BIOINFORMATICS

Vol. 20 no. 12 2004, pages 1955–1961
doi:10.1093/bioinformatics/bth188

A novel approach for increasing sensitivity and correcting saturation artifacts of radioactively labeled cDNA arrays

Troy D. Querec[1,†], Radka Stoyanova[2], Eric Ross[3] and Christos Patriotis[1,]*

[1]Department of Medical Oncology, [2]Division of Population Science and [3]Department of Biostatistics, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111-2497, USA

Received on July 9, 2003; revised on January 29, 2004; accepted on February 8, 2004
Advance Access publication March 25, 2004

ABSTRACT

Motivation: The radioactivity labeled DNA array platform is a robust and accurate way for a high-throughput measurement of gene expression levels in biological samples. Despite its high degree of sensitivity and reproducibility, this platform has several sources of variation. These are related to the presence of saturation effects in the array images and impede the degree of accuracy at which gene expression levels are determined.

Results: Here we describe a simple, but effective, approach for combining expression data from a series of autoradiographic exposures of variable length. This technique increases the sensitivity of this array platform by detecting low-expressed genes at longer exposures. It also improves the measurement accuracy of highly abundant genes by considering only values from the linear portion of dependency between the exposure times and gene intensities. As a result, the described approach improves the outcome of the subsequent steps of array data normalization and mining.

INTRODUCTION

The advent of cDNA arrays has created the possibility for the parallel analysis of the expression profiles of thousands of genes in individual cell populations, simultaneously (Bowtell, 1999; Debouck and Goodfellow, 1999; Duggan *et al*., 1999; Lander, 1999). The level of expression of a given set of genes within the sample corresponds to the intensity of a labeled cDNA probe synthesized from the purified messenger RNA, and bound specifically to the cDNAs of the genes included in the array. Typically, PCR-amplified cDNAs or oligonucleotides, representative of hundreds to thousands of genes, are deposited on specifically coated glass microslides, or alternatively, onto negatively charged, synthetic polymer membranes (filter arrays) (Bowtell, 1999; Cheung *et al*., 1999; Duggan *et al*., 1999; Lipshutz *et al*., 1999; Ramsay, 1998; Schena *et al*., 1998). Glass DNA arrays are usually hybridized to one or two cDNA probes labeled with different fluorescent dyes, and the hybridized gene signals are detected by scanning the array with a confocal laser scanner at corresponding wavelengths. Filter arrays, on the other hand, are probed with either a ^{32}P- or ^{33}P-labeled cDNA (Gress *et al*., 1992; Lennon and Lehrach, 1991; Zhao *et al*., 1995), and the array image is revealed through autoradiography, either by exposure to an X-ray film or by phosphor-imager scanning. Determining the intensities of the spots in the array images gives a relative quantification of the original mRNA levels in the studied sample. The accurate extraction of gene intensity values from the array image, therefore, is essential for subsequent data analysis and interpretation. Substantial effort has been dedicated to developing software for extraction and statistical analysis of gene intensities from glass and filter array images, including ScanAlyze (Eisen and Brown, 1999), ImaGene and GeneSight (BioDiscovery, Inc.), AtlasImage and AtlasNavigator (BD Biosciences Clontech), as well as ArrayExplorer©, a software developed by the authors (Patriotis *et al*., 2001).

There are numerous advantages of the radioactively labeled array platform over other alternative array technologies, of which, the most important are their higher signal detection sensitivity and superior reproducibility (Bowtell, 1999; Duggan *et al*., 1999). The increased sensitivity is the result of the nature of the radioactive label of the probe, which at a sufficiently long exposure time will 'activate' the light-sensitive X-ray emulsion. However, exposure time selected to maximize the detectability of genes with low levels of expression may impede the quantification of the highly expressed genes due to signal saturation, occurring when the radiation from these genes exceeds the maximum detection limit of the X-ray film or phosphor-imager. Hence, shorter exposure times are optimal for accurate quantification of highly expressed genes.

*To whom correspondence should be addressed.

†Present address: Emory University, GDBBS, 1462 Clifton Road, Dental Building, Suite 314, Atlanta, GA 30322, USA

Hybridization signals of highly expressed genes result in spots on the X-ray film with increased diameter, causing substantial interference to the neighboring signals by affecting their background values and, in some cases, partly or entirely covering them. Thus, there are two main types of errors introduced as a consequence of saturation:

- Bias toward underestimation of the intensities of high-intensity spots at long exposure times.
- Overlap (overshining) of spots by neighboring saturated signals (Herzel *et al*., 2001; Schuchhardt *et al*., 2000).

We have dealt, to a certain extent, with these aspects of signal saturation in our custom-developed software (ArrayExplorer©). The regions for estimating the average spot intensities are called circular scanning regions (CSRs) and their diameter is estimated automatically. Each CSR is expanded until the average intensity value in its one-pixel-thick rim reaches the sum of the previously determined average background value plus three SDs of the background noise. If this criterion is not met, then the gene spot is indicated with a flag, or set automatically to zero, if no signal is detected. Alternatively, the CSRs are expanded automatically until the spots resulting from strong signals are entirely encompassed. The user can also expand or reduce the size of the CSR or relocate appropriately its position within a square using the operational window for manual fine-tuning. Regardless of the flexibility of the software to accommodate large spots resulting from strong hybridization signals, it still cannot account for the loss of energy deposited in saturated areas of the X-ray film. For spots partly overshone by the neighboring ones, the 'pie-sectoring' option allows the estimation of the overall spot intensity on the basis of only the pixels within the non-overlapping portion of the spots. In cases when it is clear that a spot is completely covered by the large signal in the neighboring spot, the intensity of the corresponding gene is typically set to zero. Thus, the software, while flexible enough to deal with some artifacts introduced by saturation, is ineffective to quantify spots accurately in several of the cases mentioned above. In this study, we propose a method which successfully deals with the remaining sources of error in the estimation of gene intensities associated with signal saturation. The approach utilizes data from a series of different lengths of autoradiographic exposures. When the exposure times are corrected for loss of radiation due to the natural decay of the radioactive label, the measured expression levels of a given gene should be in linear relationship with the thus adjusted times of exposure. Intensities of genes, which fail this relationship, are discarded from further consideration (as in the case of saturated spots in subsequent exposures). Regression is then used to estimate more accurately the expression level of a given gene. Alternatively, underexpressed genes can be detected and quantified at longer exposure times. Thus, the proposed procedure improves detection by using the sensitivity of the longer exposures, while retaining the higher precision of exposures at shorter time points.

MATERIALS AND METHODS

Cell lines and treatments

A series of cell lines was derived from human ovarian surface epithelial (HOSE) cells, isolated from ovaries which were removed for prophylaxis of ovarian cancer-prone individuals, and with biological characteristics ranging from normal to overtly malignant. Increased *in vitro* life span of the cells was achieved by the transduction of the SV-40 large T-antigen (unpublished data). Two series of cell lines (HIO-117 and -118), each comprising three independent clones [mortal, immortal and tumorigenic (NuTu)], and one series (HIO-135) with mortal and immortal clones were included in the study. Cells were maintained in Medium199 mixed with MCDB105 medium (1:1) (Sigma) supplemented with 5% fetal bovine serum, penicillin (100 units/ml), streptomycin (100 μg/ml), L-glutamine (0.2 mM) and insulin (10 μg/ml). Individual cultures of each cell line were subjected to *in vitro* treatment with the synthetic retinoic acid derivative, Fenretinide (4-HPR; 5 μM) for 24, 48 and 72 h. Non-treated cultures (0 h) were used as controls. 4-HPR was obtained from DCPC Repository (Rockville, MD). Treated and untreated cells were harvested and used to purify total RNA, according to the procedure provided by the array manufacturer (ClonTech).

Acquisition of arrays at different exposure times

Array data was generated as follows: α-^{32}P-[dATP] was used to reverse transcribe 3–5 μg of the total RNA into cDNA following the protocol provided by the array manufacturer (ClonTech). The obtained ^{32}P-labeled cDNA probe was hybridized to a filter array containing 1176 genes (Atlas 1.2 Human Cancer cDNA Array, ClonTech, Catalog no. 7851-1), according to the manufacturer's instructions. The radioactively labeled array filter was exposed to BioMax MS film (Kodak). A typical experiment consisted of acquiring images between 12 and 96 h of adjusted exposure times, starting with an initial intermediate exposure at 48–50 h. Depending on the relative overall intensity of the cDNA probe, as judged by visual inspection of this initial exposure and comparison with other array images in the experiment, either one shorter (24 h) and two longer (72 and 96 h) exposures were obtained, or alternatively, two shorter (12 and 24 h) and one longer (72 h). The obtained autoradiographic images were then scanned with a MicroTek ScanMakerIII flatbed scanner at 16 bit/1200 dpi (25 μm) resolution and exported into bitmap-format image files. Array images obtained from individual exposures were then subjected to densitometric analysis using ArrayExplorer© (Patriotis *et al*., 2001) to extract the gene intensities. Briefly, after adjustment of the array image in Adobe Photoshop, a default grid was overlaid and aligned, so that each DNA spot fell grossly within

a grid-square. The average background noise was determined and CSRs were automatically aligned over the signal in each square and expanded to the size of each individual spot-signal, using criteria described in the introduction. The intensities were calculated as the sum of densitometric units from all pixels within each CSR, and were exported in a spreadsheet-format file for further analysis. In addition to the gene intensities, this file also contained information for gene identity, array location, flagging, and DNA and protein sequence database access. In total, 37 array datasets, including two sets of replicate array datasets were acquired with cDNA probes from 35 different RNAs.

Data analysis

Theory The most frequently used radioisotopes in this array platform, ^{32}P and ^{33}P, have relatively short half-lives (14.3 and 25.4 days, respectively). Therefore, to obtain gene intensities from successive autoradiographic exposures of the hybridized array, it is necessary to adjust the exposure time to account for the continuous loss of energy. If A_0 and A_1 are the radioactivity of an isotope at a given reference t_0 and a later time t_1, respectively, then:

$$A_1 = A_0 \mathrm{e}^{-k(t_1 - t_0)}, \qquad (1)$$

where k is the decay constant specific for a given isotope (for ^{32}P, $k = 2.02 \times 10^{-3}$ per hour, and for ^{33}P, $k = 1.137 \times 10^{-3}$ per hour).

The radioactivity of a source is a measure of the radiation, either in the form of electromagnetic waves or very high velocity particles, that carries energy through space (Sprawls, 1987). In radiology, exposure E denotes the amount of radiation delivered to a point. E is related to the amount of energy contained in the radiation and the duration over which the radiation hits a point. Thus, high-energy radiation over a short time frame can give the same exposure as low-energy radiation over a longer time interval. Given the decrease in radioactivity due to the decay of the isotopes, the exposure time needs to be adjusted to reflect the loss of energy, thus introducing the concept of the adjusted exposure time, T.

In autoradiography, optical density is the darkness of an X-ray film resulting from the chemical processing during its development. What allows one to compare the density D of the gene spots obtained from multiple exposures on film of a radioactively labeled array is the linear relationship between the adjusted exposure times and the optical density of the film across a wide range of exposures. However, it is important to note that the relationship between adjusted exposure times and optical density is non-linear for very low and very high exposures. If we assume that array exposures are conducted only within the linear portion of the film sensitivity (dynamic range), then $D \sim E \sim A$, and D can be substituted for A in Equation (1).

Assume that a radioactively labeled cDNA array is subjected to a series of m exposures, and let $\Delta t_i = t_{i_2} - t_{i_1}$ be the length (in min or h) of the i-th exposure in the series ($i = 1, 2, \ldots, m$). Let $D^j_{\Delta t_i}$ be the density of the j-th gene measured at the i-th exposure length. $D^j_{\Delta t_i}$ is both a measure of the gene's level of expression and its proportionality to the cumulative amount of radioactivity captured on the film for the exposure interval Δt_i. The measured gene intensity for each exposure interval, therefore, will not be directly proportional to Δt_i, due to the continuous radioactive decay of the isotopic label, but to an adjusted exposure time T_i, given below:

$$D^j_{\Delta t_i} \approx T_i = \int_{t_{i_1}}^{t_{i_2}} \mathrm{e}^{-kt} dt = \frac{1}{k}(\mathrm{e}^{-kt_{i_1}} - \mathrm{e}^{-kt_{i_2}}). \qquad (2)$$

It should be noted that Equation (2) is valid only when D can be assumed to be proportional to A. Conversely, this is not true for very low and high exposures, where for certain genes, such a behavior is evident by the existing non-linearity in the dependency between Ts and Ds. This provides the theoretical basis for the proposed approach for estimating the gene intensities by regressing the measured intensities from multiple exposures onto the exposure times adjusted for the isotope decay.

Generally, four different exposure times are sufficient to detect the regional linear behavior of the measured gene intensities with respect to T_is. For arrays hybridized to ^{32}P-labeled probes, the recommended exposure times range between 12 and 96 h. Using gene intensities measured at the four different exposures along the parameter T, one can estimate the actual intensities of the genes. For certain genes, the intensities are calculated on the basis of only a subset of the four exposures, where the linear relationship between Ts and Ds holds. This is necessary as the intensity of some genes is very low and, hence, undetected at the shorter exposure(s), or alternatively, very high and, consequently, saturated at the longer exposure(s). In order to determine the appropriate subset of exposures, three correlation coefficients (R, Pearson's correlation coefficient) are calculated, as follows: (1) between the Ts and the gene intensities (Ds) of all four exposures; (2) between the Ts and Ds of the first three exposures; and (3) between the Ts and Ds of the last three exposures. The set of exposures that yields the highest R-value is used for the estimation of the intensities by linear regression. The gene expression levels are then estimated at the last adjusted exposure time T in the series. If a gene is detected only during the longest two exposures, its expression level from the last one is used in the final calculations. It is clear that, using this method, we can estimate the gene intensities at any point along T, but we have selected the longest adjusted exposure time to be able to include genes, detected only at this exposure.

Method implementation

The described procedure is implemented in an EXCEL spreadsheet. The adjusted exposure times T are calculated using Equation (2) and placed in the top row of the spreadsheet

in the order from the shortest to the longest exposure. The measured gene intensities from the corresponding exposures are organized in the columns, each row thus containing the gene intensities from the four different exposures. Using the CORREL function, the three correlation coefficients R between the adjusted times and the gene intensities are calculated for each gene. R is set to zero if in each of the series there is a zero value for the gene intensity. The set of exposures with the highest correlation coefficient value are then used to calculate the intensities at the highest exposure time point, using the FORECAST function. Denote this value of the gene expression as 'estimated' in order to distinguish it from the actual measurement values. Alternatively, if the measured value of a gene is non-zeroed only in the first and second exposure [characteristic spots 'overshone' at the last two longest exposures (Patriotis *et al*., 2001)], then these two values are used for estimating the fitted line at the longest exposure time. In cases where the gene intensity is non-zero only in the longest one or two exposures, the value of the measurement from the last exposure is retained. This is to avoid the significant error introduced by regression using only two values. It should be noted that such an operation is inevitable in the case of overshone spots where only the first two measurement values are available. However, if a spot has non-zero intensity only at one of the first three exposures, then it is assumed as an artifact and is excluded from further analysis.

Experimental error

To assess the effectiveness of the procedure for reducing the variation between replicate experiments, we define a measure of experimental error, ε, as follows:

$$\varepsilon = \frac{\sqrt{\sum_{j=1}^{n}\left[\frac{(r'_j-\bar{r}_j)}{\bar{r}_j}\right]^2+\left[\frac{(r''_j-\bar{r}_j}{\bar{r}_j}\right]^2}}{n} \times 100, \qquad (3)$$

where r'_j and r''_j are the measured intensities of the j-th gene in the two replicates, and $\bar{r}_j$ is their average; n is the number of genes expressed in at least one of the arrays. This measure of experimental error, in general, has several components, including (1) error of detectability (when a gene is detected in only one of the two replicates); (2) saturation error of underestimation of the signal at the longest exposure due to saturation effects on the X-ray substrate; (3) error due to overshining and (4) measurement error. It is clear that, as defined in Equation (3), the presence of a zero and an observed value in the replicates, i.e. error of detectability, will dominate the contribution of the other error components. However, it should be pointed out that if a spot is saturated or 'overshone' in both of the replicate arrays, then the variations between the replicates will be small, but nevertheless the reported values of both the gene intensity and the experimental error will be underestimated. To investigate the magnitude of the bias in ε [Equation (3)] introduced by saturation or 'overshining', we performed the following simulations: we generated 1000 replicate array pairs, each containing 100 gene-intensity values. The true mean (on $\log_2$ scale) of the intensity of the j-th gene ($j = 1, 2, \ldots, 100$) in the n-th ($n = 1, 2, \ldots, 1000$) simulated dataset (μ_{jn}) was drawn from a normal distribution with mean 11.94 and variance 2.92. Simulated intensity values (I_{jnm}) for replicate m ($m = 1, 2$) were generated by adding an additional normal deviate with mean 0 and variance 0.7 to μ_{jn}. This last step was added to reproduce the between-replicate variation observed in the data. To simulate saturation, the intensity values of the highest 17% of the genes were truncated to 15, while the intensity of 10% of the remaining genes, selected at random, were zeroed to account for overshone spots. The parameters in this simulation were based on the data observed in one of the replicate datasets reported in this paper.

Results

To illustrate the approach, we selected data from a single array exposed at four exposure lengths: 73, 71, 18 and 68 h (Δt_i) with corresponding t_{i_1}: 0, 74, 187 and 252 h, and t_{i_2}: 73, 145, 205 and 320 h; from Equation (2), we calculate the following adjusted exposure times T_i: 67.9, 56.9, 12.1 and 38.2 h ($i = 1$–4). The measured intensities of four representative genes (labeled A–D) in the obtained datasets, with levels of expression ranging from weak to high, are presented in Figure 1 as a function of Δt_i (left panel) and T_i (right panel). The non-linearity between the non-adjusted exposure times and measured intensities is clear, while a linear relationship holds when adjusted exposure times are calculated.

As described in Materials and methods section, three R-values were calculated between the measured intensities and the Ts. The R-values for gene A were 0.996, 0.999 and 0.995, respectively, which indicate that the last time point is 'off' the line of proportionality. In this case, the expression level of gene A can be estimated at time 67.9 h (indicated in Fig. 2 with an asterisk), using the line fitted to the first three exposures. The R-values for gene C were 0.864, 0.905 and 0.983, and therefore, its expression is calculated on the basis of the three longer exposures. Similarly, all four exposure times can be used to calculate the intensity of gene B. Finally, only the last exposure time is utilized for the measurement of gene D.

To investigate the effectiveness of the procedure to reduce the experimental error, we analyzed two sets of replicate experiments carried out as described in Materials and methods section: four exposures at corresponding times were acquired and the 'estimated' intensities were calculated at the time of the longest exposure. The results, including the number of the detected genes and the measure of experimental error [Equation (3)], are summarized in Table 1. The ranges of adjusted exposure times across the two replicate sets for exposures 1–4 were 8–12, 24–30, 48–54 and 70–76 h. These are shown in the order from shortest to longest and not in the order that actual exposures were carried out. It is clear

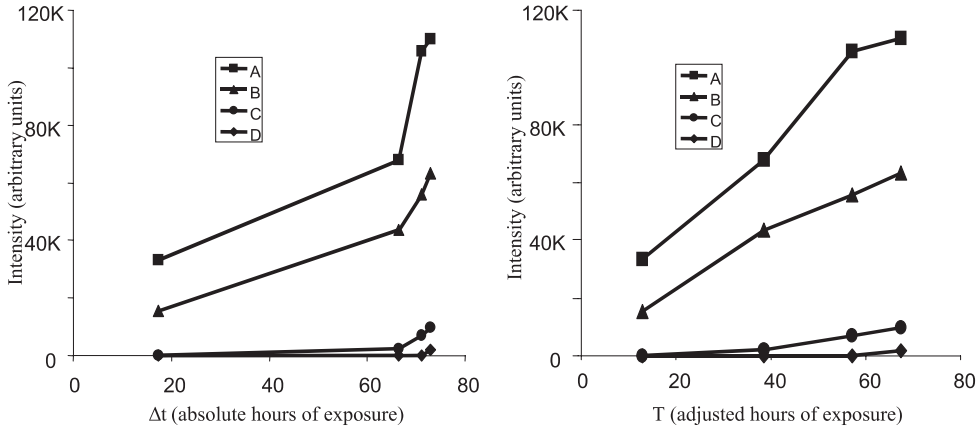

Fig. 1. Relationship between gene intensities and time of exposure. The intensities of four genes, A–D, are plotted as a function of absolute exposure time (left panel), and as a function of the adjusted exposure time (right panel).

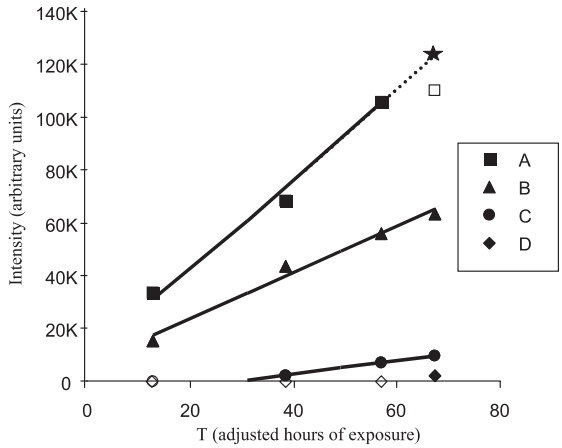

Fig. 2. Regression of gene intensities on the basis of adjusted exposure time. The intensities of the four genes in Figure 1, plotted against the adjusted exposure time with indicated lines of linear regression trough: for gene A, first three exposures; for gene B, all four exposures; and for gene C, last three exposures. The gene intensities are estimated at 67.9 h, based on the intensity values measured within the linear range (black shapes). Empty shapes depict intensity values measured within the non-linear segments of the exposures, the result of signal saturation or undetectable signal. The regression-estimated intensity value of gene A is shown with an asterisk.

from these data that the intensities estimated by the proposed procedure are superior, both in terms of the larger total number of genes detected and the smaller error between the replicate measurements. As expected, the number of detected genes is markedly increased relative to the first exposure (∼80%) and substantially increased relative to the second (∼60%) and third exposures (30%).

The average error of the first exposure relative to the error of the 'estimated' replicates is also the highest (more than double). This is mainly due to variations in the detection of low expressed genes: ∼20% of the detected genes had a non-zero value only in one of the two replicate arrays. It should be noted, though, that such a short exposure (6–12 h) is not a routine practice when the default method of analysis is used and, in our case, it was acquired for the utilization of the proposed procedure.

Another component of the error that is common when a single exposure is used for analysis is the measurement error. The latter is reduced in the proposed procedure by means of signal averaging: the final intensity calculation is based on multiple (four) measurements rather than one. Given the overall poor gene detectability in exposure 1 and the presence of significant saturation in exposure 4, this component of the error was further investigated in detail with respect to the second and third exposures only; these were carried out within the 20–50 h-range, which is also the common range utilized by the default method of single exposure-based analysis. To eliminate the contribution of the detectability and saturation errors, we selected genes with expression levels within the middle range [intensities between 5000 and 100 000 arbitrary units (a.u.)]. For these genes, in exposure 2 from replicate set 1, ε [Equation (3)] was 3.67%. The average error for the same genes in the 'estimated' replicate arrays was 3.18% (again, a smaller error than the overall average error), thus showing an ∼15% improvement of the regression-estimated values relative to the single exposure-derived intensities. A similar analysis for exposure 3 yielded an insignificant change in the error of measurement of all genes expressed within the middle range (3.48% versus 3.45%). However, it should again be noted that the overall error, as presented in Table 1, is 36% higher for the single exposure method, which is due to both lower gene detectability and, in the case of the longer exposures, overshining effects. The analysis of the measurement error in replicate 2 yielded similar results.

The difference in the error between the last exposure and the 'estimated' one is minimal. As noted in the Materials and methods section, the measurement of the experimental error

Table 1. Average number of detected genes and experimental error in single exposures and 'estimated' datasets of replicate arrays

		Exposure 1**	Exposure 2**	Exposure 3**	Exposure 4**	'Estimated'
Replicate set 1	Number of genes[a]	74	186.5	270	366	382
	Experimental error[b]	8.35%	5.66%	5.22%	*4.11%	3.85%
Replicate set 2	Number of genes	56.5	88	164	255	261.5
	Experimental error	10.01%	6.93%	8.15%	*4.91%	4.92%

[a]Estimated as an average of the number of genes expressed in the two replicate arrays.
[b]Estimated by Equation (3).
*These error values are biased due to the presence of saturated gene intensity values in the replicate datasets.
**The ranges of adjusted exposure times across the two replicate sets for exposures 1–4 were 8–12, 24–30, 48–54 and 70–76 h. These are shown in the order from shortest to longest and not in the order that actual exposures were carried out.

is quite insensitive to errors produced by saturation: a saturated spot can have a similar magnitude in the two replicates, but still be biased. Similar is the case with 'overshone' spots, which are most likely to be 'zeroed' in the highest exposure of both replicates. In the four arrays of the discussed two pairs of replicates, there were 45 (17.6%) saturated and 26 (10%) overshone spots out of a total of 260 genes found concurrently expressed in all four exposures. However, the proportion of overshone and saturated spots, as a fraction of all detected spots, is lower than in the case when only concurrently expressed genes in all exposures are considered. The mean average intensities of the saturated spots was 70 270 a.u.; the applied average correction was 5206 a.u. or about 7%; the range of corrections was from 111 to 24 315; the SD was 6158. The average correction applied for overshone spots was 117 611 a.u., with a range of 24 340–466 431, and an SD of 113 712. Given that the measurement value in the last one or two exposures in such cases is 'zeroed', the magnitude of the correction is, as expected, considerably larger than in the correction of saturated values. Our simulation study indicated that the experimental error ε [Equation (3)], assuming that 17% of the genes were saturated and 10% overshone, is underestimated by 16% relative to the error determined when the non-truncated or zeroed values were used.

The regression procedure was applied to 37 array datasets, including the two pairs of replicate arrays, each containing the gene intensities obtained from four different exposure times. The average number $\pm$SD of the detected genes was 58 ± 27, 110 ± 45, 180 ± 65, 270 ± 86 for exposures 1–4, respectively. On average, 277 ± 88 genes were detected per array from the 'estimated' data, which represent 78, 60, 35 and 2% more genes detected as compared to any of the single exposures 1–4, respectively. Due to overshining by neighboring, highly expressed genes, 2% of the genes in the array would remain undetected if only the longest exposure was taken into consideration. Finally, although the number of genes detected by the longest, fourth exposure is very close to that in the 'estimated' dataset, it should be noted that the intensity of 5–10% of the genes in the former case is erroneously determined due to saturation. We also investigated the number of genes overshone in both the third and fourth exposures. In this case, the estimated values are determined based on only two measurements and, thus, the introduced error is quite significant. However, only 16 genes in all 37 arrays were 'zeroed' in both the longest exposures, and hence, their contribution to the error in the proposed method is minimal.

DISCUSSION

In this report, we presented a procedure for the utilization of a series of autoradiographic exposures of radioactively labeled cDNA arrays in order to improve the detection of expressed genes and refine the measurement of the gene intensities. Performing gene expression analysis from multiple array exposures compensates for the inherent errors of long exposures, while increasing the number of detected genes, as compared with short exposures alone. As illustrated by our results, while the intensity of genes with an average level of expression can be linearly measured throughout the array exposure range, the measurement of highly expressed genes is often biased toward underestimation at long exposures. Conversely, the intensity of weakly expressed genes that are below the detection threshold at short exposures can be calculated from longer exposures. From a statistical point of view, to reduce the error in the estimated values, it would be desirable that one exposure is taken at a very short time (as close to zero as possible), while the rest are carried out at relatively longer times. In practice, however, such a short exposure would yield very few genes, while the longer exposures would be tainted by saturation effects. Thus, we recommend an initial exposure at an intermediate time point (48–50 h), which will allow the evaluation of the overall signal intensity of the particular cDNA probe. Depending upon that, further exposure time points are determined such that two shorter and one longer are taken if the probe intensity is too high, or alternatively, one shorter and two longer exposures.

The procedure described here also serves as an important spot-quality control procedure—genes with low correlation coefficients are flagged and investigated individually. Typically, the original array images are reviewed and the source of discrepancy identified.

The majority of the array normalization procedures determines normalization factors on the basis of averages over the behavior of the entire set of the measured genes. It is clear that, in the presence of saturated spots, these averages would be underestimated. Thus, the correction of the saturated values in the data by the proposed method will reduce the bias of the normalization factors within a series of arrays.

Finally, the described regression approach can also be applied in other types of experiments that utilize autoradiography, such as Northern, Western and Proteomic analysis from 2D gels, where individual samples may have highly variable expression profiles.

ACKNOWLEDGEMENTS

The authors would like to thank Drs S. Litwin and R. Dunbrack for critical review of the manuscript and for their suggestions for its further improvement. The work described in this report was supported by funds provided through NIH grants R29-CA73676 to C. Patriotis, P50-CA83638 (PI: R. Ozols), and Guzik Foundation Award to C.P. and R.S. C.P. is a Liz Tilberis Scholar of the OCRF, Inc.

REFERENCES

Bowtell,D.D. (1999) Options available—from start to finish—for obtaining expression data by microarray. *Nat. Genet.*, **21**, 25–32.

Cheung,V.G., Morley,M., Aguilar,F., Massimi,A., Kucherlapati,R. and Childs,G. (1999) Making and reading microarrays. *Nat. Genet.*, **21**, 15–19.

Debouck,C. and Goodfellow,P.N. (1999) DNA microarrays in drug discovery and development. *Nat. Genet.*, **21**, 48–50.

Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nat. Genet.*, **21**, 10–14.

Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Methods Enzymol.*, **303**, 179–205.

Gress,T.M., Hoheisel,J.D., Lennon,G.G., Zehetner,G. and Lehrach,H. (1992) Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm. Genome.*, **3**, 609–619.

Herzel,H., Beule,D., Kielbasa,S., Korbel,J., Sers,C., Malik,A., Eickhoff,H., Lehrach,H. and Schuchhardt,J. (2001) Extracting information from cDNA arrays. *Chaos*, **11**, 98–107.

Lander,E.S. (1999) Array of hope. *Nat. Genet.*, **21**, 3–4.

Lennon,G.G. and Lehrach,H. (1991) Hybridization analyses of arrayed cDNA libraries. *Trends Genet.*, **7**, 314–317.

Lipshutz,R.J., Fodor,S.P., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.*, **21**, 20–24.

Patriotis,P.C., Querec,T.D., Gruver,B.N., Brown,T.R. and Patriotis,C. (2001) ArrayExplorer, a program in visual basic for robust and accurate filter cDNA array analysis. *BioTechniques*, **31**, 862–872.

Ramsay,G. (1998) DNA chips: state-of-the art. *Nat. Biotechnol.*, **16**, 40–44.

Schena,M., Heller,R.A., Theriault,T.P., Konrad,K., Lachenmeier,E. and Davis,R.W. (1998) Microarrays: biotechnology's discovery platform for functional genomics [see comments]. *Trends Biotechnol.*, **16**, 301–306.

Schuchhardt,J., Beule,D., Malik,A., Wolski,E., Eickhoff,H., Lehrach,H. and Herzel,H. (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res.*, **28**, E47.

Sprawls,P. (1987) *Physical Principles of Medical Imaging*. Aspen Publishers, Inc.

Zhao,N., Hashida,H., Takahashi,N., Misumi,Y. and Sakaki,Y. (1995) High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. *Gene*, **156**, 207–213.