

# Evidence of cross-hybridization artifact in Expressed Sequence Tags (ESTs) on cDNA microarrays

Dan Handley<sup>¶,†</sup>      Nicoleta Serban\*      David Peters<sup>†</sup>      Robert O'Doherty<sup>‡</sup>  
Melvin Field<sup>§</sup>      Larry Wasserman\*      Peter Spirtes<sup>¶,||</sup>      Richard Scheines<sup>¶</sup>  
Clark Glymour<sup>¶,||</sup>

April 13, 2003

## Abstract

We present evidence of cross-hybridization artifact intrinsic to spotted single-dye cDNA microarrays as a result of cDNA containing 5'-end sequences of consecutive thymidine (dT) residues. These poly(dT) tracts result from the synthesis, via oligo (dT) primed reverse transcription, of expressed sequence tags (EST) cDNA from a polyadenylated mRNA template. Analysis of gene expression data from two separate experiments involving commercially available single-dye cDNA microarrays showed that ESTs whose reported sequences contain more than 11 consecutive 5'-end dT residues appeared to be strongly co-expressed. Our results suggest that expression data from microarray sequences containing 5' poly(dT) tracts is likely to be due to cross-hybridization rather than the actual mRNA expression. This indicates that existing data generated by cDNA microarrays may be unreliable and should be filtered to remove EST sequences containing significant poly(dT) tracts.

Keywords: *cDNA Microarray, Expressed-Tag Sequence, EST, Cross-hybridization, Poly(dT), Poly(dA), systematic source of variation.*

## 1 INTRODUCTION

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. 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 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

---

\*Carnegie Mellon University, Department of Statistics

†University of Pittsburgh, School of Public Health, Human Genetics

‡University of Pittsburgh, School of Medicine, Department of Medicine and Department of Molecular Genetics and Biochemistry

§University of Pittsburgh, School of Medicine, Department of Neurosurgery

¶Carnegie Mellon University, Department of Philosophy

||Institute for Human and Machine Cognition, University of West Florida

(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). 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 better be addressed through judicious application of statistical analysis.

In this paper we present evidence for a potentially significant source of variation in single-dye spotted cDNA microarrays that can be eliminated through an improvement in microarray design. Specifically, this paper presents evidence for a systematic artifact in spotted cDNA arrays with expressed sequence tags (ESTs) containing initial sequences of several consecutive thymidine (dT) residues. We have found that this artifact can completely mask the variation of interest in differential-expression analysis of microarray data.

## 2 RESULTS

**Data sets.** We examined data from three cDNA microarray experiments, two from experiments using single-dye spotted cDNA arrays at the University of Pittsburgh (Peters et al. unpublished data and Field et al. unpublished data) and one from two-dye (Cy3/Cy5) data published on the Stanford Microarray Database website (Boldrick et. al 2002). We make use of only normalized expression data in this paper. After log transforming the data<sup>1</sup>, intensities on each array are normalized by subtracting the median of the corresponding intensities.

**Proportion of poly(dT) tract sequences exhibiting high variability.** We looked at the variability of mRNA expression over time and experimental condition as a function of poly(dT) string length. In the TZD-treated adipocytes experiment data set, approximately 70% of the 200 sequences with the largest variance over time are prefixed by a string of consecutive 5' dT residues of length greater than 5 and are sequences derived from expressed sequence tags (ESTs). In contrast, 40% of the total number of EST's in the data had poly(dT) tracts of length at least 5. Upon examining the 500 sequences with the largest variance over experimental conditions in the expression data, we similarly found a large percentage of sequences with long poly(dT) tracts.

**Expression pattern similarity.** In the above data set, most of the sequences with large variance follow a similar expression pattern over time. Among the first 100 sequences with most variable expression profiles in the TZD-treated adipocytes experiment, only one of them<sup>2</sup> displays a pattern significantly different from the others (as shown in figure 1). This single one is prefixed by TCTTTTTCACCTCTTTATTTTTTTTAA.... 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 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 large variance, especially when they almost exclusively contain long 5' 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 artifact, rather than signal reflecting to true mRNA expression.

**Evidence of the 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 figures 4(a), 5(a), and 6(a) support this hypothesis. In each plot, 10,000 'samples', each sample containing two sequences randomly chosen from the total set of sequences, were drawn and two quantities were computed: the correlation coefficient for each sample (correlation coefficient for the expression profiles of the two sequences in the sample) and the minimum length of poly(dT) tracts corresponding to the two se-

---

<sup>1</sup>We apply the transformation and normalization only to the two single-dye spotted cDNA microarray datasets.

<sup>2</sup>This sequence has the NCBI accession number AI428396.

quences in the corresponding sample. The algorithm is displayed in Figure 2. It intends to capture 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 large we would see a high correlation coefficient. Indeed, in the plots for the two one-dye microarray data sets (figures 4(a) and 5(a)), 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 dataset, the robust regression line has a slope approximately 0 (figure 6(a)). 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 large poly(dT) tracts.

**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  $C_k$  with  $k = 0, 1, \dots$ , the proportion of variance explained by its sequences is evaluated as:

$$V_k = \frac{\sum_{g \in C_k} Var(X_g)}{\sum_k \sum_{g \in C_k} Var(X_g)}$$

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) as described in Methods section. We plotted the proportion of variance divided by the number of sequences in its category  $C_k$ ,  $\frac{V_k}{\text{size of } C_k}$ , versus the length of dT-tracts,  $k$  (figures 4(b), 5(b), and 6(b)). This average proportion of variance,  $\frac{V_k}{\text{size of } C_k}$ , estimates the explained variability proportion across experimental conditions by a sequence with the initial dT sequence of length  $k$ .

For single-dye cDNA microarray data sets, there is a noticeable trend in the average proportion of variance of those sequences versus poly(dT) tract length. The variability proportion across experimental conditions increases up to a dT sequence length of 15 or so, then decreases slowly. However, for the two-dye expression data, the averaged proportion of variance appears to be randomly distributed over the length of poly(dT) tracts. Thus, 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.

**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 boxplots of the normalized sequence expression levels from one array<sup>3</sup> separately. Each boxplot in the set of boxplots 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 boxplots for only one array(see figures 4(c), 5(c), and 6(c)). For each other measurement<sup>4</sup> in the three data sets, we've considered the same set of boxplots of the sequence categories but they are not included in this paper. For each of the three experiments, the sets of boxplots not presented in this paper showed similar trend over the length of poly(dT) tracts as the one included in the paper. That is, for the spotted one-dye cDNA data sets, in each of the set of boxplots (corresponding to one measurement or experimental condition of the Peters et al./Field et al. data set), the median level increases as a function of length of poly(dT) tract (x axis) (as in figures 4(c) and 5(c)). In contrast, the median level doesn't show any pattern over the length

<sup>3</sup>Now we consider the variability under only one experimental condition.

<sup>4</sup>Each corresponds to a different experimental condition.

of the initial poly(dT) tracts for the arrays in the two-dye microarray data set (as in figure 6(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.

**Testing for trends in the data.** The existence of a trend can be tested in multiple ways. In this case, we use the permutation trend test which evaluates the possibility of a trend by comparing the observed robust slope of the quantitative observations (here median of the intensity values measured on array  $a$  for the gene categories defined earlier) versus the length of poly(dT) tracts, to the slope of the permuted quantitative observations. According to the set of boxplots in figures 4(c), 5(c), and 6(c), it suffices to check for a linear trend only.

The hypothesis setting is:

$H_0$ : For one array  $a$ , the median of the gene categories is not linear over the length of poly(dT) tracts vs.

$H_1$ : It has a linear increasing trend.

We chose to represent a category of sequences by the median value for the intensities in the category because the median reflects the behavior of a “typical” intensity of the corresponding length of poly(dT) tract. For example, the sequences with no initial dT sequence at the experimental time 0 (control time) in the TZD-treated adipocytes experiment has typically the intensity value of 0.8427 (which is the median for this category).

**P-values.** Table 1 shows the estimated p-values for each measurement in each data set. At the significance level<sup>5</sup>  $\alpha = 0.01$ , the null hypothesis,  $H_0$ , is rejected in all the 20 measurements for the first spotted cDNA microarray data set and for 6 out of 7 measurements for the second spotted cDNA data set. None of the p-values is less than  $\alpha = 0.01$  for the 5 arrays in the two-dye data. This indicates that the measurements on spotted cDNA data sets follow an increasing trend over the length of poly(dT) tracts. This test explains once again the relationship between the measured expression levels in single-dye cDNA data and the long poly(dT) tracts while there is no evidence for such an increasing trend over the poly(dT) tracts length for two-dye data. To note that we have evidence against non-linearity of the intensity over the length of initial poly(dT) tracts even for control measurement (no drug treatment at time 0) in TZD-treated adipocytes experiment.

### 3 DISCUSSION

**Expressed sequence tags.** 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 [Bonaldo]. The clones are often deposited into the I.M.A.G.E consortium [<http://image.llnl.gov>] where they can be obtained through several distributors for use in microarray manufacture. Because of cloning and one-pass sequencing large numbers of sequences, ESTs are expected to have a relatively high reported sequence error rate (3%)[Wolfsberg & Landsman], and are therefore considered most useful for purposes such as gene discovery. ESTs are therefore considered mainly a mass screening device. 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 remain. This initial 5'-end poly(dT) sequence can be readily identified in these sequences as reported in GenBank.

---

<sup>5</sup>Here, the p-values are not corrected for multiple hypothesis testing.

The fact that some of the EST sequences have residual 5'-end poly(dT) gives rise to the 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 (ds) 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). This could generate a considerable artifact signal.

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 non-specific 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 hybridization (albeit cross-hybridization) and not non-specific hybridization, we believe that compensatory stringency adjustments is 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 either labeled sequences containing 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 I.M.A.G.E. 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 non-optimal) is to simply 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 I.M.A.G.E. 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. We believe the evidence presented here warrants further study of the artifact as well as suggesting that manufacturers of spotted cDNA microarrays take prudent measures to reduce or eliminate its impact. In the meantime, we suggest researchers consider re-analyzing 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 (Affymetrix, Amersham) which would not suffer from this sort of artifact.

**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, classification analysis, etc.

For example, in cluster analysis, the procedures applied to estimate the number of clusters are directly affected by the presence of noise in the data. In Fridlyand & Dudoit (2001), 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.

Additionally, we showed that there is a strong relationship between the intensity values and the length of poly(dT) tracts. Thus the latter is a confounding variable which will greatly affect the statistical analysis.

## 4 METHODS

### 4.1 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 troglitazone (TZD)(Peters, et al. unpublished data). Cells were harvested in 5 ml of Trizol (Invitrogen Corp., Carlsbad, CA) 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  $\mu$ L total RNA from each sample.

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

cDNA mouse filter arrays (GF400, GeneFilters Microarrays; Research Genetics, Carlsbad, CA) containing 5184 oligonucleotides, including 192 dots representing total genomic cDNA, were used. The filter arrays were prehybridized in 5mL Microhyb solution (Research Genetics), 5 $\mu$ L Poly(dA) (0.5g/mL, Research Genetics), and 5 $\mu$ L Cot-1 Human DNA (denatured at 99°C for 3 minutes, 0.5g/L, Invitrogen Corp.) for 2 hours at 42°C. The probe was then added to the hybridization buffer and incubated for 16 hours at 42°C. After hybridization, the arrays were washed twice at 50°C in 60mL 2X sodium citrate (SSC) buffer containing 1% Sodium dodecyl sulfate (SDS) for 20 minutes and once at 55°C in 0.5X SSC buffer containing 1% SDS for 15 minutes. Post-hybridization filter arrays were exposed to a PhosphorImager screen (Molecular Dynamics, Sunnyvale, CA) and the images scanned using a Storm phosphorimager (Molecular Dynamics). Signals were quantified using ImageQuant software (Molecular Dynamics).

### 4.2 Cerebral vascular tissue experiment

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). 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 M SAH CSF collected post-hemorrhage from both patients 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., Carlsbad, CA).

cDNA probes were prepared as in the previous experiment except that only 2 $\mu$ g RNA was used for reverse transcription.

Seven oligonucleotide filter arrays (GF211, GeneFilters Microarrays; Research Genetics, Carlsbad, CA) were used. The filter arrays prehybridized was done in identical fashion as discussed above.

### 4.3 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 [Boldrick]. Methods are published on a separate web supplement (<http://genome-www.stanford.edu/hostresponse/mandm.shtml>).

### 4.4 Statistical methods

**Variability metric.** We quantified the variability over experimental conditions for a single sequence by simply computing the sample variance of the observations:

$$\Delta_i = \frac{1}{m} \sum_{t=1}^m (X_{it} - \bar{X}_i)^2 \text{ where } \bar{X}_i = \frac{1}{m} \sum_{t=1}^m X_{it}.$$

In this context,  $X_{it}$  is the normalized mRNA expression level for sequence  $i$  at the experimental time  $t$ .

**Testing for trend in the data.** Trend test evaluates the possibility of a trend by comparing the observed robust slope of the quantitative observations<sup>6</sup> to the slopes of the permutations of the observations.

**Hypothesis setting.** The null hypothesis is defined as  $H_0$ : *For one array  $a$ , the median of the gene categories is not linear over the length of poly(dT) tracts* vs. the alternative  $H_1$ : *It has a linear increasing trend.*

**P-value estimation.** The p-value is determined by permuting the median values for measurement  $a$  as follows:

*Step 1:* First find the observed slope  $\beta_{obs}$  by robust regression of the intensity medians in each gene category ordered by the length of poly(dT) tracts:  $(M_{ag_0}, \dots, M_{ag_{40}})$  on length of poly(dT) tracts  $k = 0, 1, \dots, 40$ .

*Step 2:* Permute the medians  $(M_{ag_0}, \dots, M_{ag_{40}})$  and fit a robust regression of the permuted observations  $(M_{ag_{\tau(0)}}, \dots, M_{ag_{\tau(40)}})$ <sup>7</sup> on length of poly(dT) tracts. Define the new slope.

Then repeat the data permutation for  $B$  times obtaining the new slopes:  $\beta_b$  for  $b = 1, \dots, B$ .

*Step 3:* The p-value is computed as the averaged number of slopes obtained from permuted quantitative observations which are larger than the observed slope:

$$p - value = \frac{1}{B} \sum_{b=1, \dots, B} I(\beta_b > \beta_{obs})$$

<sup>6</sup>Here the observations are the medians of the intensity values measured on array  $a$  for gene categories defined earlier.

<sup>7</sup> $\tau$  is a permutation of  $0, \dots, 40$ .

**Applying the permutation trend test.** For each measurement (array) in the three data sets, the permutation trend test is applied with 10,000 permutation of the observations.

#### **ACKNOWLEDGMENTS**

The authors would like to thank Jay Kadane 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**

Boldrick J.C., Alizadeh, A.A., Diehn, M., Dudoit, S., Long Liu, C., Botstein, D, Staudt, L.M., Brown, P.O., Relman, D.A. (2002) . *Stereotyped and specific gene expression programs in human innate immune responses to bacteria.* Proc Natl Acad Sci. 99(2): 972-7.

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

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*, Cambridge University Press.

Fridlyand, J., Dudoit, S. (2001). *Applications of resampling methods to estimate the number of clusters and to improve the accuracy of clustering methods*, technical report #600.

Schena, M. (1999). *DNA Microarrays: A Practical Approach*, Oxford University Press.

Wolfsberg, T.G., Landsman, D. (2001). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, pp. 286, Second Edition ed. Andreas D. Baxevasis Wiley-Interscience.

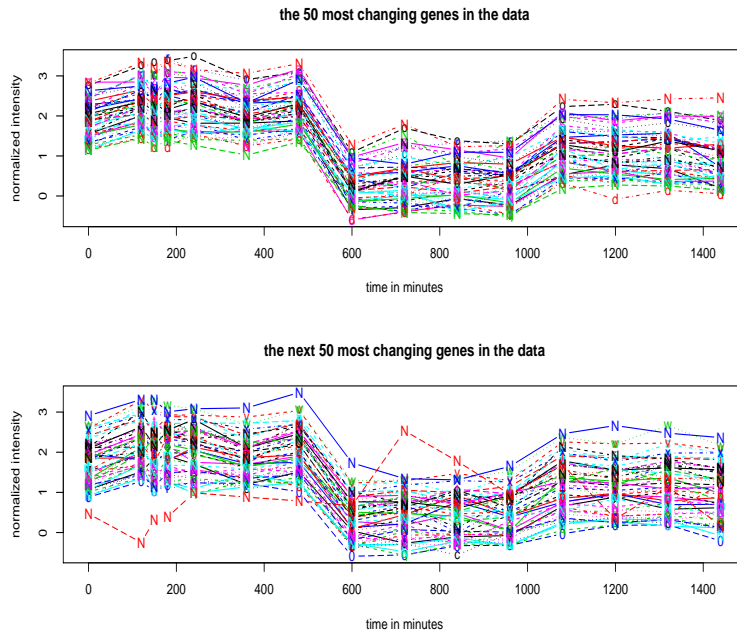
#### **WEB SITE REFERENCES**

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

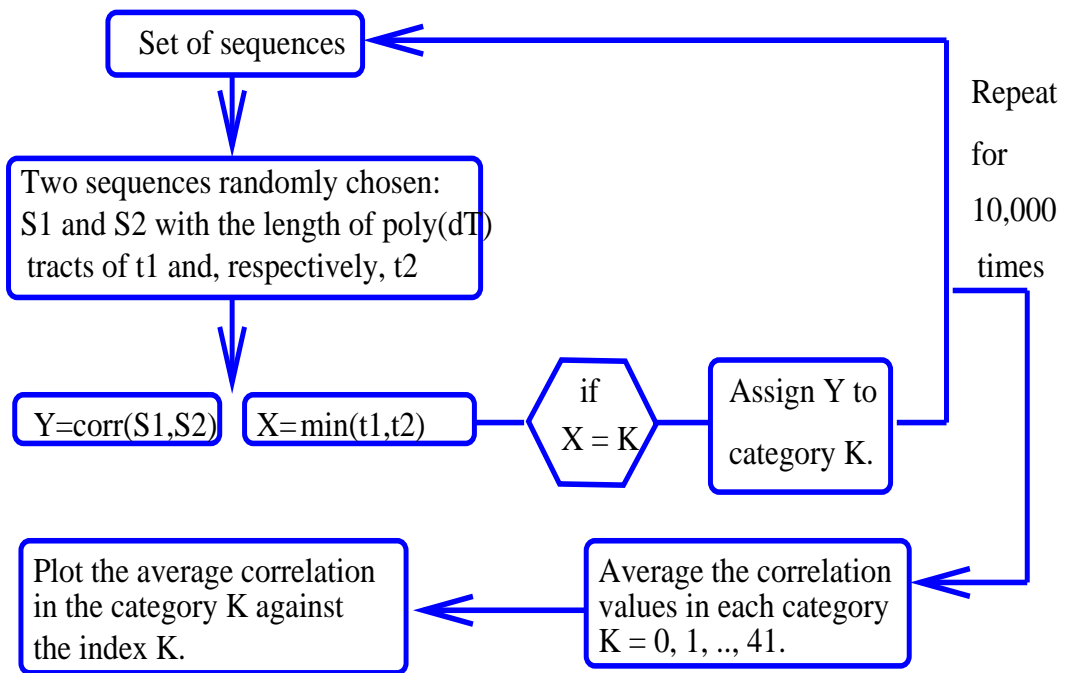
<http://genome-www.stanford.edu/hostresponse/mandm.shtml>, Boldrick et al. Methods web supplement

<http://image.llnl.gov/>, I.M.A.G.E Consortium

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



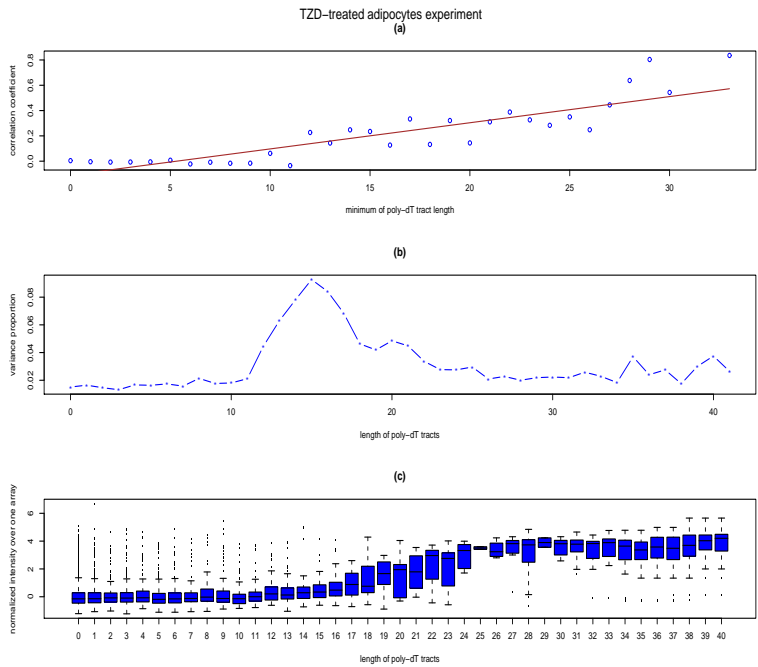
**Figure 1:** The time series plots for the 100 sequences (50 in the upper plot and 50 in the bottom one) with the largest variance under variance underexperimental conditions for the expression data of treated TZD-treated adipocytes.



**Figure 2:** This figure displays the algorithm which aims to capture the relationship between the correlation of the expression profiles of the two sequences randomly chosen and the minimum length of the poly-dT tracts of the two sequences.

IDENTIFIERS	
dbEST Id:	2281424
EST name:	mu92d07.x1
GenBank Acc:	AI465535
GenBank gi:	4319565
CLONE INFO	
Clone Id:	IMAGE:653005 (3')
Source:	IMAGE Consortium, LLNL
DNA type:	cDNA
SEQUENCE	<p>TTTTTTTTTTTTTTTACAGGGAAACAGCATTTTTAATGTTTTATTTT  TCACTTGTGAAAAATATATATAATATATCTTCCACATACAGAAGAGTC  CTGCAGCTTGAGTCAGAGGAAGCTGAAAGAAAGGCACATACAGGGA  GCAGATCTTCCATACAGTTTTTTCAGTTAAACCAGCATTTCAGGGCACA  GCAAGTGACAACAAAAGCCCAGGCTGCCTGTGCACACGACTCTGAA  GAGAAACATCACAGACAACCCTTAGGTCTACCATGAATGGTTTCTAT  TTAATAGACCTATCAGACCACCCGGAACAAATTGATACCTGAACAA  AGACACACCAGGGCAAGACAAAT</p>

**Figure 3:** Example EST sequence obtained from NCBI GenBank whose expression variance is among the 200 most variable. This sequence has a poly(dT) tract of length 15.



**Figure 4:** (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  $k = 0, 1, 2, \dots, 40$ ; (c) set of boxplots from only one measurement in the expression data of treated TZD-treated adipocytes, each boxplot corresponding to a category of genes  $C_k$  containing the normalized

intensities on the log scale for the genes in  $C_k$ .

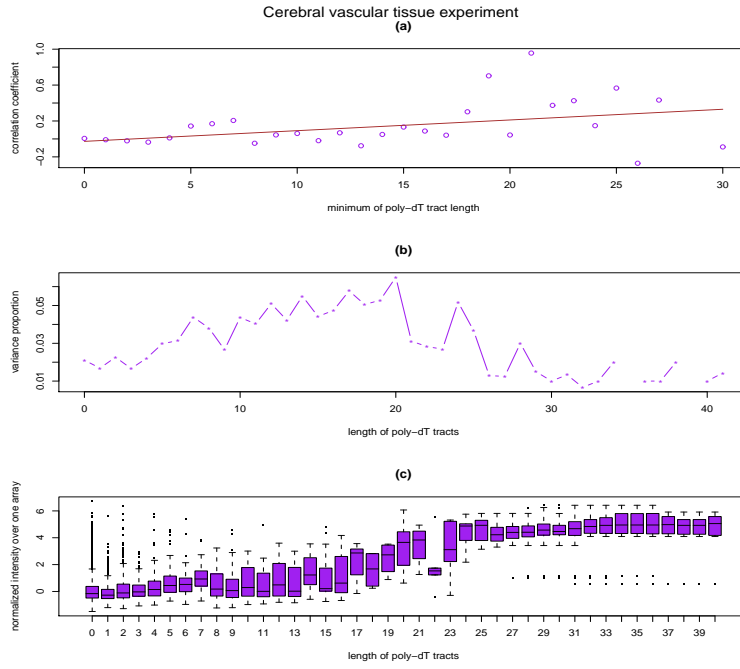


Figure 5: The same as in Figure 3, but for the cerebral vascular tissue experiment.

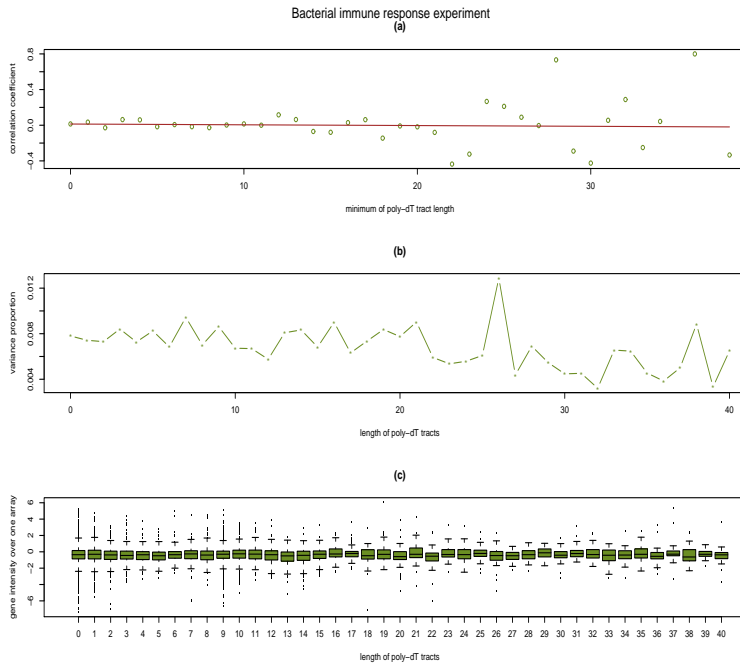


Figure 6: The same as in Figure 4, but for the two-dye data set.

TZD-treated adipocytes experiment	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0027	0.0001	0.0003	0.0011
	0.0006	0.0001	0.0005	0.0001	0.0001	0.0001	0.0001	0.0006	0.0001	0.0001
cerebral vascular tissue experiment	0.0032	0.017	0.0032	0.0001	0.0001	0.0062	0.0066	-	-	-
bacterial immune response experiment	0.0487	0.0985	0.6529	0.0423	0.7412	-	-	-	-	-

**Table 1:** *Estimated p-values permutation trend test applied to 20 arrays in the first single-dye spotted cDNA experiment, to 7 arrays in the second single-dye spotted cDNA experiment, and only 5 measurements in the two-dye(Cy3/Cy5) experiment*