Gene expression

Large scale data mining approach for gene-specific standardization of microarray gene expression data

Sukjoon Yoon$^{1,2,*}$, Young Yang$^{1,2}$, Jiwon Choi$^1$ and Jeeweon Seong$^1$

$^1$Department of Biological Sciences and $^2$Research Center for Women’s Diseases (RCWD), Sookmyung Women’s University, Hyochangwongil 52, Youngsan-gu, Seoul, Republic of Korea, 140-742

Received on May 22, 2006; revised on September 7, 2006; accepted on September 30, 2006
Advance Access publication October 10, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT
Motivation: The identification of the change of gene expression in multifactorial diseases, such as breast cancer is a major goal of DNA microarray experiments. Here we present a new data mining strategy to better analyze the marginal difference in gene expression between microarray samples. The idea is based on the notion that the consideration of gene’s behavior in a wide variety of experiments can improve the statistical reliability on identifying genes with moderate changes between samples.

Results: The availability of a large collection of array samples sharing the same platform in public databases, such as NCBI GEO, enabled us to re-standardize the expression intensity of a gene using its mean and variation in the wide variety of experimental conditions. This approach was evaluated via the re-identification of breast cancer-specific gene expression. It successfully prioritized several genes associated with breast tumor, for which the expression difference between normal and breast cancer cells was marginal and thus would have been difficult to recognize using conventional analysis methods. Maximizing the utility of microarray data in the public database, it provides a valuable tool particularly for the identification of previously unrecognized disease-related genes.

Availability: A user friendly web-interface (http://compbio.sookmyung.ac.kr/~lage/) was constructed to provide the present large-scale approach for the analysis of GEO microarray data (GS-LAGE server).
Contact: yoonsj@sookmyung.ac.kr

1 INTRODUCTION

One of the most popular uses of DNA microarrays is the comparison of differences in gene expression under two distinct experimental conditions (treated versus untreated samples, diseased versus normal tissue, mutant versus wild-type organisms, etc.) (Breitling et al., 2004). In this type of experimental setup, a major challenge is the identification of those genes whose expression is significantly different between two conditions (Aittokallio et al., 2003). Many sophisticated statistical methods have been tested, in attempts to achieve a more reliable identification of differentially regulated genes (Huber et al., 2002; Irizarry et al., 2003; Yang et al., 2001; Yang et al., 2002). In fact, many of the existing statistical methods for microarray analysis have been developed by using datasets, in which changes in gene expression are abundant, with many genes having a high magnitude of change that far exceeds the observed variability in expression (Ramaswamy et al., 2001). However, for many physiological and metabolic conditions, the changes in gene expression are often moderate compared to the array-wide variability in expression, thus leading to modest $P$-values, such that existing statistical models often miss most of the real changes (Mootha et al., 2003).

Therefore, we sought to develop an analytical approach that provides experimental biologists with a more thorough understanding of the statistical significance of any list of genes produced under conditions where the real changes are of modest magnitude but the expression level is still high in both experimental conditions. An important issue is associated with the normalization of the relative expression of genes across a series of microarray experiments (Colantuoni et al., 2002). The normalization across arrays has been extensively used to minimize systematic variations in specific samples (Bolstad et al., 2003; Gautier et al., 2004; Huber et al., 2002; Irizarry et al., 2003; Yang et al., 2002). The selection of appropriate controls for normalization has been proposed for comparisons of expression levels across samples (Yang et al., 2002). A set of controls (microarray sample pool) with minimal sample-specific bias over a large intensity range was introduced to aid in intensity-dependent normalization. In Affymetrix GeneChip arrays, each gene is represented by a set of 11–20 pairs of probes (perfect match and a mismatch), and the intensities for each probe set is summarized by the log scale robust multi-array analysis (RMA) (Irizarry et al., 2003). It has been reported (Li and Wong, 2001) that variation of a specific probe across multiple arrays could be considerably smaller than the variance across probes within a probe set. RMA method effectively accounted for this strong probe affinity effect, and consequently improved the ability to detect differentially expressed genes between samples.

Since using multiple arrays for normalization had improved the detection of differentially expressed genes between samples, we attempted a further development on gene-specific, multi-array standardization method using a large collection of expression data available from the public database. Our focus was to better understand the biological relevance of detected difference of gene expression, rather than improving the fluorescence intensity normalization that many of existing methods (e.g. RMA method) focused on. With the rapid increase of microarray expression data in the public database over the past few years, it has become possible to monitor the general expression level of a gene in diverse biological samples under various conditions. Through the creation

---

$^*$To whom correspondence should be addressed.
Large scale gene-specific standardization of array data

2 METHODS

2.1 Microarray gene expression data acquisition

Microarray gene expression data were obtained from the Entrez Gene Expression Omnibus (GEO) ftp site (ftp.ncbi.nlm.nih.gov/pub/geo) (Barrett et al., 2005; Edgar et al., 2002). A set of single channel microarray data on normal and breast cancer cell lines were analyzed for the demonstration. The GEO Accession no. GDS817, which includes six samples (GSMs) is a collection of microarray experiments for the comparison of gene expression between breast cancer and normal epithelial cell lines (Fig. 1). The two samples in GDS817 are experiments done with the breast cancer cell line, HCC622. Another two involve a normal epithelial cell line. We analyzed the difference in gene expression between these two types of cell lines. The GDS817 experiments were carried out with the Affymetrix U95A DNA chip (GEO Accession no. GPL91), which includes 12651 probesets. GDS817 contains expression data records for 12625 probesets. Thus, our analysis was limited to this subset of probesets in GPL91. In addition to GDS817, a total of 72 additional GDSs sharing a common platform, GPL91, were retrieved from the current version of GEO for the gene-specific GEO-wide standardization procedure. In summary, expression data for 12625 genes in a total of 1850 GSMs from 73 GDSs, which share a common platform, GPL91, were retrieved and analyzed in this study.

2.2 Gene specific large-scale analysis of gene expression (GS-LAGE)

It is a standard practice to correct for foreground intensities by background subtraction (Edwards, 2003). For single channel experiment data deposited in GEO, it is assumed that the values were submitted as normalized (scaled) signal count data [e.g. log (signal—background) transformation]. However, in order to calculate the mean and standard deviation of the expression level of a gene using a large collection of datasets (GDSs) that were prepared and deposited by different research groups, the individual array data from GEO...
needed to be re-standardized to achieve a common scale. Thus, we first carried out within-array Z-transformation on each of all collected GEO samples by using the mean expression level and its standard deviation on the array basis, as below.

\[ u_i = \frac{x_i - \mu_{\text{GSM}}}{\sigma_{\text{GSM}}}, \]

where, \( x_i \): the expression level of the \( i \)th probeset (gene) recorded in the given GEO sample (GSM), \( \mu_{\text{GSM}} \): the mean expression level of all genes in the given GSM, \( \sigma_{\text{GSM}} \): the standard deviation of all genes in the given GSM. \( u_i \): the standardized intensity of the \( i \)th gene in the given GSM. This GSM-based standardization process was repeated for all 1850 GSMs sharing the GPL91 platform, to permit comparisons between samples to be made. We next calculated the mean expression level and its standard deviation of each gene across arrays using the Z-transformed expression data from 1850 GSMs.

\[ \mu_j,\text{GPL} = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \]

\[ \sigma_{\text{GPL}} \]

\[ u_j,\text{GPL} \]

\[ n: \text{the total number of GSM in the given GPL (in this case, } n = 1850), \]

\[ \mu_j,\text{GPL} \text{ is thus the GPL-wide mean of } i \text{th gene. The Z-transformed expression level } (u_j) \text{ in a specific assay (GSM) was next re-scaled by the model established from the GPL-wide mean and standard deviation.} \]

\[ v_i = \frac{u_i - \mu_j,\text{GPL}}{\sigma_j,\text{GPL}} \]

\[ \sigma_{\text{GPL}}: \text{the standard deviation of GPL-wide expressions of } i \text{th gene.} \]

The expression level, \( v_i \), of a gene thus represents the re-scaling based on the observation of its expressional behavior in a large collection of diverse experiments. It is assumed that the expressional variation of a gene in the database follows the normal distribution, \( N(\mu_{j,\text{GPL}}, \sigma_{j,\text{GPL}}) \) when \( n \) is large. To confirm this assumption, we implemented an iterative procedure to remove outliers from the final \( \mu_{j,\text{GPL}} \) and \( \sigma_{j,\text{GPL}} \) calculations. An outlier was defined as an expression level of which distance to the original \( \mu_{j,\text{GPL}} \) is three times greater or smaller than the original \( \sigma_{j,\text{GPL}} \). Then, final \( \mu_{j,\text{GPL}} \) and \( \sigma_{j,\text{GPL}} \) values were used for chi square goodness of fit test between observed distribution and its ideal normal distribution. Genes whose expressional variation is significantly deviated from normal distribution (\( P < 0.05 \) in the chi square test) were removed from the second standardization (\( v \) calculation).

In this study, we compared \( u_i \) and \( v_i \) in identifying changes in gene expression between normal and breast cancer cells. \( u_i \) is assumed to be a quantity provided by the original contributor, which only considers the expression data within the GDS for the preparation. The difference in normalized gene expression between normal and breast cells was calculated as below.

\[ \Delta u_i = \frac{u_i^+ - u_i^-}{\text{avrt}(|\Delta u|)} \]

where \( u_i^+ \): the standardized expression level of the \( i \)th gene in normal cells, \( u_i^- \): the standardized expression level of the \( i \)th gene in breast cancer cells, \( \text{avrt}(|\Delta u|) \): the absolute average difference between \( u_i^+ \) and \( u_i^- \). \( \Delta u_i \) represents the relative change in gene expression in comparison to the average change in the given assay. On the other hand, the relative change in gene expression was calculated using the \( \Delta v_i \) value where the intensity level was re-scaled based on the gene-specific behavior in a wide variety of experiments.

\[ \Delta v_i = \frac{v_i^- - v_i^+}{\text{avrt}(|\Delta v|)} \]

where \( v_i^+ \): the standardized expression level of the \( i \)th gene in normal cells, \( v_i^- \): the standardized expression level of the \( i \)th gene in breast cancer cells, \( \text{avrt}(|\Delta v|) \): the absolute average difference between \( v_i^+ \) and \( v_i^- \).

### 2.3 Validation using an Affymetrix spike-in study dataset

For the validation study, a dataset from the spike-in study by Affymetrix was retrieved from Affycomp website (http://affycomp.biostat.jhsph.edu/). In this dataset, Human cRNA fragments matching 16 probesets on the HG-U133A GeneChip were added to the hybridization mixture of the arrays at concentrations ranging from 0 to 1024 pM. The same hybridization mixture, obtained from a common tissue source, was used for all arrays. The details of the spike-in data are found in the literature (Cope et al., 2004; Irizarry et al., 2003) and the website (http://affycomp.biostat.jhsph.edu/). The fluorescence intensities were normalized by RMA method. These RMA data were used for further analysis by the present two-step standardization method after log2 transformation.

### 2.4 Gene expression data analysis by GEO-provided tools

For comparison with the present method, we also analyzed the gene expression data between breast cancer and normal cell lines using GEO-provided analytical tools. Three methods were used to generate lists of probesets which showed significant higher expression levels in the breast cancer cell line than the normal epithelial cell line. The options were appropriately selected as below to include similar number of entries in the final lists. Here, \( 'A' \) represents gene expression data on the breast cancer cell line (HCC1954) and \( 'B' \) represents gene expression data on normal epithelial cell line.

1. One-tailed t-test (A > B) (0.010 significance level) selected a total of 254 probesets.
2. Query mean group A versus B by values (4-fold higher) selected a total of 266 probesets.
3. Query mean group A versus B by ranks (3-fold higher) selected a total of 204 probesets.

### 3 RESULTS

Since the downloaded data for the demonstration were the normalized and combined Affymetrix data from GEO (GPL91), the MA plot of test samples from GDS817 was already in a good shape (Fig. 2A). The present two-step standardization procedure was applied to these datasets in order to achieve a better resolution in detecting the modest expressional difference in genes with relatively small DB-wide standard deviation, \( \sigma_{\text{GPL}} \). We first carried out within-array Z-transformation on each of all collected samples by using the mean expression level and its standard deviation on the assay basis. The resulting scatter plot (Fig. 2B) is same as the MA plot (Fig. 2A). After the first step, array-specific standardization, the difference in gene expression between normal and cancer cells showed a symmetrical distribution along the diagonal axis (Fig. 2B). After \( u_i^+ \) and \( u_i^- \) are re-scaled via gene-specific normal distribution, i.e. \( N(\mu_{j,\text{GPL}}, \sigma_{j,\text{GPL}}) \) for the \( i \)th gene, the relatively large deviation in gene expression between normal and breast cancer cells were shifted to the high intensity region (Fig. 2C). This observation suggests that genes with large \( \mu_{j,\text{GPL}} \) have more variation in \( \sigma_{j,\text{GPL}} \) than those with small \( \mu_{j,\text{GPL}} \).

To further investigate the difference between Figure 2B and C, we plotted the mean (\( \mu_{\text{GPL}} \)) and standard deviation (\( \sigma_{\text{GPL}} \)) of each gene expression in 1850 experiments (Fig. 3). These mean and standard deviation were used for the second standardization (\( v \) calculation). It has been well recognized that the variance of the measured spot intensities increases with their mean. The standard deviation increases roughly linearly with the mean (Huber et al.,...
2002). The present plot of the DB-wide analysis also indicates that expressional variation (i.e. $\sigma_{\text{GPL}}$ in the plot) is increased as $\mu_{\text{GPL}}$ increase. In addition, the plot shows that the vertical distribution of $\sigma_{\text{GPL}}$ becomes wider as the $\mu_{\text{GPL}}$ increases. This relatively large variation in $\sigma_{\text{GPL}}$ values in the high $\mu_{\text{GPL}}$ region contributed to the difference of the plot shape between Figure 2B and C. For a given $\mu_{\text{GPL}}$ Value, a large $\sigma_{\text{GPL}}$ resulted in a small $\nu$ value, while a small $\sigma_{\text{GPL}}$ value resulted in a large $\nu$ in the second standardization. This analysis confirms that the present model, based on $N(\mu_{\text{GPL}}, \sigma_{\text{GPL}})$ provides an additional resolution particular to recognizing the expressional difference of genes with relatively small $\sigma_{\text{GPL}}$ and high expressional intensity, i.e. large $\mu_{\text{GPL}}$.

Figure 2A and B shows that the single step standardization actually represents the original normalization of Affymetrix GeneChip data. However, the second step standardization generated a substantial shift in the distribution from that of single step standardization of gene expression. From a biological point of view, a large expressional change between specific samples for a gene which has a large $\sigma_{\text{GPL}}$ may be less meaningful than a moderate expressional change for a gene having a consistent expression in the database (low $\sigma_{\text{GPL}}$). Typical array-specific standardization and a consequent comparison of the intensity of gene expression between samples lack this kind of biological consideration. In this sense, our two-step analysis provides a unique tool for identifying additional genes with moderate changes between samples that are not highly prioritized by conventional assay-specific standardization methods.

We compared the performance of the two methods (i.e. $\Delta u'$ versus $\Delta v'$ scorings) in evaluating differences in gene expression between normal and breast cancer cells (Fig. 4A). A significant number of genes with low and high expression levels were evaluated differently by these two methods. The result shows that the utility of two-step standardization method is on identifying those genes in which the difference of expression between samples is underestimated by the single step, array-specific standardization method. On the upper boundary region of the distribution shown in Figure 4A, the $\Delta v'$ calculation gives up to 2-fold higher estimation than the $\Delta u'$ calculation for gene expression difference between samples. Among 10792 test genes from GPL91, 90.5% showed the discrepancy of less than 1.0 between two methods (i.e. $|\Delta v' - \Delta u'| < 1.0$), while 9.5% of genes showed the discrepancy of greater than 1.0 (Fig. 4B).
A total of 2.4% of genes showed the discrepancy of greater than 2.0 between two methods.

We also compared the result of the present method with that of RMA method. RMA method is developed for better normalization of fluorescence intensity data by accounting for probe affinity effect, while our present method is for providing better biological insight on interpreting the expressional difference of genes in the database that are assumed to have already been properly normalized. We thus applied the present method to the data that were already normalized by RMA method. For this comparative analysis, a dataset from the spike-in study by Affymetrix was used. Human cRNA fragments matching 16 probesets on the HGU95A GeneChip were added to the hybridization mixture of the arrays at concentrations ranging from 0 to 1024 pM. The same hybridization mixture, obtained from a common tissue source, was used for all arrays (See Methods section for details). The fluorescence intensity data were normalized by RMA method, and then the present two-step method was applied to the normalized data. Observed concentrations are comparatively plotted against nominal concentration (Fig. 5). In this analysis, the observed intensities are averaged at each nominal concentration value, resulting in a single mean curve. Since the log2 scale was applied to the concentrations, observed concentrations should be linear in true concentrations. We therefore fit a simple linear model to the scatterplot data and report the $R^2$ coefficient. The result shows

$$R^2 = 0.99$$

Fig. 5. Average observed log2 intensity plotted against nominal log2 concentration for each spiked-in gene for all arrays in Affymetrix spike-in experiment. The resulting slope estimates are plotted against average log intensity across all concentrations. LAGE represents the present two-step standardization method.

Table 1. Prioritization of test spike-in genes by the two-step standardization

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>$\Delta$RMA</th>
<th>$\Delta v_0$</th>
<th>$\mu_{GPL}$</th>
<th>$\sigma_{GPL}$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>407_at</td>
<td>1.03</td>
<td>1.77</td>
<td>-0.33</td>
<td>0.11</td>
<td>0.89</td>
</tr>
<tr>
<td>37777_at</td>
<td>1.04</td>
<td>1.69</td>
<td>-0.33</td>
<td>0.11</td>
<td>0.99</td>
</tr>
<tr>
<td>40322_at</td>
<td>1.24</td>
<td>1.52</td>
<td>-0.28</td>
<td>0.11</td>
<td>0.33</td>
</tr>
<tr>
<td>1708_at</td>
<td>1.59</td>
<td>1.34</td>
<td>-0.31</td>
<td>0.09</td>
<td>0.80</td>
</tr>
<tr>
<td>38734_at</td>
<td>1.55</td>
<td>1.26</td>
<td>-0.31</td>
<td>0.14</td>
<td>0.90</td>
</tr>
<tr>
<td>36311_at</td>
<td>1.08</td>
<td>1.10</td>
<td>-0.32</td>
<td>0.12</td>
<td>0.98</td>
</tr>
<tr>
<td>1024_at</td>
<td>0.80</td>
<td>1.01</td>
<td>-0.29</td>
<td>0.10</td>
<td>0.41</td>
</tr>
<tr>
<td>39058_at</td>
<td>0.93</td>
<td>0.98</td>
<td>-0.12</td>
<td>0.13</td>
<td>0.76</td>
</tr>
<tr>
<td>684_at</td>
<td>0.92</td>
<td>0.92</td>
<td>-0.26</td>
<td>0.13</td>
<td>0.84</td>
</tr>
<tr>
<td>36085_at</td>
<td>1.12</td>
<td>0.65</td>
<td>-0.24</td>
<td>0.09</td>
<td>0.98</td>
</tr>
<tr>
<td>1597_at</td>
<td>0.64</td>
<td>0.24</td>
<td>-0.03</td>
<td>0.43</td>
<td>0.14</td>
</tr>
<tr>
<td>1091_at</td>
<td>0.34</td>
<td>0.24</td>
<td>-0.09</td>
<td>0.19</td>
<td>0.56</td>
</tr>
<tr>
<td>33818_at</td>
<td>0.39</td>
<td>0.04</td>
<td>0.58</td>
<td>0.59</td>
<td>1.00</td>
</tr>
<tr>
<td>SD</td>
<td>0.36</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For each gene, observed log2 ratio change between 1 and 2 pM were compared. RMA values were calculated using the retrieved raw data. $\Delta$RMA and $\Delta v_0$ were normalized by the average of 14 data, respectively. $P$-value indicates the statistical confidence of the $N(\mu_{GPL}, \sigma_{GPL})$. 546_at, 36202_at and 36889_at were removed from this analysis because of non-normal distributions of their expressional level in GEO ($P < 0.05$).

Table 2. Top ranking genes (probesets) by $\Delta v_0$ score in comparison with GEO analyses

<table>
<thead>
<tr>
<th>Probesets in GPL91</th>
<th>Top 100 probesets by $\Delta v_0$ score overlapped with GEO analyses</th>
<th>Un-overlapped GEO analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrence</td>
<td>10 791</td>
<td>78</td>
</tr>
<tr>
<td>Average $\mu_{GPL}$</td>
<td>$-0.09$</td>
<td>$-0.27$</td>
</tr>
<tr>
<td>Average $\sigma_{GPL}$</td>
<td>0.21</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Top-ranked 100 probesets were selected based on the $\Delta v_0$ score.

*Uniques probesets identified by $\Delta v_0$ scoring that were not found in the lists by GEO methods (See Methods section for the detail about GEO-provided methods).
that the present two-step standardization (LAGE) data has a similar $R^2$ coefficient with the original RMA data. This confirms that the additional standardization by $N(\mu_{GPL}, \sigma_{GPL})$ does not change the linearity of the original data. However, the rank order of test genes by observed differential expressions ($\Delta u'$ and ARMA) between samples showed a large disagreement between the present method and RMA method (Table 1). When expression difference between samples was compared among 14 genes, the RMA method provided a more consistent performance than the present method. These 14 test genes showed large variations in their $\mu_{GPL}$ and $\sigma_{GPL}$ values. The variation in ARMA among 14 genes is random, i.e. no correlation with their $\sigma_{GPL}$. However, our present method gives additional weights to the expression difference of those genes (e.g. 407_at) that have relatively small $\sigma_{GPL}$, while it gives a low significance on the expression difference of genes with a large $\sigma_{GPL}$ (e.g. 33818_at). We believe that this $\sigma_{GPL}$-dependent prioritization of gene expression difference improves the identification of previously unknown disease-related genes from database search.

For a demonstration of the usefulness of the present method, we compared the performance of the two-step method with those methods provided by GEO website (see Methods section for the detail) in evaluating differences in gene expression between normal and breast cancer cells. A total of 10791 probesets were first ranked based on the $\Delta u'$ score. Then, a total of 100 top-ranked probesets were investigated if they were also prioritized by GEO-provided analytical methods. As a result, 22 probesets on the top ranking list on the $\Delta u'$ score were unique and not found in the GEO analysis reports, while 78 probesets were found in both $\Delta u'$-ranking list and GEO analysis reports (Table 2). These two sets of genes (un-overlapped and overlapped with GEO lists) commonly showed relatively low $\mu_{GPL}$ and $\sigma_{GPL}$ in comparison with those of total 10791 probesets in GPL91. However, the un-overlapped set of probesets showed a lower GPL-wide expression variation (average $\sigma_{GPL} = 0.10$) than the overlapped set (average $\sigma_{GPL} = 0.12$). This result shows that the consideration of DB-wide expression variations ($\sigma_{GPL}$ values) has contributed to the identification of 22 additional probesets that are hardly prioritized by other methods.

A total of 22 probesets (actually 20 different genes) that were exclusively found in the top ranking list by the two-step method needed to be further analyzed to determine if there are breast cancer-related genes that were not identified by GEO analysis reports. A subset of 12 probesets that showed large difference in the ranking between $\Delta u'$ and $\Delta v'$ are listed (Table 3). From a literature search, we found that 3 of these 12 genes were specifically elevated in cancer-related cells. For example the MRE11 (gene ID: AF073362)—Rad50–NBS1 complex is a cell cycle check point protein and tumor cells have defects in the cell cycle check point protein. It has been known that two components of the MRE11–Rad50–NBS1 complex, RAD50 and NBS1 are breast cancer susceptibility genes associated with genomic instability (Heikkinen et al., 2006) and the MRE11 gene is mutated in an ataxia-telangiectasia-like disorder (Stewart et al., 1999). Insufficient information was available to determine if other nine genes exclusively found on the top ranking list of $\Delta v'$ scoring are associated with breast cancer pathogenesis in the current NCBI database. Table 3 shows that the difference of expression intensity between normal and breast cancer cell lines for these genes was estimated to be at least 1.5-fold higher in $\Delta v'$ measure than in $\Delta u'$ measure. It can be concluded that relatively low $\sigma_{GPL}$ values compensate the moderate difference in expression between samples and consequently rank genes in a different order. Further experimental study remains to confirm the association of the selected nine genes with breast tumors.

The relative merit of the present method depends on its ability to successfully identify genes that are differentially expressed, while avoiding classifying highly fluctuating genes (i.e. genes with large $\sigma_{GPL}$) as being differentially expressed (i.e. their false positive or Type I Error rate). Since the false positive rate increases exponentially as the rank goes to the bottom (Norris and Kahn, 2006), medium-level fold changes (moderate $\Delta v'$) in gene expression were usually not considered for further experimental validation. In this sense, this new approach can enrich the hit list of genes in which expression difference between normal and breast cancer cells were moderate.

For public access to this two-step standardization method for GEO gene expression data, we constructed a user friendly web-based database, GS-LAGE (Gene Specific Large-scale Analysis of Gene Expression), which includes all single channel microarray experiments listed on the GEO database (Fig. 6). It can

---

**Table 3.** Cancer-related genes exclusively identified by the present $\Delta v'$-scoring method

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>$\Delta u'$ (rank)</th>
<th>$\Delta v'$ (rank)</th>
<th>$\mu_{GPL}$</th>
<th>$\sigma_{GPL}$</th>
<th>Gene symbol, name</th>
</tr>
</thead>
<tbody>
<tr>
<td>D49354</td>
<td>4.1 (133)</td>
<td>7.1 (50)</td>
<td>-0.24</td>
<td>0.09</td>
<td>Enhancer protein in HSP70</td>
</tr>
<tr>
<td>AF073362</td>
<td>4.7 (101)</td>
<td>6.7 (53)</td>
<td>-0.27</td>
<td>0.10</td>
<td>MRE11A, Meiotic recombination 11 homolog A</td>
</tr>
<tr>
<td>AF016370</td>
<td>4.1 (134)</td>
<td>6.4 (64)</td>
<td>-0.19</td>
<td>0.09</td>
<td>U4/U6 small nuclear ribonucleoprotein</td>
</tr>
<tr>
<td>D87682</td>
<td>3.4 (208)</td>
<td>6.3 (65)</td>
<td>-0.29</td>
<td>0.08</td>
<td>KIAA0241 gene</td>
</tr>
<tr>
<td>Y07566</td>
<td>4.3 (121)</td>
<td>6.3 (67)</td>
<td>-0.21</td>
<td>0.10</td>
<td>RIT protein (Ras-like GTPase)</td>
</tr>
<tr>
<td>L03840</td>
<td>4.1 (138)</td>
<td>6.2 (68)</td>
<td>-0.23</td>
<td>0.10</td>
<td>Fibroblast growth factor receptor 4 (FGFR4)</td>
</tr>
<tr>
<td>AF038925</td>
<td>3.6 (191)</td>
<td>6.0 (73)</td>
<td>-0.27</td>
<td>0.09</td>
<td>Jak2 kinase</td>
</tr>
<tr>
<td>L42452</td>
<td>3.9 (153)</td>
<td>5.9 (75)</td>
<td>-0.30</td>
<td>0.10</td>
<td>Pyruvate dehydrogenase kinase (PDK3)</td>
</tr>
<tr>
<td>D85131</td>
<td>3.7 (174)</td>
<td>5.7 (80)</td>
<td>-0.23</td>
<td>0.10</td>
<td>Myc-associated zinc-finger protein</td>
</tr>
<tr>
<td>AF091242</td>
<td>4.0 (143)</td>
<td>5.5 (89)</td>
<td>-0.29</td>
<td>0.11</td>
<td>ATP sulfurylase/APS kinase 2</td>
</tr>
<tr>
<td>U84720</td>
<td>3.9 (148)</td>
<td>5.5 (91)</td>
<td>-0.10</td>
<td>0.11</td>
<td>mRNA export protein (RAE1)</td>
</tr>
<tr>
<td>U49089</td>
<td>3.2 (243)</td>
<td>5.3 (97)</td>
<td>-0.28</td>
<td>0.09</td>
<td>Neuroendocrine-dlg (NE-dlg)</td>
</tr>
</tbody>
</table>

Among the top100 probesets ranked based on $\Delta v'$ score, a total of 22 probesets were un-overlapped with the selections by GEO analyses (See the Method section for the detailed procedure of GEO analyses). The listed genes represents 12 probesets which shows a large rank discrepancy between $\Delta u'$ and $\Delta v'$. 

---

**References:**

- Norris, T. D., and Kahn, W. A. (2006). Medium-level fold changes (moderate $\Delta v'$) in gene expression were usually not considered for further experimental validation.
- Stewart, A. L., et al. (1999). Insufficient information was available to determine if other nine genes exclusively found on the top ranking list of $\Delta v'$ scoring are associated with breast cancer pathogenesis in the current NCBI database. Table 3 shows that the difference of expression intensity between normal and breast cancer cell lines for these genes was estimated to be at least 1.5-fold higher in $\Delta v'$ measure than in $\Delta u'$ measure. It can be concluded that relatively low $\sigma_{GPL}$ values compensate the moderate difference in expression between samples and consequently rank genes in a different order. Further experimental study remains to confirm the association of the selected nine genes with breast tumors.
- Heikkinen, J. T., et al. (2006). Insufficient information was available to determine if other nine genes exclusively found on the top ranking list of $\Delta v'$ scoring are associated with breast cancer pathogenesis in the current NCBI database. Table 3 shows that the difference of expression intensity between normal and breast cancer cell lines for these genes was estimated to be at least 1.5-fold higher in $\Delta v'$ measure than in $\Delta u'$ measure. It can be concluded that relatively low $\sigma_{GPL}$ values compensate the moderate difference in expression between samples and consequently rank genes in a different order. Further experimental study remains to confirm the association of the selected nine genes with breast tumors.

---

**Large scale gene-specific standardization of array data**

- In this sense, this new approach can enrich the hit list of genes in which expression difference between normal and breast cancer cells were moderate.
- For public access to this two-step standardization method for GEO gene expression data, we constructed a user friendly web-based database, GS-LAGE (Gene Specific Large-scale Analysis of Gene Expression), which includes all single channel microarray experiments listed on the GEO database (Fig. 6). It can
be accessed via http://compbio.sookmyung.ac.kr/~lage/index.html. It provides comparative values of $\Delta u_0$ and $\Delta v_0$ for each gene between user-selected experimental samples. It will provide a valuable tool for the in silico identification of previously unknown specific (or differential) gene expression patterns in disease-related samples.

ACKNOWLEDGEMENTS

The authors appreciate helpful and stimulating discussions with Dr. Young Ju Suh. This work was supported by the SRC/ERC program of MOST/KOSEF (R11-2005-017-01003-0) and by grant No.R01-2006-000-10515-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

Conflict of Interest: none declared.

REFERENCES


Fig. 6. Schematic view of GS-LAGE. GEO data in GS-LAGE (Gene Specific Large-scale Analysis of Gene Expression) are periodically updated and simultaneously annotated with BLAST search results.