Journal of Biochemistry and Molecular Biology, Vol. 36, No. 6, November 2003, pp. 558-564

JOURNAL OF Biochemistry and Molecular Biology

Determining a Detectable Threshold of Signal Intensity in cDNA Microarray Based on Accumulated Distribution

Xia Gao[†,§], Xuping Fu[§], Tao Li[‡], Jian Zi[†], Yao Luo, Qing Wei, Erliang Zeng[‡], Yi Xie, Yao Li* and Yumin Mao*

State Key Laboratory of Genetic Engineering, Institute of Genetics, School of Life Science, Fudan University, Shanghai 200433, P. R. China

[†]Physics Department, Fudan University, Shanghai 200433, P. R. China

[‡]Shanghia Biostar Genechip Istitute, Shanghai 200092, P. R. China

Received 27 March 2003, Accepted 4 June 2003

In microarray data mining, one of the key problems is how to handle weak signals. Based on a bent piecewise linear accumulated distribution generally found in the microarray data, a new detectable threshold finding method is proposed to filter genes with unreliable information in this paper. More reliable and reproducible data is produced for the subsequent data mining.

Keywords: Accumulated distribution, cDNA microarray, Low intensity spots

Introduction

cDNA microarray analysis has become the most widely used technique for the study of gene expression patterns on a genomic scale (Schena *et al*., 1995 and 1996). One of the most critical steps in microarray experiments is to accurately assess the expression ratios between the sample and reference in a dual colour experiment, because most subsequent data analysis, such as cluster analysis (Eisen *et al*., 1998), depends heavily on the accuracy of these ratios (Yang *et al*., 2001). The reliability of ratios is subject to signal intensity and noise caused by background and non-specific hybridization. The ratio is found to be accurate if the signal intensity is in or above the moderate level. However, if the signal intensity is relatively low, the ratio will fluctuate as the noise will mask or bias the intensity.

It has also been noticed that the data based on a single array may not be reliable and may contain many uncertainties (Lee *et al*., 2000). A Gordian knot of handling uncertainties is indistinguishability between weak signals and noises arising from non-specific hybridization of the labelled samples to elements printed on the microarray, print-tip effects, slide inhomogeneities and variability in RNA isolation, purity, labelling and detection (Tseng *et al*., 2001; Bilban *et al*., 2002; Yang *et al*., 2002). Although many of the weak signals are often ignored; nevertheless, the weak signals from rare transcripts do contain valuable information and should be kept for further analysis (Kooperberg *et al*., 2002). This paper intends to stress the importance and necessity of distilling useful information from weak signals.

Repeating the experiment, as one strategy dealing with uncertainties, may not be cost-effective and it further increases data complexity. Yet its indispensability can not be neglected. However, one way to complement experimental repetition is to adopt a proper threshold. Signals lower than the threshold are considered as invalid data, whereas those higher than the threshold are effective. Several threshold finding methods have been reported, such as the arbitrary fluorescence intensities (Sakai *et al*., 2000), relative errors in cy3/cy5 ratios (Tseng *et al*., 2001), certain signal-to-background ratios (Cunningham *et al*., 2000), and the value of mean (after local background subtraction) plus two standard deviations of the negative reference sample (Epstein *et al*., 2001). In this paper, a new method that aims to help determine the detectable threshold of signal intensities based on accumulated distribution is introduced.

Accumulated Distribution Based Threshold Method

Data employed Three sets of cDNA microarray data were

[§]These two authors contribute equally to this paper.

*To whom correspondence should be addressed.
Tel: 8621-65633936, Fax: 8621-65633919
E-mail:yaoli@fudan.edu.cn

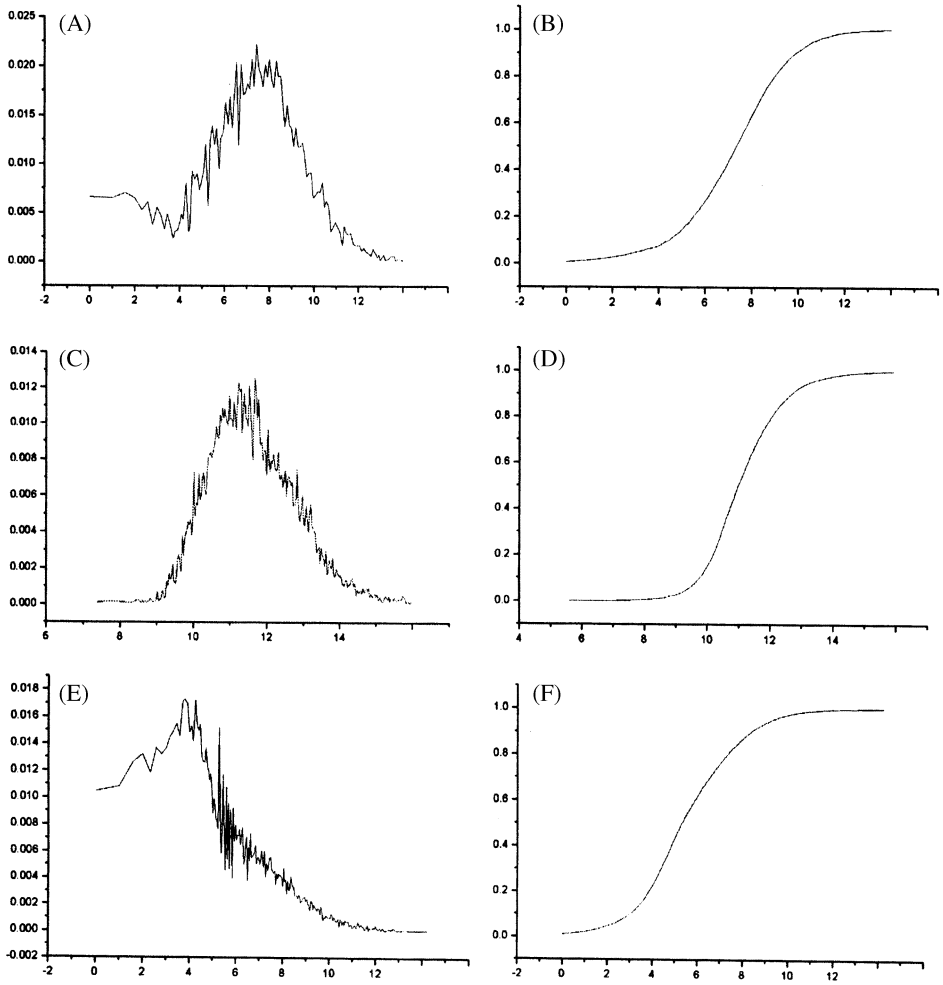

Fig. 1. Graphic representation of spot intensity distribution function and accumulated distribution function of different types of microarrays. The left hand plots show the density distribution and the right hand ones show the accumulated distribution. What the X-axis of graph represents is the logarithm value of signal intensity with the base of 2. The Y-axis is the ratio of the number of spots below corresponding intensity vs. the total number. (A) and (B) show the distribution of channel 1 intensity of cdc 15 010 min array in yeast cell cycle regulated category. This array contains 7,267 valid spots and the image analysis software is ScanAlyze. (C) and (D) are plotted by using the data of the homemade chip including 13,416 valid spots. The image analysis software is GenePixPro 4.0. (E) and (F) show the distribution of channel 2 intensity of Blood B cells; anti-IgM 24 h || 1c8n057 array in Diffuse large B-cell lymphoma. This array contains 18,433 valid spots and the image analysis software is ScanAlyze.

used to explore the distribution of spot intensities: that of the study in yeast *Saccharomyces cerevisiae* focusing on identification of cell-cycle regulated genes (Spellman *et al.*, 1998), of the study in cancer of diffuse large B-cell lymphoma (DLBCL) (Alizadeh *et al.*, 2000), and of homemade chips. The arrays of cell-cycle and diffuse large B-cell lymphoma contain about 8,000 and 18,000 spots, respectively. The image analysis software is ScanAlyze. Homemade arrays contain 14,112, 12,800, 10,368 and 8,464 spots. The image analysis software includes GenPixPro and ArrayPro. For arrays with 10,368 spots, nine experiments representing tissue-specific expression profiles were carried out. Each experiment was repeated three times.

Description of the method A typical microarray spot intensity density distribution is heavily skewed with most spots having low intensities. A lognormal provides a good approximation to the bulk of the microarray data with a disagreement in two side tails (Hoyle *et al.*, 2002). Accumulated distributions are from the intensity distribution (Fig. 1). As Fig. 1 shows, the density distribution is unimodal. The accumulated distribution obtained from a unimodal density function shows a smooth curve picture that is similar to the standard normal accumulated distribution plot.

One common feature of the accumulated distributions is that the shape of the curve is fairly stable in spite of variation of spot number of an array and the image analysis software. Through our experiment, we have observed that different spot numbers affects the scale of vertical axis. Varied signal intensity brought by different experiment and data analysis

Table 1. Results of straight lines curvefitting in the central part of accumulated distribution

Chip name		t-value of line fitting in central part
10,384-spot	Channel 1	0.119
	Channel 2	0.351
12,800-spot	Channel 1	0.275
	Channel 2	0.469
14,112-spot	Channel 1	0.380
	Channel 2	0.220
cdc15 015 min	Channel 1	0.575
	Channel 2	0.699
cdc15 015 min	Channel 1	0.583
	Channel 2	0.679
cdc15 070in	Channel 1	0.384
	Channel 2	0.472
Blood B cells;anti-IgM 6h	Channel 1	0.661
	Channel 2	0.674
Blood B cells;anti-IgM 24h	Channel 1	0.175
	Channel 2	0.685
Blood B cells	Channel 1	0.205
	Channel 2	0.610

First three chips are homemade. In the linear fitting process of the central section, the points with accumulated function values between 0.2 and 0.8 are selected. If the t-value of fitting is below 0.95, the central part has no significant difference from a line.

software impacts on the scale of horizontal axis. However, the plots of accumulated distribution are quite similar in shape. In addition, the plot can be approximately divided into three parts. The top section is a straight line parallel to X-axis. The central part shows a definitely linear character proved by t-statistics of regression (Table 1). The bottom of plot is a slightly slanted line with a linear character sometimes distorted by the uneven local background. For convenience, what is indicated above can be summarized as the bent piecewise linear accumulated distribution function.

According to the piecewise linear curve, the genes were divided into three groups respectively: high intensity spots, varied intensity spots in a great range, and low intensity spots. There is no question that a certain proportion of non-expressed genes always exists in all kinds of tissues or cell samples; therefore genes contributed to the bottom part of the curve are probably non-expressed genes. However, the non-expressed genes and failed genes generally show oscillatory intensity due to noise effect. The problem that subsequently arises is that a definite separation cannot be defined if the intensities of expressed and non-expressed genes are all close to the noise level. Therefore, the inflexion point of the piecewise linear curve is chosen as the appropriate threshold (T_D) (Fig. 2). If the threshold is lower than T_D, more false positive genes will be brought in. If slightly higher than T_D, more significantly expressed genes will be eliminated dramatically.

In summary, the method takes signal and background intensity distribution together into consideration. Using this method, almost all non-expressed genes and failed genes with intensities below the threshold would be filtered out, while most low abundance genes would be kept as true signal spots though they are close to the turning point. The method described in this section is called as AD method in subsequent discussion.

Validation of the Method

Data reliability proved by primer experiment In order to see whether or not failed spots affect the T_D, the primer hybridization experiment was carried out. Primer hybridication experiment was performed as the quality control for each batch of array. 0.02 nmol of universal primer for PCR amplification, which labeled with Cy3 at the 5'-end, was hybridized against the arrays. The signal intensity represents the DNA quantity of each corresponding spot. The failed spots, which mainly attribute to failure in PCR or

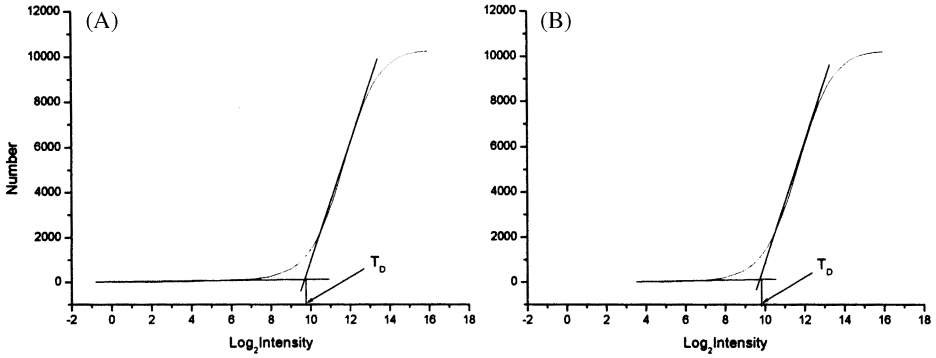

Fig. 2. Determination of T_D. The bottom and the central parts of accumulated distribution function curve are picked out for linear fit. Two linear fitting lines intersect at one point. The intensity corresponding to the intersection point is considered as the threshold (T_D). (A) is a typical microarray accumulated distribution function; (B) is accumulated distribution function of intensities which failed genes are excluded according to the primer experiment data. There is no change in both T_Ds.

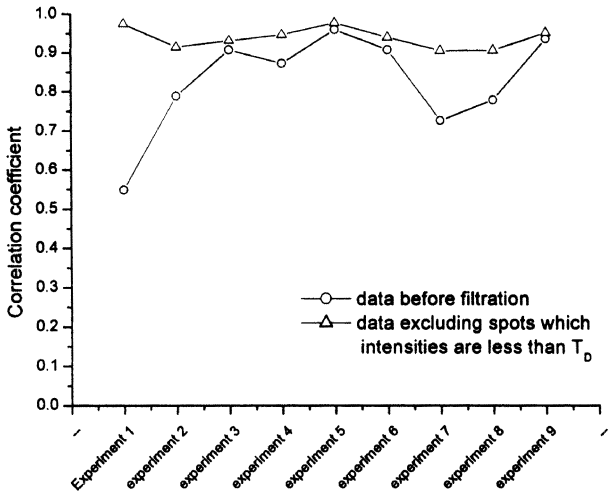

Fig. 3. A correlation coefficient diagram for the homemade replicate microarrays.

hybridization, can be identified with this method. Usually the same batch of array has the same group of failed spots. Utilizing the distribution function diagram, T_Ds were calculated before and after filtrating failed spots. The result shows that no change occurs in the value of T_Ds (Fig. 2), which means that failed spots have no effect on the method. In fact, the failed genes identified by the primer hybridization experiment are exactly plotted into the bottom part of distribution function. With this method, most of the failed spots can be distinguished from the useful data and would not be falsely taken into subsequent analysis.

Improvement of data reproducibility after filtering The method was applied in the homemade replicate microarrays. Correlation coefficient (R) and consistence rate (CR) were used to evaluate reproducibility of replicate experiments before and after filtration.

In the repeated experiments, R value between log-transformed ratios from two replicates of three are calculated. The initial data and filtrated data using AD method are compared. R was calculated as Pearson correlation coefficient.

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}}$$

x and y denote ratio values of replicates. As Fig. 3 shows, R value of initial data is about 0.8. After using the method, R value, which increase to above 0.95, is becoming higher and steadier.

CR is then considered to evaluate the method. In two replicates, genes differentially expressed in both chips are denoted as common differentially expressed genes, including consistent differentially expressed and contradictory differentially expressed genes. Here the gene with ratio above 2 or below 0.5 is considered as a differentially expressed gene. CR is the proportion of the number of consistent differentially expressed genes vs. the total number of common differentially expressed genes.

$$CR = \frac{N_d - 2 \times N_{cd}}{N_d}$$

N_d and N_{cd} denote the total number of common differentially expressed genes and contradictory differentially expressed genes in the replicates, respectively. Since a gene cannot contradictory differentially express in duplicate, CR value could evaluate the reproducibility of replicates in a microarray experiment. The CR value should be 1 theoretically. However it deviates from 1 before filtration, which is owing to the noise or uncertainty in a microarray experiment. After filtration, it is much closer to 1 in that spots after filtration mainly include expressed genes which background intensity has little effect on the estimation of ratios.

Obviously, discarding the spots filtered by the threshold greatly improve data consistency between duplicate, which can be proved by R and CR values.

Comparison of three determining methods Many other methods of finding the threshold have been reported. In

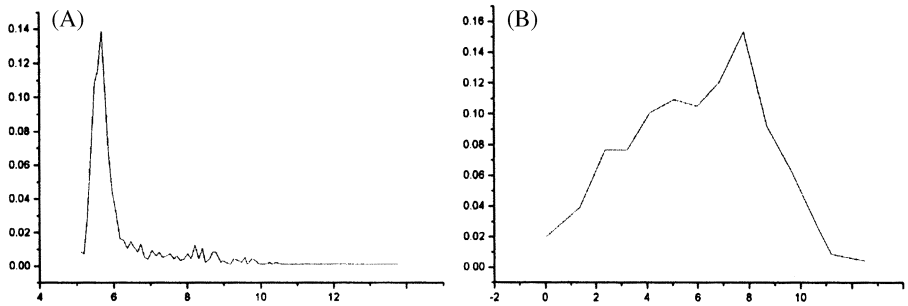

Fig. 4. Graphic representation of negative spots intensity density distribution function. The two arrays used to plot are all in yeast cell cycle regulated category. The left is from cdc15 080 min array, which the number of valid negative spots is 967. It shows that the negative spots can be approximated by a normal distribution when the number is high enough. The right is from cdc15 070 min array, which the number of valid negative spots is 457. The distribution is seriously distorted owing to many invalid negative spots.

Table 2. Comparison of three methods

Chip name		BSD method	NSD method	AD method
cdc15 030 min	Threshold in channel 1	950	989	581
	Threshold in channel 2	584	2614	631
	Valid genes number	2354	906	4529
	HQSLR	96.15%	65.38%	100%
cdc15 050 min	Threshold in channel 1	1013	721	534
	Threshold in channel 2	990	1194	564
	Valid genes number	1104	1395	3338
	HQSLR	89.87%	91.14%	100%
cdc15 080 min	Threshold in channel 1	509	1209	573
	Threshold in channel 2	273	2677	312
	Valid genes number	6338	298	5448
	HQSLR	100%	35.46%	100%
Blood B cell	Threshold in channel 1	398	332	173
	Threshold in channel 2	1747	2676	595
	Valid genes number	771	424	2969
	HQSLR	61.15%	47.11%	100%
Blood B cells; anti-IgM 24 h ||1c8n057	Threshold in channel 1	87	221	83
	Threshold in channel 2	710	1140	280
	Valid genes number	1051	512	5216
	HQSLR	70.49%	60.65%	100%
Blood B cells ||1c8n076	Threshold in channel 1	629	443	298
	Threshold in channel 2	225	367	152
	Valid genes number	1905	3074	7850
	HQSLR	67.46%	83.33%	100%

HQSLR represents rate of high quality spots left.

general, the value of mean background plus two standard deviations (BSD) is considered as an intensity threshold to identify background noise and true signal. Besides, mean blank or negative controls plus two standard deviations (NSD) is used as the minimum detectable value.

In NSD method, the number of valid negative controls in an array is too little to be normally distributed (Fig. 4). Furthermore, the NSD method is subject to the controls themselves, such as the selection of controls, the quality of controls. Sometimes negative control spots show high signal intensities because of pin contamination and cross hybridization, which will strongly affect the threshold value. For BSD method, the background derived from surface fluorescence upon laser excitation and non-specific combination of probes can be approximated by a normal distribution. This property can be readily assessed by the histogram of any background region of microarray images. If the local variance in background were significant, the threshold value would be very high. Many significantly expressed genes will be excluded by NSD method or BSD method. AD method is based on the range of the whole array intensity. Many valuable genes with significant biological function will be preserved for the down-stream analysis.

It is clear that when the signal-to-noise ratio is low, the intrinsic variation in the data is high and confidence in the accuracy of the data is low. When a spot gives significant intensities (at least 3 times the background, where signal is the mean intensity level of the spot and background is the mean local background.), the measured fluorescent intensity can be assumed to reflect the signal intensity. We denote these spots as high quality spots, which are significant and accurate in the subsequent analysis.

Here thresholds are calculated with the three methods mentioned above by using the data from yeast cell cycle regulated category and that from cancer of diffuse large B-cell lymphoma. After filtration by each method, number of valid genes and rate of high quality spots left (HQSLR, equal to high quality spots left subtract total high quality spots) are calculated (Table 2). As Table 2 showed, the thresholds of AD method are generally lower and the numbers of valid genes are much larger than those of the other two methods. In addition, the HQSLRs are all 100% in AD method while they are much lower in the other two methods. It means that most significant genes will be preserved to the following analysis. In most cases, the disadvantages of NSD methods, such as lack of enough valid negative control spots, the influence on control selection and quality, will result in the high threshold and much smaller number of significant genes. Therefore, only BSD method and AD method are compared in following text.

Self-self experiments (same sample vs. same sample) were

Table 3. Comparison of two methods by self-self experiment

Chip name	Background method				Distribution method			
	Threshold in channel 1	Threshold in channel 2	Valid genes number	FPR	Threshold in channel 1	Threshold in channel 2	Valid genes number	FPR
Self1	895	454	3290	0.97%	514	608	4732	1.37%
Self2	702	551	3807	0.35%	552	632	4246	0.38%
Self3	192	605	4595	1.04%	430	586	4478	0.41%
Self4	414	377	5639	3.11%	464	463	5081	2.28%
Self5	200	832	3790	0.77%	454	622	4609	0.79%
Self6	192	1002	3805	1.18%	498	663	5067	1.43%

FPR represents false positive rate.

then used in comparison. Theoretically, the ratios of Cy5 to Cy3 should be 1 across all spots in self-self experiments. There are numbers of genes with ratios higher than 2 or lower than 0.5 which are called as false positive genes. Proportion of false positive genes out of the total genes is calculated as false positive rate (FPR) (Table 3). Most of the false positive genes are located in lower signal intensity range before filtration. It is shown in Table 3, AD method preserves more valid genes with low intensity than BSD method without obvious increase of FPR, or keeps the similar number of valid spots as BSD method with significantly decrease of FPR.

At last, we calculated the differentially expressed genes of the array of the cdc15 030 min array of yeast cell regulated category after filtering with BSD method and AD method, respectively. Here a gene with fold change above 3 or below 0.333 can be considered differentially expressed. Seven genes are found in BSD method but not found in AD method, where none is obviously functioning in the process of cell cycle. And 34 genes are found in AD filtering method but not found in BSD filtering method, where at least 4 genes (YPL256C (G1/S transition of mitotic cell cycle), YOR058C (mitotic anaphase B), YJR066W (meiosis), YDL127W (cell cycle) are obviously functioning in cell cycle.

Therefore, AD method relatively can decrease more noise and miss less useful information comparing with traditional methods. It adapts well to the large-scale microarrays, but not applicable to arrays with small number of spots, where the failed genes and non-expressed genes would not be enough to form a line-like bottom. As the spot number increases, the curve bottom becomes longer and smoother, and the threshold from this method becomes more accurate.

Discussion

In cDNA microarray experiments, a lot of spots with low signal intensities are vulnerable to background and noise biases. It is important to determine an effective threshold, with which one can clearly distinguish the low abundance genes from background. Common piecewise linear accumulated distribution is generally found in cDNA microarrays data, so a new threshold determining method for gene expression intensity based on the accumulated distribution is proposed in the paper. Compared with previous methods, it takes the overall signal intensity and background into consideration. Using this method, the reproducibility and reliability of microarray experiments are greatly increased, and more valuable genes with significant biological function are preserved for further analysis.

Normalization is necessary to eliminate bias between samples before data mining. Many widely used normalization methods, such as all-gene normalization, LOWESS normalization (Yang *et al.*, 2002), are subject to almost all the spot intensities in the array. The inconsistence of every threshold finding method lies in low intensity signals, which impact or normalization slightly. Therefore, normalization will not be greatly influenced after filtration by the different threshold finding method.

A major goal of microarray experiments is to find the genes differentially expressed between samples (Ko *et al*., 2002; Lee *et al*., 2002). Ratio values have a systematic dependence on intensity, which most commonly appears as a deviation from zero for low intensity spots (Quackenbush, 2002). Our new threshold method can discard such spots and remove such intensity dependent effects in ratio values, so that the accuracy of ratio estimates of microarray experiments increase. At present, clustering techniques as a kind of effective data mining tool has been developed and applied to identify groups of genes with similar expression patterns. The ratio value is the data basis of clustering. More precise ratio is propitious to obtain more signifcant clusters to explain biological process.

The identification of valuable weak signals from background or noise is a common issue in microarray technology. Every method has its own advantages and disadvantages. Despite the merits, our method cannot apply well to microarrays with relative small number of spots. In addition, a few false positive genes may be brought in when more significant genes are analyzed. Our proposed AD method attempts to answer - without increasing FPR - the challenge called forth during the preservation of the low intensity spots. The proposed AD method may appear better capable of dealing with weak signals, yet it does not

completely resolve the interference of weak signals - the comprehensive solution may lie with the deepening of methodology and the improvement of experiment technology.

Acknowledgments This work was supported by a grant 2002AA2Z2002 from the National High Technology Research and Development Program of China (863 Program).

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

Bilban, M., Buehler, L. K., Head, S., Desoye, G. and Quaranta, V. (2002) Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer. *BMC Genomics* **3**, 19.

Cunningham, M. J., Liang, S., Fuhrman, S., Seihamer, J. J. and Somogyi, R. (2000) Gene expression microarray data analysis for toxicology profiling. *Ann. N. Y. Acad. Sci.* **919**, 52-67.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868.

Epstein, C. B., Hale, W. 4th and Butow, R. A. (2001) Numerical methods for handling uncertainty in microarray data: an example analyzing perturbed mitochondrial function in yeast. *Methods Cell Biol.* **65**, 439-52.

Hoyle, D. C., Rattray, M., Jupp, R. and Brass, A. (2002) Making sense of microarray data distributions. *Bioinformatics* **18**, 576-584.

Ko, J., Na, D. S., Lee, Y. H., Shin, S. Y., Kim, J. H., Hwang, B. G., Min, B. I. and Park, D. S. (2002) cDNA microarray analysis of the differential gene expression in the neuropathic pain and electroacupuncture treatment models. *J. Biochem. Mol. Biol.* **35**, 420-427.

Kooperberg, C., Fazzio, T. G., Delrow, J. J. and Tsukiyama, T. (2002) Improved background correction for spotted DNA microarray. *J. Comput. Biol.* **9**, 55-66.

Lee, M. -L. T., Kuo, F. C., Whitmore, G. A. and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* **97**, 9834-9839.

Lee, S. W., Kwak, H. B., Lee, H. C., Lee, S. K., Kim, H. H. and Lee, Z. H. (2002) The anti-proliferative gene TIS21 is involved in osteoclast differentiation. *J. Biochem. Mol. Biol.* **35**, 609-614.

Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.* **32 Suppl**, 496-501.

Sakai, K., Higuchi, H., Matsubara, K. and Kato, K. (2000) Microarray hybridization with fractionated cDNA: enhanced identification of differentially expressed genes. *Anal. Biochem.* **287**, 32-37.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. and Davis, R. W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* **93**, 10614-10619.

Spellman. P. T., Sherlock, G. Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9**, 3273-97.

Tseng, G. C., M. K., Rohlin, L., Liao, J. C. and Wong, W. H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29**, 2549-2557.

Yang, M. C. K., Ruan, Q. G., Yang, J. J., Eckenrode, S., Wu, S., McIndoe, R. A. and She, J. X. (2001) A statistical procedure for flagging weak spots greatly improves normalization and ratio estimates in microarray experiments. *Physiol. Genomics* **7**, 45-53.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data; a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15.