Published online 18 March 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 8* ***e61***
doi:10.1093/nar/gkp168

Microarray labeling extension values: laboratory signatures for Affymetrix GeneChips

Yun-Shien Lee[1,2], Chun-Houh Chen[3], Chi-Neu Tsai[4], Chia-Lung Tsai[5], Angel Chao[5,*] and Tzu-Hao Wang[2,5,*]

[1]Department of Biotechnology, Ming Chuan University, [2]Genomic Medicine Research Core Laboratory, Chang Gung Memorial Hospital, Tao-Yuan, [3]Institute of Statistical Science, Academia Sinica, Taipei, [4]Graduate Institutes of Clinical Medical Sciences, Chang Gung University and [5]Department of Obstetrics and Gynecology, Lin-Kou Medical Center, Chang Gung Memorial Hospital and Chang Gung University, Tao-Yuan, Taiwan

Received November 20, 2008; Revised February 24, 2009; Accepted March 2, 2009

ABSTRACT

Interlaboratory comparison of microarray data, even when using the same platform, imposes several challenges to scientists. RNA quality, RNA labeling efficiency, hybridization procedures and data-mining tools can all contribute variations in each laboratory. In Affymetrix GeneChips, about 11–20 different 25-mer oligonucleotides are used to measure the level of each transcript. Here, we report that 'labeling extension values (LEVs)', which are correlation coefficients between probe intensities and probe positions, are highly correlated with the gene expression levels (GEVs) on eukayotic Affymetrix microarray data. By analyzing LEVs and GEVs in the publicly available 2414 cel files of 20 Affymetrix microarray types covering 13 species, we found that correlations between LEVs and GEVs only exist in eukaryotic RNAs, but not in prokaryotic ones. Surprisingly, Affymetrix results of the same specimens that were analyzed in different laboratories could be clearly differentiated only by LEVs, leading to the identification of 'laboratory signatures'. In the examined dataset, GSE10797, filtering out high-LEV genes did not compromise the discovery of biological processes that are constructed by differentially expressed genes. In conclusion, LEVs provide a new filtering parameter for microarray analysis of gene expression and it may improve the inter- and intralaboratory comparability of Affymetrix GeneChips data.

INTRODUCTION

Microarrays, particularly Affymetrix GeneChips, have become one of the most widely used high-throughput methods for functional genomic studies (1–4). The most common application of Affymetrix GeneChips has been to study mRNA as a method of measuring transcriptome activity. Microarrays have been used in numerous studies as a powerful tool for characterizing gene expression profiles; for classification of tumors versus normal tissues, primary versus metastasized tumors, prognosis of cancer patients and drug responses of patients, although traditional methods, such as observation of clinical features (such as tumor size, staging and lymph node metastases), are still mainstream parameters for clinicians to follow (1–3,5). There is no doubt that DNA microarrays represent a potential technology that can be used as a predictive tool or cancer biomarkers (6). However, critical concerns have been raised regarding the reliability and consistency of microarray results for both clinical and academic applications, since reports used to show little consistency among lists of differentially expressed genes by different commercial gene expression chips (7–9).

Affymetrix Genechips are designed so that gene expression is probed using a set of 11–20 different 25-mer oligonucleotide probe-pairs, including perfect-match (PM) and mismatch (MM) probes within each probe-pair. The integration of expression levels for each of the 11–20 PM and MM probe-pairs quantifies the expression of a particular gene to one value. Currently, cross-laboratory comparison of microarray data is still a challenge for scientists, although reports have shown highly consistent similarities of 85–90% in interlaboratory comparison of differential expression gene profiles using the Affymetrix Genechip (10). To resolve the remaining inconsistencies, several reports have developed robust and reproducible protocols for quality control of microarray techniques (11,12).

Factors contributing to the variations of Affymetrix microarray data from different laboratories include RNA quality, RNA labeling, hybridization process and data analysis tools (13). For eukaryotic specimens, the RNA labeling processes include cDNA synthesis

*To whom correspondence should be addressed. Tel: +886 3 3281200 extension 5402; Fax: +886 3 3288252; Email: knoxtn@cgmh.org.tw
Correspondence may also be addressed to Angel Chao. Email: angel945@cgmh.org.tw

and cRNA synthesis in gene chips. The cDNA synthesis begins with the binding of reverse transcriptase and a poly(T) primer oligonucleotide annealing to the poly(A) tail at the 3′ end of a mRNA, generating cDNA according to mRNA templates in the presence of dNTPs. The cRNA is then synthesized with a T7 primer, and biotin-labeled nucleotides are incorporated into *in vitro* transcripts. For prokaryotic cells, random primers are used to replace the poly(T) primer during cDNA synthesis, and then the cRNA is labeled with biotin-labeled substrates (http://www.affymetrix.com). RNA quality is usually determined by Bioanalyzer (Agilent) which measures the 28S/18S rRNA signal ratio (14). However, calculations based on area measurements are compromised by the hazy definitions of start and end points of peaks (14). Therefore, to estimate whether RNA-labeling efficiency or RNA quality affect the microarray results, the signal intensity ratio of the 3′ probe set over the 5′ probe set (3′/5′ ratio) is often used to evaluate labeling efficiency of genes (Microarray Suite, Affymetrix, Santa Clara, CA, USA) (14–17). Nevertheless, the exact mechanism and proof of RNA degradation sites are still lacking (18).

In this study, we tested the use of labeling extension values (LEVs) to re-evaluate the performance of Affymetrix GeneChips in more than 2000 publicly available microarray data deposited in Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) (19). To focus on interlaboratory comparison, Affymetrix GeneChips data sets from Microarray Quality Control (MAQC) (20,21) Project and NIH Neuroscience Microarray Consortium were analyzed. To our surprise, LEVs could be used to identify a 'laboratory signature' that may reflect the systemic variations that are unique to each laboratory.

MATERIALS AND METHODS

Data sets

A total of 2414 .cel files were downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) for this study. In addition, four data sets were analyzed. (i) MAQC Brain Dataset (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/docs/MAQC_Main_Study_Guidance.doc) included four RNA reference samples analyzed by three laboratory sites with five replicates per site using HG-U133 Plus 2.0 GeneChips. The four RNA samples were Stratagene's Universal Human Reference RNA (SUHRR), Ambion's Human Brain Reference RNA (HBRR), 25% HBRR/75% SUHRR and 75% HBRR/25% SUHRR. (ii) MAQC Rat Toxicogenomics Dataset (20) included six RNA samples analyzed by two laboratory sites with six replicates per site using Rat Genome 230 2.0 GeneChips. Samples were isolated from rat liver and kidney after each was exposed to aristolochic acid, riddelliine and comfrey. Unexposed tissues were used as controls. (iii) MAQC Tumor Dataset (21) included five biological replicates of two RNA samples analyzed by two laboratory sites using HG-U133 plus 2.0 Chips. The samples consisted of five colorectal adenocarcinomas and five normal colonic tissues. (iv) GSE2004 Dataset (NIH Neuroscience Microarray Consortium) included three replicates of four RNA samples analyzed by two laboratory sites using U133A Chips. Samples consisted of normal kidney, spleen, liver and a Universal RNA.

Gene expression levels (GEVs) and LEVs

The GEVs of each probe set for individual datasets were normalized using its robust multiarray average (RMA) (22,23), and the RMA measures were computed using the Methods for Affymetrix Oligonucleotide Arrays R package (3) that is freely available on the World Wide Web (http://www.bioconductor.org). The LEVs were defined as the Pearson's linear correlation coefficient between the probe position and base 2 logarithm of probe intensity for each gene. The probe intensities were extracted from the .cel files, and the probe interrogation position information for each type of chips was extracted from the .probe_tab files that can be downloaded from Affymetrix World Wide Web site. For example, the probe information of U133A was extracted from http://www.affymetrix.com/Auth/analysis/downloads/data/HT_HG-U133A.probe_tab.zip. The LEV was calculated using MATLAB Version 7.4 and Bioinformatics Toolbox Version 2.5. Since the scientists of Affymetrix Genechips designed the probe sets to be used mostly at the 3′ region (www.affymetrix.com/support/technical/manual/comparison_spreadsheets_manual.pdf), we first filtered out the probe sets, in which the distances between the probes with maximum position and minimum position were <300 nt, to avoid the inclusion of probe sets with very short length. By doing that, we have removed the probe sets that could not represent whole transcripts, including 3% in HG-U95, 17% in HG-U133 and 19% in HG-U133 Plus 2 chips, respectively. The source code and retrieved datasets for LEV are available in Supplementary Data 1 and Supplementary Data 2, respectively. In addition, a standalone LEV program that can be used to generate LEVs can be free downloaded from: http://www.mcu.edu.tw/department/biotec/en%5Fpage/LEV/, and thus researchers can examine the effect of removing the genes with LEVs higher than a set threshold on subsequent analyses.

Gene filtering, randomization analysis and pathway analysis

Dataset GSE10797 was downloaded from the GEO (24). The Jaccard similarity coefficient measures the similarity and diversity of sample sets, and it is defined as the size of the intersection divided by the size of the union of the sample sets. The randomization procedure for filtering-out genes and the Jaccard coefficient were calculated using MATLAB Version 7.4 and Bioinformatics Toolbox Version 2.5. Pathway analyses of differentially expressed genes were carried out using MetaCore Analytical Suite (GeneGo Inc., St Joseph, MI, http://www.genego.com) (4,25). MetaCore is a web-based computational platform designed for systems biology and drug discovery. It includes a curated database of human protein interactions and metabolism; thus, it is useful for analyzing a cluster of genes in the context of regulatory networks and signaling pathways.

Multidimensional scaling (MDS) analysis

For each data-set, 50% of the total genes studied were filtered out because of their low GEVs or LEVs variance. Then, a two-dimensional MDS (26) was applied to explore the interspecimen and interlaboratory variation structure. With Matlab Statistics Toolbox Version 6.0, the ALSCAL algorithm developed by Takane *et al.* (27) was used to perform nonmetric MDS analyses with interarray Pearson correlation as input proximity matrices. Nonmetric MDS was chosen since the correlation measurement does not possess metric properties. The ALSCAL algorithm first converted the input correlation proximities into distances using the following transformation,

$$\delta_{rs} < \delta_{uv} \Rightarrow f(\delta_{rs}) < f(\delta_{uv}) \text{ for all } 1 \leq r, s, u, v \leq n,$$

where $(\delta_{rs}, \delta_{uv})$ are two original input proximities (correlations), and $(f(\delta_{rs}), f(\delta_{uv}))$ are transformed distances. Thus, the transformation in nonmetric MDS represents only the ordinal properties of the data. In our case study, *f* transforms the correlation coefficients into distance measurements with the same order as their original ranks in the correlation matrix.

Reposition of the probes on HG-U95 chip

The sequence-verified probe position information of HG-U95 chip was obtained from http://lungtranscriptome.bwh.harvard.edu/pseqdatabase.html. After we performed sequence matching analyses, the positions of 4404 probe sets were verified from 12250 probe sets annotated by Affymetrix.

Reverse transcription of eukaryotic RNA in HG-U133A chips using random hexamer

To mimic cDNA synthesis using random hexamer primers in Affymetrix prokaryotic RNA reverse transcription, the RNA specimens extracted from four human cancer cell lines (MDAH2774, PK, PN and TOV112D) were reverse transcribed to cDNA using random hexamer primers. The cDNA products were then fragmented by DNase I and labeled with terminal transferase and biotinylated GeneChip® DNA Labeling Reagent at the 3′ ends. The labeled cDNAs were subsequently hybridized to HG-U133A. RNA labeling and Affymetrix array hybridization were performed according to the manufacturer's protocol.

RESULTS

Distribution of the LEV for different probe positions with distinct probe intensities in various types of microarrays

The application of LEV was demonstrated using a Human HG-U133_Plus_2 chip: chip #GSM147099 in the series #GSE6400 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE6400). LEVs were calculated to be 0.34, 0.15, 0.52, –0.52, 0.21 and 0.39 in actin, beta (ACTB), aldolase A, fructose-bisphosphate (ALDOA), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), phosphoglycerate kinase 1 (PGK1), lactate dehydrogenase A (LDHA) and ribosomal protein S27a (RPS27A), respectively (Figure 1). These genes are the six most commonly used reference housekeeping genes in real-time QPCR (28). LEVs in the Affymetrix chips may be affected by the design of probes in each probe set, including nucleotide sequence or GC content of the probes. Within the same chip, among the 54675 probe sets, 43346 probe sets had LEV information and the LEVs ranged from –1 to + 1, showing a normal distribution with a minor positive skew with a median of 0.074 (Figure 2).

LEV is highly correlated with GEV in eukaryotes but not in prokaryotes

Since gene expression level, RNA quality, and labeling efficiency may also affect LEV, correlations between LEVs and GEVs in Affymetrix GeneChips were analyzed. Taking chip #GSM147099 as an example, the Spearman's rank correlation coefficients (Spearman's rho = 0.23) between GEVs and LEVs are shown in a scatter plot (Figure 3A). Spearman's rank correlation coefficient (the Kendall's tau correlation coefficient gives similar results) was used because of the nonlinear relationship observed (see Figure 3 for an example). Furthermore, all Spearman's rho of the 12 chips in series #GSE6400 are shown in box plots (Figure 3B). LEVs were correlated with RNA intensities in GSE6400 with a median Spearman's rho of 0.26 (all of *P*-values ~0) (Figure 3B). The GEO data sets consisted of 152 U133A chips (from nine different laboratories) showed that LEVs were correlated with GEV in eukaryotes in HG-U133_Plus_2 chips with a median Spearman's rho of 0.27 (all of *P*-values ~0) (Figure 3B).

In addition to Human HG-U133 Plus 2 chips, the aforementioned correlations exist in other eukaryotic chips such as ATH1-121501 (Arabidopsis), *Caenorhabditis elegans*, Drosophila 2, HG-U133A 2 (Human), HG-U133A (Human), RAE230A (Rat), Rat230-2, Rice, Soybean, *Xenopus laevis*, Yeast-2 and zebra fish chips. In contrast, the significant correlation between LEV and GEV was not found in prokaryotic cells, such as *Staphylococcus aureus*, Pae G1a (*Pseudomonas aeruginosa*), *Escherichia coli* 2 and *E. coli* ASv2. As an example, the scatter plots between GEV and LEV of two eukaryotes (HG-U133-2 and RAT230-2) and two prokaryotes (Pae G1a and *E. coli* ASv2) chips are shown in Figure 4. The box plots of correlation coefficients between GEVs and LEVs in all of the analyzed data sets are shown in Figure 5, with the amount of chips indicated in parentheses. These results indicated the LEVs are significantly correlated with GEVs in eukaryotes, but not in prokaryotes.

Our analyses detected low correlation coefficients between LEVs and GEVs in three data sets derived from DrosGenome1 chips, YG-S98 and HG U95Av2. These three old chips (indexed 29-Jan-2002 in GEO) may have mis-annotations in probe positions (29). After the probe sets of the HG U95Av2 chips were repositioned and the database was reconstructed using sequence-matched probes, higher correlation coefficients between GEVs and LEVs were found (Figure 5, red arrow). Similarly, correlation of LEVs and GEVs in DrosGenome2 and Yeast 2 (Figure 5, blue arrows) were increased when compared with the previous versions. Furthermore, when reverse transcriptions using random hexamer for eukaryotic RNA were tested, the correlations between GEV and LEV were dramatically decreased (Figure 5, green arrow).

LEVs differentiate the results obtained from different laboratories on the same specimens

We analyzed LEVs of the results of replicated specimens (GEO accession GSE2004 Dataset) that are U133A chips data generated by two different laboratories. By using GEVs, each of four tissue samples (normal kidney, liver, spleen and Stratagene Universal RNA) were well separated from other three RNA samples (Figure 6A, left panel). Strikingly, analysis of LEVs was able to differentiate the results on replicated specimens obtained by TGen laboratory and Children's National Medical Center in Washington, DC (Figure 6A, marked in red versus blue in the right panel, respectively). Based on these analyses, we propose that LEVs in highly variable genes can represent the 'laboratory signature' of different laboratories. The stress scores (badness measurement of MDS model fitting) of MDS analyses for GEV or LEV were 0.073 and 0.15, respectively, both were well below the commonly acceptable threshold of 0.2. The same results were obtained when three more MAQC data sets were

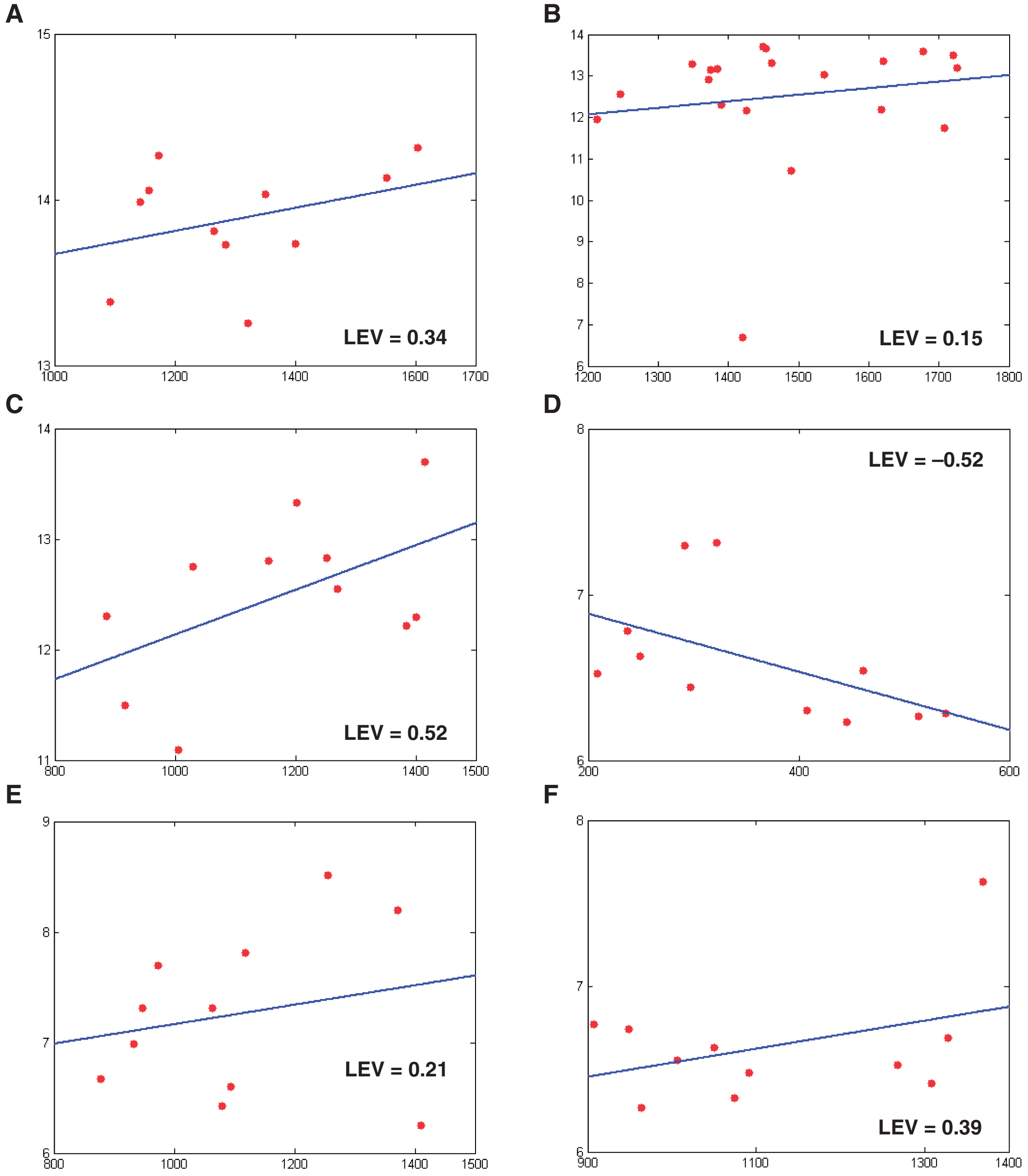

Figure 1. RNA degradation plots of six housekeeping genes. The horizontal axis indicates the probe position, and the vertical axis indicates the value of probe intensity in the base 2 logarithm. (**A**) GAPDH (212581_x_at), (**B**) ACTB (AFFX-HSAC07/X00351_3_at), (**C**) ALDOA (200966_x_at), (**D**) PGK1 (244597_at), (**E**) LDHA (206894_at) and (**F**) RPS27A in Human HG-U133_Plus_2 chip GSM147099 (of GSE6400). The labeling extension values (LEVs) are defined as the Pearson's linear correlation coefficient of the base 2 logarithm of probe intensity and probe position. Values of LEV are indicated.

analyzed by LEVs, supporting the notion that LEVs can be used as 'laboratory signatures' to classify data sets from different laboratories (Figure 6B–D). As shown in Figure 6C and D, the difference between tissue types was greater than the difference found in different laboratory sites in both LEVs (right panel) and GEVs (left panel). Nevertheless, LEVs could clearly differentiate the results obtained from different laboratories (right panels of Figure 6C and D).

Filtering out the genes with highly differential LEV improves the comparability between different laboratories

To test whether elimination of the genes with highly differential LEV could decrease the variations and improve comparability of data sets derived from different laboratories, we analyzed the correlation coefficients of fold change between two laboratories that were assayed on the same specimens. For example, fold changes between Aristolochic acid liver and Comfrey liver in the MAQC Rat Toxicogenomics Dataset were calculated separately in two laboratories. The Pearson correlation coefficient of fold change between these two laboratories was 0.785 for 22213 genes. When filtering out 5244 of the highly differential LEV genes (genes having *t*-test *P*-value <0.00001), the correlation coefficient was increased 0.03 to become 0.815 for 16969 genes (indicated by a red arrow in Figure 7). To examine whether the increment was statistically significant, we compared the removal of 5244 high-LEV genes with a random deletion of 5244 genes. This permutation process of a random deletion was repeated 1 000 000 times, and Pearson correlation coefficient of fold change was calculated each time. The distribution of the increment of interlaboratory correlation by random deletion of 5244 genes were mean = −0.00016 and median = 0.00028 (Min = –0.044, Q1 = −0.0040, Q3 = 0.0041, Max = 0.019, SD = 0.0060). The difference between the removal of the 5224 genes with high LEVs (improvement of interlaboratory correlation of 0.03) and random deletion of 5224 genes (mean increment of interlaboratory correlation of –0.00016) was highly significant ($P < 10^{-6}$).

We then carried out the same procedure for all 112 comparisons in the four MAQC data sets (Supplementary Data 3). The original correlations of comparisons were generally good (mostly >0.9), hence filtering out the genes with highly differential LEV only resulted in modest improvements of the interlaboratory correlations (Figure 7). Nevertheless, with a paired Student *t*-test, the increment of the 112 comparisons was highly significant ($P < 6 \times 10^{-7}$). These results strongly support the notion that elimination of genes with highly differential LEV can improve interlaboratory comparability of microarray data.

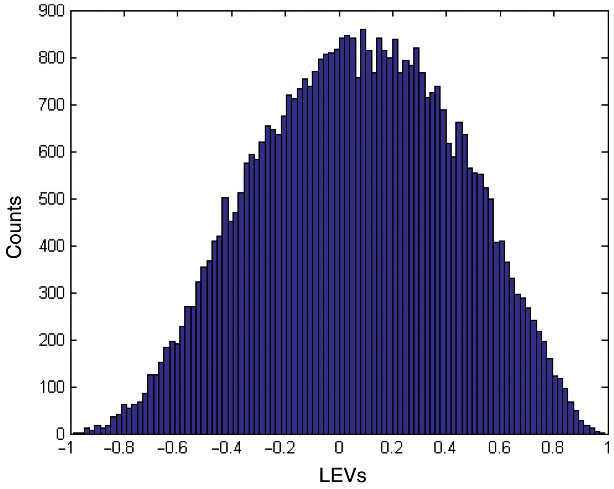

Figure 2. Histogram of the labeling extension values (LEVs) of 43346 probe sets in Human HG-U133 Plus2 chip #GSM147099 (of series #GSE6400), the median and mean of LEVs were 0.074 and 0.068, respectively.

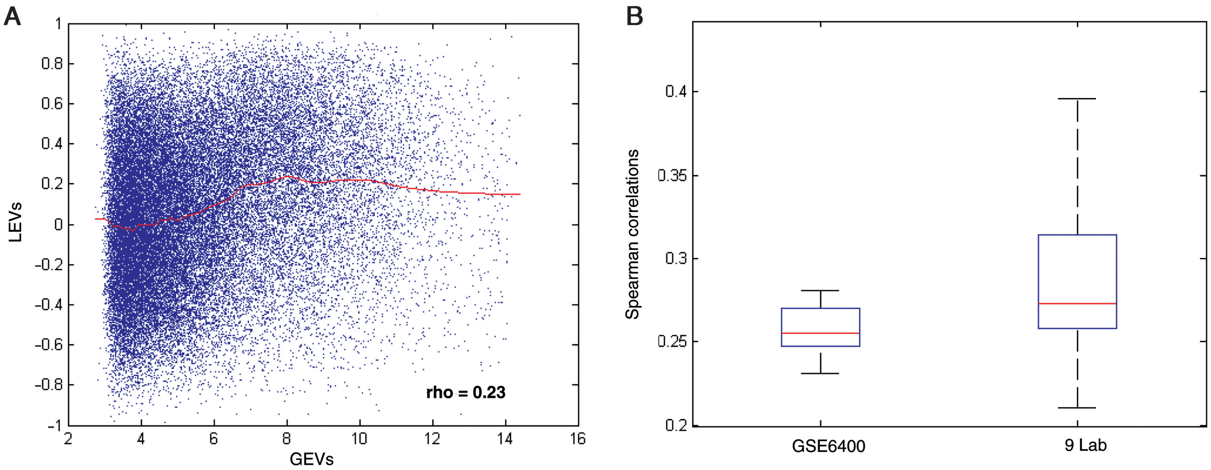

Figure 3. (**A**) Scatter plot of the gene expression levels (GEVs) and labeling extension values (LEVs) in Human HG-U133 Plus2 chip #GSM147099 (of series #GSE6400) with the Spearman's rho correlation coefficient at 0.23. The red line represents the Lowess smooth curve of GEVs and LEVs. (**B**) Box plots of the Spearman's rho correlation coefficients between LEVs and GEVs are shown as medians at 0.25 and 0.27 for 12 chips (in series #GSE6400) and 152 chips from nine laboratories, respectively. Main body of a box-plot stands for first quantile (Q1), median (M) and third quantile (Q3) while two whiskers represent Q1−1.5(Q3-Q1) and Q3 + 1.5(Q3-Q1), respectively.

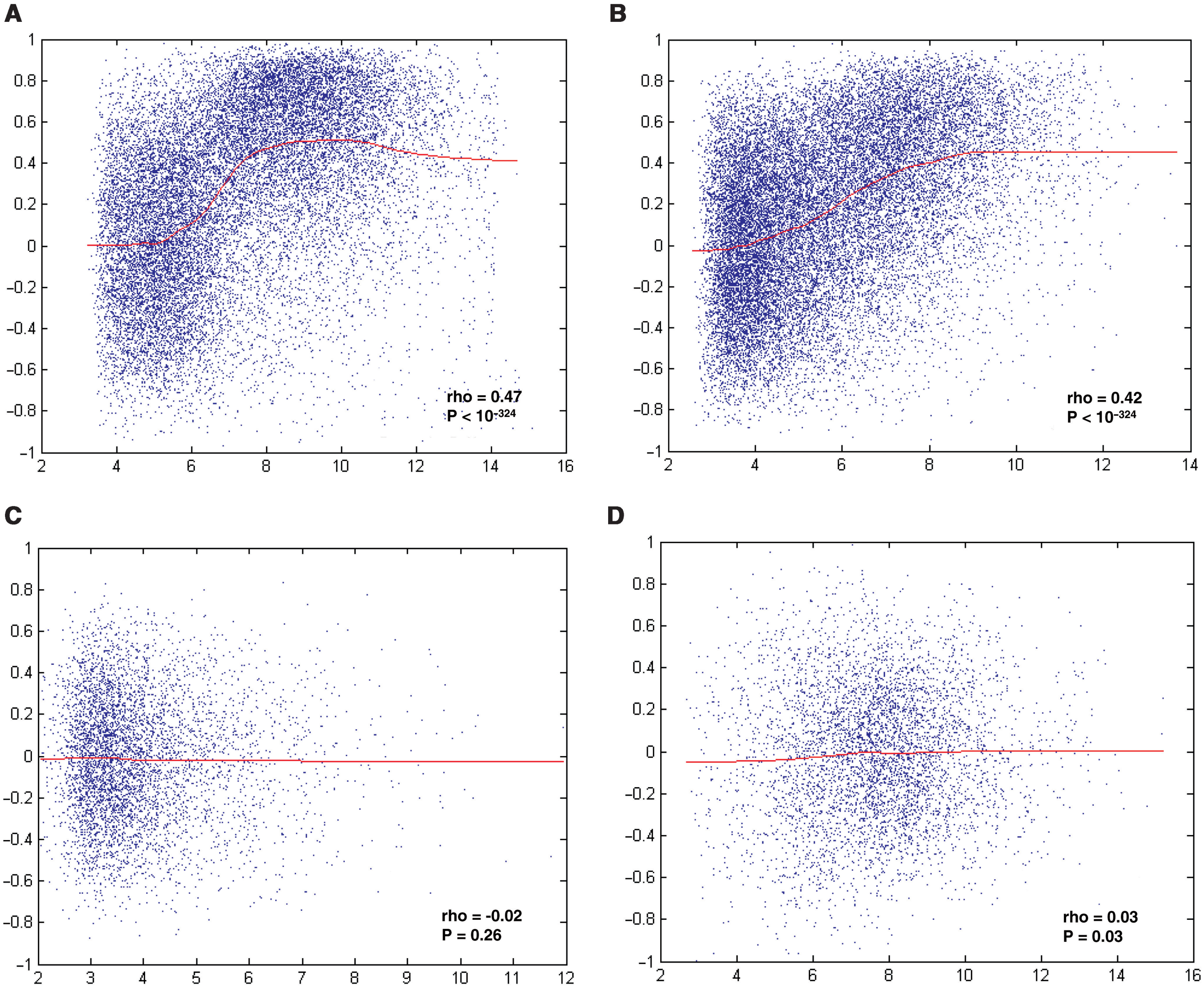

Figure 4. Scatter plots of the gene expression levels (GEVs, in horizontal axis) and labeling extension values (LEVs, in vertical axis) in (**A**) HG-U133-2 chip #GSM52623 (of the series #GSE2723), (**B**) Rat230-2 chip #GSM53 GSE287099 (of the series #GSE2870), (**C**) Pae G1a chip #GSM92164 (of the series #GSE4026) and (**D**) *E. coli* ASv2 chip #GSM18235 (of the series #GSE1121). The red line represents the Lowess smooth curve of GEVs and LEVs. The Spearman's rho correlation coefficient between GEVs and LEVs are also indicated. Correlations between GEVs and LEVs were highly significant ($P \sim 0$) in eukaryotic specimens (A, B), but less significant in prokaryotic ones (C, D).

Filtering out the genes with highly variable LEV improves homogeneity of gene expression profiles within the same class of specimens

To test whether removal of the genes with highly variable LEV could improve the comparability among subjects, we re-analyzed the data set GSE10797, in which RNA samples were prepared from breast cancer epithelium (CE) and cancer stroma (CS) of 28 subjects after tissues were isolated by laser capture microdissection (24). For each subject, differential expression of genes was shown as $\log_2$ (*CE*/*CS*). To measure the intersubject similarity, Pearson correlation coefficient was calculated for every pair of subjects across all of $\log_2$ (*CE*/*CS*)'s. This intersubject correlation was computed for all of 378 ($= C_2^{28} = 28^*27/2$) possible pairs of subjects, and the averaged correlation coefficient of 378 pairs was 0.137 for the complete set of 18 729 genes. The genes with highly variable LEV were defined as those genes with top-ranked LEVs in the CE set or the CS set. After filtering out top quarter, top two-quarters and top three-quarters of the genes with highly differential LEV, the averaged correlation coefficients were increased to 0.151, 0.168 and 0.187, respectively (Figure 8A). To test for a statistical significance, permutation processes of random deletion of the same number genes for each proportion (1/4, 1/2 and 3/4) were done for 10 000 times. The corresponding observed increments of 0.013, 0.031 and 0.050 were significant (all $P < 0.00001$) for filtering out portions of 1/4, 1/2 and 3/4, respectively.

We also calculated Jaccard coefficients of top 300 $\log_2$ (*CE*/*CS*) genes to test whether the removal of genes with high LEVs could be used as a filtering procedure to

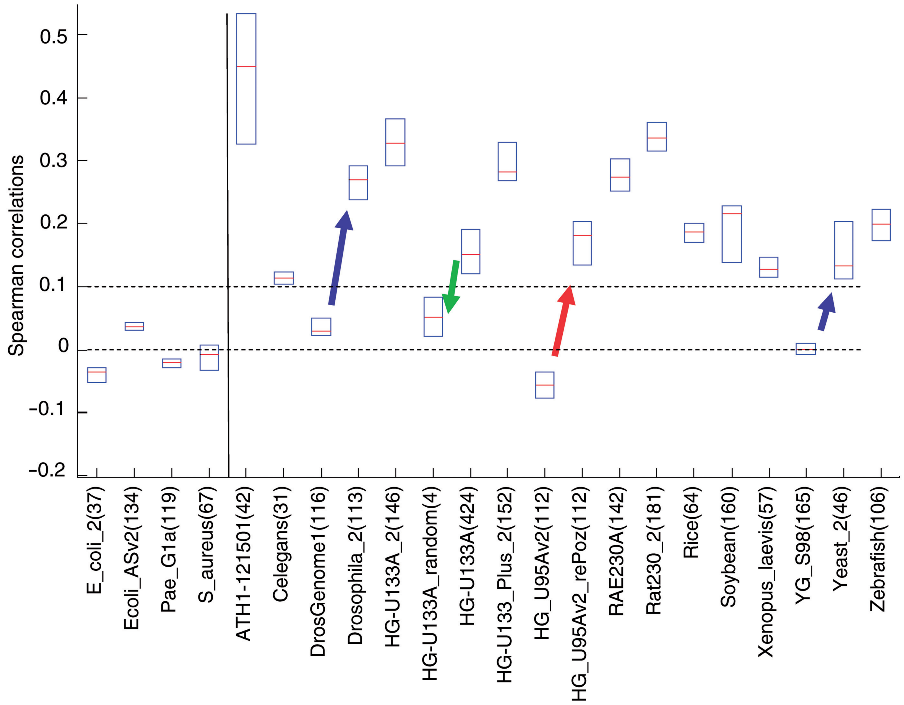

Figure 5. Box plots of the Spearman's rho correlation coefficients between gene expression levels (GEVs) and labeling extension values (LEVs) in three species (four chip types) of prokaryotes (left panel) and 10 species (16 chip types) of eukaryotes (right panel). The total number of chips analyzed for each chip type is indicated in parentheses. Detailed chip information is available in Supplementary Data 2. The red arrow indicates the increment of correlation coefficients after probe positions were re-annotated in the HG-U95Av2 chips. Blue arrows indicate the increase of correlation coefficients when probe positions were corrected from old to new versions in the Drosophila chips (DrosGenome1 to Drosophila 2) and in yeast chips (YG-S98 to yeast2). The green arrow indicates the decrease of correlation coefficients when the random primers were used to replace the oligo(T) primer during cDNA synthesis in four eukaryotic RNAs. The blue box represents the lower quartile and upper quartile values, and the red line represents median. Only boxes (without whiskers) with Q1, median and Q3 are plotted for easier comparison between chip types.

improve the comparability between any pairs of subjects within the same laboratory. The averaged Jaccard coefficient of 378 pairs of subjects was 3.9% for the complete data set of 18 729 genes. After filtering out the top 1/4, 1/2, and 3/4 portions of the genes with highly differential LEV, the averaged Jaccard coefficients were increased to 5.4%, 8.1% and 13.8%, respectively (Figure 8B). These increments were statistically significant (all $P < 0.00001$) when compared with those derived from permutation processes. These results indicate that filtering out high-LEV genes improves intersubject comparability, increasing the homogeneity of gene expression profiles within the same class of specimens.

To examine whether filtering out the high-LEV genes would reduce the number of differentially expressed genes and affect subsequent functional analyses of biological functions, we compared the main biological processes exerted by the differentially expressed genes before and after filtering out the genes with highly variable LEV. In the GSE10797 data set, Casey and associates found oxidative phosphorylation to be highly expressed in CE by GeneSifter^{TM} analyses (24). In our analyses, among the 1036 gene that were 2-fold upregulated in CE, 220 genes remained after filtered out three-quarters of genes with highly variable LEV. We are delighted to find oxidative phosphorylation to be the top-ranked pathway in functional network analysis of these 220 genes using Metacore algorithm (4). Similarly, Casey *et al.* (24) found ECM remodeling and cell adhesion to be the top biological processes as the CS signatures in the GSE10797 data set. After we filtered out the top 3/4 genes with high LEV, ECM modeling and cell adhesion remained as the most important biological processes. Our results show that filtering out genes with high LEVs did not compromise subsequent functional analyses of biological processes in this tested data set.

DISCUSSION

Affymetrix GeneChips provide the opportunity to unravel changes in gene expression profiles under a myriad of physiological, pathological and pharmacological conditions. Collaborative groups of the MAQC project have systematically analyzed several replicated specimens and revealed promising results regarding the consistency of microarray data between laboratories and across platforms (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/). In this study, all of their Affymetrix data sets were downloaded from GEO of NCBI, and RNA expression profiles were rebuilt from the original raw data of GeneChips using RMA or dChip (30). From analyses of these replicated data sets and additional data from more than 2000 chips, we have demonstrated that LEVs are significantly correlated with

GEVs in eukaryotes, but not in prokaryotes. To our surprise, results of the same specimens from different laboratories could be clearly differentiated by their LEVs (Figure 6A–D, right panels), although the RNA expression profiles shown by GEVs were quite consistent (Figure 6A–D, left panels).

Based on current mRNA degradation models, most mRNAs in eukaryotes undergo decay by the deadenylation-dependent pathway. In the first step, the poly(A) tail has to be removed by a deadenylase activity, followed by two mechanisms that degrade the mRNA: either decapping followed by 5′→3′ decay or a 3′→5′ decay (31). Once the mRNA poly(A) tail is removed, reverse transcription reaction will not proceed, resulting in no detectable signal on the GeneChips. Several studies have addressed the impact of RNA degradation on gene expression profiles and developed models to improve the reliability and efficiency of microarray data (14–17). It is suggested in the Affymetrix website that data with deviated 3′/5′ ratios may reflect poor quality of input mRNA (32). In the RNA degradation plots of several housekeeping genes such as β-actin and GAPDH, higher 3′/5′ ratios are considered to be the result of the following conditions: RNA degradation, incomplete conversion to the first stranded cDNA, or low labeling efficiency (16). Ryan *et al.* (18) proposed that, within a particular study, the outlying chips that

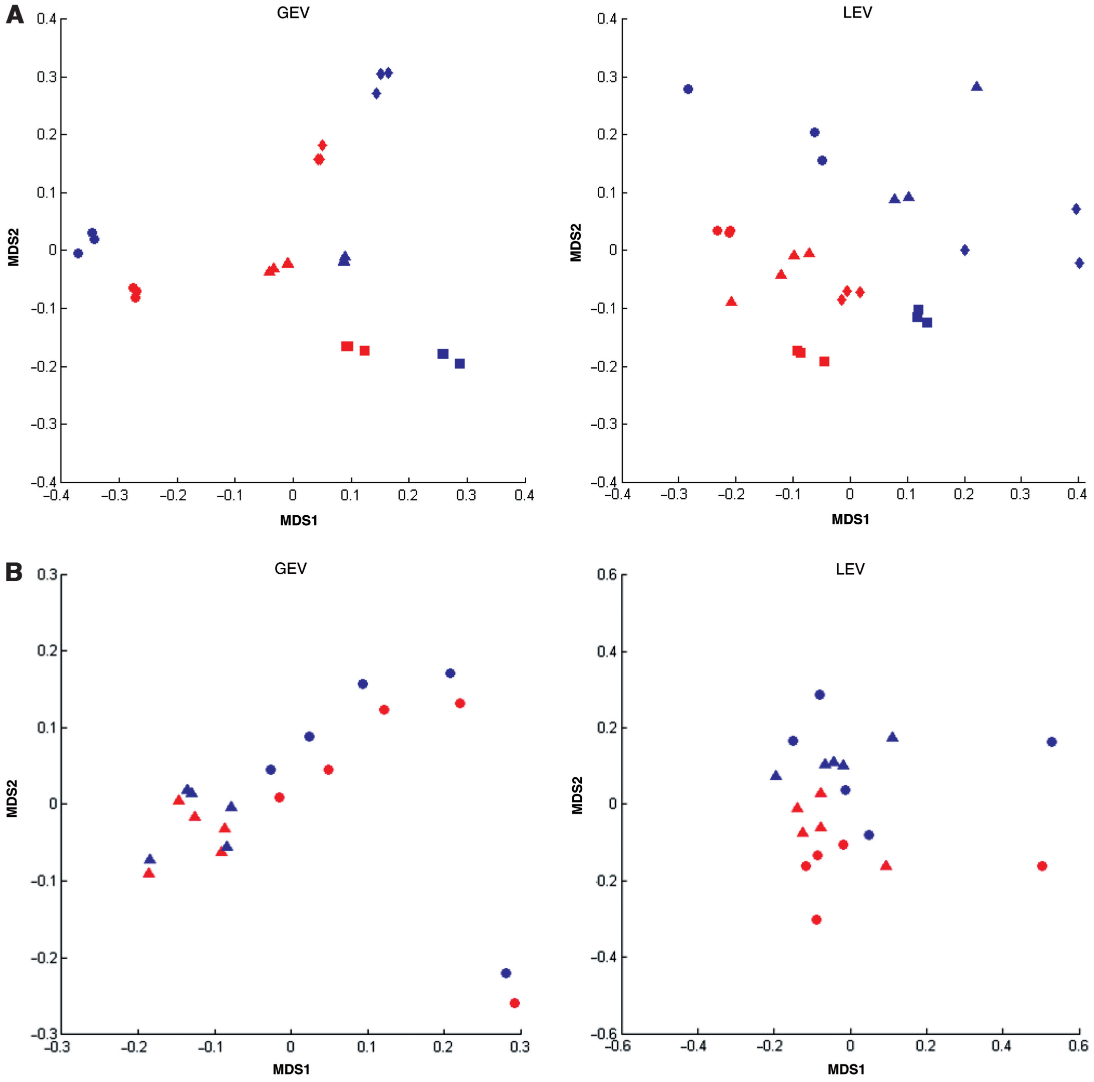

Figure 6. Multidimensional scaling (MDS) analyses of gene expression values (GEVs, left panel) and labeling extension values (LEVs, right panel). (**A**) GSE2004 Data set (triangle: kidney, circle: liver, square: spleen, diamond: Universal RNA). Two laboratory sites are labeled with different colors (red and blue). (**B**) MAQC Tumor Dataset (triangle: normal colonic tissues, circle: colorectal carcinomas). Two laboratory sites are labeled with different colors (red and blue). (**C**) MAQC Rat Toxicogenomics Dataset (triangle: aristolochic acid-treated kidney, circle: control kidney, square: aristolochic acid-treated liver, diamond: comfrey-treated liver, plus: control liver, star: riddelliine-treated liver). Two laboratory sites are labeled with different colors (red and blue). Inserts stand for zoomed-in views where between-laboratory effect can be observed as a secondary structure. (**D**) MAQC Brain Dataset (triangle: SUHRR, circle: HBRR, square: 25% HBRR: 75% SUHRR, diamond: 75% HBRR: 25% SUHRR).The four RNA samples were Stratagene's Universal Human Reference RNA (SUHRR), Ambion's Human Brain Reference RNA (HBRR), 25% Brain/75% SUHRR, 75% Brain/25% SUHRR. Three laboratory sites are labeled with different colors (red, blue and green). Inserts stand for zoomed-in views where interlaboratory differences can be observed as a secondary structure.

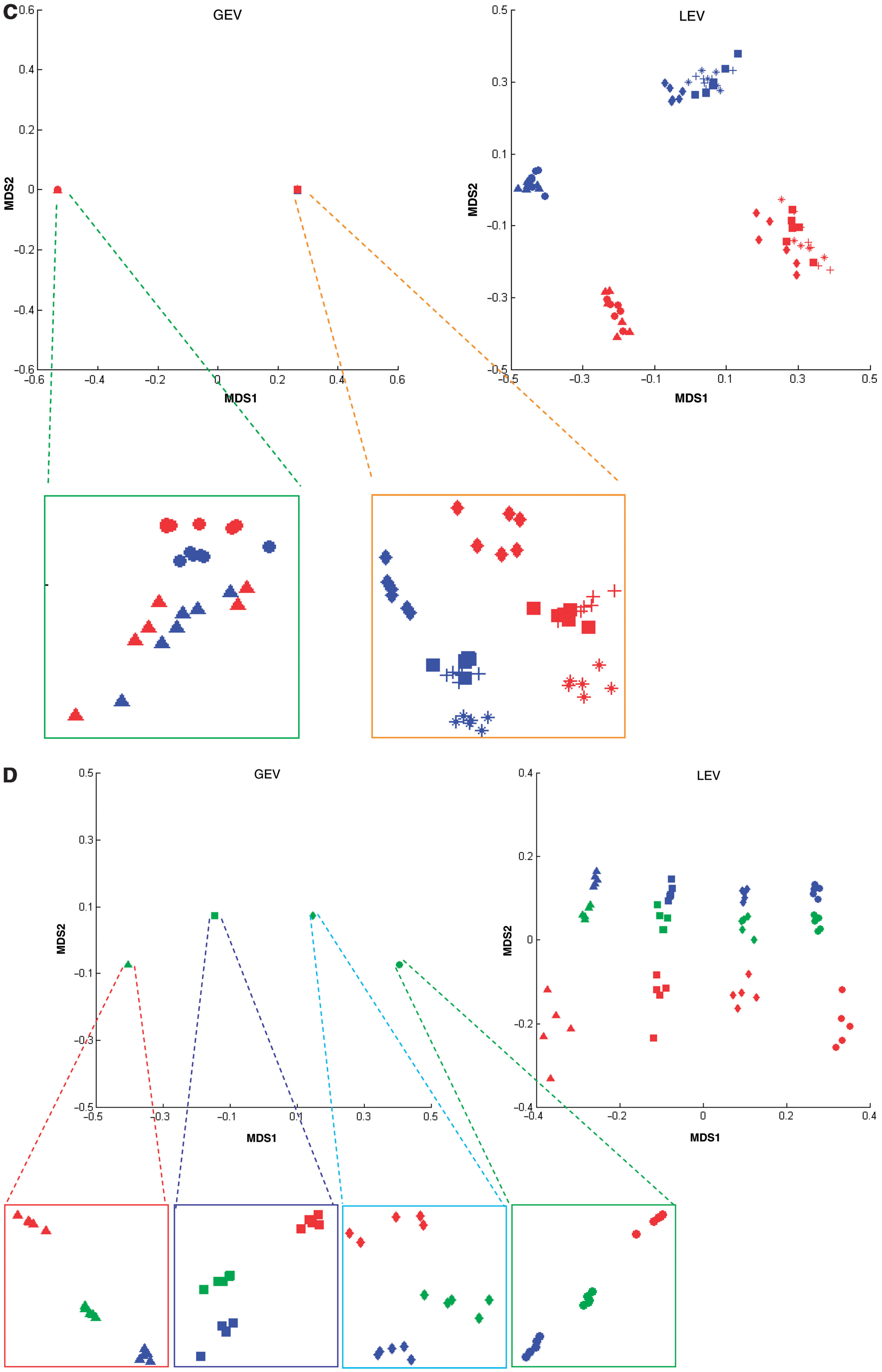

Figure 6. Continued.

were detected by RNA degradation plots were considered 'flagged chips' and should be excluded from further analysis. Penland *et al.* (33) lso suggested the transcript 3′/5′ ratio as an indicator of RNA quality control. In addition to paying attention to the quality of input RNA, our results further emphasize the importance of analyzing labeling efficiency in microarray experiments, which can be readily done by analyzing LEVs.

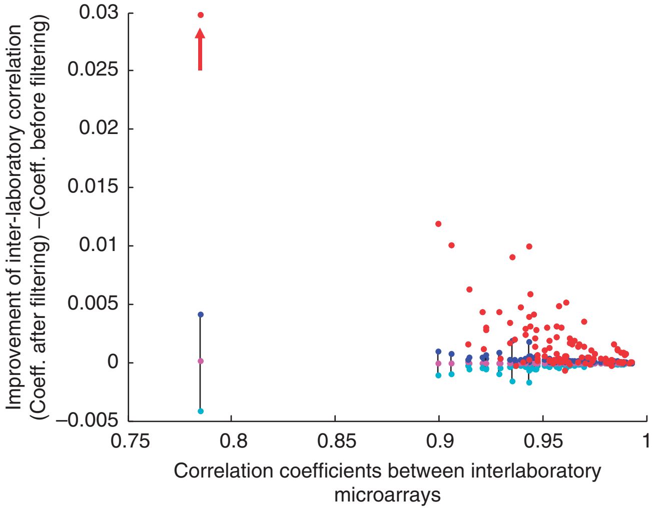

Figure 7. Filtering out the genes with highly differential LEV improves the interlaboratory comparability. Improvement of interlaboratory correlations was greater in the genes with poorer correlations, as demonstrated by the red dot and indicated by the red arrow. The horizontal-axis indicates the correlation coefficients of fold change ranking between two samples for all of possible comparisons. The vertical-axis indicates the changes of corresponding correlation coefficients after filtering out the highly differential LEV genes from the results by different laboratories. Detailed comparison information is available at Supplementary Data 3. Each line describes the distribution of the increments of interlaboratory correlation calculated from a permutation process of repeatedly random deletions of 5244 genes for 1 000 000 times. Q1s, medians, and Q3s are indicated as cyan, magenta and blue dots, respectively.

In Affymetrix genechips, prokaryotic reverse transcription is done by using random primers since the poly(A) tail normally used for eukaryotic reverse transcription is lacking. Our results show that prokaryotic mRNA as well as four test eukaryotic mRNAs that were intentionally processed with random hexamers have extremely low LEVs (Figure 5). These results support our notion that LEVs are results of uneven reverse transcription from the poly(A) tail in highly expressed genes (Figure 9, left panel).

As illustrated in Figure 9, the intensities of reverse transcribed cDNA that are extended from the oligo-dT primer vary with the amount of mRNA in the specimens, resulting in the corresponding labeled cRNAs of different lengths. Abundant mRNA increases the cRNA intensity of each probe from the 3′ to 5′ ends of the mRNA, resulting in high GEVs. Depending on the starting conditions, such as the activity of reverse transcriptase, the concentration and purity of deoxynucleotides and the structure of mRNA, the synthesis processes are not equal among different laboratories but may remain fairly consistent within each laboratory, resulting in a unique laboratory signature for each laboratory. Demonstrated in Figure 6, LEVs can clearly reveal the laboratory signature.

In addition to the classification of gene expression profiles made by GEVs (Figure 6A–D, left panels), the use of LEVs could clearly divide the four replicated data sets according to the laboratories where microarray experiments were performed, leading to the identification of 'laboratory signatures' (Figure 6A–D, right panels). The Affymetrix GeneChips are designed to standardize the

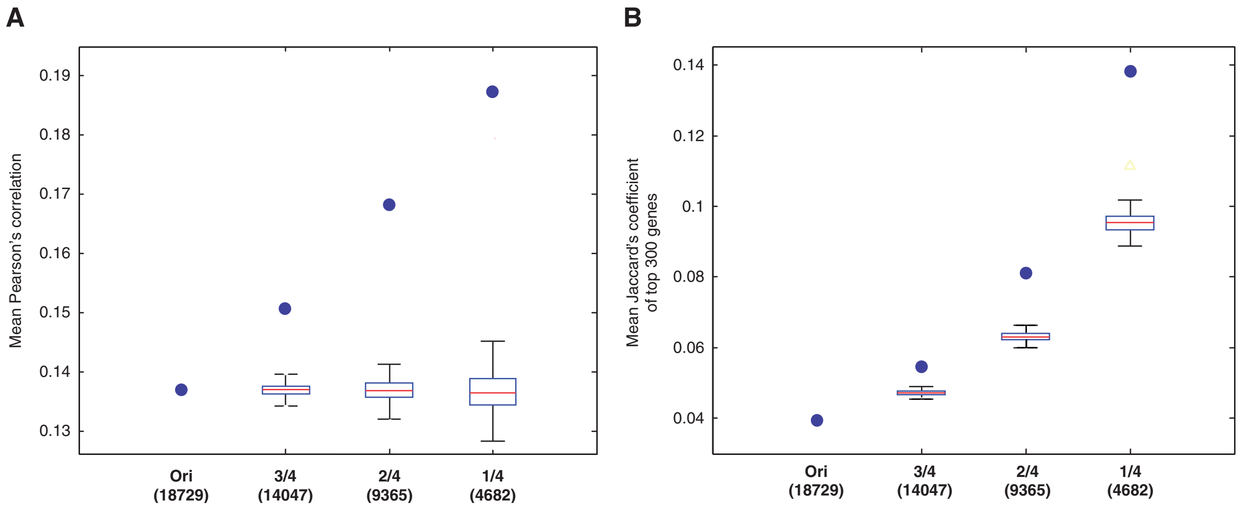

Figure 8. Filtering out the genes with highly differential LEVs improves the intersubject comparability between breast cancer epithelium (CE) and cancer stroma (CS) of 28 subjects. Improvement of inter-subject correlations was validated by (**A**) Pearson correlation coefficients and by (**B**) Jaccard coefficients of 300 top differentially expressed genes. Averaged intersubject Pearson correlation coefficients (blue dots in **A**) and Jaccard coefficients (blue dots in **B**) of all of 378 (= pairs of subjects are computed when none, 1/4, 1/2 or 3/4 portions of highly differential LEV genes were filtered out. The number of genes left for each portion of filtering is indicated in parentheses. Distributions for averaged Pearson coefficients (**A**) and averaged Jaccard coefficients (**B**) from a permutation of 100 000 random filtering of the same number of genes for each filtering portion are presented as box-whisker plots. In each plot, whiskers indicate the maximal value (upper) and minimal value (lower), box edges indicate Q3 (upper) and Q1 (lower), and the horizontal red bar indicates the median value.

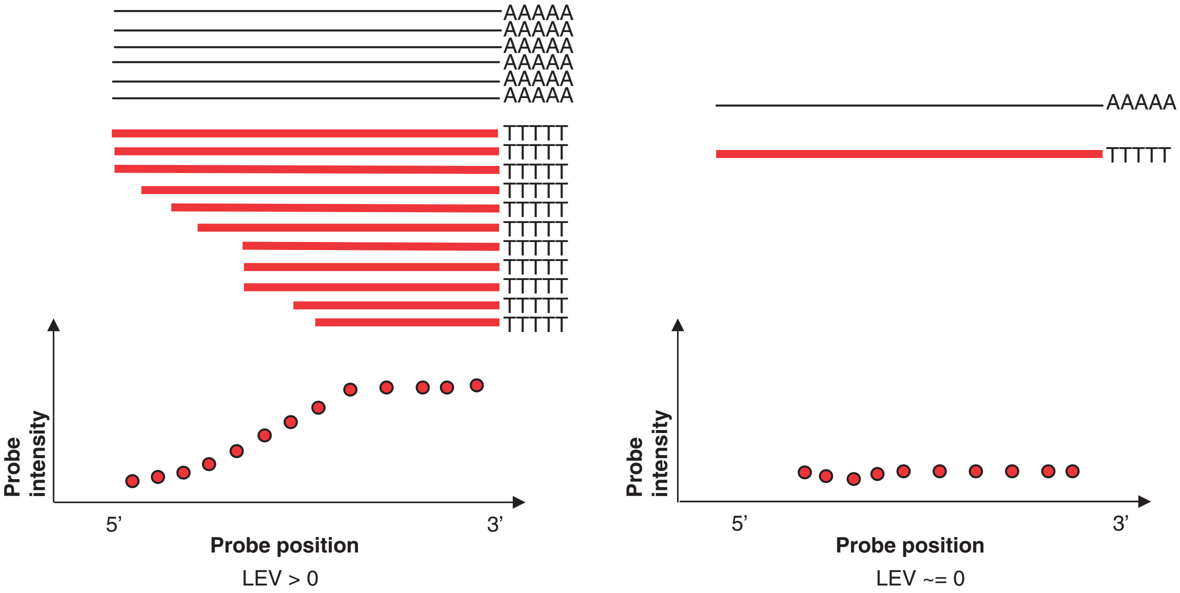

Figure 9. Proposed mechanisms for the positive correlation between labeling extension values (LEVs) and gene expression levels (GEVs) in eukaryotic RNAs. In the reverse transcription process, oligo dT primer is first annealed to the poly(A) tail of mRNA, and then cDNA synthesis proceeds. Although the reactions are generally considered to complete thoroughly, they may not be as even and efficient as previously thought, resulting in gradient intensities from high levels at 3′ end of mRNA to low levels at 5′ end of mRNA. The uneven intensities of probes within a probe set are more markedly in the genes with higher copy numbers (left panel) and less notable for those with low copy numbers (right panel). Therefore, in the genes with high GEVs (abundant copies), the gradient probe intensities, which are represented as LEVs, become positive and higher.

entire hybridization process including reagents, transcription process and labeling efficiency. Theoretically, it should derive the same results when performed in different laboratories. However, our results show that this goal has not been always achieved. Many of the steps of Affymetrix analyses of gene expression, from cDNA transcription, to cRNA labeling, to final hybridization, can introduce variability. Some types of systemic variation have been unknowingly added in each site, and these variations can be characterized by using LEVs, but not by GEVs. Our results indicated that filtering out the genes of highly deviated LEV improves the comparability among laboratories (Figure 7), improves comparability among subjects within a single laboratory (Figure 8A), and increases homogeneity of gene expression profiles within the same class of specimens (Figure 8B). Importantly to note, filtering out the genes with highly variable LEV may not interfere with functional analysis of biological process on differentially expressed genes.

CONCLUSIONS

The Affymetrix platform has been used to identify genes that are predictive of patient responses to treatment or distinguish differences between diseased and control groups (1–3). To minimize the nonconcordance of inter-laboratory measurement of Affymetrix GeneChips results, LEVs are shown to be useful in identifying ‘laboratory signatures’, which represent the systemic variations uniquely generated in each laboratory during the microarray procedures from probe labeling, hybridization, to signal detection. The use of LEVs as a filtering parameter also improves inter- and intralaboratory comparability of gene expression profiles, without compromising subsequent functional analyses of biological networks on them.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Ueng-Cheng Yang (National Yang-Ming University), Dr. Hwa-Chien Chen and Dr. Shu-Jen Chen (Chang Gung University) for useful comments, Fong-Yee Chiu for technical assistance, and Shih-Yee Mimi Wang (University of Illinois College of Medicine) for English editing.

FUNDING

NSC 95-2311-B-130-001 (Y.S.L.), NSC-96-3112-B-001-10 017 and NSC 97-3112-B-001-020 in the National Research Program for Genomic Medicine (to Y.S.L. and C.H.C.); NSC97-2314-B-182A-023-MY2 (to A.C.); and NSC-96-2314-B-182A-124 (to T.H.W.) from National Science Council, Taiwan; as well as CMRPG360951 (to A. C.) and CMRPG340463 (T.H.W.) from Chang Gung Memorial Hospital. Funding for open access charge: Chang Gung Memorial Hospital, Taiwan.

Conflict of interest statement. None declared.

REFERENCES

1. Chang,J.C., Wooten,E.C., Tsimelzon,A., Hilsenbeck,S.G., Gutierrez,M.C., Elledge,R., Mohsin,S., Osborne,C.K., Chamness,G.C., Allred,D.C. *et al.* (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet*, **362**, 362–369.
2. 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. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
3. Ramaswamy,S. (2004) Translating cancer genomics into clinical oncology. *N. Engl. J. Med.*, **350**, 1814–1816.
4. Tsai,M.S., Hwang,S.M., Chen,K.D., Lee,Y.S., Hsu,L.W., Chang,Y.J., Wang,C.N., Peng,H.H., Chang,Y.L., Chao,A.S. *et al.* (2007) Functional network analysis of the transcriptomes of mesenchymal stem cells derived from amniotic fluid, amniotic membrane, cord blood, and bone marrow. *Stem Cells*, **25**, 2511–2523.
5. Chao,A., Wang,T.H., Lee,Y.S., Hong,J.H., Tsai,C.N., Chen,C.K., Tsai,C.S., Chao,A.S. and Lai,C.H. (2008) Analysis of functional groups of differentially expressed genes in the peripheral blood of patients with cervical cancer undergoing concurrent chemoradiation treatment. *Radiat. Res.*, **169**, 76–86.
6. Wang,T.H. and Chao,A. (2007) Microarray analysis of gene expression of cancer to guide the use of chemotherapeutics. *Taiwan J. Obstet. Gynecol.*, **46**, 222–229.
7. Tan,P.K., Downey,T.J., Spitznagel,E.L. Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.
8. Marshall,E. (2004) Getting the noise out of gene arrays. *Science*, **306**, 630–631.
9. Thalamuthu,A., Mukhopadhyay,I., Zheng,X. and Tseng,G.C. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.
10. Shi,L., Reid,L.H., Jones,W.D., Shippy,R., Warrington,J.A., Baker,S.C., Collins,P.J., de Longueville,F., Kawasaki,E.S., Lee,K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.
11. Klebanov,L. and Yakovlev,A. (2007) How high is the level of technical noise in microarray data? *Biol. Direct.*, **2**, 9.
12. Stafford,P. and Brun,M. (2007) Three methods for optimization of cross-laboratory and cross-platform microarray expression data. *Nucleic Acids Res.*, **35**, e72.
13. Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Hughes,J.E., Snesrud,E., Lee,N. and Quackenbush,J. (2000) A concise guide to cDNA microarray analysis. *Biotechniques*, **29**, 548–550, 552–544, 556.
14. Auer,H., Lyianarachchi,S., Newsom,D., Klisovic,M.I., Marcucci,G. and Kornacker,K. (2003) Chipping away at the chip bias: RNA degradation in microarray analysis. *Nat. Genet.*, **35**, 292–293.
15. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
16. Archer,K.J., Dumur,C.I., Joel,S.E. and Ramakrishnan,V. (2006) Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models. *Biostatistics*, **7**, 198–212.
17. Heber,S. and Sick,B. (2006) Quality assessment of Affymetrix GeneChip data. *OMICS*, **10**, 358–368.
18. Ryan,M.M., Huffaker,S.J., Webster,M.J., Wayland,M., Freeman,T. and Bahn,S. (2004) Application and optimization of microarray technologies for human postmortem brain studies. *Biol. Psychiatry*, **55**, 329–336.
19. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
20. Guo,L., Lobenhofer,E.K., Wang,C., Shippy,R., Harris,S.C., Zhang,L., Mei,N., Chen,T., Herman,D., Goodsaid,F.M. *et al.* (2006) Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.*, **24**, 1162–1169.
21. Lin,G., He,X., Ji,H., Shi,L., Davis,R.W. and Zhong,S. (2006) Reproducibility Probability Score–incorporating measurement variability across laboratories for gene selection. *Nat. Biotechnol.*, **24**, 1476–1477.
22. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
23. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.
24. Casey,T., Bond,J., Tighe,S., Hunter,T., Lintault,L., Patel,O., Eneman,J., Crocker,A., White,J., Tessitore,J. *et al.* (2009) Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. *Breast Cancer Res. Treat.*, **114**, 47–62.
25. Nikolsky,Y., Ekins,S., Nikolskaya,T. and Bugrim,A. (2005) A novel method for generation of signature networks as biomarkers from complex high throughput data. *Toxicol. Lett.*, **158**, 20–29.
26. Chen,C.H. and Chen,J.A. (2000) Interactive diagnostic plots for multidimensional scaling with applications in psychosis disorder data analysis. *Stat. Sin.*, **10**, 665–691.
27. Takane,Y., Young,F.W. and De Leeuw,J. (1977) Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimals scaling features. *Psychometrika*, **42**, 7–67.
28. Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.*, **19**, 362–365.
29. Mecham,B.H., Klus,G.T., Strovel,J., Augustus,M., Byrne,D., Bozso,P., Wetmore,D.Z., Mariani,T.J., Kohane,I.S. and Szallasi,Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.*, **32**, e74.
30. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
31. Garneau,N.L., Wilusz,J. and Wilusz,C.J. (2007) The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.*, **8**, 113–126.
32. Lamendola,D.E., Duan,Z., Yusuf,R.Z. and Seiden,M.V. (2003) Molecular description of evolving paclitaxel resistance in the SKOV-3 human ovarian carcinoma cell line. *Cancer Res.*, **63**, 2200–2205.
33. Penland,S.K., Keku,T.O., Torrice,C., He,X., Krishnamurthy,J., Hoadley,K.A., Woosley,J.T., Thomas,N.E., Perou,C.M., Sandler,R.S. *et al.* (2007) RNA expression analysis of formalin-fixed paraffin-embedded tumors. *Lab. Invest.*, **87**, 383–391.