reams et al., 2009). At the data analysis level, the improvement of the detection of differential expression is currently the most common aim of microarray experiments (Hong and Breitling, 2008). Most of the gene selection methods provide the ranking of the genes that are sorted out based on differential expression from highest to lowest (Mukherjee et al., 2005; Tusher et al., 2001). Shaik and Yeasin (2007) have presented a unified framework for the robust selection of genes from microarray experiments while using R-test to convert ranks into *P* values.

Numerous statistical procedures including *t*-test (Notterman et al., 2001), analysis of variance (Bushel et al., 2002), Pearson correlation (Bouras et al., 2002), Mann Whitney *U* test (Kihara et al., 2001) and Wilcoxon signed rank test (Khan, 2004, 2005) have been used for comparison of microarray data. Mootha et al. (2003) initially proposed Gene Set Enrichment Analysis (GSEA) using the Kolmogorov–Smirnov statistic to quantify the degree of enrichment of a set of genes in the entire range of the strength of associations with the phenotype. GSEA was later modified by Subramanian et al. (2005). Although GSEA is not a method for testing self-contained null hypotheses via subject sampling, it is the most widely-used method of gene-set analyses (Dinu et al., 2008). Compared to the independent *t*-test, Limma software (based on paired sample *t*-test) (Smyth, 2004) has consistently shown real

*Abbreviations:* ANCOVA, Analysis of covariance; ANOVA, Analysis of variance; BGA, Between group analysis; GEI, Gene Expression Index; GSEA, Gene set enrichment analysis; ROC, Receiver operating characteristic; SAM, Significance analysis of microarrays; SAM-GS, Significance analysis of microarrays for gene sets.

* Department of Biochemistry, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia. Tel.: +966 1 4675859.
*E-mail address:* khan_haseeb@yahoo.com.

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.09.101

improvement, in particular on small sample sizes (Jeanmougin et al., 2010). Other studies have also shown satisfactory performance of Limma software (Jeffery et al., 2006; Kooperberg et al., 2005). However, the validity of various statistical methods for two-group comparison of gene signatures has not been critically evaluated using specially-designed simulated data representing variable degrees of similarities/differences. Simulation studies have been recommended for evaluating performances of methods under certain conditions of microarray experiments (Dinu et al., 2008). This study was aimed to evaluate the applicability of commonly used statistical methods for two-group comparison of expression data using the simulated as well as real gene signatures. A novel algorithm with software support is presented for robust and comprehensive interpretation of gene signatures.

## 2. Methods

### 2.1. Simulated and real microarray gene expression data

A novel gradation of fold-change with different color codes (Table 1) was used to specifically design six pairs of expression data representing various degrees of similarity/differences as shown in Table 2. Among them, the two groups in Pair-4 are not significantly different whereas the groups in Pair-6 possess the maximum difference (Table 3). The real expression data of published signatures including ovarian carcinoma (Wang et al., 1999), ulcerative colitis (Dooly et al., 2004), leukemia (Golub et al., 1999) and adenocarcinoma (Notterman et al., 2001) were also analyzed. The characteristics of these real signatures have been summarized in our earlier report (Khan, 2005). For comparison of pathway oriented gene sets, the microarray data expressed during Caco-2 cell differentiation were used (Mariadason et al., 2002). We selected 5 of the 25 predefined functional categories and the filtered expression data of the respective gene sets during Caco-2 cell differentiation were used for comparisons.

### 2.2. Computational method and theory of Gene Expression Index (GEI)

The formula used for computation of GEI score is given below:

$$\mathrm{GEI} = \sum \frac{N_{i(0\to t)} S_{j(0\to 1)}}{N_t} \times 100$$

**Table 1**
A new grading system for differential expression of gene signature.

| No. | Fold change | Score |
|---|---|---|
| 1 | < 0.03125 | 1.0 |
| 2 | ≥ 0.03125 and < 0.0625 | 0.8 |
| 3 | ≥ 0.0625 and < 0.125 | 0.6 |
| 4 | ≥ 0.125 and < 0.25 | 0.4 |
| 5 | ≥ 0.25 and < 0.50 | 0.2 |
| 6 | ≥ 0.50 and ≤ 1.5 | 0.0 |
| 7 | > 1.5 and ≤ 3.0 | 0.2 |
| 8 | > 3.0 and ≤ 6.0 | 0.4 |
| 9 | > 6.0 and ≤ 12.0 | 0.6 |
| 10 | > 12.0 and ≤ 24.0 | 0.8 |
| 11 | > 24.0 | 1.0 |

The fold-change scale at serial numbers 1–5 is gradation of down-regulated genes, 6 (green) is a normal range and 7–11 are gradation for up-regulated genes.

Where, $N_i$ is the number of genes with Score $S_j$. The subscript $i$ may vary between 0 and total number of genes in the signature and $j$ may vary between 0 (minimum score) and 1 (maximum score). $N_t$ is the total number of genes in the signature. First, all the ratios of expression data are categorized according to a logical scale (Table 1) to get the respective $N_i$ and $S_j$ values. The percent contributions of each set of genes (genes with same expression score) are computed and then summed up to get GEI score using the above equation.

### 2.3. Software design

This software has been developed in Microsoft Excel platform due to Excel's flexibility, universal availability, and macro-based automation. Moreover, the spreadsheet layout of Excel is perfectly suitable for storing and analyzing microarray data as well as developing microarray analysis software (Khan, 2004, 2005). The data selection is controlled by input box to allow the users to select the paired expression values from any place of the worksheet (Fig. 1). The software then utilizes Excel's worksheet formula function together with a macro subroutine to compute GEI scores (Fig. 2). This automation renders the entire methodology more convenient, faster and error-free as compared to manual procedure. The percent contribution of norm-regulated (green), down-regulated (blue) and up-regulated (red) genes is also shown in a color-coded bar for quick review of two-group comparison.

### 2.4. Installation of software add-in

Open the Excel program. In the 'Tool' menu of Excel workbook, click on 'Add-Ins' and then click on 'Browse'. Locate the drive and double click on 'GEI Install'. The message "Copy GEI to …" will appear, click 'Yes'; the appearance of 'GEI' on the menu bar indicates the proper installation of the Add-in.

### 2.5. Running the program

For computing GEI score, enter the gene expression data of different groups in separate columns. Prior to activating the 'Input' window, ensure that the data to be compared reside in adjacent columns. Once the data entry has been completed, click the 'GEI' button on the menu bar to pop-up the 'Input' window (Fig. 1). Select the range of data (numbers only) without including the header row (if any) as shown in Fig. 1. Now clicking the 'OK' button executes the software and a comprehensive report is displayed (Fig. 2).

### 2.6. Statistical comparisons

In this study, diverge but small data sets were chosen for simple and clear simulation of microarray signatures as well as understanding the critical problems associated with their statistical comparisons. All these six pairs of simulated data (Table 2), four real signatures (Dooly et al., 2004; Golub et al., 1999; Notterman et al., 2001; Wang et al., 1999) and pathway-characterized microarray expression data (Mariadason et al., 2002) were subjected to statistical comparisons. For real gene signatures, the comparisons were performed between control and disease groups. For pathway oriented expression data, we randomly selected 5 of the 25 predefined pathway categories (Mariadason et al., 2002) and the filtered gene sets data were used for statistical comparisons between day 0 and day 21 for the assessment of these functional groups on Caco-2 cell maturation and differentiation. The statistical package SPSS (Version 10) was used for conducting Mann–Whitney *U* test, Kolmogorov–Smirnov test, Kruskal–Wallis test, Wilcoxon signed-rank test, Sign test, Friedman test, Kendall *W* test and paired sample *t*-test. All the above mentioned statistical methods are nonparametric except paired-sample *t*-test, which is a parametric analog of Wilcoxon signed rank test. GEI scores were computed as mentioned above using the software reported in this paper (Fig. 1).