



Analysis of matched mRNA measurements from two different microarray technologies

Winston Patrick Kuo^{1, 2, 3, *, †}, Tor-Kristian Jenssen^{2, 4, †},
Atul J. Butte^{1, 3}, Lucila Ohno-Machado^{2, 3} and
Isaac S. Kohane^{1, 3}

¹Children's Hospital Informatics Program and Division of Endocrinology, Department of Medicine, Children's Hospital, ²Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, ³Harvard University and MIT, Division of Health Sciences and Technology, Cambridge, MA 02139, USA and ⁴Department of Computer and Information Science, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

Received on April 9, 2001; revised on September 29, 2001; accepted on October 25, 2001

ABSTRACT

Motivation: The existence of several technologies for measuring gene expression makes the question of cross-technology agreement of measurements an important issue. Cross-platform utilization of data from different technologies has the potential to reduce the need to duplicate experiments but requires corresponding measurements to be comparable.

Methods: A comparison of mRNA measurements of 2895 sequence-matched genes in 56 cell lines from the standard panel of 60 cancer cell lines from the National Cancer Institute (NCI 60) was carried out by calculating correlation between matched measurements and calculating concordance between cluster from two high-throughput DNA microarray technologies, Stanford type cDNA microarrays and Affymetrix oligonucleotide microarrays.

Results: In general, corresponding measurements from the two platforms showed poor correlation. Clusters of genes and cell lines were discordant between the two technologies, suggesting that relative intra-technology relationships were not preserved. GC-content, sequence length, average signal intensity, and an estimator of cross-hybridization were found to be associated with the degree of correlation. This suggests gene-specific, or more correctly probe-specific, factors influencing measurements differently in the two platforms, implying a poor prognosis for a broad utilization of gene expression measurements across platforms.

Contact: wpkuo@mit.edu

INTRODUCTION

Several technologies for large-scale assays of gene expression are now available. Serial Analysis of Gene Expression (SAGE; Velculescu *et al.*, 1995), photolithographically synthesized oligonucleotides probe microarrays (Lockhart *et al.*, 1996) and microarrays using cDNA probes (Schena *et al.*, 1995) are among the most common. Gene expression from cells in different states measured by mRNA abundance is correlated to protein level (Gygi *et al.*, 1999) and may provide considerable information on the biological roles of the measured genes in various cellular processes (Lockhart and Winzeler, 2000). The process of comparing gene expression information from a broad range of conditions has proved to be a useful approach to discovering gene function (Hughes *et al.*, 2000). Although the high-throughput microarray technologies are very efficient, they are also quite expensive, and as different laboratories may use different platforms to profile the same genes, one would like to combine expression measurements across technologies within a single analysis (Aach *et al.*, 2000).

In theory, the sequencing-based SAGE method directly measures actual levels of mRNA (Velculescu *et al.*, 1995), while microarray platforms measure mRNA abundance indirectly (through hybridization of labeled targets to specific probes). Oligonucleotide microarray technology has been shown to have good correlation with SAGE (Ishii *et al.*, 2000) and Northern blots (Harkin *et al.*, 1999; Lee *et al.*, 1999). Expression levels measured by cDNA microarrays are usually reported as a ratio of the signal from a target mRNA sample relative to one from a co-hybridized reference mRNA sample (Schena *et al.*, 1995; Shalon *et al.*, 1996). This method complicates a direct comparison with results from other technologies as the

*To whom correspondence should be addressed.

† These authors contributed equally to this work.

reported ratios depend on the chosen mRNA reference, as a high level of expression in a sample does not necessarily result in a high ratio. On a smaller scale, fold changes in mRNA levels measured with cDNA microarrays have been validated against other expression assays, including Northern blots (Heller *et al.*, 1999) and quantitative PCR (Iyer *et al.*, 1999).

In this study, we analyzed mRNA measurements from two large-scale studies measuring gene expression in cancer cell lines. The publicly available data had been measured with Stanford-type cDNA microarrays (<http://genome-www.stanford.edu/nci60>; Ross *et al.*, 2000) and Affymetrix oligonucleotide microarrays (<http://www.genome.wi.mit.edu/MPR>; Butte *et al.*, 2000). To date, there has been no other large-scale analysis of reproducibility between spotted cDNA microarrays and microarrays with synthesized oligonucleotides, the two most commonly used microarray platforms.

METHODS

Gene expression data

Details about the publicly available gene expression data sets have been described previously (Butte *et al.*, 2000; Ross *et al.*, 2000). Of the 60 cell lines from the standard panel of cancer cell lines from the National Cancer Institute measured with both technologies, we matched 56 pairs of experiments by cell line name. Two of the cell lines had been measured more than once on both platforms. There were three replicates of the MCF7 cell line and two replicates of the K-562 cell line. For these cell lines, pairs were created by randomly selecting one replicate from each technology. The cDNA microarrays contained 9703 cDNA probes while the Affymetrix HU6800 arrays contained 7245 probe sets. From the GenBank accession numbers provided for the cDNA probes and sequences supposedly represented on the Affymetrix chips we downloaded sequence data from GenBank and used stand-alone BLAST (<http://www3.ncbi.nlm.nih.gov/BLAST/>) to match corresponding sequences. For each cDNA probe sequence, we found the best matching probe set, i.e. sequence represented by a probe set. From all pairs of cDNA probe and best-matching probe set, we removed matches with a score higher than $1E-50$. This threshold value was chosen as pairs of sequences with homology similarity lower than $1E-50$ appeared to belong to the same UniGene cluster.

From the cDNA microarrays we used the RAT2N, MRAT, CH2D, and CH1D measurements in our comparisons with the average differences reported from the Affymetrix arrays. Details about these measurement variables can be found in the ScanAlyze Manual (<http://rana.lbl.gov/manuals/ScanAlyzeDoc.pdf>) and the Affymetrix Microarray Suite 4.0 User Guide.

We created two data sets with filtered data. The first filtered data set contained 463 genes that had a sequence match in the oligonucleotide data and had been measured with high variance across the cell lines in the cDNA data. The second filtered data set contained 2158 genes that had a sequence match in the oligonucleotide data and were found to be particularly well measured in the cDNA data. Further details on the filtering criteria used have been described (Ross *et al.*, 2000).

Sequence analyses

The same GenBank sequence data used in matching cDNA probes with probe sets was also used to compute GC-content and probe lengths. Within the cDNA microarrays, BLAST was used to compute the number of cross-matches for each spotted probe. The number of cross-matches was defined as the number of other cDNA probes spotted on the microarray that were found to match with a BLAST score less than $1E-10$. For sequences represented by probe sets on the oligonucleotide microarray, we defined the number of cross-matches as the number of other represented sequences that were found to match with a BLAST score less than $1E-10$. UniGene clusters for cDNA probes and for sequences represented on the oligonucleotide arrays were determined by searching UniGene (Build version #126) for the GenBank accession numbers.

Correlation computations

We computed Pearson linear correlation coefficients and Spearman rank-order correlation coefficients for genes, cell lines, and across all 162 120 matched pairs of measurements. Assuming normally distributed data, hypothesis tests of correlation coefficients can be computed via the Student's t distribution. With n being the number of observations, the variable $t = r\sqrt{\frac{(n-1)}{(1-r^2)}}$, where r is Pearson's linear correlation coefficient, is distributed as Student's t with $n - 2$ degrees of freedom. Equivalently, for a given α , critical values of Pearson's r can be found

as $r = \sqrt{\frac{t_\alpha^2}{[(n-2)+t_\alpha^2]}}$, where t_α is the critical value of the Student's t distribution with $n - 2$ degrees of freedom

and α as the level of significance. For Spearman's ρ , the variable $t = \rho^* \sqrt{\frac{(n-2)}{(1-\rho^2)}}$, is distributed as Student's t with $n - 2$ degrees of freedom. A similar inverse relationship can also be found. In a one-tailed test with $\alpha = 0.01$, the critical values are 0.3076 and 0.3102 for Pearson's r and Spearman's ρ , respectively. Using a Bonferroni-type correction for multiple tests we obtain an adjusted α of $0.01/2895$, or approximately $3.4542 E-06$. The corresponding critical values for Pearson's r and Spearman's ρ are 0.557 and 0.561 respectively.

As the data were distinctly non-normally distributed,

we randomized the data and estimated the empirical distribution of the correlation coefficients. The data were randomized by selecting from the 2895 matched pairs, a pair at random and then randomly permuting the columns in both data sets. We computed linear and rank-order correlations for one million such randomly generated pairs. This randomization was carried out for the RAT2N, MRAT, and CH2D cDNA data. Unique sets of random pairs were used for each variable.

Cluster analysis

Hierarchical clustering was performed using the Statistics toolbox in MATLAB (The Math Works, Natick, MA), using Euclidean distance as the measure of similarity and average linkage clustering. By ‘horizontally’ cutting the derived trees (dendrograms) at various levels we derived cluster memberships for each of the cell lines or genes clustered. From the clusters derived in this way we calculated binary pair vectors, where each element corresponded to a unique pair of cell lines (or genes) and had value one if both cell lines were clustered into the same cluster and zero otherwise. From two such pair-vectors, where one was derived from clusters obtained from the oligonucleotide data and the other from clusters obtained from cDNA data, we computed the Jaccard (Yeung *et al.*, 2001) and Kappa (Landis and Koch, 1977) coefficients as a measure of the similarity of the clusters. These coefficients were computed for all possible numbers of clusters, i.e. from one to the number of cell lines (or genes).

RESULTS

Correlations

As a measure of overall consistency between the two platforms we calculated the correlation using all 162 120 pairs of matched measurements. The scatter plot in Figure 1a as well as the correlation coefficients in Table 1 shows that there is little correlation between the CH2D values and the average difference values. The 162 120 pairs of data points are presumed to originate from the same gene measured in the same cell line but the difference in measurement scales may offset the linear correlation. However, since the rank-order correlation is poor, it seems unlikely that a global translation of expression measurements across the two platforms can be made, not even a non-linear transformation. Figure 1b shows that there is practically no correlation between the RAT2N values and the oligonucleotide measurements. As can be seen in Table 1, this was also true for the alternative ratio, MRAT. The CH1D values represent the mRNA reference co-hybridized with the target cell line and were expected to have little or no correlation with the oligonucleotide values. Surprisingly, these values were

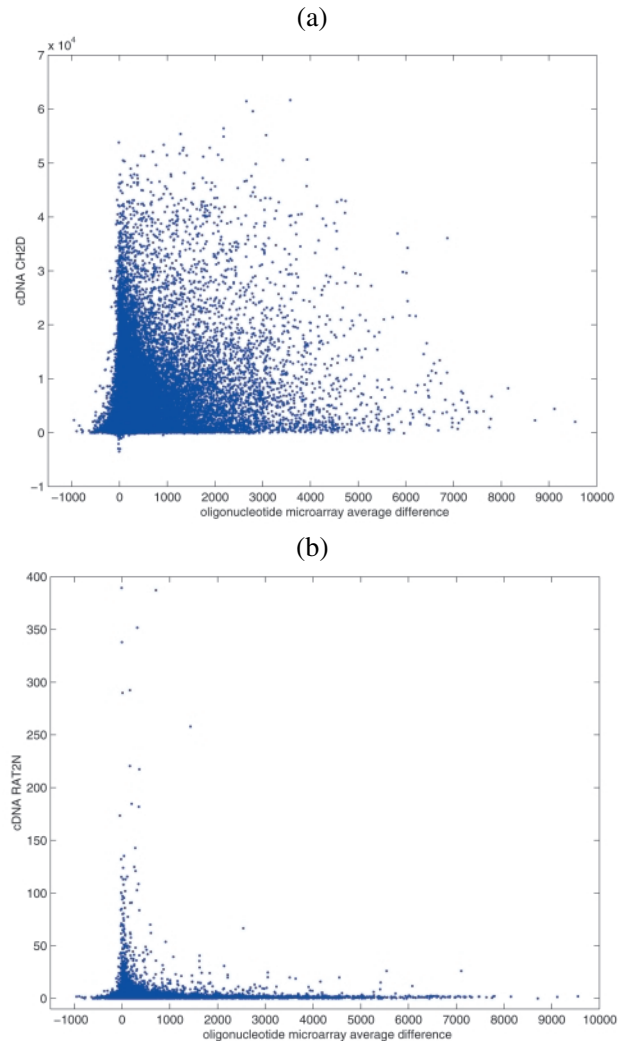


Fig. 1. Scatter plots of cDNA microarray and oligonucleotide results for 162 120 pairs of matched measurements. The overall correlation between the CH2D values and average difference was poor (a), with a linear correlation of $r = 0.328$, but substantially better than that of RAT2N values and average differences (b), with linear correlation of $r = 0.0326$. Note that three extreme outliers in the oligonucleotide data have been removed when creating these plots.

slightly correlated with the oligonucleotide data almost at the same level as the CH2D values, representing the target cell lines.

Because array-specific variation such as hybridization conditions and scan settings may severely offset the overall correlation, we calculated the correlation within each cell line. As with the overall correlations, the ratios RAT2N and MRAT had no correlation with the oligonucleotide data. For the CH2D values, the cell line correlations ranged from 0.2137 to 0.4508 for the Pearson coefficient and from 0.2824 to 0.5116 for the

Table 1. Correlation of oligonucleotide and cDNA microarray measurements

cDNA	Pearson (<i>r</i>)			Spearman (ρ)		
	Overall	Cell lines	Genes	Overall	Cell lines	Genes
RAT2N	0.0326	0.0599 (0.0482)	0.278 (0.310)	0.2214	0.225 (0.0651)	0.253 (0.280)
MRAT	0.0634	0.0944 (0.0512)	0.283 (0.307)	0.2328	0.2379 (0.0644)	0.253 (0.278)
CH2D	0.328	0.3457 (0.0478)	0.2083 (0.255)	0.4237	0.4376 (0.0666)	0.2316 (0.192)
CH1D	0.277	0.2902 (0.0435)	-0.0195 (0.151)	0.3464	0.3585 (0.0537)	-0.0222 (0.157)

The 'overall' columns show Pearson linear and Spearman rank-order correlation coefficients for all 162 120 pairs of measurements. The 'cell lines' columns show average linear and rank-order correlations calculated over all 56 pairs of matched cell lines, with the standard deviations in parenthesis. The 'genes' columns show the average and standard deviations of linear and rank-order correlations for the 2895 genes calculated across the 56 cell lines.

Spearman coefficient; average correlations and standard deviations are shown in Table 1. The low overall and cell line correlations indicate spot specific differences in measurement between the two technologies. We therefore calculated correlation coefficients for the 2895 matched genes across the 56 cell lines. Average correlations and standard deviations are shown in Table 1. For the genes, the ratios RAT2N and MRAT had better correlation than the CH2D values, and the control CH1D did not show correlation with the oligonucleotide data. We tested for significance both using the theoretical distribution assuming normally distributed data and using a randomization experiment calculating empirical distributions for the correlation coefficients. A summary of these results is shown in Table 2.

Gene features possibly influencing the correlation

To determine gene- or probe-related features associated with the degree of correlation, we calculated correlation coefficients for stratified subsets of the matched genes. In these analyses, we used only the RAT2N and CH2D data from the cDNA arrays. For the genes, the correlations calculated with the RAT2N data was strongly correlated with the correlations calculated with the CH2D data, so associations found with one variable generally would carry over to the other. Thus, unless explicitly stated otherwise, statements about correlations generally regard results found with both cDNA data measurements. We grouped genes based on GC-content, sequence length, and an index of intra-chip sequence similarity for both cDNA probes and sequences represented by oligonucleotide probe sets and examined correlations in the groups. The clearest trend was found for GC-content. Lower GC-content was associated with higher (average) correlation and higher GC content was associated with lower correlation (Figures 2a and b). However, the variation within each group was too high for these differences to be statistically significant. The probes spotted on the cDNA arrays ranged in length from 100 to 753 bases. We found reduced correlation for probe sequences of length greater than 600, both linear and rank-order (Figure 2d,

group 5). The sequences represented by probe sets on the Affymetrix arrays ranged in length from 214 to 302 250 bases. We observed lower correlations, both linear and rank-order, for the subset of sequences with 10 000 bases or more (Figure 2c, group 5).

As an estimator of possible cross-hybridization within each array type, we calculated sequence similarity between sequences spotted or represented on the arrays. We defined an intra-array cross-match to be a sequence similarity with BLAST score less than $1E-10$. For the cDNA arrays we found 1079 sequences with one or more such cross-matches. We observed a tendency for genes having probes with higher numbers of cross-matches to have lower correlation with the oligonucleotide data (Figure 2f). For the Affymetrix chips we did not have the actual probe sequences used on the arrays, only the sequences supposedly represented. Calculating cross-matches for these sequences did not reveal any clear association with gene correlation (Figure 2e).

Other studies have indicated that reliable measurements are harder to obtain for low-abundance transcripts (Bittner *et al.*, 2000). In order to see whether transcript abundance could be related to the between-platform correlations, we analyzed average correlations across groups of genes grouped by average signal across the 56 arrays. Using intensity data from both technologies we found lower correlations from genes where the average intensity was low (Figures 2g and h). For the 129 genes with average (across the 56 cell lines) CH2D value below 100, the average linear correlation was only 0.0321 and the average rank-order correlation was 0.020. Including all genes with an average CH2D value below 500, the average linear correlation was still as low as 0.104, about one third of the overall average. Similar patterns were also observed for low average difference in the oligonucleotide microarrays. For the high mean-intensity groups, the average correlations were clearly elevated compared to the overall averages.

Table 2. Genes with significant correlation in comparison between cDNA and oligonucleotide data

cDNA	Standard test				Bonferroni adjusted			
	Pearson (r)		Spearman (ρ)		Pearson (r)		Spearman (ρ)	
	t	e	t	e	t	e	t	e
RAT2N	1263	1150 (0.350)	1180	1176 (0.312)	663	22 (0.938)	483	487 (0.560)
MRAT	1261	1171 (0.346)	1185	1181 (0.312)	663	0 (0.997)	486	369 (0.608)
CH2D	883	765 (0.347)	834	833 (0.311)	339	0 (0.992)	211	190 (0.573)

The levels of significance where $\alpha = 0.01$ in the simple test and $\alpha = 0.01/2895$ in the test adjusted for multiple testing. The ‘ t ’ columns show the number of genes with significant correlation in the Student’s t -test. The ‘ e ’ columns show the number of genes with significant correlation as found by calculating the empirical distribution in randomized data. Critical values of the empirical distributions are given in parentheses. The critical values in a one-sided Student’s t -test are 2.397 and 4.979, respectively, for the two values of α . For the linear correlation, this translates to critical values for r of 0.308 and 0.557, respectively, and for the rank-order correlation, this translates to critical values for ρ of 0.310 and 0.561, respectively.

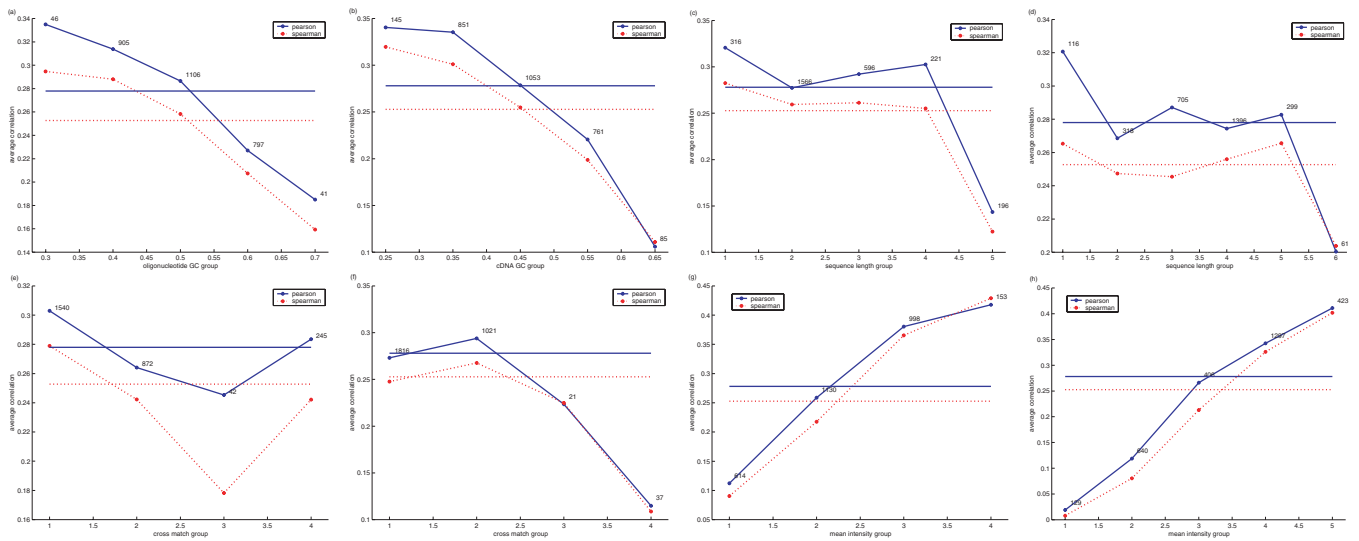


Fig. 2. Gene correlations between cDNA microarray RAT2N data and average difference from oligonucleotide microarrays. Figures (a), (c), (e), and (g) show average correlations in groups of genes based on GC-content, sequence length, intra-chip cross-matches, and mean expression from the oligonucleotide arrays. Figures (b), (d), (f), and (h) show average correlation in groups of genes based on GC-content, sequence length, intra-chip cross-matches, and mean expression from the cDNA arrays. The groups of genes were created by manual discretization. The groups based on GC-content in sequences represented on the oligonucleotide arrays (a) were created by cutting at 0.35, 0.45, 0.55, and 0.65. The groups based on GC-content in probes on the cDNA arrays (b) were created by cutting at 0.3, 0.4, 0.5, and 0.6. The groups based on length of sequences represented on the oligonucleotide arrays (c) were created by cutting at 1000, 3000, 5000, and 10000. The groups based on length of probes on the cDNA arrays (d) were created by cutting at 200, 300, 400, 500, and 600. The groups based on cross-matches of sequences represented on the oligonucleotide arrays (e) were created by cutting at 0.1, 10, and 50 (i.e. the first group contained only sequences with zero cross-matches). The same cuts were used for cross-matches of probes on the cDNA arrays (f). The groups based on average intensity on the oligonucleotide arrays (g) were created by cutting at 5, 50, and 500. The groups based on average intensity on the cDNA arrays (h) were created by cutting at 100, 500, 1000, and 5000. Common to all discretizations was that the upper limit was included in the bin, e.g. in (g), the second group of genes with average intensity between 500 and 1000 included genes with average intensity of 1000. In all plots, the horizontal lines represent the average correlations across all genes. For each group of genes created by discretization, the number of genes included in the group is shown in the plot.

Data filtering

In order to assess whether data filtering could improve the correlation, we created two filtered data sets (see Section **Methods**); one filtered data set containing 463 genes with high variance in the cDNA data and one containing 2158

genes presumed to be particularly well measured with the cDNA microarrays. Table 3 gives a summary of correlations computed from these two data sets. The main effect of the variance filtering was to remove genes with lower correlation, as the average correlations for the genes

increased. In both the filtered data sets, the overall and cell line correlations were slightly higher for the ratios and slightly lower for the CH2D data.

Cluster analysis

The results above indicate systematic differences in the detection of gene expression. As many microarray studies draw their conclusions from clustering analyses, cross-technology preservation of clustering results may be more interesting than preservation of the raw measurements. Cluster analyses should be similar if intra-technology relationships between measurements are similar. To assess whether this was the case in the present data sets, we performed hierarchical clustering of the cell lines for the whole data set with 2895 genes and the two filtered data sets. We also clustered genes in one of the filtered data sets. For these analyses we used the RAT2N and CH2D data from the cDNA arrays and, as earlier, the average differences from the oligonucleotide arrays. Figure 3 shows plots of the Jaccard and Kappa coefficients measuring cluster similarity between clusters computed from oligonucleotide data and clusters computed from the RAT2N data; the results from comparisons with CH2D data were similar (data not shown). From the findings it is evident that there is little concordance between the clusters resulting from the oligonucleotide data and the clusters resulting from the cDNA data. (Note that a Kappa coefficient above 0.4 indicates some concordance between clusters.) For the cell lines, we see that the clusters computed from the filtered data corresponding to cDNA well measured genes had some concordance for the number of clusters lying between 2 and 10, while the variance filtered data did not show higher concordance. No concordance was found in clusters of genes from the variance filtered data.

DISCUSSION

The existence of several platforms for measuring gene expression makes consistency and reproducibility across technologies important issues. Being able to use data interchangeably across platforms would be very beneficial, as this would potentially reduce the need to duplicate experiments. This study suggests that data from spotted cDNA microarrays could not be directly combined with data from synthesized oligonucleotide arrays. It seems unlikely that these data can be normalized or transformed to a common standardized index of gene expression. Clearly our study and observations are limited to these particular data sets. One important limitation is that the experiments were carried out independently in two different laboratories using different materials and protocols. Although the originating cell lines were the same, they had been grown independently and mRNA samples and hybridization targets had been prepared separately.

Associated with the above, the equipment and protocols for hybridization, washing, scanning and image analysis are possible sources of variation between replications of experiments also when carried out on a single platform. Further experiments where each of these factors can be controlled, isolating systematic differences between these two microarray technologies are clearly in place. Nevertheless, there are important differences between the two technologies and we believe that the low correlations observed, at least partially, are due to differences in the hybridization conditions and fluorescence measurement strategies inextricably linked to the platforms.

Our results indicate that sequence features may affect expression measurements differently in the two platforms. Base composition has a known effect on hybridization, given the lower stability of A–T versus G–C pairs (Southern *et al.*, 1999; Stomakhin *et al.*, 2000). Thus, the GC-content is an important factor because probes with higher GC-content tend to have increased hybridization stability due to the higher proportion of stable base pairs (Herwig *et al.*, 2000). Probe sequences with sequence homologies to other probes increase the potential for non-specific hybridization (Schadt *et al.*, 2001). The lower average correlation for cDNA probes with many cross-matches, suggests that cDNA probe similarity may cause inaccurate expression measurements on cDNA microarrays (Kane *et al.*, 2000). Spotting more specific cDNA probes of the unique regions from a set of genes and reducing the length of the probe sequences could reduce the potential for cross-hybridization. Probe length may also affect the degree of non-specific hybridization, as was in part observed on the cDNA arrays. On the Affymetrix platform, the probes have a fixed length, but the offset in correlation for probe sets on the oligonucleotide arrays representing longer sequences suggests similar effects for longer targets. In our results, we found the correlations to be more dependent on features of the cDNA probes than on features of the sequences represented on the oligonucleotide arrays. This indicates that the Affymetrix strategy using probe sets with short fixed-length probes including Perfect Match (PM) and Mismatch (MM) probes may give more homogeneous hybridization conditions. However, it should be emphasized that since the exact sequences of the probes used on the oligonucleotide arrays are not known (Affymetrix proprietary information) this added layer of abstraction makes it impossible to state whether the observed relative invariance in correlation is an indication of these measurements being more accurate.

The lack of overall and cell line correlations between RAT2N and MRAT data with the oligonucleotide array data was not surprising, as the ratios also depend on abundance of each mRNA species in the reference sample. It is commonly claimed that representing gene

Table 3. Correlations for filtered data

cDNA	Pearson (<i>r</i>)			Spearman (ρ)		
	Overall	Cell lines	Genes	Overall	Cell lines	Genes
(a) 463 variance filtered genes						
RAT2N	0.0676	0.1253 (0.1033)	0.4201 (0.3607)	0.2813	0.2838 (0.0918)	0.3175 (0.3077)
MRAT	0.1192	0.1839 (0.1286)	0.4315 (0.3599)	0.306	0.3072 (0.0996)	0.3199 (0.3081)
CH2D	0.2613	0.2771 (0.131)	0.3897 (0.327)	0.3261	0.3314 (0.1003)	0.2997 (0.287)
(b) 2158 well measured genes						
RAT2N	0.0905	0.107 (0.0521)	0.3378 (0.3079)	0.2645	0.2701 (0.0671)	0.3122 (0.281)
MRAT	0.112	0.1164 (0.0508)	0.3379 (0.3065)	0.2708	0.2774 (0.0665)	0.3101 (0.2799)
CH2D	0.3183	0.3375 (0.0503)	0.2415 (0.2558)	0.4173	0.4345 (0.0672)	0.2326 (0.2338)

The formats of these tables are similar to that of Table 1. The first filtered data set with 463 matched genes was obtained by intersecting the matched data with a list of genes having high variance across the cell lines in the cDNA data (a). The second filtered data set with 2158 matched genes was obtained by intersecting the matched data with a list of genes being well measured in the cDNA data (b).

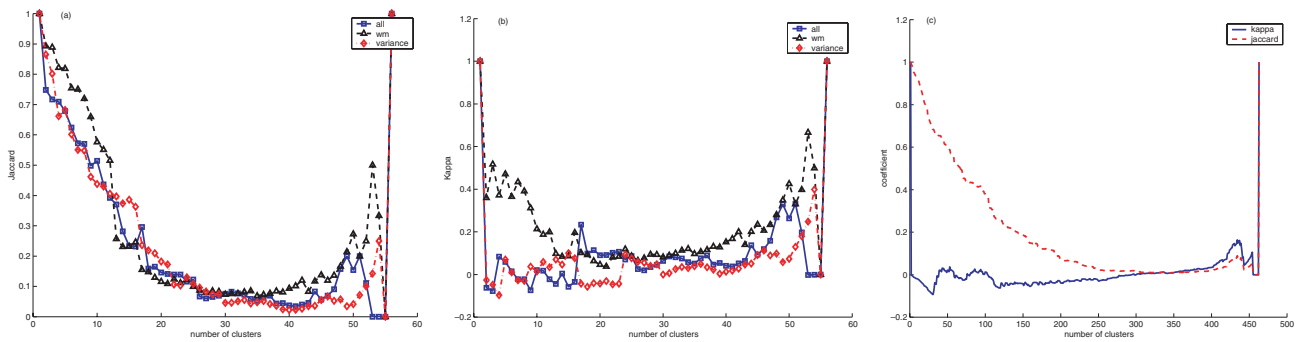


Fig. 3. Concordance of clusters from hierarchical clustering. Jaccard (a) and Kappa (b) coefficients were computed for clusters obtained from hierarchical clustering of cell lines using oligonucleotide data and RAT2N data from the cDNA arrays. The coefficients range from zero to one (y-axis) and are plotted by the possible number of clusters generated from the dendrograms (x-axis). Cell line clusters were generated from data sets containing all 2895 genes and filtered data sets containing 2158 (wm; supposedly representing well measured spots on the cDNA arrays) and 463 genes (variance; representing spots with high variance across the cell lines on the cDNA arrays). Little concordance was observed also in the filtered data sets. Genes in the filtered data sets were also clustered from the oligonucleotide and RAT2N data. Jaccard and Kappa coefficients for the clusters calculated from the data sets with 463 (c) genes are shown.

expression measurements as ratios result in more reliable (and thus more reproducible) data as the competitive co-hybridization to a reference mRNA sample may eliminate noise due to spot-variation in hybridization and signal detection (Schena *et al.*, 1995; Shalon *et al.*, 1996; Bittner *et al.*, 2000). This is in part supported by the higher correlation of genes observed between ratios and oligonucleotide data, than for the target signal alone. However, analysis of the replicated experiments revealed that internally in the cDNA data the CH2D values had higher correlation than the RAT2N values, both linear and rank-order (data not shown). Moreover, supposing that the lack of correlation was primarily due to random noise one would expect higher correlation comparing means (or medians) across replicates measured with one technology with means across replicates from the other technology.

For the MCF7 cell line, the correlations between the averages of intensity measurements across the replicates were not better than 0.388 and 0.457, for the linear and rank-order, respectively (correlations for the medians were 0.385 and 0.4851, respectively). In other words, just slightly above the average correlations between the individual replicates. This supports there being systematic differences in the two technologies, and as the cluster analyses revealed discordance in the computed clusters we anticipate that different biological relationships would be detected in gene expression analyses on the two platforms. This was further supported by the results from the filtered data.

Without an independent third source of data, we are not in a position to determine which of these two platforms is the better. We believe that there are issues in both

technologies that should be resolved before it is possible to transform measurements from either technology to a general index of gene expression. Our results suggest that more research on the hybridization dynamics on both platforms are needed before measurements are sufficiently well understood to be utilized across technologies. New techniques for expression profiling, including optically coded beads (Chen *et al.*, 1998; Michael *et al.*, 1998), inkjet arrays (Okamoto *et al.*, 2000; Hughes *et al.*, 2001), and others (Hakala and Lonnberg, 1997; Kononen *et al.*, 1998) may complement current technologies, but will also add to the challenge of comparing and combining expression data from different platforms.

ACKNOWLEDGEMENTS

This research was supported in part by the grant 'Research Training in Health Informatics' by the National Library of Medicine, 5T15 LM07092-07. T.K.J. was supported by grant 134422/410 from the Norwegian Research Council (NFR). We thank Eivind Hovig and Mark Whipple for their helpful advice and comments.

REFERENCES

- Aach, J. *et al.* (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.*, **10**, 431-445.
- Bittner, M. *et al.* (2000) Obtaining and evaluating gene expression profiles with cDNA microarrays. In Suhai, S. (ed.), *Genomics and Proteomics*. Kluwer, New York, pp. 5-25.
- Butte, A.J. *et al.* (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA*, **97**, 12 182-12 186.
- Chen, X. *et al.* (1998) DNA optical sensor: a rapid method for the detection of DNA hybridization. *Biosens. Bioelectron.*, **13**, 451-458.
- Gygi, S.P. *et al.* (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, **19**, 1720-1730.
- Hakala, H. and Lonnberg, H. (1997) Time-resolved fluorescence detection of oligonucleotide hybridization on a single microparticle: covalent immobilization of oligonucleotides and quantitation of a model system. *Bioconjug. Chem.*, **8**, 232-237.
- Harkin, D.P. *et al.* (1999) Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell*, **97**, 575-586.
- Heller, R. *et al.* (1999) Gene chips and microarrays: applications in disease profiles, drug target discovery, drug action and toxicity. In Schena, M. (ed.), *DNA Microarrays*. Oxford University Press, New York, pp. 187-202.
- Herwig, R. *et al.* (2000) Information theoretical probe selection for hybridisation experiments. *Bioinformatics*, **16**, 890-898.
- Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109-126.
- Hughes, T.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnol.*, **19**, 342-347.
- Ishii, M. *et al.* (2000) Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, **68**, 136-143.
- Iyer, V.R. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83-87.
- Kane, M.D. *et al.* (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552-4557.
- Kononen, J. *et al.* (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.*, **4**, 844-847.
- Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.
- Lee, S.B. *et al.* (1999) The Wilms tumor suppressor WT1 encodes a transcriptional activator of amphiregulin. *Cell*, **98**, 663-673.
- Lockhart, D.J. and Winzler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827-836.
- Lockhart, D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.*, **14**, 1675-1680.
- Michael, K.L. *et al.* (1998) Randomly ordered addressable high-density optical sensor arrays. *Anal. Chem.*, **70**, 1242-1248.
- Okamoto, T. *et al.* (2000) Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nature Biotechnol.*, **18**, 438-441.
- Ross, D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.*, **24**, 227-235.
- Schadt, E.E. *et al.* (2001) Analyzing high-density oligonucleotide gene expression array data. *J. Cell Biochem.*, **80**, 192-202.
- Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.
- Shalon, D. *et al.* (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, **6**, 639-645.
- Southern, E. *et al.* (1999) Molecular interactions on microarrays. *Nature Genet.*, **21** (1 Suppl.), 5-9.
- Stomakhin, A.A. *et al.* (2000) DNA sequence analysis by hybridization with oligonucleotide microchips: MALDI mass spectrometry identification of 5mers contiguously stacked to microchip oligonucleotides. *Nucleic Acids Res.*, **28**, 1193-1198.
- Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science*, **270**, 484-487.
- Yeung, K.Y. *et al.* (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309-318.