

# Significant Gene Selection Using Integrated Microarray Data Set with Batch Effect

Ki Yeol Kim<sup>1</sup>, Hyun Cheol Chung<sup>2,3,4,5</sup>, Hei Cheul Jeung<sup>4</sup>, Ji Hye Shin<sup>4</sup>, Tae Soo Kim<sup>4,5</sup> and Sun Young Rha<sup>3,4\*</sup>

<sup>1</sup>Oral Cancer Research Institute, Yonsei University College of Dentistry, <sup>2</sup>Department of Internal Medicine, Yonsei University College of Medicine, <sup>3</sup>Brain Korea 21 Project for Medical Science, Yonsei University College of Medicine, <sup>4</sup>Cancer Metastasis Research Center, Yonsei University College of Medicine, <sup>5</sup>Yonsei Cancer Center, Yonsei University College of Medicine, 134 Shinchon-Dong, Seodaemun-Ku, Seoul 120-752, Korea

## Abstract

In microarray technology, many diverse experimental features can cause biases including RNA sources, microarray production or different platforms, diverse sample processing and various experiment protocols. These systematic effects cause a substantial obstacle in the analysis of microarray data. When such data sets derived from different experimental processes were used, the analysis result was almost inconsistent and it is not reliable. Therefore, one of the most pressing challenges in the microarray field is how to combine data that comes from two different groups. As the novel trial to integrate two data sets with batch effect, we simply applied standardization to microarray data before the significant gene selection. In the gene selection step, we used new defined measure that considers the distance between a gene and an ideal gene as well as the between-slide and within-slide variations. Also we discussed the association of biological functions and different expression patterns in selected discriminative gene set. As a result, we could confirm that batch effect was minimized by standardization and the selected genes from the standardized data included various expression patterns and the significant biological functions.

**Keywords:** genomic data, integration, batch effect, bioinformatics

## Introduction

In microarray technology, many diverse experimental features can cause experimental biases including RNA sources and quality, microarray production or different platforms. Additionally, samples can be collected and processed at different institutions and assayed using different array hybridization protocols. These systematic effects causing diverse variations present a substantial obstacle to the analysis of microarray data. Due to the limited numbers of microarray experiments, however, sometimes the indication to use whole data regardless of platforms is increasing. When such data sets derived from different experimental processes were used, the result of analysis was often inconsistent with little reliable information. Therefore, one of the most pressing challenges in the microarray field is how to combine the data that comes from the two different groups.

Most existing studies that have analyzed multiple independently collected microarray data sets have focused on the differential expressions, comparing two or more similar data sets to find the genes that distinguish different groups of samples (Breitling *et al.*, 2002; Rhodes *et al.*, 2002; Yuen *et al.*, 2002; Choi *et al.*, 2003; Detours *et al.*, 2003; Ramaswamy *et al.*, 2003; Sorlie *et al.*, 2003; Xin *et al.*, 2003). Another type of comparison is exemplified by a study that examined the variability of expression for the individual gene in several human and mouse data sets (Lee *et al.*, 2002). These studies have exploited the availability of multiple data sets to identify more robust sets of genes than would be found using a single data set.

Recently, the integration of the data sets before the significant gene selection has been introduced using a method by correcting systematic bias of the data sets, Singular Value Decompositions (SVDs) in the yeast cell cycle experiments (Alter *et al.*, 2000), and in a data set containing many soft tissue tumors (Nielsen *et al.*, 2002). It has been suggested that SVD is an inappropriate method when the magnitude of the systematic effect variation is similar to other components of variations, although SVD is a method to find directions of large variation for removal of systematic effects (Benito *et al.*, 2004). Distance Weighted Discrimination (DWD), a modified form of SVM for the adjustment of systematic effects eliminated source effect and showed the good performance (Benito *et al.*, 2004). However, it still could

\*Corresponding author: E-mail rha7655@yumc.yonsei.ac.kr, Tel +82-2-2228-8050 Fax +82-2-362-5592  
Accepted by 31 May 2006

not correct some problem such as dispersion of different data sets. Moreover, previous studies have not been discussed the biological significance for evaluation of integration method.

Here, we suggested a method which effectively integrates the different experimental features and selected the discriminative gene set from the integrated data set using the unique gene selection method, and finally discussed biological significance of selected gene set.

## Materials and Methods

### Data sets

A microarray data set used in this study was consisted of 5 experimental groups to understand the molecular mechanism of tumor angiogenesis *in vitro*. The HUVEC (Human Umbilical Vein Endothelial Cells) were obtained from ATCC (American Tissue Culture Collection, USA) and cultured following the recommended guidelines. The HUVEC were cultured in 5 different experimental conditions including 1) with serum, 2) upon the matrigel as an extracellular matrix, 3) with co-culturing with YCC-3 gastric cancer cells, which is established from the ascites of Korean advanced gastric cancer patient in the Cancer Metastasis Research Center (CMRC, Yonsei University College of Medicine, Seoul, Korea), 4) without serum, and 5) combination of matrigel and co-culturing with YCC-3 cells.

The cDNA microarray was performed using human cDNA chips (CMRC-GT, Seoul, Korea) with 17664 genes in 2 batches, which were spotted in 2 different time with the same cDNA clones. cDNA microarray experiments were performed in triplicates following the established protocol of CMRC in a reference design with the Yonsei Reference RNA of 11 cancer cell line pooled RNA (CMRC, Seoul, Korea, Kim *et al.*, 2005). This data set included missing entries in the range of 349 to 785 for each experiment, and 16466 genes without missing entries were used for further analysis.

### Data normalization

We normalized expression intensities so that the intensities or log-ratios have similar distributions across a series of arrays. The method used in this study is that the MAD (median-absolute-deviation) scale estimator is replaced with the median-absolute-value and the A-values<sup>1)</sup> are normalized as well as the M-values<sup>2)</sup>.

1)  $A = \log_2 R - \log_2 G$

Within-slide normalization transforms expression values to make intensities consistent within each array and between-slide normalization transforms expression values to achieve consistency between arrays. Normalization between arrays is usually, but not necessarily, applied after normalization within arrays. We applied between-slide normalization to expression data because there were different dispersions between arrays after within-slide normalization. We executed normalization by using 'limma' library of R package (<http://www.r-project.org>).

### Standardization of expression data

In order to reduce the bias which can be occurred in each gene expression from different batches, we standardized gene expression ratio to a mean  $\pm$  s.d. of  $0 \pm 1$  in each batch respectively. The standardized expression ratio  $Z_{ij}$  is calculated as following.

$$Z_{ij} = \frac{(X_{ij} - \bar{X}_i)}{\sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2}}$$

where  $X_{ij}$  is the expression level of  $i^{th}$  gene in  $j^{th}$  experiment.

$\bar{X}_i$  is the mean expression level of  $i^{th}$  gene.  
Denominator is standard deviation of expression levels of  $i^{th}$  gene.

$N_i$  is the number of experiments of  $i^{th}$  gene.

### Significant gene selection

We selected discriminative genes that differently expressed in various experimental groups after the normalization and standardization of microarray data set. The parametric statistical methods, including t-test, were not appropriate to the current data set because the replications of data were not sufficient enough to assume any specific distribution of data. Even nonparametric method, requiring at least 5 replications of data (Kanji, 1993), was not applicable. Therefore, we defined a new measure for the discriminative gene selection, based on the variation rather than a mean difference between slides.

$$\sum std(x_{ik}) / std(\bar{x}_k)$$

where  $\bar{x}_k$  is mean expression of  $k^{th}$  experimental group.

2)  $R$  (Red Intensity) =  $(R_{\text{foreground}} - R_{\text{background}})$ ,  
 $G$  (green Intensity) =  $(G_{\text{foreground}} - G_{\text{background}})$ ,  
 $M$  =  $(\log_2 R + \log_2 G) / 2$

$std(\bar{x}_k)$  is between-slide variation.  
 $std(x_{ik})$  is within-slide variation of  $i^{th}$  gene in  $k^{th}$  experimental group.

The small value of this measure suggests that the expression values of the  $i^{th}$  gene have small variation in each experimental group and large difference between different experimental groups at the same time. Therefore it implies that the  $i^{th}$  gene is discriminative.

In the selection of significant genes which have clear expression pattern-over expressed or under expressed - in a specific experimental group, we defined the ideal genes expressed in a specific experimental group. When we defined new measure, we considered the distance between these ideal genes and each gene, in addition to variation of expression values of a gene. It resulted that the selected genes with smaller within-slide closed to expression levels of ideal genes. We could define three ideal genes in the case of three experimental groups as following.

$$Igene_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, Igene_2 = \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, Igene_3 = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

} Experimental group 1  
 } Experimental group 2  
 } Experimental group 3

The new metric, MM, was defined as following and calculated for each gene.

$$MM(i) = dist(g(i), Igene) \times \sum std(x_{ik}) / std(\bar{x}_k)$$

where  $g_i$  is the expression of  $i^{th}$  gene.

$dist(g(i), Igene)$  is distance between each  $i$ th gene and ideal target genes.

A gene with small MM is considered as discriminative and it is over-expressed or under-expressed in experimental groups.

We evaluated the selected gene set by classification accuracy and used Random Forest algorithm (RF, Breiman, 2001). We used RF program in R package (<http://www.r-project.org>) with the following steps.

- (1) Generate n datasets of bootstrap samples  $\{B_1, B_2, \dots, B_n\}$  by allowing repetition of the same sample.
- (2) Use each sample  $B_k$  to construct a Tree classifier  $T_k$  to predict those samples that are not in  $B_k$ , called out-of-bag (OOB) samples. These

predictions are called out-of-bag estimators.

- (3) Final prediction is the average of out-of-bag estimators over all bootstrap samples and we get average of them which is overall classification error (OOB error).

### Annotation of selected genes

We investigated the significance of biological functions of selected discriminative genes that classified five experimental groups accurately. These genes were separated into five gene clusters by k-means<sup>3)</sup> clustering method under the assumption that they have five different expression patterns because they classified five experimental groups exactly. We used EASE (Expression Analysis Systematic Explorer, <http://david.niaid.nih.gov/david/>) to analyze the significance of biological functions for five gene clusters.

### Results and Discussion

Fig. 1 shows the result of within and between-slide normalization of microarray data.

The last six boxplots in Fig. 1b had larger inter-quartile range (IQR) than the other box plots. It means there were larger variations in the 9-15<sup>th</sup> experiments than the other experiments when only within-slide normalization was applied to data set. Those experiments were processed experimentally in different batches and this problem was removed by between-slide normalization (Fig. 1c).

M.YCC3 and woSerum experimental groups were experimentally processed in the same batch differing from the remainder groups. When we applied unsupervised clustering method with whole data set, three experimental groups (HUVEC, Matrigel, YCC3) and two experimental groups (woSerum, M.YCC3) were separated and groups in same batch were fastened together. Therefore, we could confirm that some batch effect exists in the data set (Fig. 2a). With standardization of data, HUVEC and woSerum experimental groups from different batches were fastened together and other experimental groups were well intermingled (Fig. 2b), suggesting that the batch effect is minimized.

As a next step, we selected the discriminative genes that classified different experimental groups accurately by the proposed method. Only three most significant genes had 0% OOB (one out of bag) error. Table 1 shows the summarization of these selected genes.

We noted that the OOB error did not increase in

3) A non-hierarchical clustering method which divides the data set into k groups. k is pre-defined.

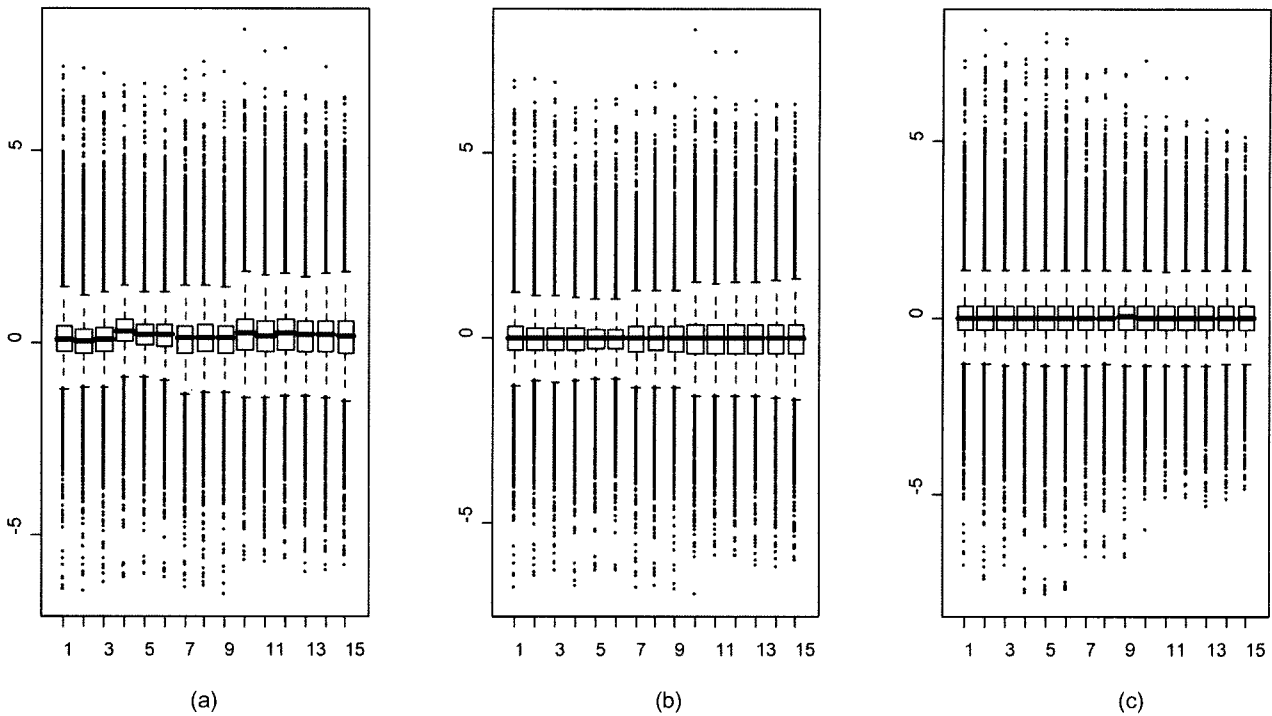


Fig. 1 Boxplots of raw data, after within-slide, and between-slide normalized data. (a) is boxplot of the M-values from 15 experiments. (b) is boxplot of the same arrays after within-slide normalization to equalize the median absolute value for each array. (c) is boxplot after between-slide normalization.

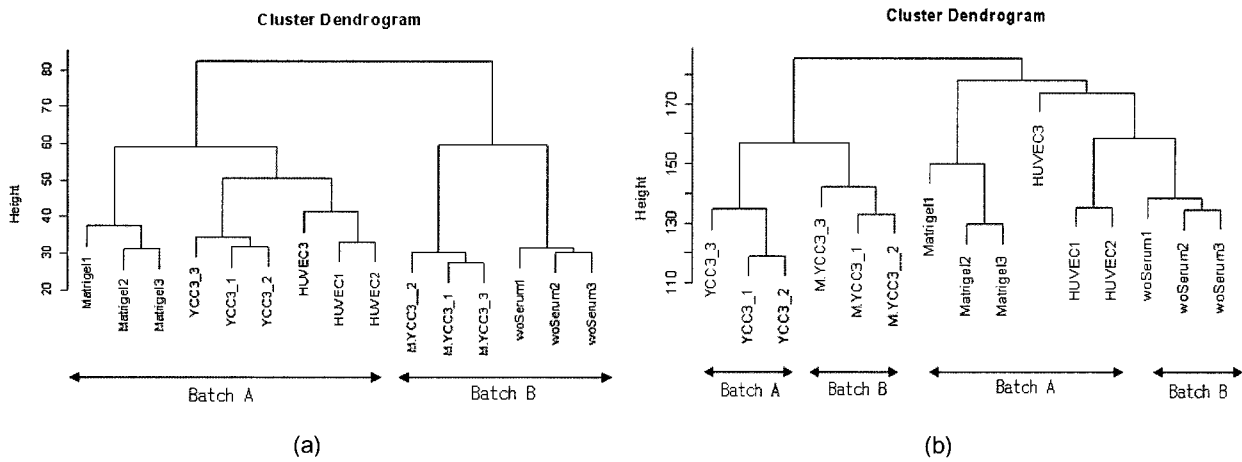


Fig. 2. Hierarchical clustering analysis of raw data (a) and standardized data (b). HUVEC1-3: HUVEC in the conventional culture condition with serum, Matrigel1-3: cultured HUVEC in Matrigel, YCC3\_1-3: Co-cultured HUVEC with YCC-3, woSerum1-3: HUVEC cultured without Serum, M.YCC3\_1-3: cultured HUVEC in Matrigel and co-cultured with YCC-3. The numbers behind experiment labels represent the numbers of replicates in the experiment.

classification even though we increased genes to 200 by ranking approach<sup>4</sup>) (Data is not shown). When we investigated expression patterns of top 100 genes, the

experimental groups and gene groups were clustered together showing no more batch effects (Fig. 3). By using new defined metric for gene selection, top 100 genes had similar expression values in the same experimental group and the clear difference between the

4) One approach for gene selection which select genes sequentially from a gene that have the least measure

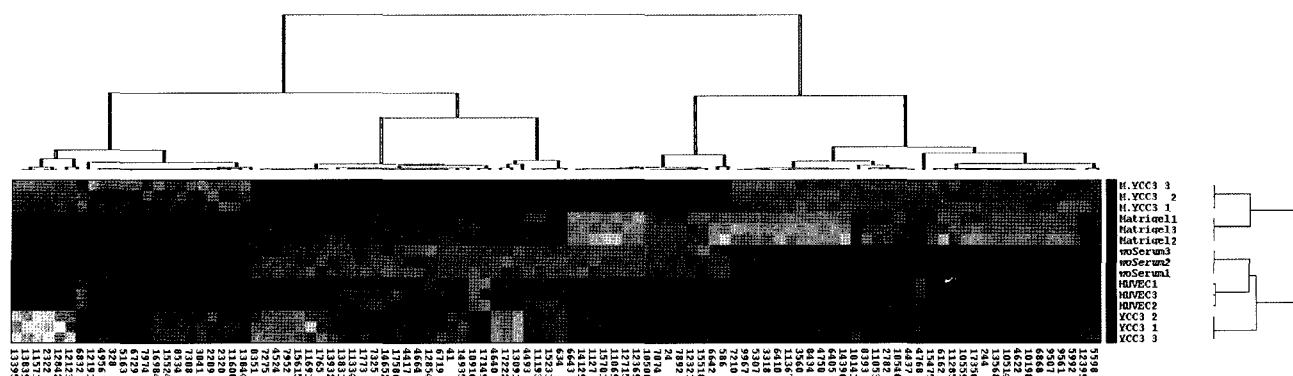


Fig. 3. Cluster analysis of top 100 genes. Five experimental groups were classified accurately with top 100 genes and genes were clustered into five groups with different expression patterns.

Table 1. Summary of top three discriminative genes selected by the proposed method

	Gene description
AA458634	diaphorase (NADH/NADPH) (cytochrome b-5 reductase)
T49159	plasminogen activator inhibitor, type II (arginine-serpin)
A1524212	ESTs, Weakly similar to ALU7_HUMAN ALU SUBFAMILY SQ SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens]

groups. Even though, the clustering result using whole gene set was slightly different from the one using top 100 genes, it showed more specific patterns. Therefore we confirmed that small set of genes might be more effective to classify different experimental groups.

We analyzed biological interpretations of five gene clusters by k-means clustering method and explored expression patterns of each cluster (Table 2). Five clusters showed the different expression patterns and were expected that such different expression patterns have different biological functions.

In addition, we did significance analysis of biological functions using EASE and used top 1000 genes as background genes. EASE calculates the degree of significance, EASE score, of biological function including selected gene set. Fisher exact test can be used for this purpose, however, we offered EASE score. We should put whole gene set as background to calculate EASE score but used top 1000 genes because of run time error possibly due to huge data size. Therefore, if we used whole gene set as background, EASE score might be decreased than the values shown in Table 3. Top 100 genes were concerned with various biological functions corresponding to three categories of biological process, molecular function, and cellular component. We also investigated the association of expression patterns and

corresponding biological functions for each gene clusters

Biological functions in the first gene cluster were mainly related to molecular function. All of the biological functions of the second gene cluster were about biological processes, which were highly significant. The third and the fourth gene clusters had not include significant biological functions but those genes were relatively significant in cellular component and biological process respectively. The fifth gene cluster included genes that had biological functions related to Carboxylic acid metabolism and Organic acid metabolism. Meanwhile, we observed some irregularities. As in Table 3, the same gene sets had highly significant two biological functions; one is Regulation of cell cycle and Mitotic cell cycle in the second gene cluster and the other is Carboxylic acid metabolism and Organic acid metabolism in the fifth gene cluster.

AA458634, T49159, and A1524212, which were selected as the most significant genes, were clustered into the first and the fourth gene clusters. While AA458634 was concerned as in highly significant function of biological process, T49159 was related to the cell death but not concerned to any significant biological function. From this, we confirmed that the significant biological functions are caused by the interactions of genes or gene sets, not by the several most significant genes.

We selected seven genes with 0% OOB error from non-standardized data set. It is relatively larger gene set comparing to the result from the standardized data set and any of them were not consistent with three genes selected from standardized data set (data not shown). From the functional annotation analysis using DAVID (Database for Annotation, Visualization and Integrated Discovery, <http://apps1.niaid.nih.gov/david/>), 72.6% of top 100 genes selected from the standardized data set included in any categories of biological functions but 64.9% in the

**Table 2.** Summary of top five genes for each gene cluster

Gene cluster	Gene ID	Gene description
1	AA458634	diaphorase (NADH/NADPH) (cytochrome b-5 reductase)
	T49159	plasminogen activator inhibitor, type II (arginine-serpin)
	AA283090	CD44 antigen (homing function and Indian blood group system)
	N33920	diubiquitin
	AA599127	superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))
2	AA278384	cell division cycle 2, G1 to S and G2 to M
	AA262212	KIAA0008 gene product
	R16712	anillin
	AA598610	mesoderm specific transcript (mouse) homolog
	AA873060	leukemia-associated phosphoprotein p18 (stathmin)
3	AA857098	collagen, type V, alpha 2
	R36467	transforming growth factor, beta 1
	AA461456	collagen, type V, alpha 2
	N47717	fatty acid binding protein 5 (psoriasis-associated)
	AA877213	cytochrome P450, subfamily XXIV (vitamin D 24-hydroxylase)
4	AI524212	ESTs, Weakly similar to ALU7_HUMAN ALU SUBFAMILY SQ SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens]
	AA481519	potassium voltage-gated channel, shaker-related subfamily, beta member 3
	AI347124	hypothetical protein, expressed in osteoblast
	AA410188	hypothetical protein, expressed in osteoblast
	AA465166	cyclin L ania-6a
5	AA894927	asparagine synthetase
	AW055062	phospholipase A2, group IVC (cytosolic, calcium-independent)
	H26184	CCAAT/enhancer binding protein (C/EBP), beta
	R41787	cadherin 13, H-cadherin (heart)
	AA664040	tryptophanyl-tRNA synthetase

non-standardized data set did. Also, the result of cluster analysis of top 100 genes was unbalanced in non-standardized data set. Almost half of the genes were clustered into the second gene cluster and only eight and five genes were clustered into the fourth and the fifth gene clusters, respectively. Gene sets expressed in three experimental conditions- Matrigel, YCC3, and M.YCC3-included significant biological functions, while gene sets expressed in the other experimental conditions did not in the non-standardized data set (The least EASE scores were 0.49 and 0.34). Though the significances of biological functions of the third and the fourth gene clusters in standardized data set were low, five gene clusters had relatively significant biological functions. This means that we could select more discriminative genes representing various expression patterns, which can be interpreted as various representative biological functions from the standardized data set.

Most selected discriminative genes from the non-standardized data were clustered into the same gene cluster meaning that the selected genes had redundant expression patterns. On the other hand, the selected genes from the standardized data included various expression patterns and the significant biological functions. It suggests that the problem of redundancy in gene selection was solved by the standardization.

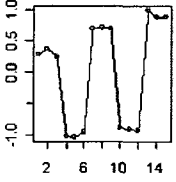
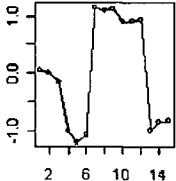
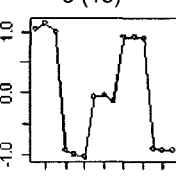
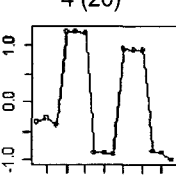
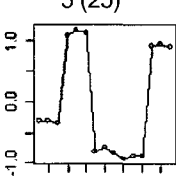
Before discriminative gene selection, we applied standardization to expression ratios respectively in each batch to control batch effect. As a result, we could confirm that batch effect was minimized using dendrogram<sup>5)</sup>. The selected discriminative genes in this study were selected by the expression pattern but not by the magnitude of expression ratio. However this method can not be appropriate when different batches include biologically significant differences because the standardization transforms the expression values into a relative score of a gene expression in each batch. Hence, currently we are on the study about investigating more efficient method that is useful in such condition.

### Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Cancer Metastasis Research Center (CMRC) at the Yonsei University College of Medicine, and the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, KRF-2005-005-J05901).

5) Tree diagram frequently used to illustrate the arrangement of the clusters produced by a clustering algorithm.

Table 3. Summary of biological functions of five gene clusters.

Gene cluster* (# of genes)	System & category	Gene ID	EASE score**
1 (19) 	Molecular function Glycosaminoglycan binding	AA283090, AI268937 AI889554, R45640	0.00297
	Biological process Response to chemical substance	AA458634, AI268937 AI889554, W46900	0.00926
2 (18) 	Biological process Regulation of cell cycle	AA278384, AA284072 AA454094, AA598974 AA774665, AI932735 H59204, R16712	0.000458
	Biological process Mitotic cell cycle	AA278384, AA284072 AA454094, AA598974 AA774665, AI932735 H59204, R16712	0.000671
3 (18) 	Cellular component Extracellular matrix	AA857098, H95960 R75635	0.0751
	Molecular function Protein serine/threonine kinase activity	AA460152, AA487034 AA683077	0.0803
4 (20) 	Biological process Organismal physiological process	AA464417, AA490996 AA862371, AA985421 AI431726, N25945	0.0777
	Cellular component Integral to membrane	AA454597, AA464417 AA663439, AA862371 AA985421, N25945	0.155
5 (25) 	Biological process Carboxylic acid metabolism	AA171606, AA664040 AA894927, AI015679, AW055062	0.0125
	Biological process Organic acid metabolism	AA171606, AA664040 AA894927, AI015679 AW055062	0.0125

\* Plots in the first column show the expression patterns of standardized data for each gene cluster.

\*\* Small EASE score means that the corresponding biological function is significant.

## References

- Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97, 10101-10106.
- Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C.M., and Marron, J.S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* 20, 105-114.
- Breiman, L. (2001). Random Forests. Berkeley, CA, Statistics Department, University of California 1-33.
- Breitling, R., Sharif, O., Hartman, M.L., and Krisans, S.K. (2002). Loss of compartmentalization causes misregulation of lysine biosynthesis in peroxisome-deficient yeast cells. *Eukaryot. Cell* 1, 978-986.
- Choi, J.K., Yu, U., Kim, S., and Yoo, O.J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19, 184-190.
- Detours, V., Dumont, J.E., Bersini, H., and Maenhaut, C. (2003). Integration and cross-validation of high-throughput

- gene expression data: Comparing heterogeneous data sets. *FEBS Lett.* 546, 98-102.
- EASE (Expression Analysis Systematic Explorer). <http://david.niaid.nih.gov/david/>
- Kanji, G.K. (1993). *100 Statistical Tests*. (London, Thousand Oaks, New Delhi, SAGE publication).
- Kim, T.M., Jeong, H.J., Seo, M.Y., Kim, S.C., Cho, G., Park, K.H., *et al.* (2005). Determination of genes related to gastrointestinal tract origin cancer cells using a cDNA microarray. *Clin Cancer Res.* 11, 79-86.
- Lee, P.D., Sladek, R., Greenwood, C.M., and Hudson, T.J. (2002). Control genes and variability: Absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* 12, 292-297.
- Nielsen, T.O., West, R.B., Linn, S.C., Alter, O., Knowling, M.A., O'Connell, J.X., Zhu, S., Fero, M., Sherlock, G., Pollack, J.R., Brown, P.O., Botstein, D., and van de Rijn, M. (2002). Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* 359, 1301-1307.
- R: A language and environment for statistical computing. <http://www.R-project.org>.
- Ramaswamy, S., Ross, K.N., Lander, E.S., and Golub, T.R. (2003). A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* 33, 49-54.
- Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D., and Chinnaiyan, A.M. (2002). Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* 62, 4427-4433.
- Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., *et al.* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* 100, 8418-8423
- Xin, W., Rhodes, D.R., Ingold, C., Chinnaiyan, A.M., and Rubin, M.A. (2003). Dysregulation of the annexin family protein family is associated with prostate cancer progression. *Am. J. Pathol.* 162, 255-261.
- Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J., and Sealfon, S.C. (2002). Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* 30, e48.