

Short Technical Report

Correcting for Signal Saturation Errors in the Analysis of Microarray Data

BioTechniques 32:330-336 (February 2002)

**L.-L. Hsiao, R.V. Jensen¹,
T. Yoshida, K.E. Clark, J.E.
Blumenstock¹, and S.R.
Gullans**

Brigham and Women's Hospital,
Harvard Medical School, Bos-
ton, MA, and ¹Wesleyan Univer-
sity, Middletown, CT, USA

ABSTRACT

A variety of technical errors have arisen in data analysis when using cDNA or oligonucleotide microarrays. One of the most insidious problems is the saturation of the hybridization signal of high-abundant transcripts. This problem arises from the truncation of the laser fluorescence signal. When the hybridization signal on the microarray is very strong, this truncation can result in serious consequences that may not be readily apparent to the user. As an illustration of this problem, two subclasses of normal human tissue samples (six liver and six lung samples) were analyzed with GeneChip[®] probe arrays to evaluate the patterns of expression for approximately 7000 human genes. Five of these data sets were found to suffer from signal truncation. This caused several tissues to be incorrectly classified using hierarchical clustering. To rectify this problem so that the gene expression data could be properly compared and clustered, we developed a "filtering" procedure that identifies a subset of genes least affected by the signal saturation. This filtering procedure can be obtained at www.hugeindex.org.

INTRODUCTION

DNA and oligonucleotide microarrays are gaining widespread acceptance as powerful tools for genome-wide gene expression analysis. Rigorous statistical analysis and pattern recognition capability are needed now more than ever to extract useful biological information from these data. Currently available biostatistics tools mainly focus on "high-level analysis" (1,6,8,11) that assumes from the outset that the data are free of instrumental defects. Less work has been done on modeling and correcting for systematic errors in the large-scale gene expression measurements (2,5).

Using Affymetrix microarrays (Santa Clara, CA, USA), we have identified a variety of technical errors that can arise in data analysis and give incorrect results when one uses higher-level clustering algorithms. One of the greatest problems is the saturation of the hybridization signal of abundant transcripts. This problem arises from the truncation of the laser fluorescence measurements when the hybridization signal exceeds a maximum intensity. When the hybridization signal for certain probes on the microarray is very strong, this truncation has two serious consequences that may not be readily apparent to the user. First, the probe sets for the most abundant transcripts will be bounded above by a maximal hybridization signal. Then, when the Affymetrix software rescales the data to a mean "target intensity", these maximum signals are all shifted to relatively low values. Second, when both the perfect

match and control mismatch signals of the probe sets for highly abundant transcripts are truncated, the resulting "average difference" computed by the Affymetrix software as a measure of the expression level will be artificially small. The proper way to avoid signal truncation is to lower the sensitivity of the array scanner, and Affymetrix has been routinely adjusting the detector photomultiplier tube (PMT) voltage to remedy these problems. However, a significant amount of data may have been collected before adjustment that suffer from saturation defects, which may only be apparent long after the original samples have been processed.

As an illustration of this problem and its resolution, two subclasses of normal human tissue samples (six liver and six lung samples) were analyzed with Affymetrix GeneChip[®] probe arrays (GeneChip HuGeneFL) to evaluate the mRNA expression patterns for approximately 7000 human genes. Five of these data sets were found to suffer from signal truncation. To rectify this problem so that the gene expression data could be properly compared and clustered, we developed a filtering procedure that identifies a subset of genes least affected by the signal saturation problem.

MATERIALS AND METHODS

Tissue Specimens

Discarded human liver and lung specimens were obtained from 12 individuals, including six males and six fe-

Table 1. The Filtering Process

1. Plot histograms of gene expression levels for all samples (An Excel macro program for plotting histograms of expression data is available at www.hugeindex.org).
2. An abrupt termination of the expression distribution indicates that no genes have expression levels beyond the termination point. This suggests electronic clipping/truncation problems.
3. Identify the highest expression level for each sample. This expression level represents the possible level of signal truncation.
4. The smallest high expression level among all samples will be chosen as the truncation point for filtering the data from all samples.
5. A subset of genes with expression levels greater than the target intensity (target intensity = 100) and less than the truncation point is identified for further analysis.

males with average ages of 69.5 and 63.5 years, respectively. These specimens were provided by tissue banks and had been obtained in the course of surgical procedures (Massachusetts General Hospital and Brigham & Women's Hospital, Boston, MA, USA) with appropriate Institutional Review Board (IRB) consent. Hematoxylin-stained slides were generated from each specimen and reviewed by a pathologist. Only those with normal histological examinations were included in this study.

RNA Preparation for Hybridization

Total RNA was isolated using TRIzol[®] solution (Invitrogen, Carlsbad, CA, USA). As previously described (6, 10), 7 µg total RNA was started for ds-DNA synthesis using the SUPERSCRIPT[™] Choice System (Invitrogen) and a T7-(dT)-24 primer (Geneset Oligos, La Jolla, CA, USA). The cDNA was purified using Phase Lock Gel[™] (Eppendorf-5 Prime, Westbury, NY, USA). In vitro transcription was performed to produce biotin-labeled cRNA using a BioArray HighYield RNA Transcript Labeling Kit (Affymetrix), according to the manufacturer's instructions. The biotinylated RNA was cleaned with the RNeasy[®] Mini kit (Qiagen, Valencia, CA, USA).


Hybridization of RNA to High Density Oligonucleotide Microarrays

Biotinylated cRNA (20 µg) was fragmented and hybridized to microarrays containing oligonucleotide probe sets representing approximately 7000 known human genes (GeneChip HuGeneFL)

using the protocol described previously (6). Briefly, the hybridization mixture was incubated at 99°C for 5 min, followed by incubation at 45°C for 5 min. The hybridization was then carried out at 45°C for 16–18 h. After washing, the array was stained with streptavidin-phycoerythrin (Molecular Probes, Eugene, OR, USA) and amplified with biotinylated anti-streptavidin antibody (Vector Laboratories, Burlingame, CA, USA). The intensity of all of the features of the microarrays were captured and examined for artifacts using Affymetrix GeneChip software version 4.0 according to standard procedures (7). The GeneChip software was used to generate quantitative gene expression values measured by the average difference between the hybridization intensity and the perfect match and mismatch probe sets. The raw expression levels were then multiplied by a scaling factor to make the mean expression level on the microarray equal to a target intensity of 100. This scaling is automatically performed by the Affymetrix software to normalize the gene expression levels to allow comparison between any two samples.

Quality Control of Samples

To initially assess total RNA degradation, a portion of the RNA from each sample was dissolved on a 1% agarose/formaldehyde gel using standard procedures. The samples with unsatisfactory quality were discarded. Each probe array contains several prokaryotic genes that serve as hybridization controls for RNA spiked into experimental samples. In addition, before hybridization to ex-



perimental arrays, the quality of cRNA was assessed using test arrays (Test2 gene arrays; Affymetrix) designed to compare the relative expression levels of several housekeeping genes, including β -actin and GAPDH, using oligonucleotide probes complementary to both the 3' and 5' ends of gene products. According to the manufacturer's instructions, when the ratio of the average difference of the 3' end to the 5' end of gene products is equal or less than 3, the cRNA quality is deemed satisfactory. Data that failed to meet this criterion were excluded from analysis.

Data Analysis

A hierarchical clustering algorithm (AGNES) in the statistical package SPLUS (9) was used to classify all 12 samples according to the relative variation in gene expression patterns. The gene hybridization intensities (from GeneChip software) were appropriately scaled to a target intensity of 100 to facilitate the comparison of the data from all arrays. A Microsoft® Excel® macro program was created for plotting histograms of expression data (available at www.hugeindex.org). To minimize the effects of unreliable expression levels of genes with small values for the average difference, we restricted our analysis to genes whose expression levels exceeded the target intensity in at least one of the 12 samples. Since the majority of genes on the arrays are called "Absent" by the Affymetrix software, most of the expression levels are at or below the target intensity (100) and can be considered to be background noise (2). The cluster analysis was initially restricted to the subset of 628 genes that met this criterion.

RESULTS AND DISCUSSION

Identification of Signal Truncation

We used oligonucleotide microarrays (GeneChip and HuGeneFL) to evaluate the mRNA expression patterns of approximately 7000 human genes for samples of two normal human tissues (six liver and six lung samples) (3). A scatter plot of gene expression levels for approximately 7000 genes was used to assess the gene expression correlations be-

tween pairs of samples of the same tissue type. Under identical scaling and normalization processes, a comparison of two lung samples (Lung001 and Lung002) prepared from different individuals using old generation chips (low signal sensitivity) revealed a 45° linear scatter-plot distribution (Figure 1A), as was expected. However, a comparison of Lung002 with a different sample (Lung014) using one of the new-generation chips with enhanced signal sensitivity clearly revealed signal saturation in the Lung014 data. Figure 1B shows the signal saturation as an apparent plateau of expression level in Lung014 at a value of about 1400. This was caused by the truncation of the laser fluorescence measurement at the maximal intensity and the subsequent application of the manufacturer's normalization algorithm, leading to a decrease in the calculated expression level. Interestingly, when two samples that manifest truncation are compared with each other (data not shown), there are no obvious signs of saturation, aside from lower than usual maximum expression levels. Using the scatter plots to compare pairs of samples, we identified three liver samples (Liver008, Liver009, and Liver010) and two lung samples (Lung014 and Lung018) containing truncated signals.

Signal Truncation Affects Accurate Sample Classification

We first selected a subset of 628 genes that exhibit expression levels greater than the target intensity of 100 in any one of the 12 samples (six liver and six lung samples). This selects the genes that are most likely to be called "Present" by the Affymetrix software algorithm and that exhibit the least variability (2). We subjected this subset of genes to a hierarchical clustering algorithm (4,9) to group the tissue samples. Instead of neatly separating the data into one liver and one lung cluster, the three liver (LI008–LI010) and two lung (LU014 and LU018) samples that had truncated signals were put into a third subclass (Figure 2A). The hierarchical clustering of these data led to the incorrect conclusion that there were three subclasses of tissue samples. Furthermore, the truncated signals can distort the calculation of fold changes for individual genes.

MICROARRAY *Technologies*

Filtering Genes with Electronically Truncated Signals

To correct this problem, we developed a stringent filtering process to identify the subset of genes that were not affected by truncation in any of the samples. Our approach was to first survey the distribution pattern of gene expression for each sample using a histogram of the average differences. Samples that

did not have truncated signals yielded a histogram showing the gene expression distribution in a bell-shaped curve (Figure 1C) with a long tail extending to high expression levels. However, samples that contained truncated signals showed an abrupt termination at the expression level of truncation (Figure 1D). For example, the distribution of expression levels for Lung014 showed a sharp drop at an expression level of 1400 (Fig-

ure 1D). After surveying the 11 other liver and lung samples, expression-level distributions were also sharply truncated for liver008, 009, and 010 and lung018 at levels of 3100, 1900, 2600, and 2900, respectively, while those for liver002, 004, and 005 and lung001, 002, 004, and 005 extended smoothly to values greater than 10000. These truncation levels (after software normalization) differ from chip to chip because of variations in the

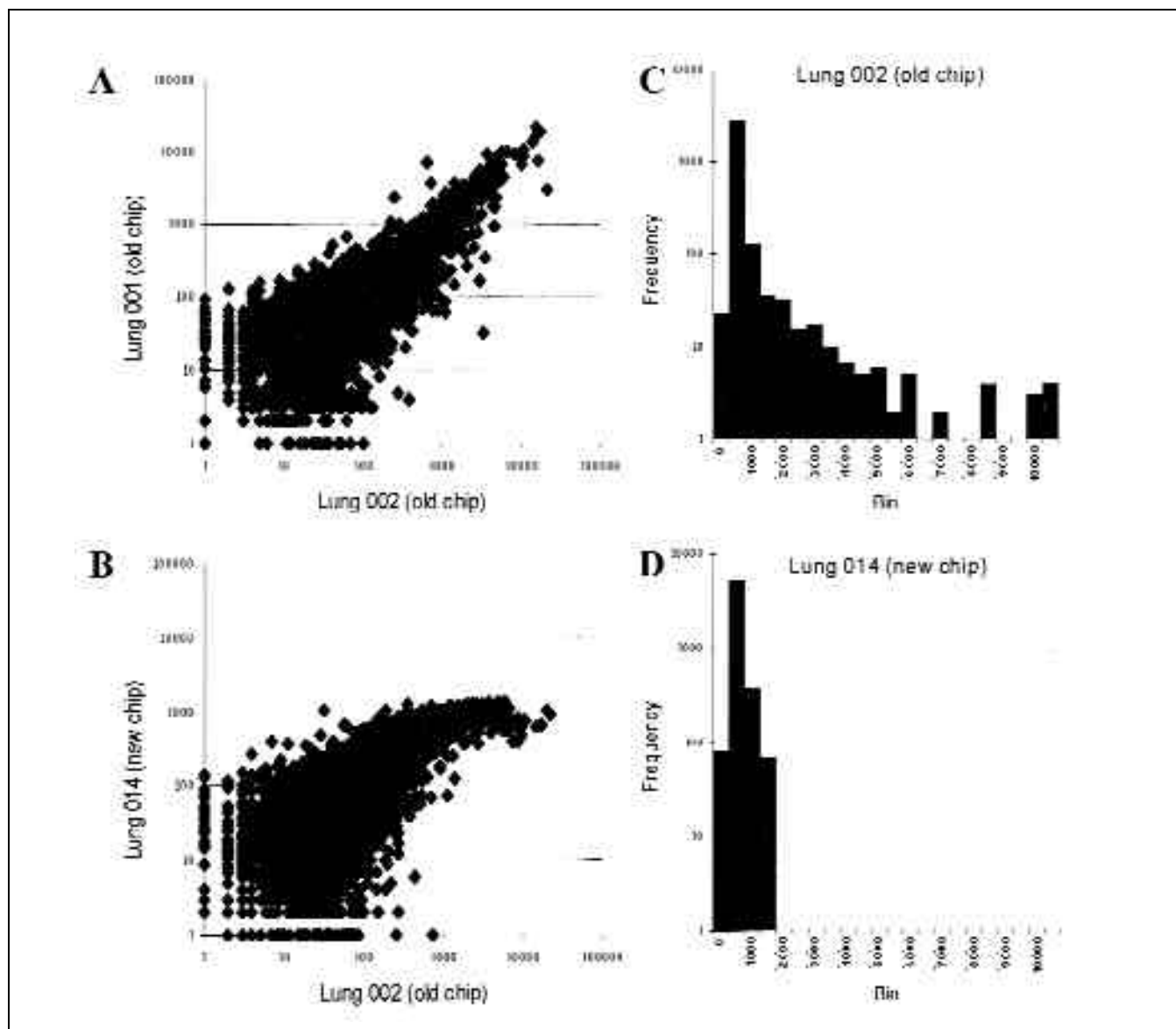


Figure 1. Scatter plots (A and B) and histograms (C and D) of gene expression levels for approximately 7000 genes using old-generation versus new-generation Affymetrix chips. (A) Lung001 and Lung002 were lung samples from two different individuals run on the old-generation chips that showed a 45° linear scatter plot distribution. (B) Lung002 and Lung014 were lung samples from two different individuals, with Lung002 run on the old-generation chip and Lung014 on the new-generation chip. The gene expression levels of Lung014 showed a plateau pattern at level of approximately 2000. (C) The histogram for the expression levels for Lung002, which did not have truncated signals, showed a bell-shaped curve. (D) The histogram for the expression levels for Lung014, which contained truncated signals, revealed an abrupt termination of the expression distribution at the level of truncation. The X-axis represents the gene expression level in bins for 100 bins of width 100 between the expression levels of 0 and 10000, and the Y-axis represents the frequency of gene expression on a logarithmic scale.

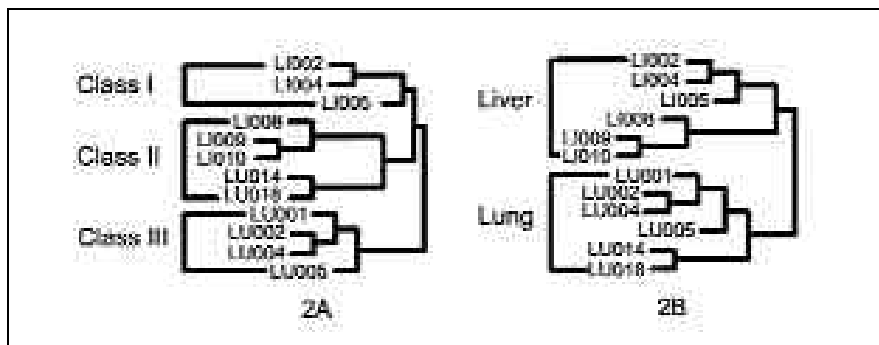


Figure 2. Hierarchical clustering of 12 normal human tissue samples. (A) Using a subset of 628 genes that include those with saturated signals, the cluster analysis divided the tissues into three clusters, suggesting that the samples contain three subclasses instead of two. (B) After applying the filtering procedure to the data, a subset of 376 genes that exclude those with saturated signals, the samples were clustered correctly into two subclasses. Liver tissue includes LI002, LI004, LI005, and LI008-010, and Lung tissue includes LU001-2, LU004-5, LU014, and LU018.

RNA quality and the hybridization of the samples. Therefore, we concluded that genes in any sample with expression levels above the minimum truncation level of 1400 could not be reliably compared between all of the samples and must be excluded from the data set. Together with the requirement of expression levels greater than the target intensity of 100 in all samples, this filtering procedure generated a subset of 376 genes. A hierarchical clustering analysis applied to this subset of genes revealed the correct classification of the two tissue classes, liver and lung (Figure 2B).

In summary, we propose a stringent filtering procedure (Table 1) for oligonucleotide microarray data to identify the most reliable gene expression measurements. By this method, the genes with extremely high and low expression are excluded, leaving those of intermediate expression levels. Here, we report two observations: (i) the genes with intermediate expression levels are the least variable and are less likely to suffer truncation problems and (ii) the genes with intermediate expression levels maintain "sample-specific" expression patterns. Using this subset of filtered genes for hierarchical clustering analysis, we were able to exclude the defective data and correctly group the tissues. This procedure for identifying saturation problems and correcting the data set should be essential when such data are used for class discovery, as in cancer studies. It also eliminates erroneous identification of differentially expressed genes related solely to truncation arti-

facts. In particular, this procedure will be a useful tool for comparing data generated from pre- and post-scanner adjustments and after other signal enhancing improvements. Finally, we note that these signal saturation problems are not unique to Affymetrix GeneChip data and should be evaluated and corrected when using any microarray gene expression platform.

REFERENCES

- Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96:6745-6750.
- Butte, A.J., J. Ye, H.U. Haring, M. Stummvoll, M.F. White, and I.S. Kohane. 2001. Determining significant fold differences in gene expression analysis. *Pac. Symp. Biocomput.*, p. 6-17.
- Hsiao, L.-L., F. Dangond, T. Yoshida, R. Hong, R.V. Jensen, J. Misra, W. Dillon, K.F. Lee et al. 2001. A compendium of gene expression in normal human tissues. *Physiol. Genomics* 7:97-104.
- Kauffman, L. and P.J. Rousseeuw. 1990. *Finding Groups in Data*. Wiley & Sons, New York.
- Li, C. and W.H. Wong. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98:31-36.
- Lockhart, D.J., H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14:1675-1680.
- O'Dell, S.D., S.R. Bujac, G.J. Miller, and I.N. Day. 1999. Associations of IGF2 ApA

RFLP and INS VNTR class I allele size with obesity. *Eur. J. Hum. Genet.* 7:821-827.

- Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907-2912.
- Venables, W.N. and B.D. Ripley. 1997. *Modern Applied Statistics with S-Plus*. Springer Publishing Company, New York.
- Warrington, J.A., A. Nair, M. Mahadevappa, and M. Tsyganskaya. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics* 2:143-147.
- Wodicka, L., H. Dong, M. Mittmann, M.H. Ho, and D.J. Lockhart. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15:1359-1367.

The first two authors contributed equally to this manuscript. We thank Dr. Raphael Bueno and Mohammed Miri for assistance with tissue acquisition. We also thank Nathan Best and Kent Auerbach for technical assistance and Web site design. This work was supported by the Merck Genome Research Institute. In addition, support was provided by the National Institutes of Health grant nos. DK-36031 and DK58849 to S.R.G. and DK09987 to L.-L.H.

Received 25 June 2001; accepted 13 November 2001.

Address correspondence to:

Dr. Steven R. Gullans
65 Landsdowne St.
Cambridge, MA 02139, USA
e-mail: sgullans@rics.bwh.harvard.edu

For reprints of this or
any other article, contact
Reprints@BioTechniques.com