Research

Within the fold: assessing differential expression measures and reproducibility in microarray assays

Ivana V Yang*, Emily Chen*, Jeremy P Hasseman*, Wei Liang*, Bryan C Frank*, Shuibang Wang*, Vasily Sharov*, Alexander I Saeed*, Joseph White*, Jerry Li*, Norman H Lee*, Timothy J Yeatman† and John Quackenbush*

Addresses: *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. †H. Lee Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, FL 33612, USA.

Correspondence: John Quackenbush. E-mail: johnq@tigr.org

Published: 24 October 2002

Genome ***Biology*** 2002, **3(11)**:research0062.1–0062.12

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/11/research/0062

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 22 May 2002
Revised: 28 August 2002
Accepted: 19 September 2002

Abstract

Background: 'Fold-change' cutoffs have been widely used in microarray assays to identify genes that are differentially expressed between query and reference samples. More accurate measures of differential expression and effective data-normalization strategies are required to identify high-confidence sets of genes with biologically meaningful changes in transcription. Further, the analysis of a large number of expression profiles is facilitated by a common reference sample, the construction of which must be carefully addressed.

Results: We carried out a series of 'self-self' hybridizations in which aliquots of the same RNA sample were labeled separately with Cy3 and Cy5 fluorescent dyes and co-hybridized to the same microarray. From this, we can analyze the intensity-dependent behavior of microarray data, define a statistically significant measure of differential expression that exploits the structure of the fluorescent signals, and measure the inherent reproducibility of the technique. We also devised a simple procedure for identifying and eliminating low-quality data for replicates within and between slides. We examine the properties required of a universal reference RNA sample and show how pooling a small number of samples with a diverse representation of expressed genes can outperform more complex mixtures as a reference sample.

Conclusion: Analysis of cell-line samples can identify systematic structure in measured gene-expression levels. A general procedure for analyzing cDNA microarray data is proposed and validated. We show that pooled reference samples should be based not only on the expression of individual genes in each cell line but also on the expression levels of genes within cell lines.

Background

DNA microarray analysis has become the most widely used technique for the study of gene-expression patterns on a genomic scale [1,2]. Differential microarray co-hybridization assays measure the relative gene expression of paired query and reference samples, and the power of microarray analysis

comes from identification of informative patterns of gene expression across multiple experiments. Achieving both these objectives is facilitated by using a common reference sample that provides a baseline expression measure for each gene, enabling normalization and comparison of independent experiments.

Pooled RNA derived from cell lines is a commonly used reference sample. To provide optimal coverage of genes spotted on the array, reference samples are often constructed from a large number of cell lines from a variety of tissues. One example is the universal human RNA reference commercially available from Stratagene [3]. This reference consists of equimolar quantities of RNA from ten human cancer cell lines representing ten different tissues (B cells, breast, brain, cervix, liver, lipocytes, macrophage, skin, T cells and testis).

Another challenging technical consideration in microarray analysis is the cutoff value used to distinguish differential expression from natural variability in the data. A cutoff of twofold up- or down-regulation has been chosen to define differential expression in most published studies [1,2]. However, little has been done to evaluate the accuracy of the technique and assess the confidence levels for various fold-level changes in expression ratios. In addition, microarray analysis is a complex, multistep technique involving array fabrication, labeling, hybridization and data analysis, and many laboratories have developed a variety of protocols for each of these steps [4,5]. Studies by a number of groups using a range of protocols and including many different RNA samples will give a better picture of how reliable microarrays are for elucidating gene-expression profiles.

In this study, we evaluate the performance of cDNA microarrays using our current laboratory and data-analysis protocols and derive a new intensity-dependent approach to identifying differentially expressed genes. RNA from 19 different human cancer cell lines, the Stratagene universal reference RNA, and RNA isolated from a tumor specimen were assayed in a series of 'self-self' hybridizations on a 19,200-element cDNA array (containing 9,600 elements spotted in duplicate). Statistical analysis of the ratios of Cy5/Cy3 fluorescence intensities among this large number of distinct samples provides insight into the variation inherent in expression ratios extracted from cDNA microarrays. We assess reproducibility of array experiments by analyzing replicates, using both clones spotted in duplicate on the same array and in triplicate hybridization assays. In addition, we use the methodology developed in this study to compare expression in a colon and an ovarian cell line to identify tissue-specific genes.

Self-self hybridization results for individual cell lines were also analyzed to determine the composition of an optimal reference pool consisting of RNA derived from cell lines. Our underlying hypothesis was that pooling a large number of cell lines might not necessarily improve the overall gene representation. Although some cell lines express significantly more genes than others, not all cell lines express all genes at the same levels. Consequently, mixing cell lines may dilute rare transcripts so that their representation in the final RNA pool is below the detectable limit. Results of our analysis will aid in the future assessment of gene-expression patterns and the design of reference RNA samples.

Results and discussion

Cell line self-self hybridizations

To evaluate the natural variability present in spotted cDNA microarray data across a number of distinct RNA samples, we carried out a set of self-self assays in which aliquots of the same RNA are separately labeled with Cy3 and Cy5 dyes and co-hybridized to a single microarray. Poly(A)$^+$ RNA was prepared from 19 human carcinoma cell lines (2 brain, 5 colon, 10 ovarian, 1 pancreatic, and 1 testicular) and Stratagene's universal human reference RNA. RNA was reverse transcribed into cDNA using random hexamer primers in the presence of 5-aminoallyl-dUTP; Cy3 and Cy5 dyes were covalently coupled to the incorporated aminoallyl linkers in a subsequent labeling reaction.

For each cell line, Cy3- and Cy5-cDNA samples were co-hybridized to a spotted microarray containing 19,200 human cDNA clones (9,600 clones printed in duplicate). Hybridization results were analyzed to determine relative expression levels for each printed element, and hybridization intensity data were first normalized globally using an iterative mean-$\log_2$(ratio)-centering approach. Briefly, $\log_2$(ratio) values were calculated for each array element and the mean of the distribution was calculated. Ratios were adjusted such that the mean $\log_2$(ratio) for the entire collection of genes was set to zero(or a corresponding average ratio of 1). As outliers can significantly influence this process, these were identified and excluded and the process repeated until convergence. This process results in an average Cy5/Cy3 $\log_2$(ratio) of zero for the spots analyzed in each microarray.

We and others have noted that the $\log_2$(ratio) values often have a systematic dependence on intensity most often observed as a deviation from zero for low-intensity spots. Locally weighted linear regression (lowess) [6] has been proposed as a normalization method for microarray assays [7,8] that can remove intensity-dependent dye-specific effects in the $\log_2$(ratio) values.

In this procedure, fluorescence intensities are measured from both channels for all elements on the array and the $\log_2$(Cy5/Cy3) ratios for each spot are represented as a function of the $\log_{10}$(Cy5*Cy3) product intensities. Scatterplots of such data, referred to as an 'R-I plot' (for ratio-intensity), can reveal intensity-specific artifacts in the measurement of the ratio, which tend to occur most notably

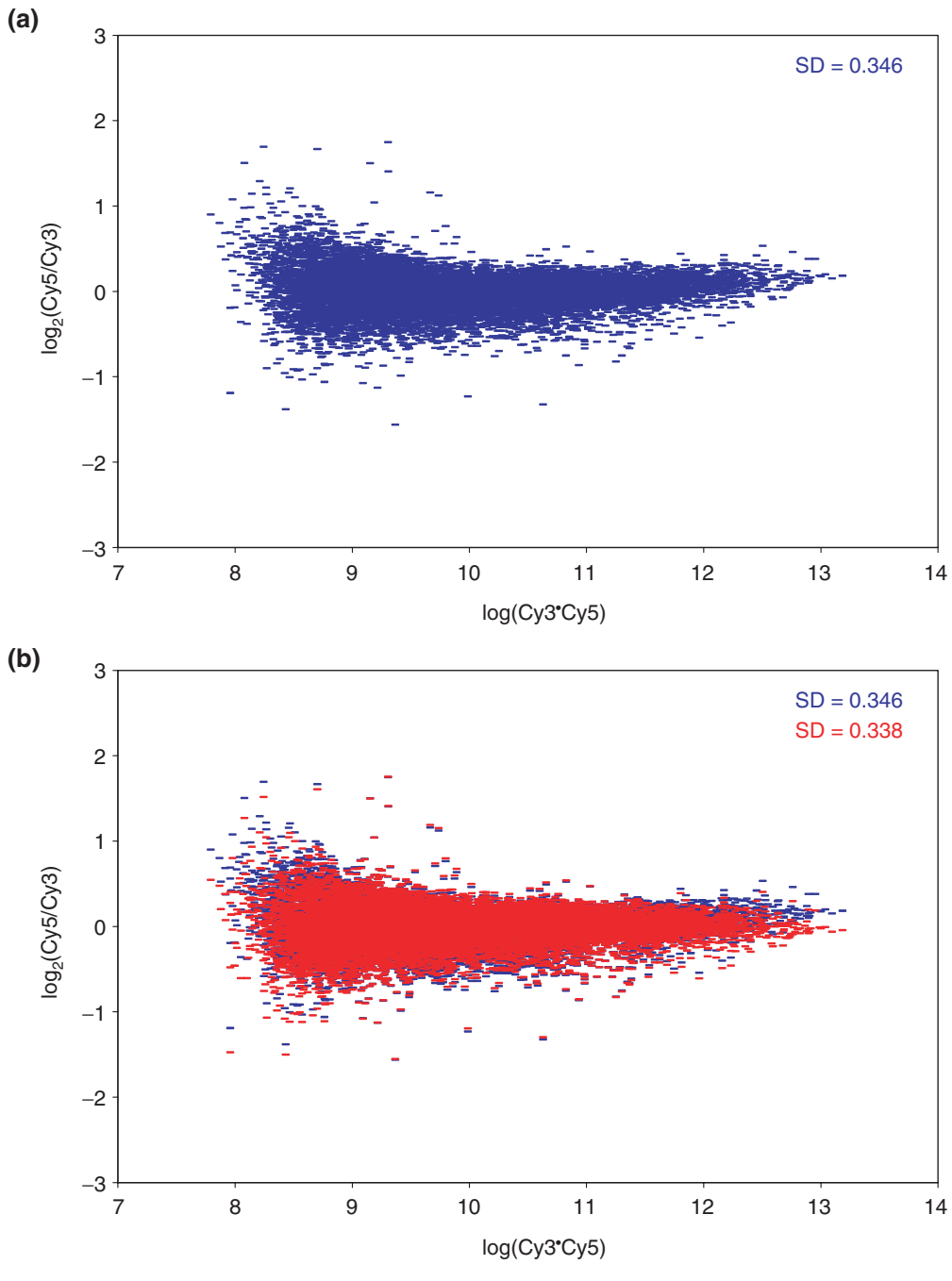

Figure 1 *(continues on the next page)*
Self-self hybridization of the KM12L4A cell line. **(a)** R-I (ratio-intensity) plot for a self-self hybridization of the KM12L4A cell line before lowess correction. **(b)** The same dataset, showing the effect of lowess correction (red) relative to the uncorrected data (blue). Lowess removes the intensity-dependent curvature that is evident in the uncorrected data and in the process, reduces the SD in the dataset. **(c)** Similar plots for all 30 self-self hybridizations performed in this study, including the SD for the dataset before (blue) and after (red) the application of the lowess correction.

Figure 1 *(continued from the previous page)*

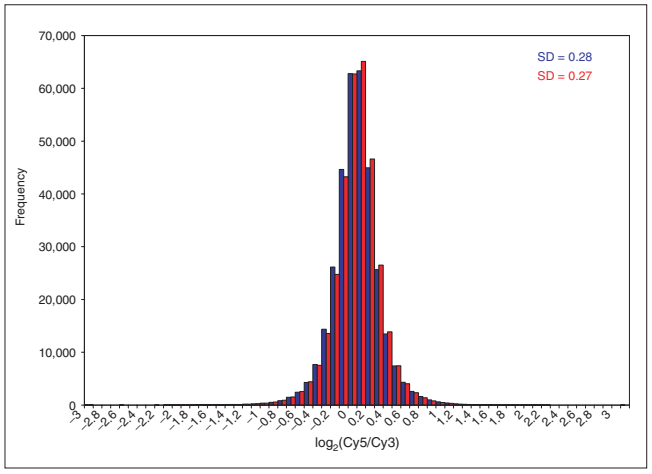

Figure 2
A combined histogram of the $\log_2$(expression ratio) measured for all array elements across all 30 hybridizations used in this study both before (blue) and after (red) application of lowess correction.

for weakly fluorescing arrayed elements. For a self-self hybridization of the type analyzed here, one expects a mean $\log_2$(Cy5/Cy3) ratio of zero for each element in the array independent of intensity. Lowess enables deviations from this expected behavior to be detected and corrected by performing a local weighted linear regression for each data point in the R-I plot and subtracting the average ratio from the experimentally observed ratio. A representative R-I plot for the cell line KM12L4A in Figure 1 shows a slight upward curve of the $\log_2$(ratio) for low-intensity data points (shown in blue) which is removed after application of the lowess correction, producing a balanced distribution of expression ratios around zero independent of intensity (shown in red). Similar plots, including a calculated standard deviation (SD) both before and after lowess correction, are shown in Figure 1c. The most important feature of the data is the narrow distribution of ratios in individual array experiments regardless of the RNA sample assayed in each experiment. This is also true of the entire dataset, as can be seen in the histogram of combined expression ratios from 30 experiments, shown in Figure 2.

These results are comparable to those obtained in a recent study of 30 self-self hybridizations on 10,000-element cDNA arrays [9]. Yue *et al.* [9] created histograms for three sets of data, each consisting of 10 independent array experiments, using mRNA derived from human placenta, brain and heart tissue. In one of their data sets, a 2-SD limit of 1.25-fold was observed; combined, 99.5% of the data points were within ± 1.4-fold. A small number of distinct samples (3) in combination with a large number of replicates per sample (10) most probably accounts for the slightly tighter distribution of expression ratios than that observed in our study.

The second important aspect of our dataset is the number of 'outliers'; approximately 5% of the data points in each array fall outside of 2 SD from the mean (data not shown). While this is not unexpected, it suggests that the distribution of $\log_2$(ratio) values does not deviate badly from normal. This is of a particular significance because it should enable detection of genes with low levels of differential expression at high confidence in future studies. Combining all data from all 30 hybridizations, the 2-SD limit is equivalent to 1.46-fold change (± 0.546 $\log_2$(ratio)), only 18,221 out of 332,601 points, or 5.5% of the expression ratios, fall outside a 2-SD threshold. Consequently, if differential expression is defined as greater than 2 SD from the mean, genes with fold changes greater than 1.5 can be classified as up- or down-regulated at approximately 95% confidence level.

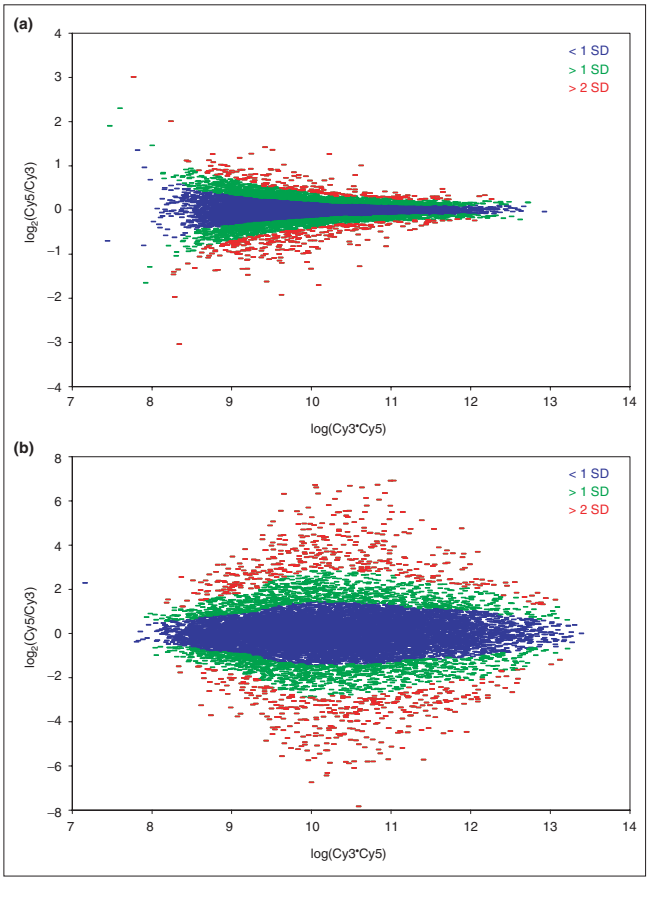

Figure 3
Intensity-dependent calculations of SDs described in the text show distinct patterns that depend on how closely related are the samples being compared. **(a)** The 'tadpole' pattern seen in the self-self hybridization of RNA samples from the KM12L4A cell line is characteristic of RNA samples derived from similar sources are compared. **(b)** RNA samples from very different samples show a characteristic 'eye' pattern, with greater diversity of expression for genes expressed at intermediate levels, as seen in this co-hybridization of a Cy5-labeled PA-1 (ovary) with a Cy3-labeled CaCO2 (colon) RNA sample.

Good-quality array data can only be generated using well developed and extensively tested laboratory and data-analysis protocols. Our group has been working continuously to improve our laboratory procedures and implement new data-analysis methods. One of the changes recently adopted by ours and many other laboratories is indirect cDNA labeling. As first-strand cDNA synthesis is identical for both samples, incorporation of nucleotides with an aminoallyl linker followed by the covalent coupling reaction to Cy dyes reduces bias for incorporation of one dye over the other by the reverse transcriptase. Slight dye-specific effects are present in normalized low-intensity data, probably caused by a wavelength-dependent differential response of the photomultiplier tubes in the array scanners or slight differences or nonlinearities in the quantum efficiencies of the fluorescent dyes, but these can be efficiently removed by the application of the lowess correction.

Tissue sample self-self hybridization

In our experience, hybridization using RNA derived from tissue samples is generally more problematic than that derived from cell lines. To determine whether data of comparable quality can be generated with tissue RNA, we carried out a similar expression analysis using RNA derived from a colon tumor liver metastasis. Twenty micrograms of total RNA that have been extracted using Trizol and subjected to secondary purification using Qiagen RNeasy columns were labeled with both Cy3 and Cy5 and examined in a self-self co-hybridization assay. The distribution of expression ratios is in excellent agreement with cell-line data; only 507 out of 11,743 (4.3%) of identified spots fall outside the 1.42-fold change, 2-SD limit.

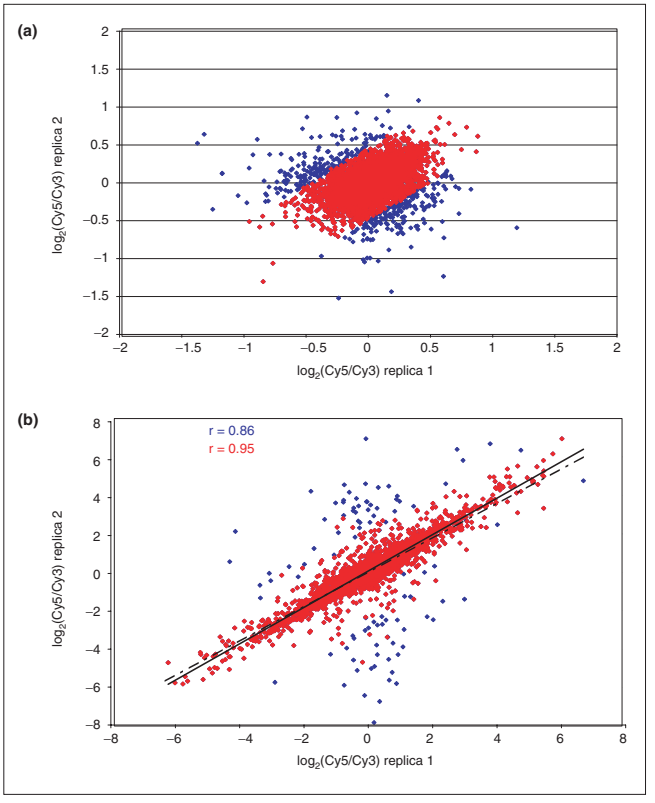

Figure 4
Replicate filtering within an array can reduce variability in the data. Scatterplots showing correlation coefficients (r) for the logarithms of the Cy5/Cy3 ratios for duplicate spots within arrays for **(a)** a self-self hybridization of RNA samples from the CaOV3 cell line and **(b)** a co-hybridization of a Cy5-labeled PA-1 with a Cy3-labeled CaCO2 RNA sample. In both cases, data before replicate filtering (blue) includes a number of outliers that are eliminated from the filtered data (red), resulting in a much better correlation between duplicate measurements

Intensity-dependent estimation of differential expression

While lowess normalization greatly reduces dye-specific artifacts that often appear for low-intensity data points, the data exhibit additional structure that can be used to evaluate patterns of gene expression. As can be seen in Figure 1a, the $\log_2$(ratio) measures generally show greater variation at lower intensities. Most published microarray studies use a single $\log_2$(ratio) threshold as a measure of differential expression. Examination of the distribution in Figure 1 suggests that this may inappropriately identify genes at intensities where the natural variation in the data would not support their selection, while at the same time causing genes to be missed in other intensities where the data suggest that much lower $\log_2$(ratio) values are statistically significant. Several mathematical models have been derived to explain this phenomenon [10-14]. One [10] discusses the use of a smoothed estimate of the SD as a function of the fluorescence intensity. The study conducted by Hughes *et al.* [11] shows how their model for estimating intensity-dependent differential expression can be used to identify biologically meaningful differential regulation at levels lower than twofold in a compendium of 300 different *Saccharomyces cerevisiae* mutants and chemical treatments.

We developed a much simpler approach to identify differentially expressed genes using an intensity-dependent calculation of a standard *Z*-score. Using a sliding window of fixed width in $\log_{10}$(Cy5*Cy3), the local mean and SD was calculated for each gene using the data in the normalized R-I plot. Figure 3a depicts the result of such a calculation for the cell line KM12L4A, with data less than 1 SD shown blue, between 1SD and 2SD in green, and greater than 2 SD in red, respectively; genes greater than 2 SD from the local mean ($Z > 2$) are identified as being significantly differentially expressed.

The shape of the R-I plot and the *Z*-score distributions are characteristic of the samples being compared in the microarray assay under consideration. For closely related samples, such as RNA from a cell line and the same line perturbed with a stressor, the R-I distribution has a 'tadpole' shape similar to that shown for our self-self hybridization, although typically with a slightly broader distribution in the $\log_2$(ratio) than that shown in Figure 3a. When two very different samples or

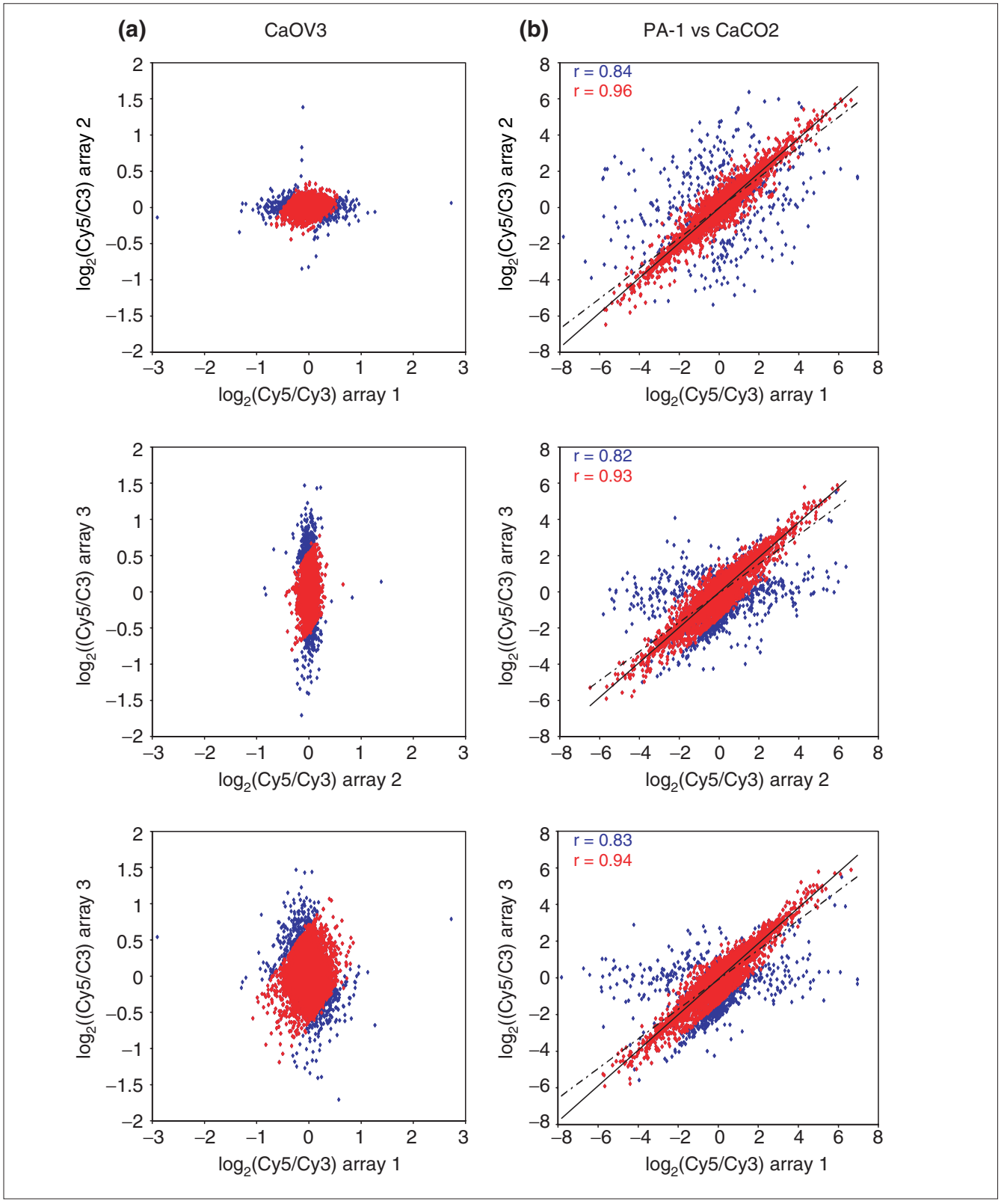

Figure 5
Replicate filtering between slides can also significantly improve data quality. Scatterplots showing correlation coefficients (r) for the logarithms of the Cy5/Cy3 ratios for duplicate spots within arrays for three arrays used to analyze independently labeled sets of **(a)** CaOV3 RNA samples (self-self hybridizations) and **(b)** PA-1 (Cy5) and CaCO2 (Cy3) RNA samples.

tissues are compared, however, the distribution generally has a characteristic 'eye' shape, with the greatest spread coming at intermediate expression levels. Figure 3b shows the results from a hybridization comparing the CaCO2 (colon) and PA-1 (ovary) cell lines.

Analysis of replicates

Expression data for clones spotted in duplicate were examined to assess reproducibility within each array. Plots of the $\log_2$-transformed ratios of the Cy5 and Cy3 intensities for the two replicates were generated for this purpose. As can be seen in the representative plot in Figure 4a (blue), $\log_2$(Cy5/Cy3) values form a 'sphere' centered around zero with the majority of data points lying between - 0.5 and + 0.5 on both axes. However, a small number of outliers is present and probably represent situations where one or both of the replica spots are of poor quality. Similar plots were obtained for the other 29 self-self hybridization assays (data not shown).

Because one cannot determine *a priori* which of the duplicates is in error, irreproducible elements should be removed before any further analysis, unless many more replicates are analyzed. The idea of outlier filtering was also incorporated into the model for the intensity-dependence of expression ratios introduced by Baggerly *et al.* [10]; we developed a simple approach to flag and eliminate questionable replicates. Ideally, Cy5/Cy3 ratios should be the same for the two replicas and $\log_2(r_1/r_2)$ (where r_1 and r_2 are Cy5/Cy3 ratios for the two replicas) should be equal to zero; replicates where this condition is not satisfied probably contain one or more bad elements. Consequently, we sought to filter out genes whose $\log_2(r_1/r_2)$ deviates greatly from this expected value of zero. We calculated the mean and SD of this $\log_2(r_1/r_2)$ for each pair of replicas in the entire array, and eliminated pairs of elements whose $\log_2(r_1/r_2)$ is greater than 2 SD from the mean. In Figure 4a, 6.3% of data points were eliminated using this outlier criterion (red), resulting in a much tighter dataset.

We also examined duplicate clone filtering in a hybridization assay of two different samples, CaCO2 and PA-1 (Figure 4b). In the case of differential expression, the correlation of the replica ratios is expected to be linear. The correlation increases from r = 0.86 (blue; dashed line) to r = 0.95 (red; solid line) after 1.7% of outliers are filtered out using the method outlined above. The substantial increase in the correlation coefficient suggests that our filtering procedure efficiently removes elements with at least one unreliable $\log_2$(Cy5/Cy3) value.

A similar approach can be used to identify questionable replicate spots on separate slides. To show this, five cell lines (CaOV3, HCT-116, KM12L4A, NT2/D1, and SW480) were selected and assayed in triplicate. In addition, expression in CaCO2 and PA-1 was compared in triplicate hybridizations. RNA from the same isolation was used for replicate hybridizations, thus allowing us to concentrate on technical and not biological replication and to explore more thoroughly the systematic errors that can arise. Replicate slides were subject to filtering in pairs (arrays 1 and 2; 2 and 3; 1 and 3) using a process analogous to that applied to within-slide replicates; representative data are shown in Figure 5. Correlations of $\log_2$(Cy5/Cy3) values for the three pairs of arrays before (blue; dashed line) and after (red; solid line) filtering for the CaOV3 cell line self-self hybridizations are shown in Figure 5a. The same set of plots for differential expression assays of CaCO2 and PA-1 cell lines including linear fits to the data is shown in Figure 5b.

Triplicate hybridization assays enable us to assess variability among independent labeling reactions and hybridizations. As can be seen in Figure 5a, there is some variation in correlations for each pair of hybridization replicas (arrays 1 and 2, as well as 2 and 3, show better correlation than do arrays 1 and 3). When averaged over three pairs of replicas, however, the variability between arrays is comparable to within-array variation. Similar results were obtained for the remaining four cell lines (data not shown). For the case of hybridizations of two different samples (Figure 5b), the mean r values increase significantly from 0.83 (blue; dashed lines) to 0.94 (red; solid lines) after outliers in each pair of hybridizations are filtered out. These correlations are, as in the case of self-self hybridizations, similar to those seen for within-array replicas.

Our analysis suggests that replication is essential for ensuring data quality, whether spotting replicate clones on single arrays or performing multiple independent hybridization assays. Further, the replicate filtering process we describe here is a simple approach that can eliminate questionable array elements and provide higher-confidence expression measurements.

Identification of differentially expressed genes

To show how the concepts presented in this study can be applied to detect differentially regulated genes, we sought to identify sets of genes that are over- and under-expressed in ovary relative to colon by analyzing hybridization assays of CaCO2 (colon) and PA-1 (ovary) cell lines. The CaCO2 sample was labeled with Cy3 and the PA-1 sample was labeled with Cy5; genes that are expressed in colon at much higher levels will, therefore, have negative log-ratios and ovary genes will have positive log-ratios.

For this analysis, we applied the process outlined in the previous discussion as follows. First, individual arrays were normalized using mean-log-centering and lowess procedures (as outlined in the Cell line self-self hybridizations section). Next, unreliable spots among triplicate arrays were eliminated using the filtering approach discussed in the Analysis of replicates section. Then, intensity-dependent SDs on the $\log_2$(ratio) were computed for each array, as described in the

Intensity-dependent estimation of differential expression section. Array elements with ratios outside the 2-SD limit were defined as differentially regulated; 510, 543 and 523 were identified in each individual assay. Finally, the intersection of differentially expressed genes on all three arrays was taken; this resulted in a list of 324 array elements (225 genes, as some are within-array replicates), of which 157 (109 genes) have higher expression in colon and 167 (116 genes) are more highly expressed in the ovary relative to colon.

A list of differentially expressed genes with mean $\log_2$(Cy5/Cy3) values and SD calculated over three experiments is available online (see Additional data files). The small SDs illustrate the reproducibility of measurements made using the methods described here and show their ability to provide a high-confidence list of differentially expressed genes and their expression ratios.

In an attempt to validate the differential expression we observed in this assay, we compared our results to the expression patterns derived from EST libraries generated from 18 colon and 12 ovarian sources (including normal tissue, cancer tissue, and cell lines) using the digital differential display (DDD) available at the National Center for Biotechnology Information (NCBI) [15] (see Additional data files). Although these data sets are not completely complementary, we identified one gene that is significantly upregulated in the colon relative to ovary (*PLAB*, coding for prostate differentiation factor; Hs.296638; THC888554), and one gene that is overexpressed in the ovary relative to colon (24-dehydrocholesterol reductase; *DHCR24*; Hs.75616; THC932478) in both microarray and DDD data. We also identified three gene families significantly differentially expressed in both datasets. We found genes in the solute carrier families 34 (Hs.84700; DDD), 22 (THC960422; microarray) and 16 (THC863879; microarray), as well as laminin A5 (Hs.11669; DDD) and A4 (THC933239; microarray) to be more abundant in the ovary. In addition, serologically defined colon cancer antigens 28 (Hs.84700; DDD) and 33 (THC899239; microarray) were more highly represented in colon. However, we also found two discrepancies in relative abundances of one gene and one gene family between the

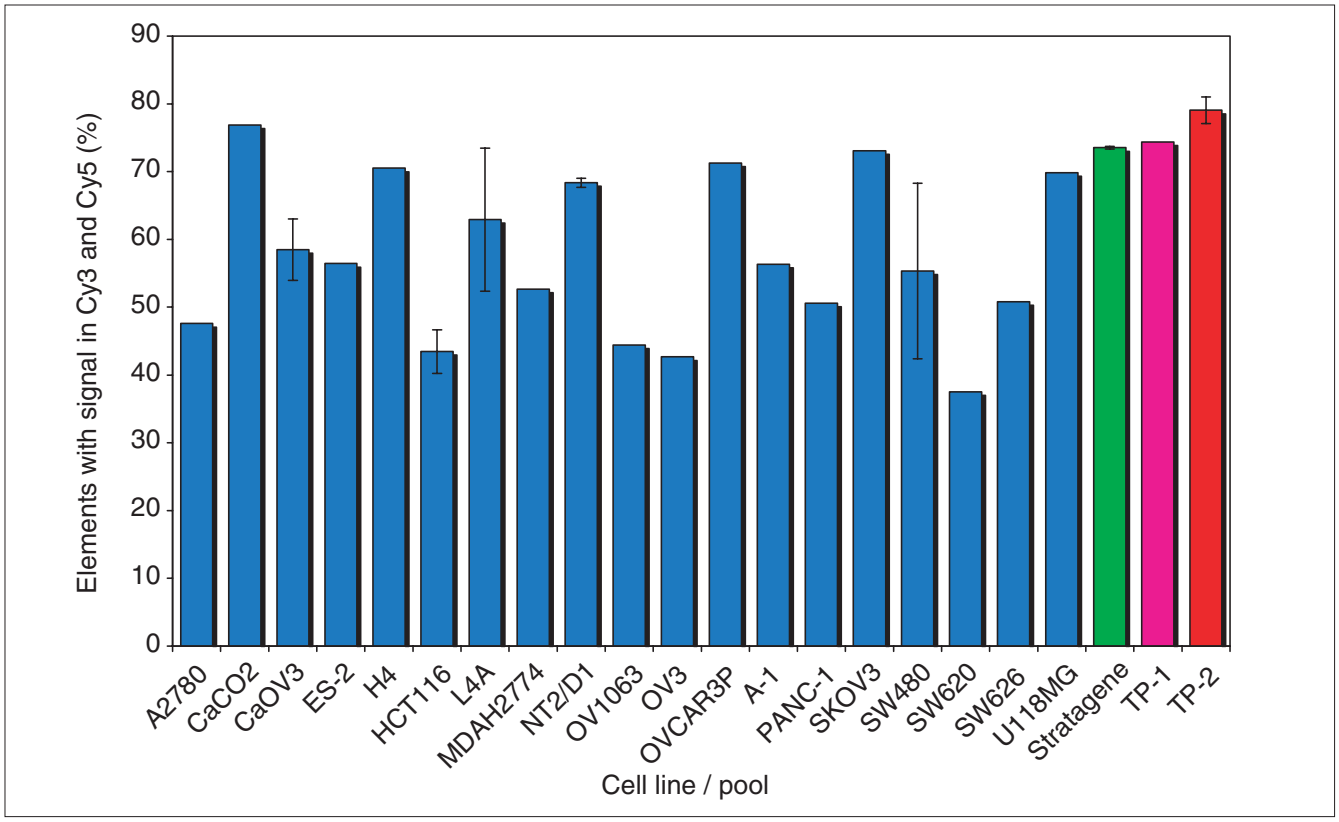

Figure 6
An ideal reference RNA sample will provide detectable hybridization above background for as broad as possible representation of the arrayed genes. The histogram shows the percentage of array elements with detectable signals in both the Cy3 and Cy5 channels for a series of self-self hybridizations representing all of the primary cell lines used in this study, the Stratagene universal reference RNA, and RNA pools created on the basis of our analysis. TP-1 consists of equal amounts of CaCO2 (colon), KM12L4A (colon), and OVCAR3 (ovary) cell lines. TP-2 consists of equal amounts of CaCO2 (colon), KM12L4A (colon), and U118MG (brain) cell lines. Mean values with 1 SD as the error bars are plotted for the samples that were assayed more than once. CaOV3, HCT-116, KM12L4A, NT2/D1, and SW480 cell lines were assayed in triplicate, and Stratagene and TP-1 pools were hybridized in duplicate.

DDD and microarray data. DDD data indicate that the protein-kinase inhibitor P58 (Hs.177574; THC906738) is more abundant in the ovary but the same gene appears to be more highly represented in the colon in our microarray study. Similarly, insulin-like growth factor binding protein 3 (Hs.77326; DDD) is upregulated in the ovary relative to colon while the opposite is true for insulin-like growth factor binding protein 6 (THC1022140; microarray).

To further validate our findings, we performed quantitative real-time reverse transcription PCR (RT-PCR) assays on 25 genes with $\log_{10}(R^*G)$ and $\log_2(R/G)$ values encompassing the entire range present in the microarray data. Expression ratios obtained by quantitative RT-PCR (see Additional data files), are in good agreement with the ratios obtained using microarrays. This concordance is especially significant for genes with low expression ratios and for transcripts expressed at low copy number, confirming the validity of our approach for obtaining a high-confidence list of differentially expressed genes. RT-PCR data also showed that the P58 protein-kinase inhibitor (Hs.177574; THC906738) is upregulated in the colon relative to ovary, which supports our array results and suggests that the discrepancy between microarrays and DDD is not due to technical limitations of microarray assays but reflects a true biological difference between the cell lines we compared and tissues used in DDD.

Reference pool selection

Comparison of query samples to a common reference sample is among the most widely used experimental designs for large-scale microarray studies [16,17]. Reference samples often consist of pools of RNA molecules derived from cell lines because cell lines provide a renewable source of large quantities of RNA. Each cell line expresses a distinct assortment of RNA species, so any particular cell line will provide measurable, baseline hybridization for only a subset of the genes on an array. Consequently, cell line RNA reference pools are typically constructed from a large number (10 or more) of cell lines derived from a variety of primary tissues. The idea underlying such a design is that by combining RNAs from diverse cell lines, one might obtain a more complete representation of the genes spotted on the array.

This approach, however, does not take into account the expression levels of individual genes in each cell line. While genes expressed at high levels will give detectable signal even when diluted by a factor of 10, many of the more rare transcripts may get diluted to below the detection limit. Therefore, we hypothesized that comparable or better representation of spotted genes might be obtained using fewer cell lines, each expressing a large number of diverse genes. To test this hypothesis, we constructed two reference pools, TP-1 and TP-2, using a naive gene-counting approach designed to include three cell lines with the greatest representation of unique genes that are also easy to grow and yield high quantities of RNA. TP-1 consists of colon cell lines CaCO2 and KM12L4A as well as the ovarian cell line OVCAR3; OVCAR3 is replaced with the U118MG (brain) cell line in TP-2.

The results of our analysis are summarized in Figure 6, which shows a histogram representing the percentage of array elements with signals in both Cy3 and Cy5 channels in self-self hybridizations. The performance of TP-1 is comparable to that of the Stratagene pool (approximately 75%), despite a large difference in the number of cell lines in each pool. This supports our hypothesis that adding more cell lines to the pool may not necessarily improve the overall gene representation because some genes are diluted below the detection limit.

Including cell lines from very dissimilar tissues may, however, give a better representation of the genes on the array. TP-2 contains a brain instead of an ovarian cell line and covers close to 80% of the spotted sequences. This improved performance relative to TP-1 is likely to be due both to the fact that brain exhibits the greatest diversity of transcripts (on the basis of based EST and serial analysis of gene expression, SAGE, analysis), and that the subset of genes expressed in brain is more disjoint with the genes observed in colon than in ovary. This illustrates that a simple pool of RNA from diverse cell lines can provide a superior reference.

Conclusion

DNA microarray assays performed using well optimized laboratory protocols can produce high-quality, reproducible data, although the use of replicates, particularly dye-reversal, or 'flip-dye' assays, are highly desirable. While global normalization of the data can help to provide meaningful expression ratios, a more sophisticated normalization using lowess allows for correction of intensity-dependent artifacts in the data. Together, the laboratory and analytical techniques described here can produce highly precise and accurate data. In the 30 assays we performed, the global 2-SD limit corresponded to an expression ratio between ± 1.33-fold and ± 1.62-fold induction or repression, with a mean of ± 1.47 (corresponding to $\log_2$(ratio) values of ± 0.41, ± 0.70, and ± 0.56, respectively), with fewer than 5% of the data points falling outside the 2 SD limit. This suggests that changes in gene expression smaller than the ± 2-fold commonly used can be reliably identified as differential expression. More careful analysis of the distribution of ratios as a function of intensity suggests that an intensity-dependent assessment of local SD of the distribution provides a better measure of statistically significant differential expression; essentially calculating a local *Z*-score. The same holds true for tissue RNA from tumor samples. Individual labeling reactions and hybridizations do not introduce significant variability, as judged from good correlations of replicate experiments. However, a small number of outliers present in replicate spots and/or arrays can be filtered out, underlining

the importance of replicates in generating high-quality expression data.

Pooled reference samples should be designed on the basis not only of representation of individual genes in each cell line but also of expression levels of genes within the cell lines. Adding more cell lines to the pool does not necessarily improve overall gene representation because rare transcripts become undetectable when diluted in a pool of a large number of cell lines.

Overall, the data presented herein, derived from a large number of RNA samples from cell lines and tissue, suggest that the laboratory and data analysis protocols we describe should allow for a more accurate and reproducible identification of differentially expressed genes than does the selection of an arbitrary, global fold-change threshold chosen independent of any measure of the natural variability in the data. We propose a model for analysis of cDNA microarray data consisting of three steps: lowess-normalization of individual arrays, followed by replica filtering to remove questionable elements, and finally estimation of the local *Z*-score for identification of differentially regulated genes.

Materials and methods

Array fabrication

cDNA clones were obtained from the Research Genetics sequence-verified human cDNA collection. Clone inserts were PCR-amplified directly from culture and purified according to a previously published protocol [5,18]. PCR from plasmid miniprep DNA was carried out for microtiter plates that had fewer than 85% single-band PCR products when amplified from culture. The overall success rate for single-band amplification was 88%. Clones were printed in duplicate from 50% DMSO onto SuperAmine slides (Telechem International) as described previously and cross-linked at 90 mJ using a Stratalinker (Stratagene).

RNA extraction

Cells were grown in a tissue culture incubator (37°C, 5% CO_2) in RPMI 1640 or DMEM medium (Life Technologies) supplemented with 10% fetal bovine serum, 200 μg/ml streptomycin and 200 U/ml penicillin. RNA was extracted with Trizol (Life Technologies) according to the manufacturer's protocol and stored at -80°C. mRNA selection from 100 μg total RNA using oligo-$d(T)_{25}$ Dynabeads (Dynal) and following the manufacturer's directions yielded 2-3 μg poly(A)$^+$-enriched RNA in 20 μl 10 mM Tris-HCl, pH 7.5. Agilent 2100 Bioanalyzer mRNA assays were carried out on 1-μl aliquots and ribosomal contamination was estimated to vary between 0 and 25%. The rest of the mRNA sample was used to prepare fluorescently labeled cDNA probes.

RNA from a colon tumor liver metastasis was extracted with Trizol (Life Technologies) and further purified on RNeasy columns (Qiagen) using the protocol supplied by the manufacturers. For each labeling reaction, 20 μg total RNA were used.

Probe preparation

First-strand cDNA synthesis was primed with 4 μg of random nonamers (New England Biolabs) or 6 μg random hexamers (Life Technologies) by heating at 70°C for 10 minutes, snap-cooling in dry ice/ethanol for 30 sec, and incubating at room temperature for an additional 5-10 min. Reverse transcription was performed in the presence of 500 μM each of dATP, dCTP, and dGTP, 200 μM 5-aminoallyl-dUTP (Sigma), 300 μM dTTP, 1x first-strand buffer, 10 mM dithiothreitol, and 400 U Superscript II (Life Technologies) in 40-μl reactions at 42°C for 3 h to overnight. Reactions were quenched by the addition of 10 μl of 0.5 M EDTA and RNA template was hydrolyzed by the addition of 10 μl of 1 M NaOH followed by heating at 70°C for 10 min. Reactions were neutralized with 10 μl of 1 M HCl, and cDNA was purified on QIAquick columns (Qiagen) according to the manufacturer's directions but substituting phosphate wash buffer (5 mM potassium phosphate pH 8.0, 80% ethanol) for buffer PE, and phosphate elution buffer (4 mM potassium phosphate pH 8.5) for buffer EB.

cDNA was lyophilized to dryness and resuspended in 4.5 μl of 0.1 M sodium carbonate pH 9.0 buffer. NHS ester (4.5 μl) of Cy3 or Cy5 dye (Amersham Pharmacia) in DMSO (dye from one tube was dissolved in 72 μl of DMSO) were added and reactions were incubated at room temperature in the dark for 1 h. Coupling reactions were quenched by the addition of 41 μl of 0.1 M sodium acetate pH 5.2, and unincorporated dye was removed using QIAquick columns. Labeling efficiency was determined by analyzing the whole undiluted sample in a spectrophotometer using a 50-μL MicroCuvette (Beckman). Total incorporated Cy dye ranged from 300-600 pmol and the ratio of unlabeled to labeled nucleotides was typically between 25 and 50.

Hybridization and image processing

Slides were prehybridized in 1.0% BSA, 5x SSC, 0.1% SDS for 45 min, washed by repeated dipping in MilliQ water twice and 2-propanol once, and air dried. Fluorescent cDNA probes were lyophilized to dryness and resuspended in 12 μl hybridization buffer (50% formamide, 5x SSC, 0.1% SDS). To combined Cy3 and Cy5 samples were added 20 μg Cot1 DNA and 20 μg poly(A)$^+$ DNA and samples were denatured at 95°C for 5 min, followed by snap cooling on ice for 1 min. Room-temperature probes were applied to a prehybridized array, covered with a glass coverslip (Fisher), and placed in a humidified hybridization chamber (Corning). Hybridizations were carried out at 42°C for 16-20 h, followed by washing in (5 min each): 1x SSC and 0.2% SDS at 42°C once, 0.1x SSC and 0.2% SDS at room temperature once, and 0.1x SSC at room temperature twice. Arrays were scanned using a GenePix 4000 dual-color confocal laser scanner (Axon).

Data were collected in Cy3 and Cy5 channels and stored as paired TIFF images.

Data analysis

Spots were identified and local background subtracted in the TIGR_Spotfinder software [19]. In the first step, a grid consisting of square cells is drawn around each array element. Spot segmentation is then performed using a histogram segmentation method that uses the distribution of pixel intensities to separate probable signal from background and a binary thresholding approach to identify spots, followed by a procedure to exclude disconnected features. Raw intensity for each element is obtained by first excluding saturated pixels and then summing all remaining pixel intensities inside the spot contour. The area outside the spot contour but inside the cell is used to calculate local background. Background per pixel is estimated as a median of the pixels in this area and is multiplied by the spot area to give an estimated spot background value. In the final step, this integrated background value is subtracted from the raw integrated spot intensity to produce the background-subtracted integrated intensities we use for further analysis. Furthermore, a quality control (QC) filter is used to remove questionable array features. Two criteria for spot rejection are a spot shape that deviates greatly from a circle and low signal-to-noise ratio. Spots for which the ratio of area to circumference deviates by more than 20% from the value for an ideal circle and spots containing fewer than 50% of pixels above the median background value are flagged and eliminated from further consideration. This approach has proved extremely robust to misidentification of the spot boundaries, and expression measures have shown it to be both reproducible and verifiable. The output data from Spotfinder is available as additional data files with the online version of this paper.

Lowess normalization was implemented using a native Java implementation in TIGR MIDAS (Microarray Data Analysis System), freely available through [19] and based on the LOCFIT package developed by Bell Labs [20]. For all normalization, the smoothing parameter was set to 33%. All additional data files and protocols used in this study are available from [21].

Quantitative real-time RT-PCR

Quantitative RT-PCR assays were performed on the ABI Prism 7900HT sequence-detection system using the Taqman Reverse Transcription Kit (Applied Biosystems) and QuantiTech SYBRGreen PCR Kit (Qiagen) using a two-step reaction with protocols supplied by manufacturers. Primers were designed in the Primer 3 Old Version [22] using default parameters with minor modifications (product size range 100-150; optimum T_m 60; minimum T_m 58; maximum T_m 61; maximum 3′ complementarity 1). Following the initial RT, 1 μl of the resulting reaction product amplified by PCR in a 20 μl reaction volume with 200 nM final primer concentration. Data were normalized to 18S ribosomal RNA Ambion QuantumRNA 18S Internal Standards Kit; the 18S primers and competimers were in a 3:7 ratio with the primers at 200 nM final concentration.

Additional data files

Spotfinder data files (tab-delimited text files) and tables (Word files) showing (1) Differentially expressed genes in PA-1 and CaCO2 cell lines, and (2) genes differentially expressed in colon and ovary as determined using digital differential display to analyze sequence abundance in EST libraries deposited in dbEST, are available with the online version of this paper.

Acknowledgements

We wish to thank Y. Wang, H. Kim, K.-Y. Kwong, M. I. Klapa, and H. Wang for helpful discussions and comments on the manuscript and J. Tsai and T. Currier for bioinformatics support for the microarray work The authors also wish to thank M. Heaney and S. Lo for database support, and V. Sapiro, B. Lee, J. Shao, C. Irwin, J. Neubrech, R. Karamchedu, M. Sengamalay and E. Arnold for computer system support. This project was supported by grants from the US National Cancer Institute, the National Heart, Lung, and Blood Institute, and the National Science Foundation.

References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
2. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci USA* 1996, **93:**10614-10619.
3. **Stratagene** [http://www.stratagene.com]
4. Eisen MB, Brown PO: **DNA arrays for analysis of gene expression.** *Methods Enzymol* 1999, **303:**179-205.
5. Hegde P, Qi R, Abernaty K, Gay C, Dharap S, Gaspard R, Hughes J, Snesrud E, Lee N, Quackenbush J: **A concise guide to cDNA microarray analysis.** *Biotechniques* 2000, **29:**552-562.
6. Cleveland W, Devlin S: **Locally weighted linear regression: an approach to regression analysis by local fitting.** *J Am Stat Assoc* 1988, **83:**596-609.
7. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15.
8. Dudoit S, Yang YH, Callow M, Speed T: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Statistica Sinica* 2002, **12:**111-139.
9. Yue H, Eastman P, Wang B, Minor J, Doctolero M, Nuttall R, Stack R, Becker J, Montgomery J, Vainer M, Johnston R: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic Acids Res.* 2001, **29:**e41.
10. Baggerly KA, Coombes KR, Hess KR, Stivers DN, Abruzzo LV, Zhang W: **Identifying differentially expressed genes in cDNA microarray experiments.** *J Comput Biol* 2001, **8:**639-659.
11. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
12. Ideker T, Thorsson V, Siegel AF, Hood LE: **Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data.** *J Comput Biol* 2000, **7:**805-817.
13. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *J Comput Biol* 2001, **8:**37-52.

14. Rocke DM, Durbin B: **A model for measurement error for gene expression arrays.** *J Comput Biol* 2001, **8:**557-569.
15. **National Center for Biotechnology Information: Digital Differential Display** [http://www.ncbi.nlm.nih.gov/UniGene/info_ddd.shtml]
16. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, *et al.*: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
17. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, *et al.*: **Diversity of gene expression in adenocarcinoma of the lung.** *Proc Natl Acad Sci USA* 2001, **98:**13784-13789.
18. Gaspard R, Dharap S, Malek J, Qi R, Quackenbush J: **Optimized growth conditions for direct amplification of cDNA clone inserts from culture.** *Biotechniques* 2001, **31:**35-36.
19. **The Institute for Genomic Research: software tools** [http://www.tigr.org/software]
20. **LOCFIT: local regression and likelihood** [http://cm.bell-labs.com/cm/ms/departments/sia/project/locfit/index.html]
21. **The Institute for Genomic Research: cancer pages** [http://cancer.tigr.org]
22. **Primer 3** [http://www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi#PRIMER_SELF_ANY]