Journal of Biochemistry and Molecular Biology, Vol. 40, No. 2, March 2007, pp. 226-231

JOURNAL OF Biochemistry and Molecular Biology

Identification of Novel Universal Housekeeping Genes by Statistical Analysis of Microarray Data

Seram Lee[1], Minjoung Jo[1], Jungeun Lee[1], Sang Seok Koh[2,*] and Soyoun Kim[1,*]

[1]Department of Chemistry, Dongguk University, Seoul 100-715, Korea

[2]LG Life Sciences Ltd./R&D Park, Daejeon 305-380, Korea

Received 9 October 2006, Accepted 14 November 2006

Housekeeping genes are widely used as internal controls in a variety of study types, including real time RT-PCR, microarrays, Northern analysis and RNase protection assays. However, even commonly used housekeeping genes may vary in stability depending on the cell type or disease being studied. Thus, it is necessary to identify additional housekeeping-type genes that show sample-independent stability. Here, we used statistical analysis to examine a large human microarray database, seeking genes that were stably expressed in various tissues, disease states and cell lines. We further selected genes that were expressed at different levels, because reference and target genes should be present in similar copy numbers to achieve reliable quantitative results. Real time RT-PCR amplification of three newly identified reference genes, CGI-119, CTBP1 and GOLGAl, alongside three well-known housekeeping genes, B2M, GAPD, and TUBB, confirmed that the newly identified genes were more stably expressed in individual samples with similar ranges. These results collectively suggest that statistical analysis of microarray data can be used to identify new candidate housekeeping genes showing consistent expression across tissues and diseases. Our analysis identified three novel candidate housekeeping genes (CGI-119, GOLGA1, and CTBP1) that could prove useful for normalization across a variety of RNA-based techniques.

Keywords: Affymetrix genechip, Genomics, Housekeeping genes, Internal control, Real-time quantitative RT-PCR

Introduction

Recently developed technologies such as microarray analysis allow researchers to determine genome-wide expression patterns, providing important insights into complex regulatory networks, enabling the identification of new or under-explored biological processes, and implicating genes in various disease processes (Schena *et al.,* 1995; Golub *et al.,* 1999; Graveel *et al.,* 2001; Hamadeh *et al.,* 2002; Lee and Thorgeirsson, 2002; Suh *et al.,* 2006; Zhong *et al.,* 2006). Another new technology, real-time reverse transcription polymerase chain reaction (RT-PCR), simultaneously measures gene expression in many different samples, providing quantitative information and good reflections of expression level changes (Gibson *et al.,* 1996; Heid *et al.,* 1996). Both of these mRNA-based strategies, as well as others such as Northern blotting and RNase protection assays, require accurate, reproducible normalization of results. Various strategies have been used to normalize gene expression data, including cell counting, quantification of total RNA and measurement of rRNA (Vandesompele *et al.,* 2002). The most common mRNA normalization strategy involves the use of internal control genes. These so-called housekeeping genes (Suzuki *et al.,* 2000) are generally stable across tissues, cells and experimental treatments, thus providing good normalization. However, although housekeeping genes are uniformly expressed in certain cell types, they can vary in others (Thellin *et al.,* 1999; Suzuki *et al.,* 2000; Warrington *et al.,* 2000), particularly in clinical samples associated with malignant diseases (Suzuki *et al.,* 2000; Khimani *et al.,* 2005). Thus, the selection of proper control genes for clinical patient samples is vital to gene expression analysis.

A variety of statistical methods and programming efforts have been employed to seek and evaluate new, stable, reference genes (Szabo *et al.,* 2004). For example, Speleman and colleagues (Vandesompele *et al.,* 2002) developed the GeNorm program, which uses geometric means to calculate the correct normalizing factor from existing housekeeping genes. Here, we utilized statistical tools, such as geometric mean, standard

Abbreviations: RT-PCR: Reverse transcription-polymerase chain reaction

*To whom correspondence should be addressed.
Tel: 82-2-2260-3840; Fax: 82-2-2268-8204
E-mail: skim@dongguk.edu

deviation and linear regression, to search a large microarray database for new, stably expressed, novel genes. We further screened for reference genes that are expressed at different levels, as it is beneficial for the reference genes and the genes of interest to be within similar ranges of expression.

Materials and Methods

Data samples. The Oncology Database of Gene Logic contains the genomic expression profiles of clinical tissue samples from many different human organs. These profiles were originally generated using high-density oligonucleotide microarray analysis (HG-U133; Affymetrix) of 281 normal tissue samples from 17 different organs, including breast (27), cervix (5), colon (26), duodenum (10), endometrium (9), esophagus (14), kidney (29), liver (21), lung (32), lymph node (5), myometrium (5), ovary (19), pancreas (19), prostate (15), rectum (18), skin (5), and stomach (22) (numbers in parentheses indicate the number of normal tissue samples analyzed). Normalized signals (expression values) were obtained using the Microarray Suite 5.0 software (Affymetrix), which deletes the largest 2% and the smallest 2% outliers, and the mean of the remaining values (trimmed mean) was used to compute the scale factor (SF = 100/trimmed mean).

Statistical data analysis. Novel reference genes exhibiting little variation across the 17 tissue sample sets were identified by comparing the geometric means of the expression values in each sample set, using the GeneExpress 2000 Software Contrast Analysis and Electronic Northern Analysis tools. Contrast analysis was used to find genes that were similarly expressed across sample sets, while electronic Northern analysis allowed us to infer the range of expression levels for each gene in each sample set (Schmitt *et al.,* 1999). Selected novel reference genes were further evaluated by linear regression analysis (Analyzing Data with Graphical Prism, GraphPad Software Inc.). Briefly, for each reference gene, the Fold Value to Minimum (FVM) of each tissue sample set was obtained by dividing the geometric mean of the sample set by the minimum among the 17 mean values. Linear regression analysis was performed using the FVMs to generate slope and R^2 values. A lower slope value indicated less variation in the expression of a given gene across the 17 tissue sample sets. For individual tissue samples, linear regression analysis was performed with FVMs calculated by dividing the expression value of each tissue sample by the minimum among the 281 expression values.

Comparison analysis. Three selected novel reference genes (CGI-119[Transmembrane BAX inhibitor motif containing 4], GOLGA1 [golgi autoantigen, golgin subfamily a, 1] and CTBP1[C-terminal binding protein 1]) were compared with three well-known, housekeeping genes (B2M[Beta-2-microglobulin], GAPD[Glyceraldehyde-3-phosphatedehydrogenase] and TUBB[tubulin beta 2A]) (Thellin *et al.,* 1999; Suzuki *et al.,* 2000), using the statistical analyses described above. Real time RT-PCR data from the different sets of clinical samples were analyzed for comparison of the six genes.

Real time RT-PCR. Sixty-seven independent tissues from different clinical samples used for microarray analysis were prepared as previously described (Kim and Kim, 2003; Kim and Park, 2005). Total RNA was reverse transcribed to cDNA using oligo(dT), and quantitative real-time PCR analysis was performed as previously described (Kim *et al.,* 2003). In brief, the templates and primer sets were mixed with 2x QuantiTect SYBR Green PCR Master Mix (Quiagen), and 20 cycles of PCR reaction were performed using a Rotor-Gene real-time PCR machine (Corbett Research, Inc.). The gene-specific primers were TUB (TTCCAGCTGACCCACTCTCT; ACAGGGCCTCGTTATCAATG), GAPD (TGCACCACCAACT GCTTAGC; GGCATGGACTGTGGTCATGAG), B2M (TGCTGT CTCCATGTTTGATGTATC; TCTCTGCTCCCCACCTCTAAG), CGI-119 (TGGTGAAACCCCGTCTCTAC; TGATCTTGGCTCA ATGCAAC), GOLGA1 (GAAACAGGACTTGGAGCAGC; ATG TTTGCCATCTCAGGTCC), and CTBP1 (TTCACCGTCAAGCA GATGAG; GGCTAAAGCTGAAGGGTTCC) (Vandesompele *et al.,* 2002). All experiments were performed at least twice.

Results and Discussion

Accurate normalization of gene expression levels is an absolute prerequisite for reliable study results, especially when investigating the biological significance of subtle differences in gene expression. As numerous studies have reported that housekeeping gene expression can vary considerably (Suzuki *et al.,* 2000; Chen *et al.,* 2002; Kim and Wang, 2003), it is important to choose the most appropriate control for a given tissue or disease. Here, we used basic statistical methods such as geometric mean, standard deviation and linear regression to identify novel reference genes from a large microarray database.

We obtained the genomic expression patterns of 281 normal tissue samples from the 17 different organs, available from the Oncology Database of Gene Logic, and used two statistical methods to screen these data for novel reference genes. First, we used mean and standard deviations. For initial correction, all expression profiling data of each gene were divided by the corresponding minimum data, which normalize the data with regards to the copy numbers of each gene. Then the mean and standard deviations of the values from each gene were calculated. Since the data had been divided by their minimum values, corrected values near 1 indicated that the genes showed little variation, as did small standard deviations. We then analyzed the data by multiplying the mean values with the corresponding standard deviations, which increased the reliability of the analysis. Thus, lower values (mean times standard deviation) indicated genes with lower variations in expression level. This is a necessary, although insufficient, characteristic for identification of a candidate reference gene. Secondly, we utilized a simple linear regression model. The data from each gene were divided by the corresponding minimum data and sorted with regards to the FVM data value as described in the statistical data analysis section of Materials and Methods. The sorted data were analyzed by a simple linear regression model (slope and R^2 value), in which the

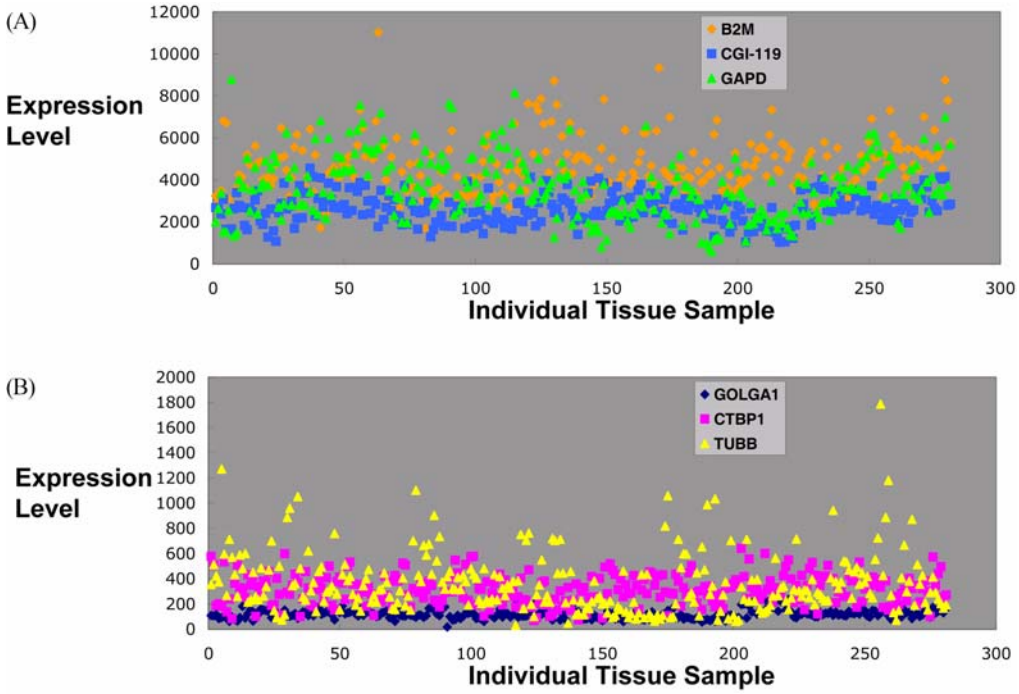


Fig. 1. Identification of novel reference genes. A and B: Expression profiles of reference genes across 281 individual tissue samples (X-axis). For better comparison, the genes are divided into 2 groups according to their expression levels (Y-axis): high expression genes (A) and low expression genes (B).

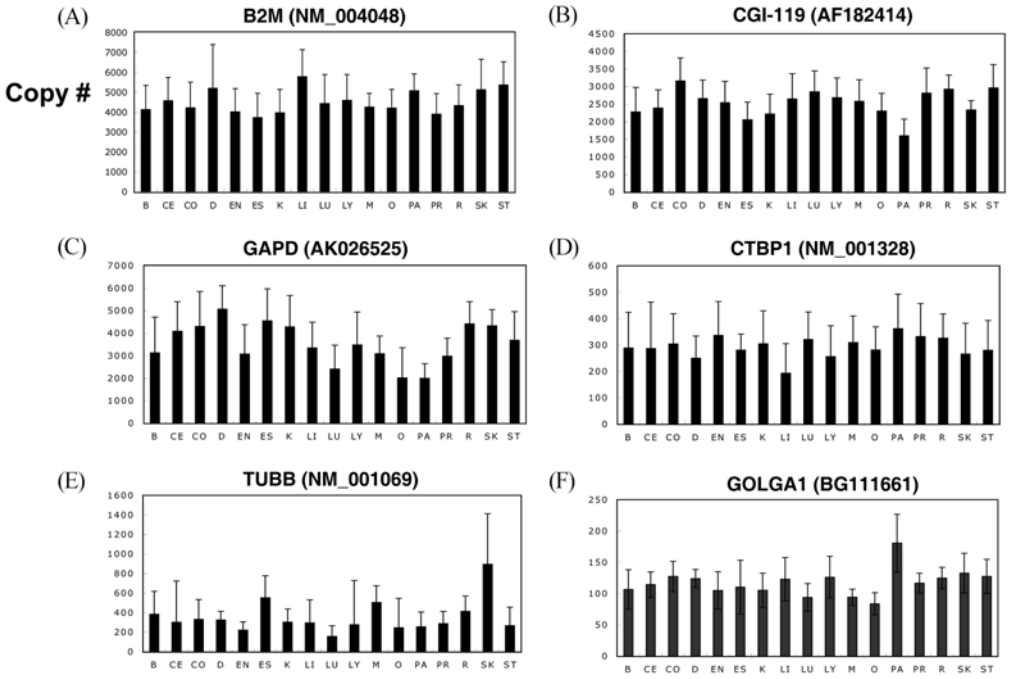
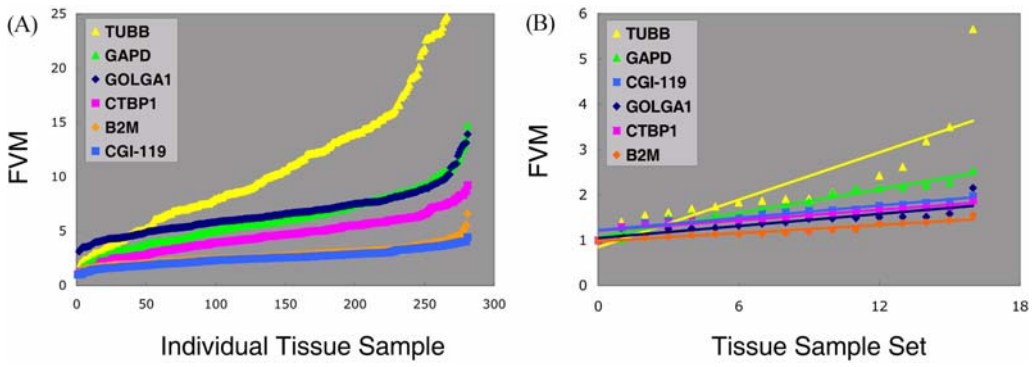


Fig. 2. A~F: Expression profiles of novel reference genes (A, B, D, and F) and the commonly used reference genes, GAPD (C) and TUBB (E), across 17 different tissue sample sets. GenBank accession numbers are indicated. The Y-axis indicates the copy number of the corresponding genes. The abbreviations used on the X-axis are: B, breast; CE, cervix; CO, colon; D, duodenum; EN, endometrium; ES, esophagus; K, kidney; LI, liver; LU, lung; LY, lymph node; M, myometrium; O, ovary; PA, pancreas; PR, prostate; R, rectum; SK, skin; ST, stomach.



(C)

Gene Symbol (Genbank)	Individual Sample FVM			Tissue FVM		
	Linear Regression	R^2	R^2/slope	Linear Regression	R^2	R^2/slope
CGI-119 (AF182414)	y = 0.0083x + 1.3701	0.9651	116.3	y = 0.0454x + 1.1724	0.9221	20.3
B2M (NM_004048)	y = 0.0093x + 1.4704	0.8875	95.4	y = 0.0299x + 0.9452	0.9244	30.9
CTBP1 (NM_001328)	y = 0.0204x + 1.6995	0.9658	47.3	y = 0.0383x + 1.1698	0.8971	25.2
GOLGA1 (BG111661)	y = 0.023x + 3.4116	0.8864	38.5	y = 0.0447x + 0.9979	0.7833	17.5
GAPD (AK026525)	y = 0.0301x + 1.8696	0.9448	31.4	y = 0.0878x + 0.9802	0.9640	11.0
TUBB (NM_001069)	y = 0.085x - 0.1564	0.7930	9.3	y = 0.1744x + 0.6689	0.6698	3.8

Fig. 3. Linear regression analysis of FVMs for the reference genes. A and B: FVMs across 281 individual tissue samples and 17 tissue sample sets, respectively. C: The slope and R^2 values for each reference gene from A and B.

value of the slope is zero when the expression levels of a given gene do not differ among tissues samples. Thus, a lower slope value for a given gene indicates its greater potential as a reference gene. In addition to the slope, linearity (R^2) can also be an important factor in determining a better reference gene with similar slope value. Here, we used the R^2/slope to identify novel reference genes, which could be characterized by higher values (R^2/slope) indicating genes with low variations in expression level.

The results of our data analysis identified several genes having low mean times standard deviation values and high R^2/slope values. Analysis of the respective copy numbers of these genes allowed us to select B2M, CGI-119, CTBP1 and GOLGA1 as possible reference genes, as they showed low expression variation across individual tissue samples and a good range of copy numbers (Fig. 1A and B). Of these, B2M is a previously known, housekeeping gene, whereas the other three have not previously been identified as housekeeping genes. Comparison of the means and standard deviations of tissue sample sets revealed that CGI-110, CTBP1 and GOLGA1 showed lower values than the commonly used housekeeping genes, GAPD and TUBB (Fig. 2A-F). Furthermore, simple linear regression of the sorted data revealed that CGI-119, B2M, CTBP1 and GOLGA1 had lower slope values than GAPD and TUBB (Fig. 3). The results of these two separate statistical methods confirmed the same three novel genes (CGI-110, CTBP1 and GOLGA1) as good candidate housekeeping genes that may be more stably expressed than the commonly used housekeeping genes, GAPD and TUBB (Figs. 1, 2 and 3).

Using the linear regression model, we further analyzed the expression profiles of the identified stable genes (GOLGA1, CTBP1, B2M and CGI-119) in the same tissues under different disease states (Fig. 3B). Although the relative rankings of their stabilities were somewhat altered, the four tested genes were more stable in the 23 different cancers than were the commonly used housekeeping genes, GAPD and TUBB (Fig. 3C). The expression levels of GOLGA1, CTBP1, B2M, CGI-119, GAPD and TUBB in the different tissues were then analyzed by linear regression to identify the most stable gene in each cancer type, regardless of the sampled tissue. GOLGA1, CTBP1, B2M and CGI-119 showed higher stabilities than GAPD and TUBB in each cancer type (data not shown).

To validate these array-based analyses, we next performed real time quantitative RT-PCR analysis. We used gene-specific primers (see Materials and Methods) to amplify cDNA from independent sets of normal, cirrhotic and cancerous liver tissues. The stabilities of GOLGA1, CTBP1 and CGI-119

Table 1. Statistical analysis of real time RT-PCR data from independent tissue samples

Gene symbol	Standard deviation				Linear regression				GeNorm	
	AVR	SD	AVR × SD	[a]Rank	Slope	R^2	R^2/Slope	[b]Rank	[c]GeNorm score	[d]Rank
CGI-119	2.178	0.720	1.568	1	0.231	0.941	4.079	1	1.2862	2
B2M	12.674	14.567	184.619	5	4.318	0.806	0.187	5	1.9292	5
CTBP1	11.238	11.435	128.507	4	3.410	0.815	0.239	4	1.4879	4
GOLGA2	2.403	1.724	4.142	2	0.453	0.634	1.399	2	1.2702	1
GAPD	3.753	2.528	9.489	3	0.811	0.942	1.162	3	1.4183	3
TUBB	65.793	86.934	5719.643	6	23.761	0.685	0.029	6	2.5844	6

[a]Genes are ranked in terms of the average (AVR) multiplied by the standard deviation (SD).
Lower AVR × SD values indicate the higher ranks and increased gene stability.
[b]Genes are ranked in terms of the R^2/slope value. Higher R^2/slope values reflect higher ranks and increased gene stability.
[c]The GeNorm score was calculated using the GeNorm Program (3).
[d]A lower GeNorm score indicates higher stability.

were high in these tissues, while B2M was less stable than GAPD in this analysis (Table 1). This highlights the limitations of commonly used housekeeping genes such as B2M, which may show disease-specific effects. We compared these two statistical methods to previously developed program, GeNorm scoring, which uses geometric averaging of multiple internal control genes. The GeNorm program results indicated that the three novel genes showed high stability using the same data set (Table 1). To further examine the tissue specific effect, we generated cDNA from several laboratory cell lines originating from different tissues, and tested the expression levels of the selected genes by real-time quantitative analysis. Again, although the relative order of the stabilities differed from those in the microarray analysis, the stabilities of GOLGA1, CTBP1 and CGI-119 were consistently better than those of the commonly used housekeeping genes, TUBB and GAPD. To be consistent with our liver tissue data (Table 1), we further validated our newly identified housekeeping genes by using GOLGA1, CTBP1 and CGI-119 as references to normalize the expression levels of genes having different ranges of copy number, and found that the use of a reference with a similar copy number yielded better normalization results (data not shown).

In summary, these results collectively indicate that statistical analysis of microarray data can be used to identify new candidate housekeeping genes showing consistent expression across tissues and diseases. Our analysis identified three new candidate housekeeping genes (GOLGA1, CTBP1 and CGI-119) that could prove useful for normalization across a variety of RNA-based techniques. Importantly, we identified these genes using the basic statistical tools of geometric mean, standard deviation and linear regression. All of these functions are available in the basic Excel package (functions STDEV, AVEG and Linear Regression), making this method available to most scientists for study-specific selection of the most optimal housekeeping genes.

Acknowledgments This work was supported in part by a MOST grant (FG03-32-01) from the 21st Century Frontier Functional Human Genome Project to S.S.K. This work was supported by KRF (Grant # D0006), MOST(protein chip program), KIEST(Grant #101-051-022) and MOHW(Grant# A050814) to S.K. This work was supported by the National Research Laboratory Program (NRL, MOST).

References

Chen, X., Cheung, S. T., So, S., Fan, S. T., Barry, C., Higgins, J., Lai, K. M., Ji, J., Dudoit, S., Ng, I. O., Van De Rijn, M., Botstein, D. and Brown, P. O. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell* **13**, 1929-1939.

Gibson, U. E., Heid, C. A. and Williams, P. M. (1996) A novel method for real time quantitative RT-PCR. *Genome Res.* **6**, 995-1001.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Graveel, C. R., Jatkoe, T., Madore, S. J., Holt, A. L. and Farnham, P. J. (2001) Expression profiling and identification of novel genes in hepatocellular carcinomas. *Oncogene* **20**, 2704-2712.

Hamadeh, H. K., Bushel, P. R., Jayadev, S., DiSorbo, O., Bennett, L., Li, L., Tennant, R., Stoll, R., Barrett, J. C., Paules, R. S., Blanchard, K. and Afshari, C. A. (2002) Prediction of compound signature using high density gene expression profiling. *Toxicol. Sci.* **67**, 232-240.

Heid, C. A., Stevens, J., Livak, K. J. and Williams, P. M. (1996) Real time quantitative PCR. *Genome Res.* **6**, 986-994.

Khimani, A. H., Mhashilkar, A. M., Mikulskis, A., O'Malley, M., Liao, J., Golenko, E. E., Mayer, P., Chada, S., Killian, J. B.

and Lott, S. T. (2005) Housekeeping genes in cancer: normalization of array data. *Biotechniques* **38**, 739-745.

Kim, J. W. and Wang, X. W. (2003) Gene expression profiling of preneoplastic liver disease and liver cancer: a new era for improved early detection and treatment of these deadly diseases? *Carcinogenesis* **24**, 363-369.

Kim, S. and Kim, T. (2003) Selection of optimal internal controls for gene expression profiling of liver disease. *Biotechniques* **35**, 456-460.

Kim, S. and Park, Y. M. (2005) Specific gene expression patterns in liver cirrhosis. *Biochem. Biophys. Res. Commun.* **334**, 681-688.

Kim, S., Shi, H., Lee, D. K. and Lis, J. T. (2003) Specific SR protein-dependent splicing substrates identified through genomic SELEX. *Nucleic Acids Res.* **31**, 1955-1961.

Lee, J. S. and Thorgeirsson, S. S. (2002) Functional and genomic implications of global gene expression profiles in cell lines from human hepatocellular cancer. *Hepatology* **35**, 1134-1143.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.

Schmitt, A. O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C. P., Hinzmann, B. and Rosenthal, A. (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* **27**, 4251-4260.

Suh, Y. J., Yang, M. H., Yoon, S. J. and Park, J. H. (2006) GEDA: new knowledge base of gene expression in drug addiction. *J. Biochem. Mol. Biol.* **39**, 441-447.

Suzuki, T., Higgins, P. J. and Crawford, D. R. (2000) Control selection for RNA quantitation. *Biotechniques* **29**, 332-337.

Szabo, A., Perou, C. M., Karaca, M., Perreard, L., Quackenbush, J. F. and Bernard, P. S. (2004) Statistical modeling for selecting housekeeper genes. *Genome Biol.* **5**, 59.

Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A. and Heinen, E. (1999) Housekeeping genes as internal standards: use and limits. *J. Biotechnol.* **75**, 291-295.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. and Speleman, F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, 34.

Warrington, J. A., Nair, A., Mahadevappa, M. and Tsyganskaya, M. (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics* **2**, 143-147.

Zhong, J., Wang, Y., Qiu, X., Mo, X., Liu, Y., Li, T., Song, Q., Ma, D. and Han, W. (2006) Characterization and expression profile of CMTM3/CKLFSF3. *J. Biochem. Mol. Biol.* **39**, 537-545.