BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 14 2010, pages 1723–1730
doi:10.1093/bioinformatics/btq279

Gene expression

Advance Access publication May 28, 2010

Global analysis of microarray data reveals intrinsic properties in gene expression and tissue selectivity

Changsik Kim[1], Jiwon Choi[1], Hyunjin Park[1], Yunsun Park[1], Jungsun Park[2], Taesung Park[2,3], Kwanghui Cho[4], Young Yang[1] and Sukjoon Yoon[1,*]

[1]Department of Biological Sciences, Sookmyung Women's University, [2]Interdisciplinary Program in Bioinformatics, Seoul National University, [3]Department of Statistics, Seoul National University and [4]Department of Bioinformatics, Soongsil University, Seoul, Republic of Korea

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: It is expected that individual genes have intrinsically different variability in the global expressional trend among them. Thus, the consideration of gene-specific expressional properties will help us to distinguish target-selective gene expression over non-selective over-expression.

Results: The re-standardization and integration of heterogeneous microarray datasets, available from public databases, have enabled us to determine the global expression properties of individual genes across a wide variety of experimental conditions and samples. The global averages and SDs of expression for each gene in the integrated microarray datasets were found to be intrinsic properties, which were consistent among independent collections of datasets using different microarray platforms. Using the gene-specific intrinsic parameters to rescale the microarray data, we were able to distinguish novel selective gene expression [cartilage oligomeric matrix protein (COMP) and Collagen X] in breast cancer tissues from non-selective over-expression, a difference that has not been detectable by conventional methods.

Availability and Implementation: The web-based tool for GS-LAGE is available at http://lage.sookmyung.ac.kr

Contact: yoonsj@sookmyung.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 12, 2010; revised on May 4, 2010; accepted on May 23, 2010

1 INTRODUCTION

DNA microarray experiments have provided a powerful tool for monitoring transcriptome-wide changes in gene expression that are associated with physiological or pathological states (Su *et al.*, 2001). Large-scale microarray experiments can serve as a resource for a systematic understanding of global expression trends in the transcriptome under various experimental conditions or in different tissues. With the recent rapid increase in microarray expression datasets available in public databases, it has become possible to virtually monitor the global expression trends of a gene in diverse biological samples under various conditions. The availability of a large collection of microarray data in the NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo; Barrett *et al.*, 2005, 2007) provides a resource for the physiome-wide analysis of gene expression on a genome scale. However, the microarray datasets available in public domains have been contributed by many different research groups using a variety of experimental and data-normalization methods. The effects of normalization strategies have been extensively evaluated elsewhere (Shippy *et al.*, 2006). For example, two different transformations of the same data may produce two differential expression of gene (DEG) lists with only ~30% agreement (Gagarin *et al.*, 2005). Thus, the heterogeneity among groups of microarray datasets requires additional standardization of individual microarray samples before integrative analysis. To this end, we have attempted to rescale and integrate heterogeneous microarray datasets from the NCBI GEO by using a two-step gene-oriented standardization procedure, gene-specific large scale analysis of gene expression (GS-LAGE; Yoon *et al.*, 2006). The analysis of the resulting integrated dataset provides insights into gene-specific global expression trends found in various microarray samples. These trends were then used to improve the statistical confidence of the marginal changes in gene expression between experiments (Yoon *et al.*, 2006).

A major challenge in microarray experiments has been the identification of genes whose expression is significantly different between samples (Aittokallio *et al.*, 2003). Many sophisticated statistical methods have been developed in attempts to achieve more reliable identification of differentially regulated genes (Huber *et al.*, 2002; Irizarry *et al.*, 2003; Yang *et al.*, 2001, 2002). However, under many experimental conditions, the changes in gene expression are often moderate, as compared to the array-wide variability in expression, resulting in modest *P*-values that prevent conventional statistical models from identifying real changes in expression (Mootha *et al.*, 2003). In fact, all existing measures of DEG from microarray fluorescence signals are optimized to have a linear correlation with the change in mRNA levels (Cope *et al.*, 2004; Yang *et al.*, 2002). Given that individual genes have unique expression levels and fluctuations (necessitated by their biological functions), a metric for the direct comparison of the changes in mRNA expression between experiments cannot provide a relevant biological interpretation based only on the significance of the observed DEGs between genes.

One gene may have an intrinsically large variation in expression in different biological samples or under different conditions, while another gene may require only small variations for its biological function. Thus, it is reasonable that the observed DEG (i.e. mRNA

[*]To whom correspondence should be addressed.

change) between experiments should be differentially interpreted among genes, based on their specific expression behavior. Our focus has been to better understand the biological relevance of detected DEGs, rather than improving the fluorescence intensity normalization, which has been the focus of many existing methods (Yoon *et al*., 2006). Through the creation of a database-wide expression profile for individual genes (i.e. probesets), which provides an estimation of the gene-specific distribution of expression levels under various experimental conditions, it is possible to rescale individual gene expression intensities from a specific assay by using the unique database-wide average expression and standard deviation (SD) values for each gene. This consideration of a gene's behavior under a wide variety of biological conditions provides a tool for interpreting DEGs among samples and allows for the prioritization of genes that are differentially expressed in specific samples. In this study, we re-standardized and integrated a recent update of human microarray datasets obtained from the NCBI GEO in order to analyze the global expression trends of individual genes in the human transcriptome. We then attempted to quantitatively determine two gene-specific expressional properties, global average expression and SD, to systematically understand the diversity of the intrinsic expression behaviors among genes. Using these gene-specific properties of global expression trends as rescaling parameters, we attempted to re-evaluate the observed gene expression data from various human samples. Using the gene-specific interpretation of observed DEGs, it is possible to distinguish tissue-selective over-expression from non-selective over-expression among genes. In this study, we applied this method to the identification of novel breast cancer-selective gene expression and confirmed these findings with quantitative polymerase chain reaction (PCR) on various human tissues.

2 METHODS

2.1 Integration of microarray datasets

To construct two independent integrations of large-scale heterogeneous microarray datasets, datasets sharing the same platforms were obtained from the NCBI GEO ftp site. Each collection consisted of microarray sample arrays created using *Affymetrix HG-U133A* Array chips (NCBI GEO ID: GPL96) and *Affymetrix U133 Plus 2.0* Array chips (NCBI GEO ID: GPL570). Each array sample was assigned a unique and stable GEO accession number (e.g. GSM), and array samples were assembled by GEO curators to form GEO DataSet records (e.g. GDS), based on the original submitter-supplied information summarizing each experiment. In this study, 5018 GSMs from 236 GDSs using the GPL96 platform and 1303 GSMs from 70 GDSs using the GPL570 platform were used to construct two independent datasets. To further validate the intrinsic properties of gene expression between different platforms, the datasets of *Illumina Sentrix HumanRef-8 Expression BeadChip* (NCBI GEO ID: GPL2700) were also obtained and processed.

For the integration, the individual sample arrays from the GEO were first z-transformed by using the average expression value and SD in each sample array (Yoon *et al*., 2006). To minimize the characteristics of heterogeneity within each collection of microarray datasets caused by systematic bias in various experimental conditions, we applied the quantile normalization technique (Bolstad *et al*., 2003), which creates equal distributions of probe intensities for all GSM samples within the same pool of datasets.

2.2 Quantile normalization

For quantile normalization, all expression values were initially ranked according to the expression values within each GSM sample. Second, average expression values for all probes with the same rank in the same pool of datasets were calculated. These average expression values constitute the 'rank-average distribution' (Supplementary Fig. 1). Since each collection of microarray datasets include thousands of diverse samples in various conditions, the 'rank-average distribution' was found mostly to follow a normal distribution. Third, the distribution of each GSM sample was fitted into the average distribution. See the study of Bolstad *et al*. (2003) for further details of quantile normalization.

Using z-transformed GSMs before quantile normalization, the deviation of the expression distribution of each GSM from the rank-average distribution in Supplementary Figure 1A was estimated by a χ^2 goodness-of-fit test. We explored the relationship between the median expression values and the χ^2 values of the GSMs (Supplementary Fig. 1B and C). It should be noted that each GSM was z-transformed as a first step before quantile normalization. Thus, it was assumed that the deviation of the distribution of each GSM from the average distribution increases as the median value deviates from zero. That is, the distribution of median expression intensities significantly deviating from zero could be considered a skewed distribution, due to systematic bias. To validate this assumption, we also performed the χ^2 goodness-of-fit test between the average distribution and the expression data of a standard sample from Affymetrix Company, known as the 'Latin Square dataset' (http://www.affymetrix.com/support/technical/sample_data/datasets.affx). This dataset was originally created to test newly developed statistical algorithms for Affymetrix GeneChip data. Thus, it was assumed that the distribution of sample arrays from this Latin Square dataset is unbiased and does not contain any systematic noise, and the distribution should have a near-zero-valued median. Based on the χ^2 goodness-of-fit test of the Latin Square dataset, the Latin Square dataset had a relatively small χ^2 value in comparison with our calculated database (DB)-wide rank-average distribution (Supplementary Fig. 1C). This observed similarity between the gene expression in the Latin Square dataset and the DB-wide rank average distribution implies that the distribution of gene expression in individual array samples follows a normal distribution, unless systematic bias is introduced. This supports the relevance of the present quantile normalization of fitting all GSMs to rank-average distributions of DB-wide gene expression and minimizing the systematic bias within each sample array.

2.3 Gene-specific interpretation of transcriptome data

Probe-specific transformation of the expression data is carried out to rescale the intensity signal of the expression data (u_i) of gene i by using the global average expression (μ_i) and SD (σ_i) of the gene (Yoon *et al*., 2006).

$$z_i = \frac{u_i - \mu_i}{\sigma_i}$$

With this transformation, similar expression data among genes can be differentially weighted depending on the genes' global average expression level and variability, represented by μ_i and σ_i, respectively.

2.4 Quantitative real-time PCR analysis

Quantitative real-time PCR (qPCR) experiments were carried out on a human tissue array panel purchased from Origene, USA. The panel includes a total of 381 cDNA samples from 18 different normal and tumor tissues (Supplementary Table 3). The PCR primers and reagents for Inhibin, Collagen VIII, COMP and Collagen X were purchased from Applied Biosystems. We selected pre-made, vendor-suggested Taqman assay systems (primers and probes), Hs01081598_m1, Hs00156669_m1, Hs00164359_m1 and Hs00166657_m1 for probing the expression of Inhibin, Collagen VIII, COMP and Collagen X, respectively. The qPCR reaction was carried out in a 384-well format with an ABI 7900HT Fast Real-Time PCR system. We followed the standard protocol for qPCR experiments provided by Applied Biosystems.

3 RESULTS

3.1 Global trends in gene expression

We separately integrated two independent collections of heterogeneous microarray datasets (i.e. GPL96 and GPL570 datasets) after re-standardization of individual array samples (see ‘Section 2’ for details of data standardization). For a comparative analysis of gene-specific expressional trends in global data, the average expression and SD of individual probe features were calculated for each of the two integrated datasets (black plots in Fig. 1A and B). The DB-wide average expression of genes was found to be widely distributed in the range −2 to 3 in both the GPL96 and GPL570 datasets. Although the GPL96 and GPL570 datasets were independent collections of array samples, they showed a similar distribution of genes in terms of both average expression and SD. To determine whether the observed diversity in the DB-wide average expression and SD among genes was random within the array-wide variability in expression, we also calculated the expected distributions of the average expression and SD of individual genes after randomly shuffling the expression data from all microarray samples (gray plots in Fig. 1A and B). The results showed that the observed diversity in the DB-wide average expression and SD (black dots in Fig. 1A and B) among genes was greater than the diversity from the random distribution (gray dots in Fig. 1A and B). More than 20% of the genes in the observed dataset were found to have a significantly higher or lower average expression than expected, based on the random distribution ($P<0.01$, see Supplementary Table 1 for details), implying that the diversity of the average expression level among genes is an intrinsic property. Furthermore, 81.8% of the genes in the GPL96 and 72.3% of the genes in the GPL570 platform showed significantly lower than expected SDs ($P<0.01$) in their expression trends (Supplementary Table 1). Since the average expression and SD of individual genes were calculated by using large collections of diverse microarray samples, they were assumed to represent global trends in gene expression. These results indicate that most genes have a consistent width of variability in the global expression trend, although the absolute expression level differs significantly among genes.

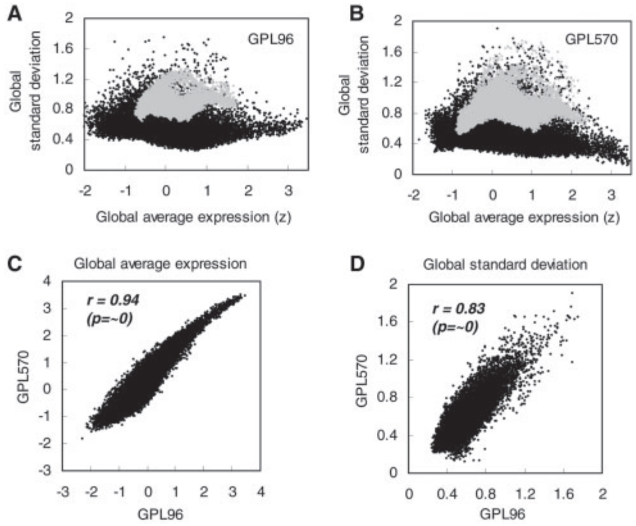

Fig. 1. DB-wide survey of global average expression and SD of individual genes. (**A**) The average expression and SD of each of the 22 215 gene probes found in the GPL96 platform were calculated using a total of 5018 array samples from 236 different datasets. (**B**) The average and SD of each of the 54 613 probes found in the GPL570 platform were calculated from a total of 1303 array samples in 70 different datasets. The gray dots in (A) and (B) represent the expected distribution calculated by 22 215 and 54 613 permutations (random shuffling) of all expression data in the GPL96 and GPL570 datasets, respectively. (**C**, **D**) A total of 22 215 gene probes commonly found in both the GPL570 and GPL96 platforms were analyzed for consistency in average expression and SD between independent collections of datasets using different platforms. (C) The average expression of individual genes for the GPL96 and GPL570 datasets. (D) The SD of gene expression for the GPL96 and GPL570 datasets.

To confirm that these two global features (average expression and SD) are intrinsic properties of a gene, we investigated whether they are conserved between different collections of datasets (i.e. GPL96 and GPL570 datasets). A total of 22 215 gene probes are commonly found between both the GPL96 and GPL570 platforms. The average expression and SD were compared between independent collections of array samples (GPL96 and GPL570 datasets; Fig. 1C and D). The results show that the calculated average expression of individual probe features is highly correlated between the GPL96 and GPL570 datasets ($r=0.94$, Fig. 1C). This confirms that the calculated DB-wide average expression of a gene probe is an intrinsically determined property of that gene, effectively representing the global average expression level of the gene. Likewise, the calculated SD of the gene expression is significantly correlated between the two independent datasets ($r=0.83$, Fig. 1D). This confirms that the observed DB-wide SD of gene expression is also an intrinsic property of genes and that it effectively represents the global variability of gene expression.

We also compared the global average and SD of a gene between datasets using different probe design (Affymetrix (GPL570) versus Illumina (GPL2700) platforms). It should be noted that arrays produced by Affymetrix are fabricated by *in situ* synthesis of 25-mer oligonucleotides, while the arrays of Illumina are produced by standard long-oligonucleotide synthesis methods, in which the oligonucleotides are attached to microbeads that are attached to microarrays using random self-assembly mechanism. Besides of the difference in the physical attachment of oligonucleotide, these two platforms are also manufactured with different probe selection and design procedure. Affymetrix uses multiple probes for each gene along with one-base mismatch probes as controls for non-hybridization. On the other hand, Illumina has the randomly generated arrays with 30 copies of the same oligonucleotide. Therefore, we have adopted the cross-platform probe annotations from the study of Barnes *et al.* (2005). They constructed the cross-platform probe annotations with their own sequence analysis by comparing each probe sequence of Affymetrix and Illumina with Human genome sequence. They analyzed the relationship of cross-platform agreement with probe location on the genome by measuring the distance between two probes as the distance between the centers of their alignments on the genome. This distance is considered as the probe-matching score between cross-platforms. Thus, the cross-platform matched probes were ranked based on the probe-matching score between cross-platforms. That is, the shorter the distance, the more agreement between cross-platform matched probes.

In this study, we observed that the correlation coefficient (r) for the global average expression and SD of genes were

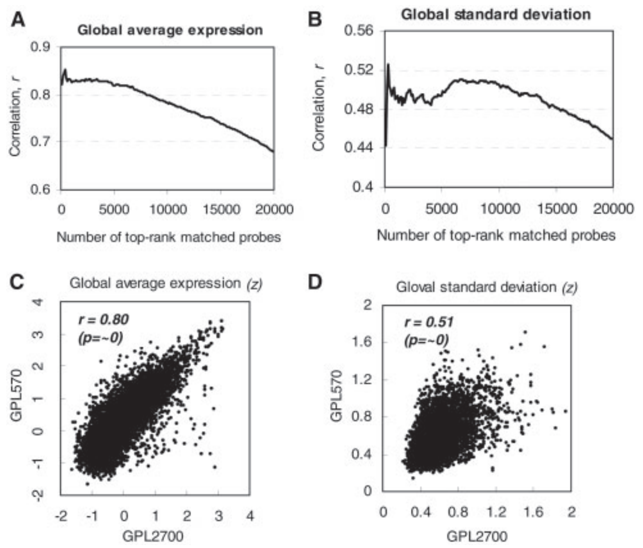

Fig. 2. Comparison of global average expression and SD of individual genes between Affymetrix and Illumina platforms. (**A**, **B**) The correlation coefficient (*r*) for the global average expression and SD of genes, respectively, is shown to increase as the similarity of gene probes between GPL570 and GPL2700 platforms increase. A total of 490 array samples in 19 different datasets were used for the calculation of global average and SD for GPL2700. (**C**, **D**) A total of 8000 top-ranked probes with high probe-matching score were selected and their global average expression and SD were plotted.

increased as the degree of agreement between cross-platform-matched probes increase (Fig. 2A and B). Thus, a total of 8000 top-ranked probes with relatively high probe-matching were selected to plot the global average expression and SD of individual cross-platform-matched probes (Fig. 2C and D). We found that there were significant correlations ($P = \sim 0$) in both the global average expression and SD between datasets from different platforms. These observations indicate that the intrinsic diversity of global trends in gene expression is platform independent. However, the intrinsic properties between cross-platform-matched probes were relatively less conserved than the ones between same-platform-matched probes due to the difference of probe selection and design.

In this integrative analysis, we were able to quantitatively determine the global expression properties of individual genes (i.e. gene-specific average expression and SD). A statistical test showed that these global expression properties were significantly diversified among genes (Fig. 1A and B). We also found that these properties were intrinsic gene-specific features, as they were consistent among independent collections of experiments using different array platforms (Figs 1C, D and 2).

3.2 Gene-specific interpretation of transcriptome data

Previous methods for the analysis of microarray data have been optimized to find a linear correlation between the observed signal intensity and the actual mRNA level. However, since different genes are found to have intrinsically different expression levels and variability (Fig. 1), necessitated by their biological functions, gene-specific behaviors in the expression data should be considered in prioritizing genes of biological significance. Thus, we attempted to rescale the intensity signal of the expression data of gene by using the global average expression and SD of the gene (see Method section from the study of Yoon *et al.*, 2006 for details). With this transformation, similar expression data among genes can be differentially weighted depending on the genes' global average expression level and variability. For example, if two genes in a biological sample are observed to have similar expression levels, a gene with an intrinsically low global expression level and variability should be interpreted as more biologically relevant (i.e. selective expression in the target sample versus global samples) than a gene with a relatively high global expression level and variability. Thus, the rescaled expression value (LAGE *z*-score) of a gene effectively represents the selectivity of the observed expression in the target sample.

Additionally, we computed a gene-specific DEG value from z_i values of two compared samples

$$\Delta z_i = \frac{u_i^- - u_i^+}{\sigma_i}$$

where u_i^+ is the expression level of the *i*-th gene in the control sample and u_i^- the expression level of the *i*-th gene in the target sample..

Conventional DEG analyses have focused on the mechanical interpretation of the observed change in mRNA expression levels between control and target samples, without any consideration of gene-specific behaviors in transcriptional activity. However, Δz_i emphasizes the biological significance of the expressional change of a gene in the target samples by rescaling the observed DEG values with the gene-specific global expression variability, σ. This transformation is based on the notion that a small DEG value for a gene with a low σ value may be more biologically significant than a large DEG value of another gene with a large σ value, given that σ represents the intrinsic variability of the gene's expression.

3.3 Breast cancer-selective gene expression

We applied the gene-specific analysis method to the identification of novel selective gene expression and cancer-specific DEGs in breast tissue. As a proof of concept, we comparatively analyzed the differential expression and selectivity (LAGE Δz-score) between known breast cancer prognostic markers and reference genes. The microarray dataset (NCBI GEO ID: GSE2034) of breast cancer tissue was used to calculate the gene-specific selective expression value for a total of 17 probes for 11 Oncotype DX breast cancer prognostic markers and 7 probes for 3 reference genes (Paik *et al.*, 2004). For those marker genes, the LAGE Δz score were compared with the differential expression levels between ER-positive and ER-negative samples (Fig. 3A). It was observed that the prognostic markers, in general, had higher selectivity score than the expression level in ER-positive samples, while the reference genes had lower selectivity score than the expression level. That is, the LAGE selectivity score provides a better resolution in distinguishing the expression of prognostic markers from that of reference genes.

We thus attempted to identify novel selective gene expression and cancer-specific DEGs in breast tissue using this method. We analyzed the microarray dataset (GDS2635) that includes both normal and cancer breast tissues. Among 54 613 transcripts (probe features), a total of 451 probe features were found to be

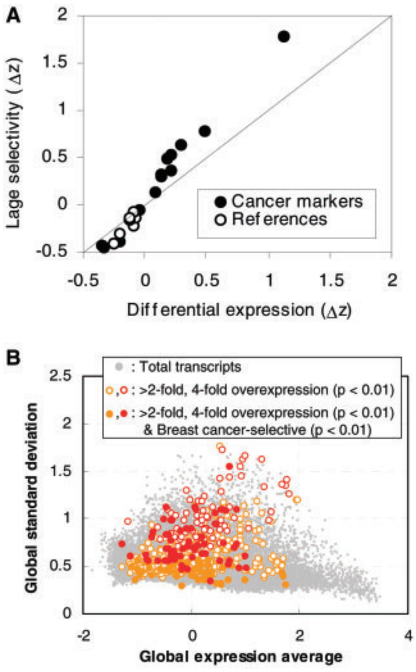

Fig. 3. Analysis of gene expression and its selectivity in breast cancer tissues. (**A**) The gene-specific selective expression (Δz) between ER-positive and ER-negative breast cancer tissues (GSE2034) was calculated for a total of 17 probesets for 11 Oncotype DX breast cancer prognostic markers and 7 probesets for 3 reference genes. (**B**) Global average expression and SD of genes with selective versus non-selective over-expression in breast cancers. All of the 54 613 transcripts (probe features) in the GPL570 platform were plotted as a gray circle in the background. A total of 451 probe features were found to be over-expressed with at least a 2-fold change and $P < 0.01$ in the t-test. Among them, 272 and 73 probe features with over 2- and 4-fold non-selective over-expression were plotted with empty orange and red circles, respectively. The 47 and 59 probe features with over 2- and 4-fold selective over-expression were plotted with filled orange and red circles, respectively. The selectivity was determined with a P value cutoff of 0.01 for Δz.

over-expressed in breast cancers with at least a 2-fold change with $P < 0.01$ from a t-test (Fig. 3B). From the gene-specific interpretation (Δz) of DEG, we found that 272 and 73 probe features with over 2- and 4-fold over-expression, respectively, were non-selective in breast cancers (P for Δz is >0.01), while 47 and 59 probe features with over 2- and 4-fold over-expression, respectively, were selective in breast cancers (P for Δz is <0.01). It is clearly shown that probe features of large DEG (>4-fold over-expression) tend to have higher global SD than those of moderate DEG (>2-fold over-expression). In addition, probe features with selective (P for Δz is <0.01) over-expression in breast cancers tend to have smaller global SD than genes with non-selective (P for Δz is <0.01) over-expression. A large SD of the expression of a gene in global samples implies a relatively non-selective over-expression in various biological samples or conditions. Since highly over-expressed genes in Figure 3B tend to have large SDs, it will be difficult to find breast cancer-specific expression among them. In fact, many existing statistical methods for microarray analysis cannot prioritize target-selective gene expression over many non-selective over-expressed genes, because they just emphasize genes with a high magnitude of change that far exceeds the observed variability in expression (Ramaswamy *et al.*, 2001). However, through the gene-specific interpretation of DEG (Δz), we could select genes of over a 4-fold change in breast cancers and a relatively small global SD, which implies specific over-expression in breast cancers.

Table 1. Gene-specific interpretation of the observed over-expression in breast cancer

Gene description	Probe ID	σ	Observed DEG	Gene-specific interpretation
Similar to cytochrome P450	236445_at	0.62	Fold change >4, $P < 0.01$ (t-statistics)	Breast cancer selectivity score, $z > 1.6$, Significant DEG ($P < 0.01$ for Δz)
Leucine-rich-repeat-containing 15	213909_at	0.60		
COMP	**205713_s_at**	**0.77**		
Fibrilin 1	235318_at	0.67		
Collagen type X	**217428_s_at**	**0.85**		
Collagen type I	202311_s_at	1.10		
Dedicator of cytokine 1	241709_s_at	0.46		
Collagen-triple-helix-repeat-containing 1	225681_at	1.10		
Inhibin, beta	**227140_at**	**1.18**		Breast cancer selectivity score, $z < 1.6$, Insignificant DEG ($P > 0.05$ for Δz)
Lysozyme	213975_s_at	1.62		
Collagen type VIII	**226237_at**	**1.20**		
Collagen type I	1556499_s_at	1.61		
Collagen type III	211161_s_at	1.72		
Collagen type VI	201438_at	1.44		

Among genes with >4-fold over-expression ($P < 0.01$ from t-statistics) in breast cancer versus normal breast tissue, breast cancer-selective (high z) gene expression and specific DEG (high Δz) between normal and cancer tissues were identified by the gene-specific rescaling methods. σ represents the global SD of the gene expression calculated from 1303 samples in the GPL570 datasets. Genes listed in bold were further analyzed for the DB-wide expression profiling in Figure 4.

For further validation, we first selected a total of 132 probe features (genes) with >4-fold over-expression and high statistical confidence ($P < 0.01$ in t-distribution) in the breast cancer tissue, as compared to normal breast tissue. We then applied the gene-specific rescaling process to the expression data of the selected genes. As a result, the expression data of eight genes showed that they were selectively expressed in breast cancer ($z > 1.6$). Additionally, the Δz values of these genes, as compared between normal and cancer tissues, were significant over the other genes ($P < 0.01$ for the Δz value) (first eight genes in Table 1). In contrast, six of the over-expressed genes were found to have relatively low LAGE selectivity ($z < 1.6$) in breast cancer and exhibited insignificant Δz between normal and cancer tissues ($P > 0.05$ for the Δz value) (last six genes in Table 1).

Before we experimentally tested the breast cancer selectivity of the expression pattern of these genes, we verified the reproducibility of the over-expression pattern by examining another microarray dataset from the integrated DB samples. We generated DB-wide expression profiles for the selected genes, using all 1303 integrated array samples from 70 different datasets. In addition to the analyzed breast cancer dataset (GDS2635) in Figure 3 and Table 1, two other datasets contained breast cancer tissues (GDS2250 and GDS2046). As a result, among the 14 over-expressed genes shown in Table 1, only four genes (Inhibin, Collagen VIII, COMP and Collagen X) were shown to have consistent over-expression patterns in breast

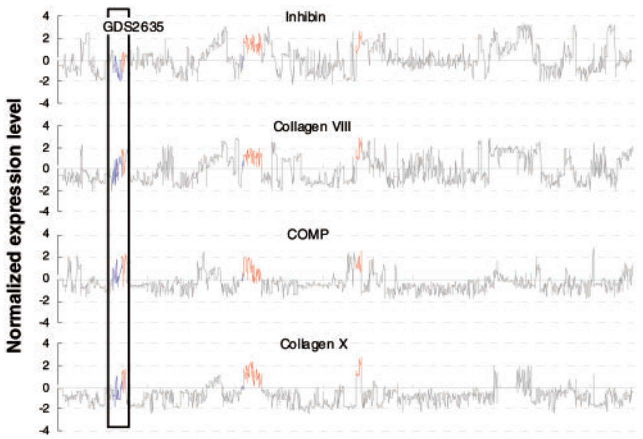

Fig. 4. DB-wide expression profiles of four genes with over-expression in breast cancer tissues. While Inhibin and Cllagen VIII showed non-selective over-expression, the expression of COMP and Collagen X in breast cancer was highly selective over other biological samples. The expression trends of selected genes were investigated in the integrated DB samples. The *x*-axis represents a total of 1303 array samples from 70 datasets sharing the GPL570 platform. The red line represents the expression level in breast cancer tissue samples, while the blue line represents the expression level in normal breast tissue. The square box represents the dataset (GDS2635) that was used for the initial gene selection in Table 1.

cancer tissues in the two additional datasets (Fig. 4). Based on this observation, we used the data of these four genes for further experimentation.

The global SD of the selected genes (σ in Table 1) is assumed to represent the intrinsic variability of the gene's expression. COMP and Collagen X were found to have low σ values and high selectivity to breast cancer (Table 1). Consistently, the DB-wide expression profile showed that they had consistently low expression in most of the DB samples and exceptionally high expression in only a few DB samples, including the breast cancer samples (Fig. 4). On the other hand, Inhibin and Collagen VIII were found to have relatively high σ values and low selectivity to breast cancer (Table 1). Consistently, their DB-wide expression profiles showed that they had high expression in many DB samples, in addition to breast cancer tissues (Fig. 4).

3.4 Selectivity analysis by qPCR

Using qPCR, we experimentally investigated the global expression trend of these four selected genes in diverse human tissue samples, including breast cancer and normal breast samples (Fig. 5). COMP and Collagen X showed the highest expression levels in breast cancer tissues among the 18 different human tissues compared. Although Inhibin and Collagen VIII also showed high expression in breast cancer tissues, they showed similar or higher expression in many other types of normal and cancer tissues. This qPCR result agrees well with our gene-specific interpretation of the microarray data shown in Table 1 and Figure 4. Using the qPCR results, the experimental selectivity score of gene expression in breast cancer was also compared among the four genes (Supplementary Table 2). The equation used for the *z*-score was again employed for the selectivity scoring of the qPCR results, and the u, μ and σ values for each gene were derived from the 381 observed qPCR expression data points. Consistent with the DB-derived z scores, the qPCR selectivity score showed that the expression of COMP and Collagen X is more selective to breast cancer than that of Inhibin and Collagen VIII.

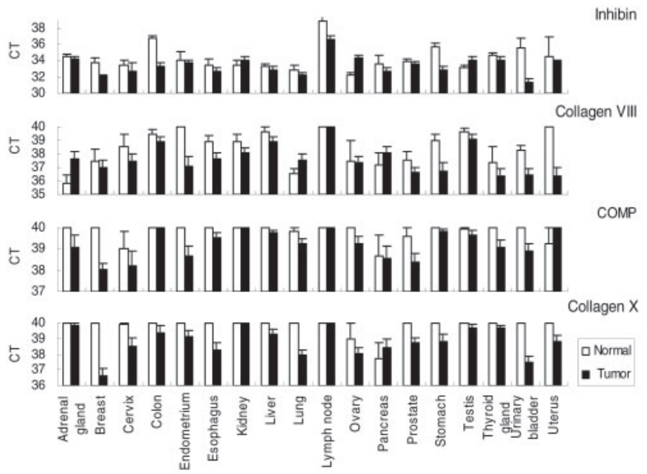

Fig. 5. The results of qPCR analysis of the expression of four selected genes in diverse human tissues. cDNAs from a total of 381 normal and tumor samples from 18 different human tissues were used for the analysis. The number of samples for each tissue type is found in Supplementary Table 3. The cycle threshold (CT) in the *y*-axis is a direct reading of the qPCR signals, indicating that a low CT value represents high gene expression.

Likewise, in the DEG (i.e. fold change) analysis between normal and cancer samples in 18 different human tissues, the DEG values of COMP and Collagen X in breast tissue were observed to be higher than the DEGs measured for other types of tissues. The DEGs of Inhibin and Collagen VIII in breast tissue were not distinguishable from the DEGs in many other tissues (Fig. 5). In this qPCR analysis, Inhibin and COMP showed a similar level of DEG (~2-fold change) in breast tissue. However, only the DEG of COMP was uniquely high in breast samples among the 18 compared tissues. Inhibin showed higher DEG values in colon, lymph node, stomach and urinary bladder samples, as compared to breast tissue. Thus, in a biological sense, it is clear that the over-expression of COMP is more specific to breast cancer than of other similarly over-expressed genes. In the analysis of global expression trends, COMP was found to have a low variability in gene expression, while Inhibin was a relatively highly variable gene (Table 1). By using this intrinsic difference between these two genes for the interpretation of a microarray dataset, we could successfully prioritize COMP ($P < 0.01 \Delta z$ value) over Inhibin ($P > 0.05 \Delta z$ value) as a breast cancer-specific gene, although both genes showed a significant DEG in breast cancer samples (>4-fold change and P <0.01 for t-test).

It has already been shown that these four genes are over-expressed in several types of human cancer (Aigner *et al.*, 1997; Bettelheim *et al.*, 1984; Liao *et al.*, 2003; Sheth *et al.*, 1984). However, the over-expression of COMP and Collagen X in breast cancer tissues has not been previously reported. Therefore, the breast cancer specificity of these two genes, as compared to other cancers, has not yet been clarified. It is obvious that the identification of tissue selectivity or disease specificity in the study of gene expression is of critical importance in accelerating mechanistic studies and/or biomarker discovery. We have demonstrated that the expression of COMP and Collagen X in breast cancer is highly selective, as compared to many other types of human tissues and cancers, despite that the level of over-expression in breast cancer was not uniquely high in comparison to that of other non-selective over-expressed genes, such as Inhibin and Collagen VIII.

4 DISCUSSION

In the protein–protein interaction network, COMP and Collagen X have relatively few interacting partners compared to Inhibin and Collagen VIII (Fig. 6), suggesting that COMP and Collagen X are topologically and functionally peripheral in the cellular network, while Inhibin and Collagen VIII are hub proteins. It has been reported that essential genes are likely to encode hub proteins and are expressed widely in most tissues (Goh *et al*., 2007; Han *et al*., 2004; Jeong *et al*., 2001). Goh *et al*. (2007) also found that the vast majority of disease genes were non-essential and without tendency of encoding hub proteins. They provided an explanation for most disease genes using an evolutionary argument, such that only disease-related mutations in the functionally and topologically peripheral regions of the cell have a higher chance of viability. Consistently, the present study shows that two hub proteins, Inhibin and Collagen VIII, are over-expressed in many different normal and cancer tissues, indicating that their expression is essential and not specific to a disease. On the other hand, two peripheral genes, COMP and Collagen X, show highly selective over-expression in breast cancers, thus suggesting that their expression is non-essential and specific to breast cancers. This analysis demonstrates that the gene-specific interpretation of transcriptome data is able to distinguish non-essential disease-related genes from essential genes with over-expression in various tissues.

In conclusion, the selective expression of COMP and Collagen X in breast cancer, observed by qPCR experiments, supports the hypothesis that gene-specific interpretation of microarray data allows highly sample-specific gene expression to be distinguished from non-selective over-expression, without large-scale experiments on various samples. In a transcriptome-wide study, such as a microarray experiment, a large number of candidate genes with significant over-expression in target samples are often identified. Thus, it is challenging to determine candidates of more direct or specific importance to a target disease or the metabolic regulation of specific samples. In this study, we have shown that gene-specific interpretation of microarray data can be used to prioritize genes with greater selective expression to specific tissues over many other over-expressed genes.

We believe that a transcriptome-wide survey of the global trends in gene expression is of critical importance to better interpreting large-scale gene expression data. Given that individual genes each have their own regulatory system for transcriptional activity, the analysis of global expression trends can provide a direct measure of a gene's intrinsic expressional behavior. The present study shows that DB-wide average and SD of gene (probe feature) expression are an effective measure of the intrinsic expressional behavior of a gene and that they can be used for better interpretation of microarray data. Some of intrinsic properties observed in this study might reflect characteristics of probes themselves. Thus, the present gene-specific rescaling method also minimizes the probe-specific variation in the interpretation of microarray data. Rapid increases in public microarray data will further improve the quality of this approach and provide a unique tool for biological interpretation of gene expression data on a transcriptome basis. We are currently expanding the integrated microarray datasets to generalize the 'gene-specific interpretation of transcriptome data' to diverse organisms and experimental conditions. All of the resources used in the present study, including the integrated microarray datasets and analysis methods, are available at http://lage.sookmyung.ac.kr.

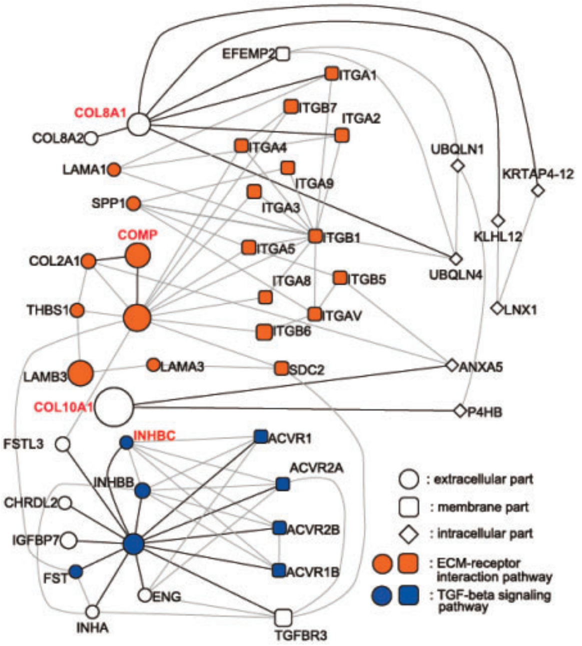

Fig. 6. Protein interaction network including proteins encoded by four selected genes. The gene symbols of Inhibin, Collagen VIII, COMP and Collagen X are presented as COL10A1, INHBA, COMP and COL8A1 (red letters) in the network. It is shown that COMP and Collagen X have relatively fewer interacting partners than Inhibin and Collagen VIII in the network. The size of individual nodes represents the level of over-expression in breast cancers versus normal tissues. Note that the information of metabolic pathways was retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg), and the information of protein interactions was retrieved from the Human Protein Reference Database (HPRD; http://www.hprd.org).

ACKNOWLEDGEMENTS

The authors thank Dr Aeri Kim at Pusan National University and Dr Dukwhan Kim at Samsung Medical Center for advice and help with the qPCR experiments.

Funding: National Research Foundation of Korea (KRF) grants funded by Korea government (MEST) (No. 2009-0076307) and SRC program (No. R11-2005-017-04003-0); National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126 to T. P.).

Conflict of Interest: none declared.

REFERENCES

Aigner,T. *et al*. (1997) Type X collagen expression and hypertrophic differentiation in chondrogenic neoplasias. *Histochem. Cell Biol.*, **107**, 435–440.

Aittokallio,T. *et al*. (2003) Computational strategies for analyzing data in gene expression microarray experiments. *J. Bioinform. Comput. Biol.*, **1**, 541–586.

Barnes,M. *et al*. (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.*, **33**, 5914–5923.

Barrett,T. *et al*. (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.*, **33**, D562–D566.

Barrett,T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

Bettelheim,R. *et al*. (1984) Immunocytochemistry in the identification of vascular invasion in breast cancer. *J. Clin. Pathol.*, **37**, 364–366.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Cope,L.M. *et al*. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.

Gagarin,D. *et al*. (2005) Genomic profiling of acquired resistance to apoptosis in cells derived from human atherosclerotic lesions: potential role of STATs, cyclinD1, BAD, and Bcl-XL. *J. Mol. Cell. Cardiol.*, **39**, 453–465.

Goh,K.I. *et al*. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Han,J.D. *et al*. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), S96–S104.

Irizarry,R.A. *et al*. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Liao,Q. *et al*. (2003) COMP is selectively up-regulated in degenerating acinar cells in chronic pancreatitis and in chronic-pancreatitis-like lesions in pancreatic cancer. *Scand. J. Gastroenterol.*, **38**, 207–215.

Mootha,V.K. *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Paik,S. *et al*. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl J. Med.*, **351**, 2817–2826.

Ramaswamy,S. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Sheth,N.A. *et al*. (1984) Circulating levels of inhibin in cancer. *Neoplasma*, **31**, 315–321.

Shippy,R. *et al*. (2006) Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat. Biotechnol.*, **24**, 1123–1131.

Su,A.I. *et al*. (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.*, **61**, 7388–7393.

Yang,M.C. *et al*. (2001) A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol. Genomics*, **7**, 45–53.

Yang,Y.H. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Yoon,S. *et al*. (2006) Large scale data mining approach for gene-specific standardization of microarray gene expression data, *Bioinformatics*, **22**, 2898–2904.