Available online at www.sciencedirect.com

Genomics 83 (2004) 1169–1175

GENOMICS

www.elsevier.com/locate/ygeno

Short Communication

Evidence of systematic expressed sequence tag IMAGE clone cross-hybridization on cDNA microarrays

Daniel Handley,[a,b,*] Nicoleta Serban,[c] David Peters,[a] Robert O'Doherty,[d] Melvin Field,[e] Larry Wasserman,[c] Peter Spirtes,[b,f] Richard Scheines,[b] and Clark Glymour[b,f]

[a] *Human Genetics, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, PA 15261, USA*
[b] *Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[c] *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[d] *Department of Medicine and Department of Molecular Genetics and Biochemistry, School of Medicine; University of Pittsburgh, Pittsburgh, PA 15261, USA*
[e] *Department of Neurosurgery, School of Medicine; University of Pittsburgh, Pittsburgh, PA 15261, USA*
[f] *Institute for Human and Machine Cognition, University of West Florida, Pensacola, FL 32501, USA*

Received 7 September 2003; accepted 18 December 2003

Available online 27 February 2004

Abstract

We present evidence of a potentially serious source of error intrinsic to all spotted cDNA microarrays that use IMAGE clones of expressed sequence tags (ESTs). We found that a high proportion of these EST sequences contain 5′-end poly(dT) sequences that are remnants from the oligo(dT)-primed reverse transcription of polyadenylated mRNA templates used to generate EST cDNA for sequence clone libraries. Analysis of expression data from two single-dye cDNA microarray experiments showed that ESTs whose sequences contain repeats of consecutive 5′-end dT residues appeared to be strongly coexpressed, while expression data of all other sequences exhibited no such pattern. Our analysis suggests that expression data from sequences containing 5′ poly(dT) tracts are more likely to be due to systematic cross-hybridization of these poly(dT) tracts than to true mRNA coexpression. This indicates that existing data generated by cDNA microarrays containing IMAGE clone ESTs should be filtered to remove expression data containing significant 5′ poly(dT) tracts.

Keywords: cDNA microarray; Expressed sequence tag; Cross-hybridization; IMAGE clone; Systematic source of variation

Spotted cDNA microarray experiments are often aimed at identifying and quantifying differential gene expression with respect to tissue type, disease state, nutritional status, or drug/toxicant exposure [1]. Two general strategies for identifying differentially expressed genes are to examine static (one time point) differences in expression level and dynamic (time course) differences. Microarray expression data typically contain many sources of variation, including unavoidable random error introduced during the performance of the experiment as well as measurement error associated with acquisition of raw intensity data from microarray image files. A large portion of the statistical analysis of microarray data involves identifying and quantifying sources of variation, with the purpose of distinguishing between experimental error (noise) and inherent variability (signal) resulting from the actual biological phenomena under study. To this end, there is an abundant literature concerning different methods for normalizing microarray data, as well as approaches for screening and filtering data (e.g., outlier removal) [2–5]. Once sources of experimental variation are identified, they can fall into one of two categories: they may be sources of variation that can be reduced or eliminated through improvements in experimental design, performance, or technology or they may be sources of variation that may be better addressed through judicious application of statistical analysis.

In this paper we present evidence for a potentially significant source of error in single-dye spotted cDNA microarrays that can be eliminated through an improvement in microarray design. Specifically, this paper presents evidence for an intrinsic, systematic cross-hybridization artifact on spotted cDNA arrays that use expressed se-

[*] Corresponding author.
E-mail address: dhandley@andrew.cmu.edu (D. Handley).

0888-7543/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2003.12.010

quence tag (EST) IMAGE clones. We have found that this artifact can completely obscure the true variation of interest in differential expression analysis of cDNA microarray data.

Results

We examined data from three cDNA microarray experiments: two experiments performed by separate investigators using commercially manufactured single-dye spotted cDNA arrays at the University of Pittsburgh (GF400, GeneFilters Microarrays; Research Genetics, Carlsbad, CA, USA; Peters et al., unpublished data, and Field et al., unpublished data) and one two-dye (Cy3/Cy5) experiment published on the publicly available Stanford Microarray Database Web site [6]. The first spotted cDNA microarray experiment consisted of a time-sequenced sampling of differential mRNA expression from 3T3L1 cultured mouse adipocytes treated with the insulin-sensitizing agent troglitazone (TZD) (Peters et al., unpublished). The second spotted cDNA microarray experiment consisted of exposing primary cultures of human middle cerebral artery (MCA) smooth muscle cells (SMC) to aneurysmal subarachnoid hemorrhage (SAH) cerebrospinal fluid (CSF) from two patients with ruptured intracranial aneurysms and measuring the resultant mRNA expression levels (Field et al., unpublished data).

Proportion of poly(dT) tract sequences exhibiting high variability

We looked at the variability of mRNA expression over time course and experimental condition as a function of poly(dT) string length. In the first experiment data set (TZD-treated adipocytes), approximately 70% of the 200 sequences with the highest variance over time are prefixed by a string of consecutive 5′ dT residues of length greater than 5 and are sequences derived from ESTs. In contrast, 40% of the total number of EST sequence expression data had poly(dT) tracts of length at least 5.

Expression variability as a function of poly(dT) tract length over all measurements

To evaluate gene expression variability with respect to length of poly(dT) tracts, we first categorized the set of sequences analyzed in the three microarray data sets by length of the initial poly(dT) tracts. For each category C_k

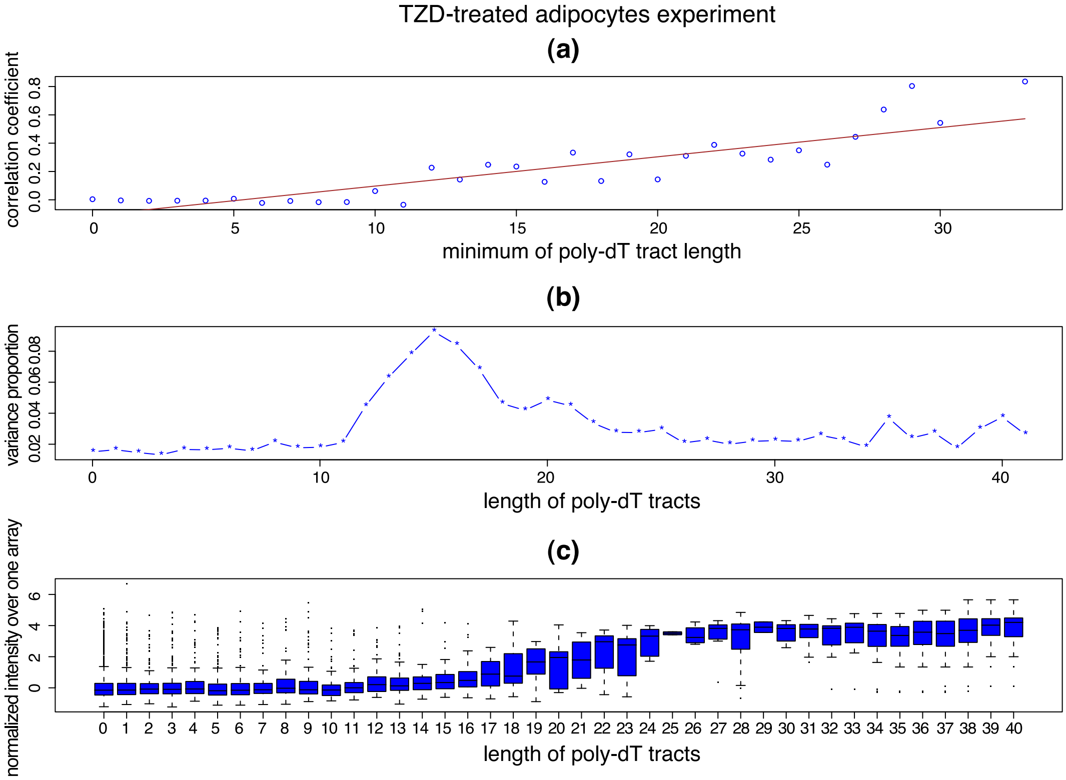

Fig. 1. (a) Minimum length of poly(dT) tracts for a pair of sequences vs the average correlation coefficient of two sequences with the corresponding minimum length of poly(dT). (b) Proportion of variance, V_k, divided by the number of genes for category C_k vs the length of poly(dT) tracts, k, with $k = 0,1,2...,40$. (c) Set of box plots from only one measurement in the expression data of TZD-treated adipocytes, each box plot corresponding to a category of genes C_k containing the normalized intensities on the log scale for the genes in C_k.

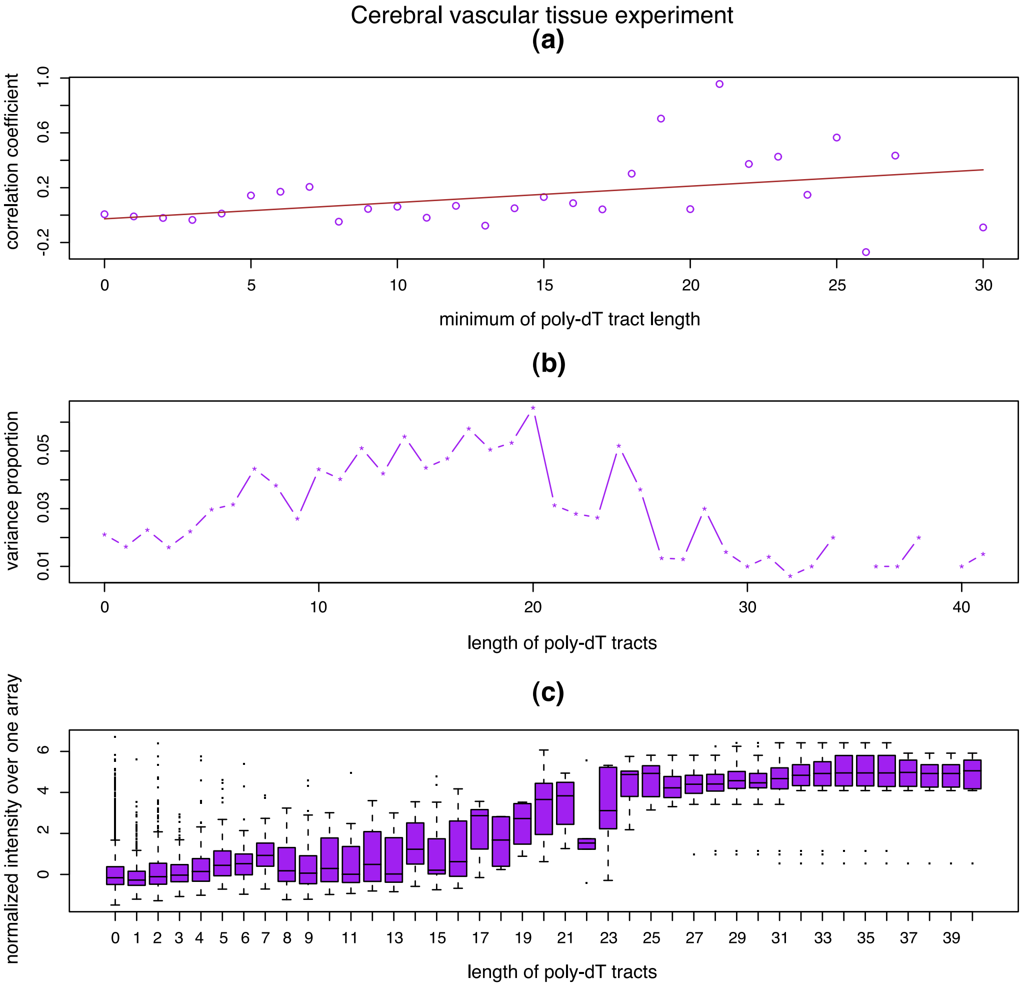

Fig. 2. The same as in Fig. 1, but for the cerebral vascular tissue experiment.

with $k=0,1,\ldots,40$, the proportion of variance explained by its sequences is evaluated as

$$V_k = \frac{\sum\limits_{g\in C_k} Var(X_g)}{\sum\limits_{k}\sum\limits_{y\in C_k} Var(X_y)},$$

where category C_k contains all the sequences with the poly(dT) tracts of length k and $Var(X_g)$ is the variance of gene g across experimental conditions (e.g., time). We plotted the proportion of variance divided by the number of sequences in its category C_k, V_k/size of C_k, versus the length of dT tracts, k (Figs. 1b, 2b, and 3b). This average proportion of variance, V_k/size of C_k, estimates the explained variability proportion across experimental conditions by a sequence with the initial dT sequence of length k.

For the two-dye expression data, the averaged proportion of variance appears to be randomly distributed over the length of poly(dT) tracts. In the case of the one-dye microarray data sets, we identified a relationship between the length of poly(dT) tracts and the variability over experimental of those sequences with initial dT sequence longer than 11 for the TZD-treated adipocytes (Peters et al.) and 3 for the cerebral vascular tissue data (Field et al.).

Time-course expression pattern similarity

Most of the sequences with high variance follow a curiously similar expression pattern over time (i.e., despite the fact that the group exhibited such high variance as a whole, the expression patterns of each individual sequence in the group were markedly similar). In addition, among the first 100 sequences with most variable expression profiles in the TZD-treated adipocytes experiment, only one of that hundred[1] displays a pattern significantly different from the others (this single sequence contains the leading sequence TCTTTTTCACCTCTTTATTTTTTTTA. . .) On the other hand, the sequences of only 9 of those 100 do not contain long poly(dT) tracts, one of them being the gene with a different pattern over time. One would be very surprised to see such a similarity in expression pattern over a time course for those sequences with high variance, especially when the

[1] This sequence has NCBI Accession No. AI428396.

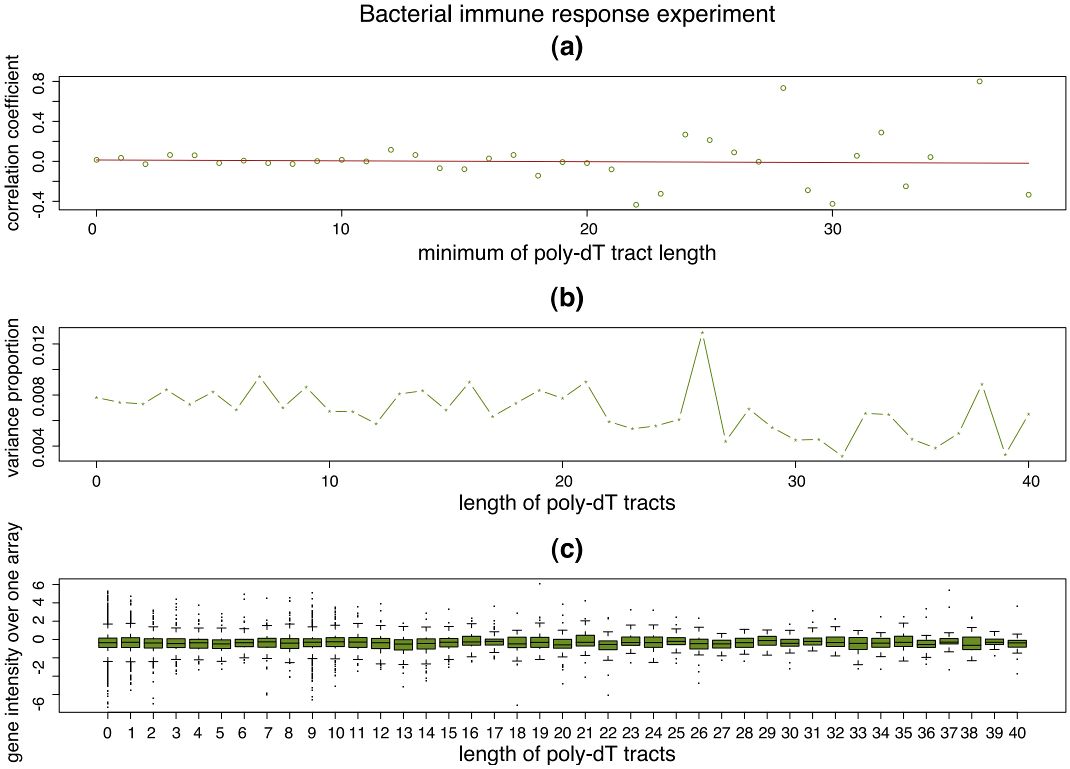

Fig. 3. The same as in Fig. 1, but for the two-dye data set.

only distinguishing feature among them we could identify is that they all contain long leading (5′-end) poly(dT) tracts. These results raise the question of whether there might be some systematic source of variation in the data, such as that which might arise from cross-hybridization between sequences containing leading poly(dT) tracts rather than signal reflecting true mRNA expression.

Evidence of time-course expression pattern similarity for sequences with long poly(dT) tracts

We suspect that the expression pattern over time is similar in all sequences with long initial poly(dT) tracts. The plots in Figs. 1a, 2a, and 3a support this hypothesis. This analysis is intended to demonstrate whether there is a relationship between the correlation of the expression profiles of the two sequences randomly chosen and the minimum length of the poly(dT) tracts for the two sequences. In the case of a strong positive relationship between the two measures (correlation and minimum of poly(dT) tracts), if the minimum dT sequence length is low we would expect to see a low correlation between the expression profiles of the two sequences and conversely, if the minimum length is high we would see a high correlation coefficient. Indeed, in the plots for the two one-dye microarray data sets (Figs. 1a and 2a), the average correlation of the expression profiles (y axis) increases with the minimum of the length of initial poly(dT) tracts (x axis); the robust regression line has a positive slope. On the other hand, for the two-dye microarray data set, the robust regression line has a slope of approximately 0 (Fig. 3a). Based on this analysis, we concluded that for the two one-dye microarray data sets there is a similarity in pattern for the expression profiles of those sequences with long poly(dT) tracts.

Expression variability as a function of poly(dT) tract length for each measurement separately

We can gain a different perspective by looking at expression variability over the poly(dT) tract length through a series of box plots of the normalized sequence expression levels from one array[2] separately. Each box plot in the set of box plots contains the intensities of the sequences in one category, C_k (defined according to the number of consecutive 5′ dT residues, k). For each of the three data sets, we present in this paper the set of box plots for only one microarray (see Figs. 1c, 2c, and 3c). The median level increases as a function of length of poly(dT) tract (x axis) (as in Figs. 1c and 2c). In contrast, the median level does not show any pattern over the length of the initial poly(dT) tracts for the arrays in the two-dye microarray data set (as in Fig. 3(c)). The median expression level is robust to outliers, thus analyzing them with respect to dT-length category minimizes the possibility the results are being significantly affected by outliers.

[2] Now we consider the variability under only one experimental condition.

Discussion

Expressed sequence tags

ESTs are short (200–500 bp) sequences derived from directionally cloned plasmid cDNA libraries. Typically, total mRNA is isolated from cells in a particular type of tissue, stage of development, pathological state (e.g., normal versus tumor), or environmental/nutritional state (e.g., heat shock). The mRNA is reverse-transcribed using an oligo(dT) sequence primed with a restriction site. The resultant cDNA is then cloned into a plasmid vector, isolated, and one-pass sequenced, with the sequence submitted to a database [7,8]. The clones are routinely deposited into Lawrence Livermore National Laboratory's Integrated Molecular Analysis of Genomes and their Expression (IMAGE) Consortium (http://image.llnl.gov) from which they can be obtained through several distributors for use in microarray manufacture. Once individual clones have been isolated, double-stranded DNA is amplified via PCR for microarray spotting. While the vector sequences derived from cloned cDNA are typically removed from these PCR products, part of the sequence complementary to the mRNA 3′ poly(A) tail sometimes remains. This initial 5′-end poly(dT) sequence can be readily identified in these sequences as reported in the GenBank nucleic acid sequence database administered by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

The fact that some of the EST sequences have residual 5′-end poly(dT) gives rise to the a priori possibility that these sequences will cross-hybridize with any complementary sequence. Since the types of cDNA microarrays discussed here are fabricated by spotting a substrate with PCR-amplified denatured double-stranded cDNA, it might be expected that any labeled cDNA sequence containing stretches of consecutive dA or dT residues may hybridize to the spotted EST cDNA sequences containing 5′ poly(dT). In principle this could generate a considerable artifact signal.

Many cDNA microarray protocols, including the one described here, involve a prehybridization step using poly(dA) oligonucleotide to block polyadenylated sites. However, the cDNA deposited on these microarrays is denatured double stranded DNA. This means it contains both poly(dT) tracts and its complementary poly(dA) sequence. The oligo(dA) added during this blocking step may bind to the poly(dT) tracts, but not to the poly(dA). In mRNA experiments, the probe is produced by reverse transcribing oligo(dT)-primed mRNA to single-stranded labeled cDNA (e.g., using 33P-dCTP). Thus, the probe may have poly(dT) tracts that will hybridize to any exposed poly(dA) tracts on the microarray. However, only the poly(dT) tracts had been blocked in the prehybridization step. If this is the source of the cross-hybridization signal we have observed, then one would wish to use oligo(dT) during the prehybridization step on the microarray, or alternatively, add oligo(dA) to the probe.

An alternative explanation for our observation might be that the phenomenon is a result of sequence similarity between the ESTs not relating to poly(dT) sequences. To test this, we performed pairwise BLAST searches between every combination of the 100 most variable sequences and found no significant sequence similarity other than the poly(dT) tracts.

One normally expects that optimizing hybridization stringency conditions, such as adjusting temperature and buffer salt concentration, minimizes nonspecific hybridization such that the resultant signal-to-noise ratio is acceptable. However, in this case using the manufacturer's recommended protocol, including optimizing stringency conditions, did not eliminate the artifact we have observed even though in both experiments background intensity on the microarray images was uniformly low. Since we believe this is an issue of legitimate true hybridization (albeit cross-hybridization) and not nonspecific hybridization, we believe that compensatory stringency adjustments are not a viable solution to the problem.

We have no evidence indicating what molecular species might be actually responsible for the cross-hybridization. Since the spotted cDNA is double stranded, the cross-hybridizing species could be labeled sequences containing either poly(dA) or poly(dT). In either case, however, this artifact would be eliminated by excising the 3′ poly(dA) tail region from the cloned cDNA at the same time the vector sequence is removed, although in practice this would be technically difficult, and performing this operation on the entire IMAGE library would constitute a major undertaking. While the phenomenon might be the subject of further study to elucidate the exact mechanism responsible for the artifact we have observed, the most immediate solution (although admittedly nonoptimal) is simply to filter out EST sequences having significant poly(dT) sequences prior to statistical analysis of expression data.

Two-dye microarrays

An alternative to using single-dye cDNA microarrays is to use two-dye spotted cDNA microarrays. In the two-dye design, we would expect that any cross-hybridization would be equally (or nearly) represented by each dye, and therefore the resulting artifactual signal components would cancel. Our analysis of differential gene expression from a two-dye microarray data set supports this conclusion.

However, in two-dye microarrays containing ESTs having significant 5′ poly(dT) sequences, we would expect increased competition for hybridization sites from the cross-hybridizing molecular species. High concentrations of a cross-hybridizing species may therefore exclude smaller relative amounts of the actual labeled molecule of interest. This would produce a higher relative variability in the signal of interest. Therefore, even though the two channels (e.g., Cy3 and Cy5) containing noise are subtracted, the resultant signal-to-noise ratio may still be diminished compared to the case in which there were no poly(dT) artifact. Even with two-dye spotted cDNA microarrays, then, it might be advantageous to remove the 3′ poly(dA) tail from the source cloned cDNA.

Implications

Experiments involving microarrays that contain PCR-amplified sequences derived from IMAGE clones are widespread in almost every area of biological and medical research. A significant cross-hybridization artifact such as that which we have observed may obscure legitimate signals and may cause researchers to miss indications of potentially important phenomena. Admittedly, what we have presented here is a preliminary indication of a potentially serious problem; we have not suggested a particular mechanism that might be responsible for the problem. However, we believe the evidence presented here is sufficient to warrant caution in interpreting cDNA microarray data, as well as further study of the source of the observed phenomenon. In the meantime, we suggest researchers also consider reanalyzing existing data after filtering out any EST sequences containing long initial poly(dT) sequences. Alternatively, researchers may wish to consider switching to synthesized oligonucleotide microarrays (e.g., Affymetrix GeneChips, Amersham CodeLink), which would be immune from this type of cross-hybridization.

Statistical issues

Because of their effects on estimated variances, significant cross-hybridization artifact in spotted cDNA microarrays would influence many statistical analyses, especially those involving analyses of variability such as ANOVA, principal component analysis, and classification analysis. For example, in cluster analysis, the procedures applied to estimate the number of clusters are affected directly by the presence of noise in the data. In Fridlyand and Dudoit [9], a simulation study on different procedures for estimating the number of clusters showed that they have a low rate of recovery of the true number of clusters when noise variables are added to the data. We have shown that in the current cDNA microarray data sets there is a strong relationship between the intensity values and the length of poly(dT) tracts. Thus, the latter is a confounding variable that may adversely affect the statistical analysis.

Methods

TZD-treated adipocytes experiment

The first spotted cDNA microarray experiment consisted of a time-sequenced sampling of differential mRNA expression from 3T3L1 cultured mouse adipocytes treated with the insulin-sensitizing agent TZD (Peters et al., unpublished data). Cells were harvested in 5 ml of Trizol (Invitrogen Corp., Carlsbad, CA, USA) and RNA was extracted according to the manufacturer's instructions. RNA integrity for each sample was confirmed on formaldehyde/formamide agarose gels prior to microarray analysis. cDNA probes for microarray analysis were synthesized from 5 μl total RNA from each sample.

Total RNA was first heat denatured in the presence of 0.1 g/L oligo(dT) for 10 min at 70°C. Reverse transcription was performed in 1 × first-strand buffer (Invitrogen Corp.) in the presence of 1.5 μl reverse transcriptase (10 U/μl SuperScript II RT; Invitrogen Corp.), 1.0 μl DTT (0.1 M), 1.5 μl dNTP mixture (dATP,dGTP,dTTP at 20 mM), 10 μl [^{33}P]dCTP (3000 Ci/mmol, 10 mCi/ml), and RNasin (8 U/sample, Promega Corp., Madison, WI, USA) for 90 min at 37°C. Each probe sample was purified by passage through a Quick Spin G-50 Sephadex Column (Roche Diagnostics Corp., Indianapolis, IN, USA) and denatured for 3 min at 99°C before use.

cDNA mouse filter arrays (GF400, GeneFilters Microarrays; Research Genetics) containing 5184 cDNA sequences, including 192 dots representing total genomic cDNA, were used. The filter arrays were prehybridized in 5 ml Microhyb solution (Research Genetics), 5 μl poly(dA) (0.5 mg/ml; Research Genetics), and 5 μl Cot-1 human DNA (denatured at 99°C for 3 min, 0.5 mg/ml; Invitrogen Corp.) for 2 h at 42°C. The probe was then added to the hybridization buffer and incubated for 16 h at 42°C. After hybridization, the arrays were washed twice at 50°C in 60 ml 2× sodium citrate (SSC) buffer containing 1% sodium dodecyl sulfate (SDS) for 20 min and once at 55°C in 0.5× SSC buffer containing 1% SDS for 15 min. Posthybridization filter arrays were exposed to a PhosphorImager screen (Molecular Dynamics, Sunnyvale, CA, USA) and the images scanned using a Storm PhosphorImager (Molecular Dynamics). Signals were quantified using ImageQuant software (Molecular Dynamics).

Cerebral vascular tissue experiment

The second spotted cDNA microarray experiment consisted of exposing primary cultures of human MCA SMC to aneurysmal SAH CSF from two patients with ruptured intracranial aneurysms and measuring the resultant mRNA expression levels (Field et al., unpublished data). Of these two patients, one developed severe symptomatic cerebral vasospasm resulting in multiple cerebral infarctions and associated permanent neurological deficits, while the other had a benign course with no evidence of cerebral vasospasm, development of neurological deficit, or infarction.

Cells were exposed to cultured SAH CSF collected posthemorrhage from both patients and were exposed to cultured MCA SMC tissue isolates. The MCA cells were then removed and RNA was extracted for cDNA microarray analysis. University and hospital institutional review boards approved the experimental protocol performed on the patient specimens. Total RNA was extracted, purified, and quantified as described above (Invitrogen Corp.).

cDNA probes were prepared as in the previous experiment except that only 2 μg RNA was used for reverse transcription. Seven oligonucleotide filter arrays (GF211, GeneFilters Microarrays; Research Genetics) were used. The filter arrays were prehybridized in a fashion identical to that discussed above.

Bacterial immune response experiment

Publicly available data were obtained from the Stanford Microarray Database (http://genome-www5.Stanford.EDU/MicroArray/SMD/). Boldrick et al. examined differential gene expression in human peripheral blood mononuclear cells in response to bacteria and bacterial products using two-dye Cy3/Cy5 microarrays [6]. Methods are published on a separate Web supplement (http://genome-www.stanford.edu/hostresponse/mandm.shtml).

Acknowledgments

The authors thank Jay Kadane, Takis Benos, and Joe Ramsey for their helpful comments. This work was supported by NASA Grant NCC2-1227 and the Copeland Fund of the Pittsburgh Foundation Grant D200-0251.

References

[1] M. Schena, DNA Microarrays: A Practical Approach, Oxford Univ. Press, London, 1999.

[2] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 4 (2002) 249–264.

[3] M. Kerr, G.A. Churchill, Analysis of variance for gene expression microarray data, J. Compu. Bio. 7 (2000) 819–837.

[4] S. Dudoit, Y.H. Yang, T.P. Speed, M.J. Callow, Statistical methods for identfying differentially expressed genes in replicated cDNA microarray experiments, Stat. Sin. 12 (1) (2002) 111–139.

[5] Y.H. Yang, M.J. Buckley, S. Dudoit, T.P. Speed, Comparison of methods for image analysis on cDNA microarray data, J. Comput. Graph. Stat. 11 (1) (2002) 108–136.

[6] J.C. Boldrick, A.A. Alizadeh, M. Diehn, S. Dudoit, C. Long Liu, D. Botstein, L.M. Staudt, P.O. Brown, D.A. Relman, Stereotyped and specific gene expression programs in human innate immune responses to bacteria, Proc. Natl. Acad. Sci. USA 99 (2) (2002) 972–977.

[7] M.F. Bonaldo, G. Lennon, M.B. Soares, Normalization and subtraction: two approaches to facilitate gene discovery, Genome Res. 6 (9) (1996) 791–806.

[8] T.G. Wolfsberg, D. Landsman, in: A.D. Baxevanis (Ed.), Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, second edition, Wiley–Interscience, New York, 2001, p. 286.

[9] J. Fridlyand, S. Dudoit, Applications of resampling methods to estimate the number of clusters and to improve the accuracy of clustering methods, Technical Report 600. University of California, Berkeley, Division of Biostatistics, 2001.

Web site references

http://genome-www5.Stanford.EDU/MicroArray/SMD/, Stanford Microarray Database.

http://genome-www.stanford.edu/hostresponse/mandm.shtml, Boldrick et al. Methods Web supplement.

http://image.llnl.govl, Lawrence Livermore National Laboratory's Integrated Molecular Analysis of Genomes and Their Expression (IMAGE) Consortium.

http://www.ncbi.nlm.nih.gov/Entrez/, National Center for Biotechnology Information, Entrez search and retrieval system.