



## Limited agreement among three global gene expression methods highlights the requirement for non-global validation

Peter M. Haverty<sup>1</sup>, Li-Li Hsiao<sup>2</sup>, Steven R. Gullans<sup>2</sup>, Ulla Hansen<sup>1,3</sup> and Zhiping Weng<sup>1,4,\*</sup>

<sup>1</sup>Bioinformatics Program, <sup>2</sup>Departments of Medicine and Neurology, Brigham and Women's Hospital, Harvard Medical School, <sup>3</sup>Department of Biology and <sup>4</sup>Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

Received on December 14, 2003; revised on June 29, 2004; accepted on July 10, 2004

Advance Access publication July 15, 2004

### ABSTRACT

**Motivation:** DNA microarrays have revolutionized biological research, but their reliability and accuracy have not been extensively evaluated. Thorough testing of microarrays through comparison to dissimilar gene expression methods is necessary in order to determine their accuracy.

**Results:** We have systematically compared three global gene expression methods on all available histologically normal samples from five human organ types. The data included 25 Affymetrix high-density oligonucleotide array experiments, 23 expressed sequence tag based expression (EBE) experiments and 5 SAGE experiments. The reported gene-by-gene expression patterns showed a wide range of correlations between pairs of methods. This level of agreement was sufficient for accurate clustering of datasets from the same tissue and dissimilar methods, but highlights the need for thorough validation of individual gene expression measurements by alternate, non-global methods. Furthermore, analyses of mRNA abundance distributions indicate limitations in the EBE and SAGE methods at both high- and low-expression levels.

**Contact:** zhiping@bu.edu

### INTRODUCTION

The ability of DNA microarrays to measure global gene expression with a single experiment has led to a paradigm shift in medical and basic biological research. Recent examples include identification of significant gene expression patterns for differentiating between histologically similar leukemias and discovering novel leukemia subtypes (Golub *et al.*, 1999), prediction of the disease outcome of childhood cancers (Pomeroy *et al.*, 2002) and identification of candidate genes involved in medulloblastoma metastasis (MacDonald *et al.*, 2001). The application of microarrays is so broad that tens of variants of the technology have been developed, in general

falling into two categories: high-density oligonucleotide arrays (HDAs) (Lockhart *et al.*, 1996) and long cDNA probe arrays (Iyer *et al.*, 1999). HDAs are constructed with sets of short synthesized oligonucleotides (probes), each set being specific to a transcript of interest. Affymetrix HDAs also pair each of their 25 nt perfect match probes with a mismatch probe, which differs from the perfect match probe only in the central nucleotide. Since mismatch probes display significantly decreased affinity for the target transcript, they are intended to estimate non-specific binding or cross-hybridization (Lockhart *et al.*, 1996). This feature of Affymetrix HDAs is intended to measure gene transcription levels in an absolute and specific manner.

Despite the impact of HDAs, their results are generally considered to require validation before interpretation. This stems primarily from the potential for inaccurate cross-hybridization between probes and unintended transcripts. In addition, even with recent cost reductions, HDA analyses are rarely performed with sufficient numbers of control and replicate experiments. Since there are often vastly more variables than data points, the statistical significance of the conclusions drawn from such studies have been difficult to assess. Quantitative RT-PCR is a common method for validating microarray results (Chuaqui *et al.*, 2002), but it is too time and resource consuming to be used for large-scale validation. It has been proposed (Chuaqui *et al.*, 2002) that comparison to large-scale, sequencing-based expression methods may be useful for validation of microarray results. Here, we test this proposal by comparing HDA data to those from two alternative global methods: expressed sequence tag (EST) Based Expression analysis (EBE) and Serial Analysis of Gene Expression (SAGE).

EBE infers gene expression levels from the abundances of ESTs representing different genes in a cDNA library. If the library is prepared without normalization or subtraction (i.e. procedures that disturb the quantification of the

\*To whom correspondence should be addressed.

initial proportions of mRNAs), the resulting population of ESTs represents a snapshot of gene expression in an RNA sample. SAGE is based on the principle that unique short sequence tags (10 bp) can potentially distinguish all transcripts in a genome (Velculescu *et al.*, 1995). During SAGE analysis, such tags are concatamerized, cloned and then sequenced.

Detailed comparison of HDAs with such other gene expression techniques has been limited, although with the rapid development of microarray technology, more and more data are becoming publicly available. Among them are HDA (Butte *et al.*, 2000) and cDNA microarray (Ross *et al.*, 2000) gene expression experiments by separate laboratories on a standard panel of 60 cancer cell lines from the National Cancer Institute (NCI 60). A subsequent study reported minimal correlation between these two datasets (Kuo *et al.*, 2002), which draws attention to the need for an independent method to resolve discrepancies. In an even more direct comparison of results from different global gene expression profiling methods, two RNA samples, prepared in the same laboratory from monocytes and induced macrophages were analyzed by HDA and SAGE. The top 50 most highly expressed genes identified by the two techniques overlapped only ~50%. Similar levels of agreement were observed for the genes with the largest positive and negative differences in expression when monocyte gene expression levels were compared with those in macrophage (Ishii *et al.*, 2000). Thus, although a large proportion of the measurements by either method are likely to be correct, additional input is necessary to determine which method is correct in the remaining cases. Yuen *et al.* (2002) performed a three-way comparison of HDA and cDNA arrays with Quantitative real-time RT-PCR (QRT-PCR) to measure the expression level differences in 47 genes in five pairs of mRNA samples from a gonadotrope cell line treated with gonadotropin-releasing hormone or vehicle. Only three pairs of samples were tested by both HDA and QRT-PCR, and one pair was tested by all three techniques. Both array techniques were able to identify most of the genes that had previously been identified as being regulated under these experimental conditions; however, both array methods generally underestimated the magnitudes of mRNA abundance changes. Finally, although the data were available, the authors did not analyze the ability of HDAs to measure absolute expression levels (Yuen *et al.*, 2002).

In this study, we explored the relative abilities of three global gene expression methods, HDA, SAGE and EBE, to measure gene expression levels. This comparison allowed us to investigate the accuracy of individual methods and the possibility of combining global gene expression datasets from different methods to generate more reliable data. We use a much larger dataset than in previous studies in order to survey the agreement across multiple tissue types and laboratories. The EBE method identifies transcripts using relatively long sequences,

which can potentially provide higher accuracy. Therefore, EBE could potentially resolve discrepancies between HDA and SAGE, observed previously (Ishii *et al.*, 2000). We show that the agreement between the methods was highly variable from gene to gene. Although agreement is considerable for some genes, a large proportion of individual mRNA abundance measurements differ dramatically between methods. In addition, our analyses highlight specific limitations in the dynamic ranges of SAGE and EBE. These findings indicate the need for gene-by-gene validation of important global gene expression measurements using non-global methods and cast doubt on the feasibility of combining global gene expression methods as a means for increasing confidence in expression results.

## MATERIALS AND METHODS

### Data collection

Data were collected from three public sources. Affymetrix oligonucleotide microarray data were obtained from the Human Gene Expression Index (HugeIndex) database (Hsiao *et al.*, 2001; Haverty *et al.*, 2002) and re-processed using Affymetrix MAS 5.0 (Affymetrix, 2001 [www.affymetrix.com/support/technical/technotes/statistical\\_reference\\_guide.pdf](http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf)). We selected 25 HDA experiments from the HugeIndex representing five human organs also available in EBE and SAGE datasets (brain, kidney, liver, lung and muscle). The 25 experiments were selected from 35 available experiments with these organ types, as some of the HDA experiments in the HugeIndex have been shown to exhibit saturation (Hsiao *et al.*, 2002). We created scatter plots for all pairs of experiments within one organ type and selected only those experiments that did not exhibit saturation. Saturation was indicated by a consistent upper threshold in one experiment's expression values relative to other experiments from the same organ type. All HDA experiments were performed using Affymetrix GeneChip® Hu6800 oligonucleotide arrays, which probe for mRNA from 6939 genes. Although more current chip designs exist, these data are the most appropriate available given the unique nature of this large study of gene expression in normal human tissue. Redundancies in microarray gene to EBE/SAGE gene associations were resolved by removing 111 genes from the dataset that were represented by multiple probe sets on the Affymetrix HDA. This was done because the level of expression reported by these multiple probe sets, usually specific to different regions of the transcript, were not identical. The tissue samples were identified as histologically normal. Additional information about these data was published previously (Hsiao *et al.*, 2001).

In order to construct the EBE dataset, a parser was written to select all libraries from NCBI's *libs.Hs* file (downloaded 4/19/01) that met the following requirements for preserving

the proportions of transcripts from different genes: (1) non-normalized and (2) non-subtracted. Libraries were also required to meet the following conditions to be appropriately compared to HDA data: (1) normal tissue, (2) primary cells, (3) unambiguously stated organ type and (4) a library size of more than 500 ESTs. A total of 23 libraries were obtained. Five of these libraries were accompanied by annotation which stated that the first round of RT-PCR used an oligo (dT) primer. The initial priming method for the remaining libraries was unavailable.

In order to link ESTs with the genes they represent, we used the UniGene (Boguski and Schuler, 1995) resource from the NCBI. We parsed a text file, data.Hs (downloaded from the NCBI FTP server 4/17/01), representing these clusters in an automated fashion. For each cluster we obtained a UniGene ID, the name of the gene represented by the cluster and the GenBank accessions of the sequences included in the cluster. Genes from the HDAs were associated with specific clusters by identifying the UniGene cluster that included the GenBank accession number of each of the genes on the HDA. We unambiguously linked 4630 of the 6939 HDA genes to a UniGene cluster. EST libraries from the five organs used in this study contained 3389 of these genes. This set of 3389 genes was used for all comparisons in this study.

SAGE experiments were selected from the NCBI's SAGEmap resource (Lash *et al.*, 2000). All available SAGE experiments performed using normal human tissue from the brain, kidney, liver, lung and muscle were downloaded. Two experiments using muscle tissue and one experiment each for the other four organs were obtained. Of the two muscle experiments available (GSM819 and GSM824), GSM824 was chosen to represent muscle tissue, as the ages of the tissue donors more closely matched those of the muscle tissue donors from the HDA studies. The age of the EBE muscle tissue donor was unavailable. All SAGE tissue samples were identified as normal in the NCBI documentation.

### Calculating HDA, SAGE and EBE expression values

We calculated the EBE levels of each gene both within each library and for each organ type. Raw expression levels represent the number of ESTs corresponding to a certain gene within a particular library. Normalized expression levels in a library, in units of tags per million (tpm), were calculated as the raw expression level for a gene in a given library divided by the sum of all raw expression levels recorded for all genes in that library, multiplied by 1 000 000 (1). Normalized expression levels in an organ type, in tpm, were calculated as the sum of raw EBE expression levels for a given gene in all libraries of that organ type, divided by the total number of ESTs for all genes in libraries of that organ type and then multiplied by 1 000 000 (2). These normalized expression levels will be referred to as the EBE expression levels and overall organ EBE expression

levels, respectively.

$$EBE = \left( \frac{\text{No. of ESTs from Gene } i \text{ in Library } j}{\text{Total no. of ESTs in Library } j} \right) \times 1\,000\,000. \quad (1)$$

$$\text{Organ EBE} = \left( \frac{\text{No. of ESTs from Gene } i \text{ in Organ } k \text{ Libraries}}{\text{Total no. of ESTs in Organ } k \text{ Libraries}} \right) \times 1\,000\,000. \quad (2)$$

HDA expression levels were those reported by the GeneChip<sup>®</sup> Software (Affymetrix Microarray Suite 5.0<sup>®</sup>). This software scaled the raw intensities so that the mean expression level on the entire microarray was equal to 100, in order to allow comparison among multiple microarray datasets. We purposefully did not perform any normalization between data from separate platforms in order to avoid disturbing the integrity of the data with additional transformations. The use of correlation as a measure to compare experiments alleviates any potential problems arising from small differences in scale. Organ HDA expression levels were calculated as the average of HDA expression levels of a given gene in all experiments within a given organ.

SAGE expression levels, in tpm, were calculated for each of the five SAGE experiments. The data obtained from the NCBI consisted of raw counts of each of the 10 nt long SAGE tags for a particular experiment as well as the total number of tags obtained. SAGE tags were associated with GenBank accession numbers from the HDA and EBE datasets using build 29 of the NCBI SAGE tag to UniGene mappings (<http://www.ncbi.nlm.nih.gov/SAGE/index.cgi?cmd=mappings>). In cases where multiple options for this SAGE tag to UniGene cluster mapping were provided, we used the mapping listed as the most reliable, because the addition of the 108 additional genes identified by non-unique tags generally improved results to a small degree. Once SAGE tags were mapped to a UniGene cluster, these UniGene cluster identifiers were used to link each tag to a GenBank accession number in the same manner used to match genes from the HDA dataset to genes in the EBE dataset. As genes were often represented by multiple sequence tags, SAGE expression levels for each gene were calculated as the sum of all normalized counts for all tags corresponding to a given accession number. Normalized counts, in tpm, were calculated for each gene (3).

$$\text{Normalized SAGE expression} = \left( \frac{\text{No. of tags for gene}}{\text{Total no. of tags}} \right) \times 1\,000\,000. \quad (3)$$

### Clustering

Hierarchical clustering and *K*-means clustering were each performed using Matlab version 6.5 (<http://www.mathworks.com>). For the *K*-means clustering, the initial clusters were created by randomly selecting five of the experiments as initial

**Table 1.** Accuracy of  $K$ -means clustering<sup>a</sup>

| Data sources   | ARI <sup>b</sup> | Number of genes <sup>c</sup> |
|----------------|------------------|------------------------------|
| HDA, SAGE, EBE | 0.41             | 1603                         |
| HDA, SAGE      | 0.82             | 149                          |
| HDA, EBE       | 0.45             | 1598                         |
| SAGE, EBE      | 0.30             | 2538                         |
| HDA            | 1.00             | 146                          |
| EBE            | 0.35             | 2334                         |

<sup>a</sup>Accuracy of clusters obtained using different datasets individually and in groups. Accuracy was judged by comparing clusters to the ideal case where each cluster contains all experiments of one and only one organ type.

<sup>b</sup>Adjusted Rand Index, a measure of clustering accuracy.

<sup>c</sup>The number of the 3389 genes being studied that had a CV over 250%.

centroids. Each  $K$ -means clustering was repeated 100 times using a new random sampling of initial centroids in order to avoid local minima. We used average-linkage method for the hierarchical clustering. Correlation was chosen as the distance metric for both clustering methods. We calculated the Adjusted Rand Index for the  $K$ -means data using the formula in Yeung *et al.* (2001). In cases where SAGE or EBE did not detect the expression level of a gene, that gene was assigned an expression level of zero in the SAGE or EBE data. Before clustering, the dataset was filtered to remove genes that showed insignificant levels of variation in signal and thus would obscure real differences between experiments. The coefficient of variation (CV) was calculated for all expression values for each gene across all experiments, and only genes with CV values greater than a chosen cut-off (250%) were used for clustering. Table 1 lists the number of genes passing this threshold for each clustering.

## RESULTS

Three sets of global human gene expression data generated with different methods were compiled. They consisted of Affymetrix GeneChip<sup>®</sup> HDA data from the HugeIndex database (Haverty *et al.*, 2002), all available SAGE data of similar sample types from the NCBI's SAGEmap resource (Lash *et al.*, 2000) and a dataset generated using EBE with all applicable EST libraries from dbEST (Boguski *et al.*, 1993). These datasets provided global gene expression information for five human organs (brain, kidney, liver, lung and muscle). In total, 25 HDA experiments, 23 EBE experiments and 5 SAGE experiments were compared over a set of 3389 genes, although not all of these genes had detectable expression levels by each method in every experiment.

### Biological replicate variations within each method

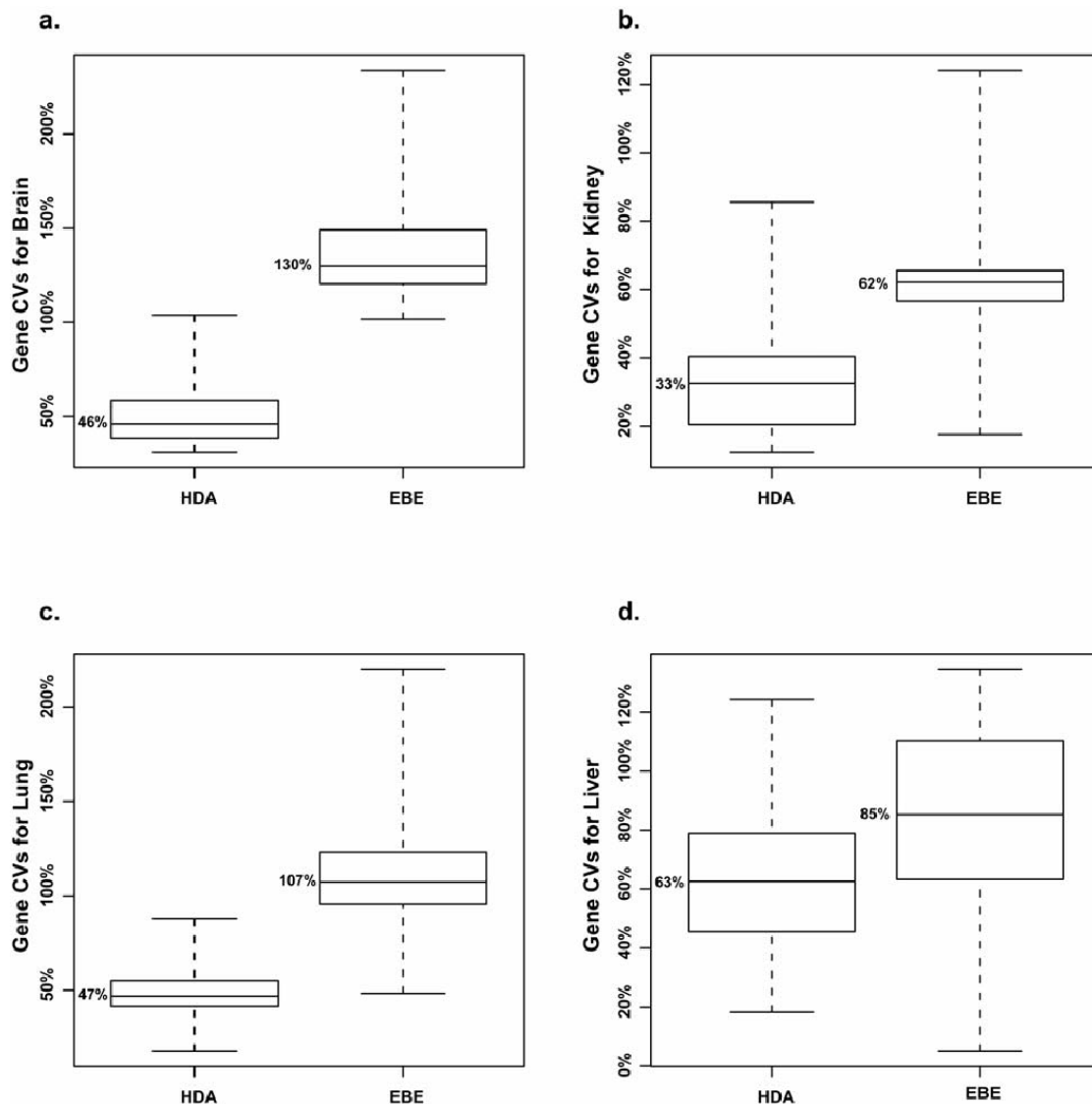
In order to assess the reliability of the platforms, we calculated the CV for the measurements obtained with each platform on a gene-by-gene basis. There was only one experiment performed for EBE on muscle and for SAGE on each organ type;

therefore, this analysis was not possible for these cases. Thus, CV was computed for HDA and EBE on brain, kidney, liver and lung. Figure 1 shows the distribution of CV values for HDA and EBE, which indicates a trend for much greater variability in EBE measurements relative to that in HDA data. We restricted the analysis to the genes of which more than half of the measurements for a given organ were non-zero, for both HDA and EBE. This restriction avoided situations where a method failed to detect the expression of a gene in multiple samples, thus giving a false impression of low variability. With this gene filter, the gene lists were reduced to 26, 22, 56 and 65 genes for brain, kidney, liver and lung, respectively. A range of other cutoffs were analyzed, resulting in similar findings both in terms of the higher level of EBE variation relative to that of HDA and of the magnitudes of the CVs.

### Correlation of single gene expression patterns between platforms

Spearman rank correlations were calculated to measure the agreement among the three platforms regarding the pattern of a gene's expression across the five organs. When multiple experiments were available for an organ type, the average expression value for each gene was used. Genes may appear to have the same pattern of expression between two methods when, in fact, the actual pattern of expression for that gene was not detected by either method. Thus, we filtered the gene lists based on the proportion of average expression values reported as zero. A range of the proportion of non-zero average expression values was tested. Figure 2 depicts the resulting distributions of correlations at two extreme cutoffs.

Figure 2a shows the distribution of correlation values when at least one of the five values (corresponding to five organ types) was required to be non-zero for each of the three datasets. Only 1996 of 3390 genes passed this filter. Here, the median correlation is very low between each pair of methods. HDA and SAGE generally show the greatest agreement, SAGE and EBE tend to agree less, and the agreement between HDA and EBE appears to be in random ( $R = 0.00, 0.10$  and  $0.33$  for HDA versus EBE, EBE versus SAGE and HDA versus SAGE, respectively). Figure 2b shows the same distributions when four or more of the five average organ expression values were required to be non-zero for each of the three methods, including 42 genes. This much smaller, but higher quality subset of data shows a trend opposite to that in Figure 2a. Here, the HDA versus EBE and EBE versus SAGE comparisons show much higher median correlation values ( $R = 0.50, 0.50$  and  $0.30$  for HDA versus EBE, EBE versus SAGE and HDA versus SAGE, respectively). The HDA versus EBE distribution is particularly tightly grouped around the region of  $R = 0.5$  indicating that a large proportion of these selected expression patterns might be considered meaningful. The HDA versus SAGE distribution is largely unchanged from that in Figure 2a.



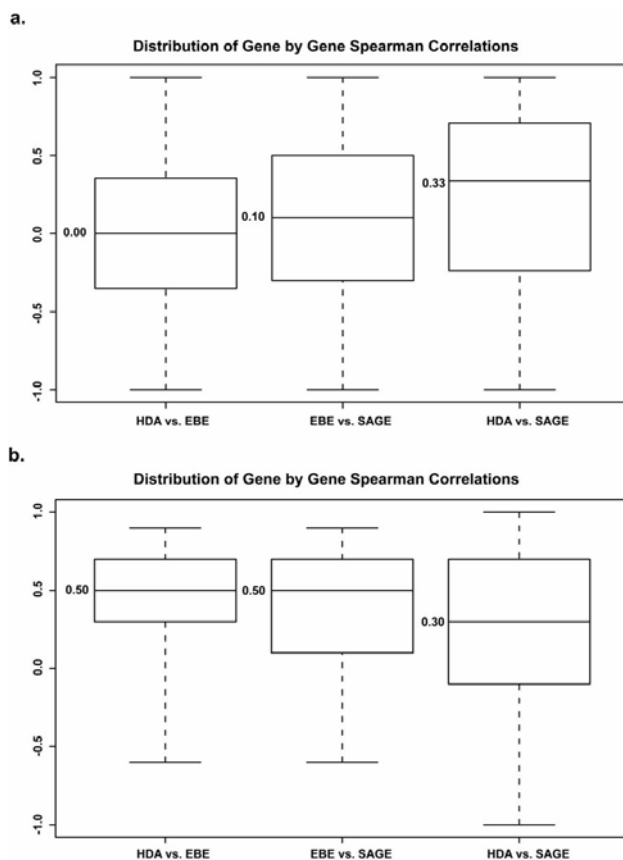
**Fig. 1.** Variability of expression measurements between replicate samples. Box and Whisker plots show the distributions of the CV for biological replicate measurements of expression for individual genes. Box and Whisker plots provide a simple description of a distribution of values by depicting the 25th and 75th percentile values as the bottom and top of a box, respectively. The median value is marked by a line within the box and the minimum and maximum values are depicted by error bars, or whiskers, protruding from the box. In order to remove spuriously low-variation examples, genes were filtered to include only those for which  $>50\%$  of measurements were greater than zero for each platform. The EBE muscle and SAGE experiments have only one replicate per organ and therefore are not included in the analysis.

### Comparison of distributions of gene expression levels reveal limitations of different methods

To investigate whether any method has a systematic bias, we also compared the distributions of all gene expression levels (after  $\log_{10}$  transformation) as measured by HDA, SAGE or EBE. In Figure 3a–c, we plot the histograms of the average expression levels for each method. The distribution of HDA measurements follows a bell curve, with roughly equal numbers of low- and high-expression genes. In contrast, both SAGE and EBE lack the low-value tail, which may indicate the

insensitivity of these techniques in detecting low-expression levels.

To directly compare the distributions of expression levels measured by two methods, we also plotted Quantile–Quantile (Q–Q) comparisons between any two methods (Fig. 3d–f) using the *R* statistical package (<http://www.r-project.org>; version 1.6.2). Each Quantile point adopts the value that is greater than or equal to the expression levels of a certain percentage (or quantile) of genes. For example, the 90% quantile point for the SAGE distributions represents the value



**Fig. 2.** Correlations of gene expression values across organ types. Box and Whisker plots show the distributions of Spearman rank correlations of single-gene expression across five organs. (a) Genes were filtered to remove examples where at least one of the five measurements was greater than zero for each of the three platforms, resulting in a list of 1996 genes. (b) Genes were filtered to remove examples where at least four of the five measurements were greater than zero for each of the three examples, resulting in a list of 42 genes.

that is  $\geq 90\%$  of all SAGE values, or 137 tpm. A Q–Q plot compares two distributions, with a diagonal line indicating a perfect correspondence. The advantage of such plots is that distributions of different units (e.g. tpm versus fluorescent intensity) can be directly compared. The SAGE–EBE Q–Q plot forms nearly a straight diagonal line, indicating the high similarity between SAGE and EBE gene expression distributions. The most obvious difference between these two distributions is the jump prior to the highest quantile, where only EBE reports very high levels of expression. This jump also exists in the HDA–EBE Q–Q plot, suggesting that identification of a few very highly expressed genes is unique to EBE. We suggest that some stage of the EST library construction, such as the growth of the cloned cDNAs in bacteria, results in a few genes appearing to have extremely high-expression values. A recent publication provides some support for the presence of extreme values on the high end

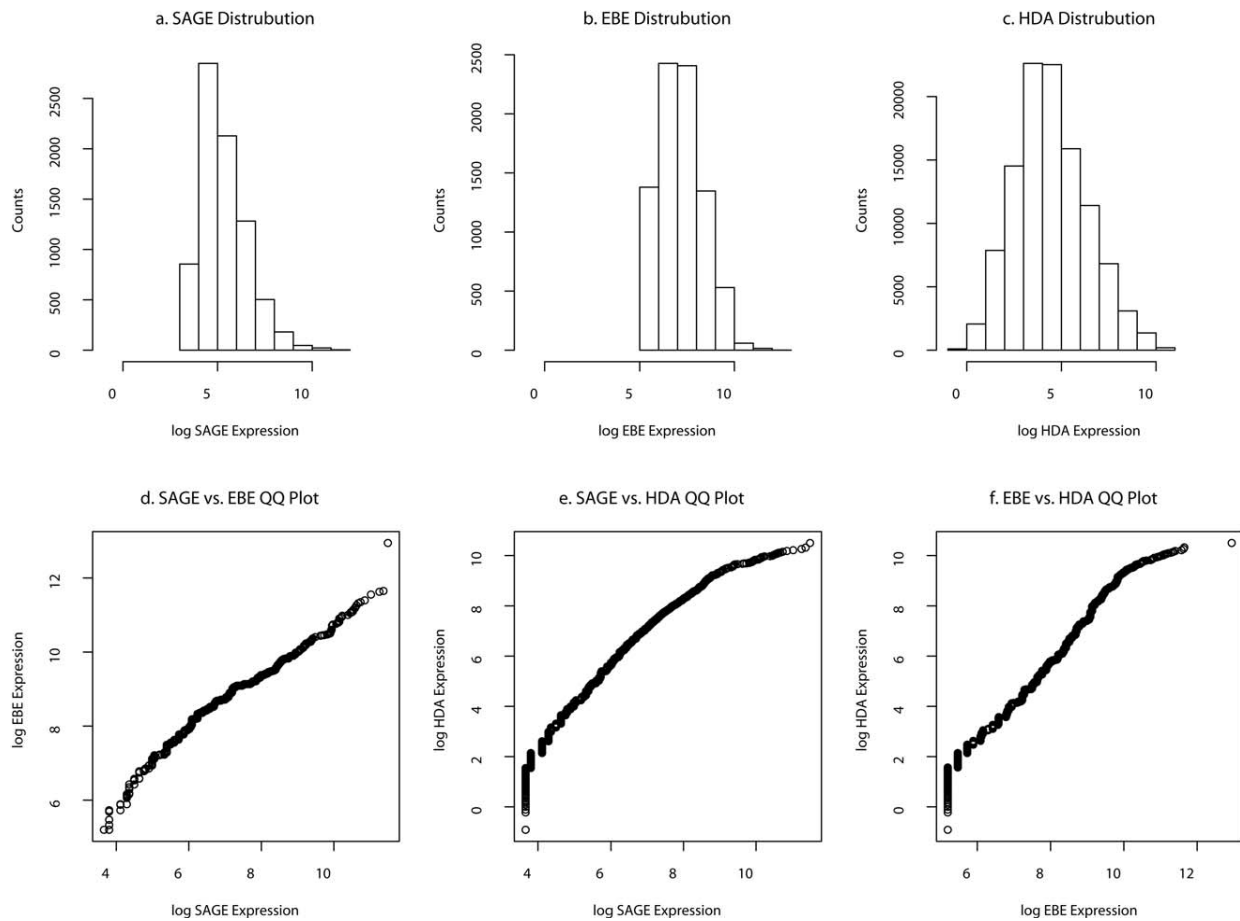
of the EBE distribution by demonstrating that statistical tests for co-expression using EBE data are more specific when it is assumed that the values follow a log–normal distribution rather than a normal distribution (Price and Rieffel, 2004).

The HDA distribution (Fig. 3c) is unique in having a much higher proportion of low values. The difference is also reflected in the HDA–SAGE and HDA–EBE Q–Q plots, where the curves substantially deviate away from the diagonal and toward the HDA axis at lower values. The difference reflects the different sampling techniques of the three methods: SAGE and EBE are sequencing based while HDA is hybridization based. Sequencing-based sampling is inherently limited by the total number of genes that can be sequenced, and may not be sufficiently sensitive to detect genes with low-expression levels.

The Q–Q plots also reveal differences in the upper expression values of the three methods. The HDA distribution shows a noticeably smaller abundance of high-expression values than do those of SAGE and EBE. We investigated whether this could be explained by redundant clusters of ESTs in UniGene, which was the basis for constructing the EBE and SAGE datasets in this study. Theoretically, UniGene redundancy could reduce apparent EBE or SAGE expression values by assigning sequences corresponding to a single gene to multiple UniGene clusters. Genes not affected by such redundancy might therefore exhibit relatively high EBE or SAGE expression values. The presence of at least one full-length mRNA sequence in a cluster should ensure that all sequences from that gene are included in that cluster. However, removing genes from the analysis that did not correspond to a cluster with a sequence annotated as ‘complete CDs’ did not significantly change the distributions (data not shown). A second explanation for the different distribution profiles at high values could be the effects of subtracting mismatch probe intensities from perfect match probe intensities in the process of calculating HDA expression values. It has been documented that overestimates of background hybridization using the current Affymetrix HDA analysis software, MAS 5.0, can lead to underestimates of large magnitude expression (Irizarry *et al.*, 2003). Interestingly, the use of background hybridization by MAS 5.0 has also been shown to contribute to the higher variability in quantifying low abundance transcripts (Han *et al.*, 2004). This variability at low expression levels is somewhat expected, as the relative magnitude of errors with respect to background fluorescence intensity increases as the signal decreases (Quackenbush, 2002).

### Co-clustering of experiments of dissimilar methodology from the same organ

We performed clustering analysis to further compare HDA, EBE and SAGE. Our goals were to judge the ability of each method in classifying gene expression patterns according to organ types and to evaluate the similarity of these classifications by the three methods. Grouping of experiments



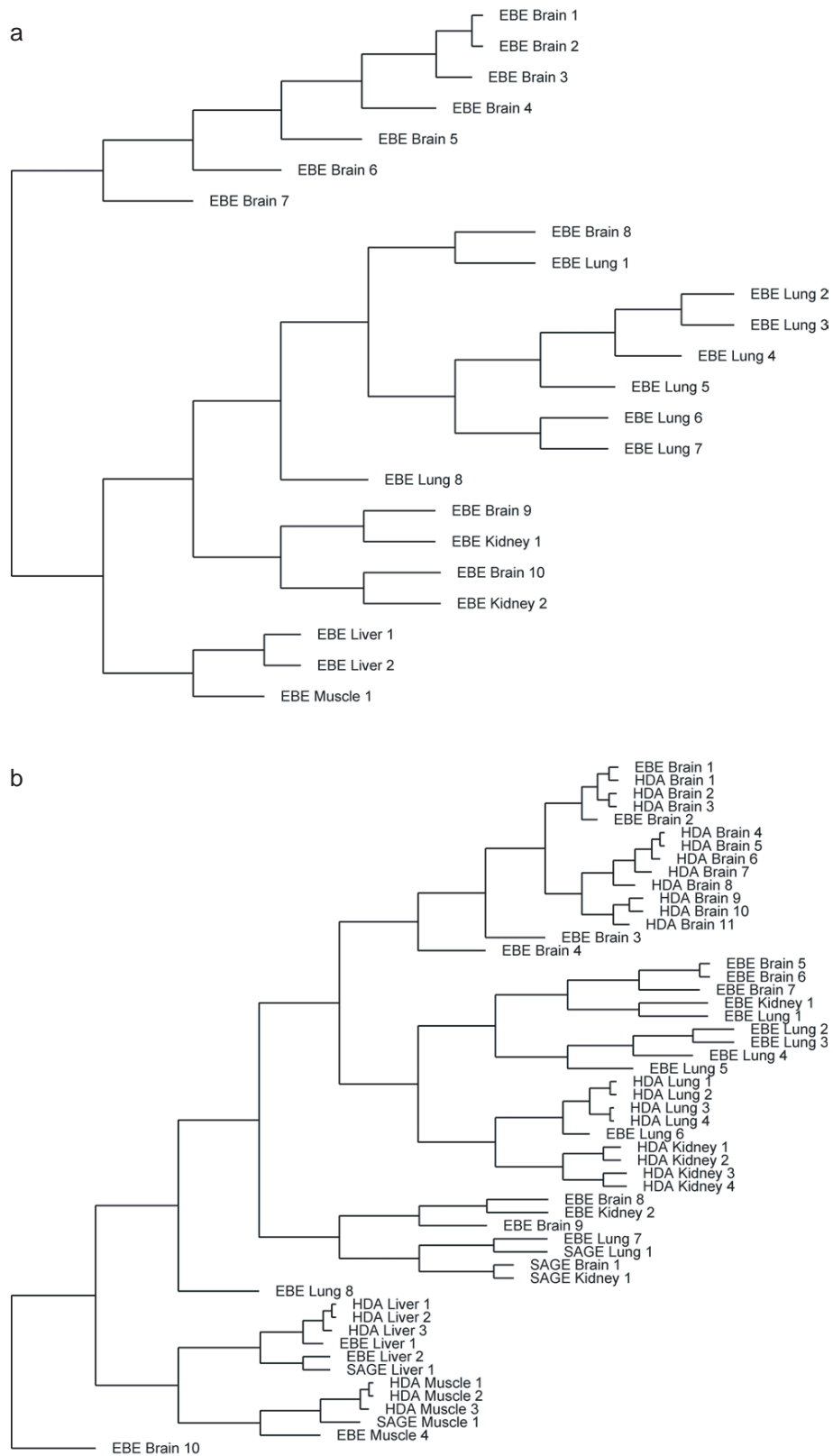
**Fig. 3.** Comparison of distributions of average values for all 3389 genes. (a–c) Distributions of  $\log_{10}$  transformed average expression values for 3389 genes for (a) SAGE, (b) EBE and (c) HDA. (d–f) Quantile–Quantile plots comparing distributions of (d) SAGE versus EBE, (e) SAGE versus HDA and (f) EBE versus HDA.

performed with one method on the same organ type would suggest high internal consistency. Clustering behavior of this type has been documented previously (Ross *et al.*, 2000; Hsiao *et al.*, 2001). We were also interested in investigating whether experiments performed using different methods on the same organ would cluster together. To the best of our knowledge, no such analysis has been performed before.

We used both  $K$ -means and hierarchical clustering techniques. Hierarchical clustering provides a visual description (a tree) of the relationships between experiments; however, it is not straightforward to convert a tree into some number of clusters (five clusters in our case, corresponding to five organ types).  $K$ -means clustering is more powerful for dividing data into distinct groups without an obvious hierarchical structure. The quality of a  $K$ -means clustering result can be measured using the Adjusted Rand Index (ARI) (Yeung *et al.*, 2001), which quantifies the deviation from the optimal clustering, defined as all experiments of each organ type exclusively belonging to a distinct cluster. ARI has an expected value of 0 for random results and a value of 1 for perfect clustering.

The ARI values for  $K$ -means clustering on various datasets are listed in Table 1. HDA data were separated perfectly by both clustering methods (indicated by an ARI value of 1; tree not shown). Although imperfect, EBE data showed clear grouping of experiments from like organs. Figure 4a presents the tree produced by hierarchical clustering of EBE data. All eight lung samples formed one cluster. Distinct clusters were also observed for samples from two other organs: the two kidney samples and the two liver samples. The only organ that was not perfectly classified was brain. Seven out of ten brain samples were grouped together. Two others were grouped with kidney samples, and brain #8 grouped with lung samples. This could reflect the inherent heterogeneity of the brain tissue. Similarly, the  $K$ -means clustering of EBE data resulted in ARI of 0.35.

The ARI value was considerably lower for EBE (0.35) than for HDA (1). Nevertheless, the  $K$ -means clusters (data not shown) and hierarchical clustering tree (Fig. 4a) indicate that the EBE samples are largely clustered correctly. The ARI scoring scheme tends to penalize the mixing of two clusters



**Fig. 4.** (a) Hierarchical clustering of all EBE experiments. The EBE experiments were clustered using average-linkage clustering as described in Materials and Methods section. (b) Hierarchical clustering of all experiments. All SAGE, EBE and HDA experiments were clustered using average-linkage clustering as described in Materials and Methods section.

even when these clusters have been correctly discriminated from samples of other types. Therefore, ARI values near 1 would be obtained only in near-ideal cases and relatively low-ARI values may still be significant.

The *K*-means clustering using two data types led to ARI between 0.30 (EBE and SAGE) and 0.82 (HDA and SAGE). The lowest ARI for EBE and SAGE experiments is unexpected given the similarities in methodology between these methods. Interestingly, adding HDA data to EBE improved the ARI from 0.35 to 0.45, indicating some agreement between these two methods. Unfortunately, we could not assess any potential ability of EBE or SAGE data to improve the clustering of the HDA data due to the already highly accurate clustering of the HDA data alone. The *K*-means clustering using all three data types led to reasonably good results (ARI = 0.41).

The tree produced by hierarchical clustering is shown in Figure 4b. Two organ types were sorted out entirely. At the bottom of the tree, we observe one cluster formed by all three HDA muscle experiments, the only EBE muscle experiment and the only SAGE muscle experiment. Similarly, all six liver experiments form another cluster (three HDA, two EBE and one SAGE). At the top of the tree, there is a large cluster that contains all 11 HDA brain experiments and 4 EBE brain experiments. For the rest of the tree, we largely see grouping within individual methods, with two exceptions: (1) EBE lung #6 is grouped with the cluster containing all four HDA lung experiments and (2) EBE lung #7 is grouped with the only SAGE lung experiment. The SAGE muscle and liver samples cluster well with the other data types and the clustering of the SAGE lung sample with an EBE lung sample may be significant (Fig. 4b), but without additional SAGE samples it is impossible to address the frequency of such positive results. Likewise, the clustering of the SAGE brain and kidney samples (Fig. 4b) may indicate an artifact of the SAGE method, but this statement requires the support of additional samples.

The same CV cutoff was used for the clustering experiments in order to put the three methods on an equal footing and further explore the effects of internal variability on the utility of the data. Similar results were obtained when CV cutoffs were selected to obtain similar numbers of genes for each method. For example, the CV cutoff had a minimal effect on the quality of the *K*-means clustering of the HDA data by itself. Reducing the CV cutoff from 250 to 100% resulted in the selection of 2499 genes, which was more in line with the 2334 genes selected for EBE using the CV cutoff of 250%. This larger group of genes also produced an ARI of 1. When clustering two or more data types together it was necessary to apply the same CV cutoff to all data in order to obtain a single set of genes with which to cluster all of the experiments. However, the *K*-means ARIs for the clustering of HDA, EBE and SAGE was in the range of 0.38–0.51 for CV cutoffs of 150–500%, resulting in 3262 to 214 genes.

## DISCUSSION

Given the immense popularity and expansive scope of global gene expression technologies, it is critical to perform cross-platform comparison. Here we compare Affymetrix HDAs with two orthogonal methods, SAGE and EBE, on normal tissue from five organ types. Our goal for this study was to explore the limit of cross-platform consistency and to evaluate the feasibility of combining data from different sources to increase the accuracy of global gene expression measurements or to validate results for a small set of genes. Our analyses show highly variable, yet for some genes considerable, levels of agreement between the methods. The large numbers of contrasting measurements for individual gene expression levels dictate that data from any one source be interpreted with great care, and that use of multiple global gene profiling methods cannot necessarily, by themselves, resolve the problem.

### Performance of the individual methods

The clustering of data from the same organ and dissimilar methods (Fig. 4b) shows that the three methods provide similar pictures of global gene expression. HDA performs somewhat better in this regard than do SAGE and EBE, indicated by the more reliable clustering of the HDA data (Table 1). One should take into account, however, that some of the apparently higher reliability of the HDA data in our study may be due to the fact that all of the HDA data originated from a single laboratory, and thus were subject to fewer sources of variability. Although this difference would appear to be confounding, in practice EBE analysis is almost exclusively performed with collections of existing data, while HDAs are generally used in specifically designed experiments. The limited number of SAGE samples makes interpretation of the SAGE results difficult. The limited availability of SAGE samples also precludes the analysis of internal variation for SAGE by CV (Fig. 1) and by clustering of SAGE samples alone. If larger collections of SAGE data are created in the future, it will be informative to address these issues.

Even if the overall picture of gene expression provided by any one method is sufficiently accurate and detailed to make general claims about the differences between two samples, the measurements of individual gene expression levels can be different in a disconcertingly large proportion of cases. Many genes show high correlation between methods, especially if proper filtering is applied to limit the analysis to meaningful EBE and SAGE results (Fig. 2b). However, it is more important to note that the spread of correlation values is quite wide in all cases indicating highly variable accuracies in at least two, if not all three, of the methods. In light of these findings, the validation of global gene expression assays is even more critically important than was believed previously. The individual gene EBE or SAGE expression data provided by resources such as GeneCards (Rebhan *et al.*, 1997, <http://www.bioinformatics.weizmann.ac.il/cards/>) or

CGAP's Digital Northern (Lash *et al.*, 2000), respectively, should certainly be interpreted with care.

### Potential for combining methods

This study addresses the possibility of sharing, combining and comparing expression data from different sources and on different platforms. If such sharing is possible, expression analyses can be made simpler and more powerful. For example, once the expression analysis of a normal tissue/cell type has been performed, the results could be used as the reference for comparison with numerous disease conditions or experimental perturbations to detect genes with altered expression. Combining methods in this global way could, in theory, serve to increase the statistical power of analysis tools by increasing the number of control replicates. Alternatively, data from another global platform could potentially be used to confirm patterns of expression detected in an important subset of genes. Data from multiple methods could also be combined to escape the limitations of any one method.

The results of our analyses suggest that HDA data could be combined with SAGE or EBE to provide additional confidence on those expression measurements for which two or more methods show high levels of agreement. Such comparisons could be used to extract a high-confidence subset of genes that describe the unique properties of a particular tissue type, for example. However, given the abundance of single gene disagreements among the three methods, combining two or more of these methods in order to confirm expression patterns in a specific set of genes is not appropriate. At this time it would be advisable to rely on more accurate non-global methods, such as QRT-PCR, for validation until such time as techniques are developed to improve the reliability of SAGE and EBE data or to discriminate high-quality data from lower quality data.

### Expression level specific limitations of SAGE and EBE

Comparisons of the distributions of gene expression values (Fig. 3) show limitations in EBE and SAGE at certain expression levels. Both EBE and SAGE show a distinct lack of low-expression values in contrast to HDA. This is interpreted as the inability of these sampling based methods to detect transcripts below a certain abundance level. In addition, EBE appears to exaggerate the expression level of the most highly expressed genes.

### Summary

We report highly variable agreement among three most widely used global gene expression methods HDA, SAGE and EBE on a gene-by-gene basis. The levels of agreement we observed are enough to make general statements about the similarities and differences among samples in a clustering setting, but emphasize the necessity of validating specific gene expression methods with gene-specific methods. At this time, sharing

data across platforms has only limited utility. Given the complications with data source diversity and tissue heterogeneity, we believe that our study has pushed the consistency among the three methods to its limit. However, existing global gene expression methods are evolving and new methods are emerging rapidly. A similar comparison approach can be used in the future to compare and calibrate diverse methods.

### ACKNOWLEDGEMENTS

We thank Joseph Szustakowski, Martin Frith, Avi Spira, Gang Liu and Kavitha Venkatesan for thoughtful discussion and Dmitriy Leyfer for introducing us to EBE. P.M.H. and Z.W. are partially supported by NSF grant DBI#0078194 and 1R01HG03110-01. U.H. is partially supported by grant #CA81157 from the National Institutes of Health. This work was supported in part by the NIDDK Biotechnology Consortium.

### REFERENCES

- Affymetrix (2001). Affymetrix Technical Note: Statistical Algorithms Reference Guide. Santa Clara, CA.
- Boguski, M.S., Lowe, T.M. and Tolstocher, C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nat. Genet.*, **4**, 332–333.
- Boguski, M.S. and Schuler, G.D. (1995) ESTablishing a human transcript map. *Nat. Genet.*, **10**, 369–371.
- Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. and Kohane, I.S. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA*, **97**, 12182–12186.
- Chuaqui, R.F., Bonner, R.F., Best, C.J., Gillespie, J.W., Flaig, M.J., Hewitt, S.M., Phillips, J.L., Krizman, D.B., Tangrea, M.A., Ahram, M. *et al.* (2002) Post-analysis follow-up and validation of microarray experiments. *Nat. Genet.*, **32** (Suppl.), 509–514.
- 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.
- Han, E.S., Wu, Y., McCarter, R., Nelson, J.F., Richardson, A. and Hilsenbeck, S.G. (2004) Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. *J. Gerontol. A Biol. Sci. Med. Sci.*, **59**, 306–315.
- Haverty, P.M., Weng, Z., Best, N.L., Auerbach, K.R., Hsiao, L.L., Jensen, R.V. and Gullans, S.R. (2002) HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Res.*, **30**, 214–217.
- Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P. *et al.* (2001) A compendium of gene expression in normal human tissues. *Physiol. Genomics*, **7**, 97–104.
- Hsiao, L.L., Jensen, R.V., Yoshida, T., Clark, K.E., Blumenstock, J.E. and Gullans, S.R. (2002) Correcting for signal saturation errors in the analysis of microarray data. *BioTechniques*, **32**, 330–332, 334, 336.

- 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.
- Ishii,M., Hashimoto,S., Tsutsumi,S., Wada,Y., Matsushima,K., Kodama,T. and Aburatani,H. (2000) Direct comparison of Gene-Chip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, **68**, 136–143.
- Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C., Trent,J.M., Staudt,L.M., Hudson,J., Jr, Broguski,M.S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.
- Kuo,W.P., Jenssen,T.K., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.
- Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
- Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
- MacDonald,T.J., Brown,K.M., LaFleur,B., Peterson,K., Lawlor,C., Chen,Y., Packer,R.J., Cogen,P. and Stephen,D.A. (2001) Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nat. Genet.*, **29**, 143–152.
- Pomeroy,S.L., Tamayo,P., Gaasenbeek,M., Sturla,L.M., Angelo,M., McLaughlin,M.E., Kim,J.Y., Goumnerova,L.C., Black,P.M., Lau,C. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442.
- Price,M.N. and Rieffel,E. (2004) Finding coexpressed genes in counts-based data: an improved measure with validation experiments. *Bioinformatics*, **20**, 945–952.
- Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, (Suppl.), 496–501.
- Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1997) GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center, Rehovot, Israel.
- Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., Van de Rijn,M., Waltham,M. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.
- Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
- Yeung,K.Y., Fraley,C., Murua,A., Raftery,A.E. and Ruzzo,W. (2001) Model-based clustering and data transformations for gene expression data. *Technical Report UW-CSE-2001-04-02*. Department of Computer Science and Engineering, University of Washington, Seattle, WA.
- Yuen,T., Wurmbach,E., Pfeffer,R.L., Ebersole,B.J. and Sealfon,S.C. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.*, **30**, e48.