Genomics & Informatics Vol. 1(2) 94-100, December 2003

Rank-Based Nonlinear Normalization of Oligonucleotide Arrays

Peter J. Park[1]*, Isaac S. Kohane[1] and Ju Han Kim[2]

[1] Children' s Hospital Informatics Program, Children' s Hospital, Harvard Medical School, Boston, MA 02115, USA

[2] SNUBI: Seoul National University Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Republic of Korea

Abstract

Motivation: Many have observed a nonlinear relationship between the signal intensity and the transcript abundance in microarray data. The first step in analyzing the data is to normalize it properly, and this should include a correction for the nonlinearity. The commonly used linear normalization schemes do not address this problem.

Results: Nonlinearity is present in both cDNA and oligonucleotide arrays, but we concentrate on the latter in this paper. Across a set of chips, we identify those genes whose within-chip ranks are relatively constant compared to other genes of similar intensity. For each gene, we compute the sum of the squares of the differences in its within-chip ranks between every pair of chips as our statistic and we select a small fraction of the genes with the minimal changes in ranks at each intensity level. These genes are most likely to be non-differentially expressed and are subsequently used in the normalization procedure. This method is a generalization of the rank-invariant normalization (Li and Wong, 2001), using all available chips rather than two at a time to gather more information, while using the chip that is least likely to be affected by nonlinear effects as the reference chip. The assumption in our method is that there are at least a small number of non-differentially expressed genes across the intensity range. The normalized expression values can be substantially different from the unnormalized values and may result in altered down-stream analysis.

Keywords: gene expression, microarray normal: Edtion, rank statistic

* Corresponding author:
E-mail peter-park@harvard.edu, Tel +1617-355-3697, Fax +1617-730-0253
Accepted 15 November, 2003

Introduction

Simultaneous measurements of genome-wide gene expression levels using DNA microarrays have become an essential tool in many areas of biology and medicine in the recent years (see, e.g., Collins(1999), and the review articles following). This high-throughput method enables researchers to obtain a global view of gene expression and has brought about a shift in the way they approach many biological problems. Microarrays have been used in many contexts, for example, to identify different types of cancer and the genes that characterize those types, e.g., Golub *et al.* (1999); Alizadeh *et al.* (2000), and to study the cell-cycle of the yeast and its response to various stimuli, e.g., DeRisi *et al.* (1997); Cho *et al.* (1998).

As more attention is directed toward detecting finer structure in the microarray data, a more careful analysis becomes crucial. For example, some genes display such dramatic changes in expression levels that any crude analysis or even a visual inspection will reveal their behavior. But to find genes with small fold changes accurately, more sophisticated analysis is necessary. One important part of such analysis is the first step of normalization (Hoffmann *et al.*, 2002). The expression levels usually have a high level of systematic variations and noise, caused, for instance, by inconsistencies in array fabrication, staining, and scanning. The problem occurs both on the individual gene level as well as on the array level. Ideally, the normalization step should address all sources of error to the extent possible, both within and between arrays. Currently, most normalization, if performed at all, is limited to a linear adjustment for each array to account for different gain in the scanning process.

There are several methods for such between-array normalization in the literature (Kepler *et al.*, 2002; Quackenbush, 2002; Kroll and Wölfl, 2002; Shmulevich and Zhang, 2002). Their primary aim is to standardize the "brightness" of chips to make expression levels across arrays comparable. When the hybridized chips are scanned to generate image files, different chips invariably result in images of different brightness. Depending on the data set, the mean expression level in each chip can vary widely. It is not unusual to have a factor of two or greater between some chips, as is the case with the data set we examine in this paper. Whether the chips should have the same brightness is not always clear. For some experiments, the expected number of highly expressed

genes differs considerably between arrays, and the brightness therefore should not be the same. However, since variability coming from the hybridization and scanning procedure seems to be larger than that from the experiments themselves and it is difficult to estimate the brightness a priori, some type of chip-to-chip normalization is usually applied.

The simplest method is to set a trimmed mean or the median of the distribution of the expression levels for each chip to be the same as that of a reference chip. A typical choice for the reference chip is the one with the median brightness. Similar linear rescaling methods include computing a scaling factor based on a linear regression fit between each sample and a reference chip or a maximum likehood estimation (Golub *et al.*, 1999; Hartemink *et al.*, 2001). For certain purposes, setting the mean 0 and variance to 1 can be a useful approach as well.

While linear rescaling is necessary, it is not sufficient. In particular, on each chip, the relationship between the true expression level and the signal intensity appears to be nonlinear outside a certain dynamic range. This problem has been observed and discussed by many researchers, but most in the context of cDNAs (Workman *et al.*, 2002; Yang *et al.*, 2002). Ramdas *et al.* (2001), for example, examined this phenomenon experimentally, and concluded that the sources of error are signal quenching associated with excessive dye concentration for flourescently labeled cDNAs on glass and nonlinear transformation by the scanner for radioactively labeled cDNAs on nylon membranes. Through a series of 'spike-in' hybridization experiments involving prokaryotic transcripts in the absence and presence of eukaryotic background, Chudin *et al.* (2001) verified that a linear relationship between the transcript abundance and signal is limited to a small range in the number of transcripts (between 1 pM and 10 pM). They argue that the linear range initially reported in lockhart may have been inflated because it was obtained using a custom array containing probe sets with more than 500 PM/MM probe pairs per gene (commercial arrays contain 16 or 20 probe pairs) and under specialized conditions different from the routine operating settings.

In this paper, we propose an algorithm for dealing with the nonlinearity. As shown in the next section, this nonlinearity can be severe at times and misleading expression levels can result if it is left uncorrected.

A more complete normalization step would involve reducing other sources of errors; we address only the nonlinearity aspect of the normalization process in this work. We believe that a overall linear scaling and the correction for nonlinearity when it is observed should be the minimal components of normalization. Our focus in this paper is on the Affymetrix GeneChip arrays, which are the most commonly used oligonucleotide arrays at present. The issues that arise for these arrays are different and often more complex than those for cDNA arrays. After an illustration of the nonlinear effects, we describe a new method in which within-chip ranks of genes across all chips are used as a basis to identify non-differentially expressed genes. Once we find these genes, we can use them to standardize the expression levels across chips.

Methods

The data set we examine is from Golub *et al.* (1999). This contains 38 arrays from leukemia patients with either ALL (acute lymphoblastic leukemia) and AML (acute myeloblastic leukemia) and possibly belonging to different subtypes within each group. The data contain 7,129 genes. We choose this set because it has been analyzed extensively as a benchmark data set, especially in the context of class prediction problems.

The basic limitations of the commonly used approaches are that they are linear transformations on the data. The more difficult part of a proper normalization is accounting for the nonlinearity between true expression values and the signal intensity. An example of this nonlinearity can be seen in Fig. 1. In this scatterplot of all genes between two arrays (4 and 20) from the leukemia data, at least one chip exhibits nonlinear behavior.

In this case, both samples belong to the same category of ALL and even the same subcategory of the B-cell type. We expect that most genes in these arrays are non-differentially expressed, falling on the line of slope one. However, we observe that the genes do not follow a linear relationship at all. It appears that a majority of the genes fall below the regression line even though it seems unlikely that their expression levels are all lower in sample 20. Among the few data sets we have studied, this nonlinearity appears frequently, sometimes in a serious manner as shown in Fig. 1. This problem is mitigated when logarithm is taken, but the effect is still clearly noticeable and should be addressed. Why it happens in some chips and not in others is not clear. Experimentally, one possible remedy is to use external controls with spiked cDNAs. Hill *et al.* (2001), for example, presents a normalization scheme based on a common pool of biotin-labeled transcripts of known concentration spiked into every hybridization. This nonlinearity is observed for cDNA arrays as well. Yang *et al.* (2001) propose one solution, fitting a local regression line through the log-ratios of the two chips. The underlying assumption for this normalization, however, is a rather strong one, that there is roughly an equal number of up-regulated and down-regulated genes at all intensity levels.

Beside the expression levels ('average difference'),

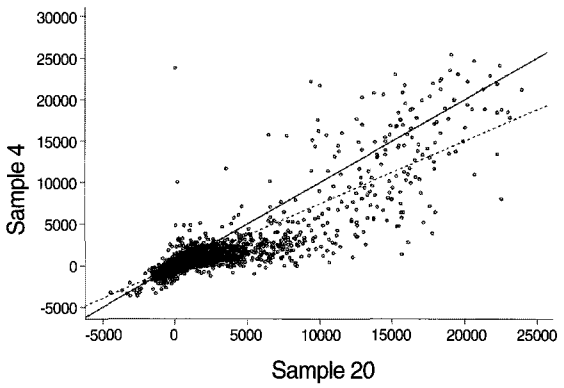

Fig. 1. An example of nonlinearity: samples 20 and 4 from the leukemia data. Solid line is the line of slope 1; dotted line is the linear regression line.

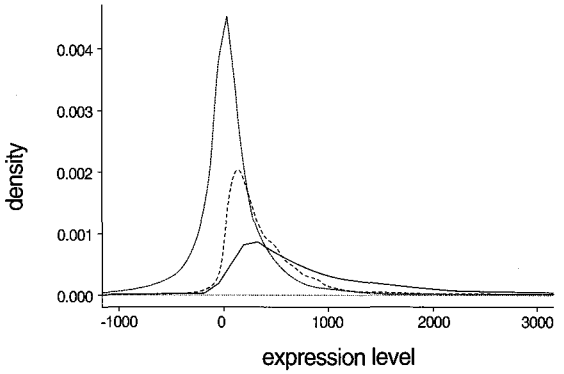

Fig. 2.: Density of the P, M, and A calls. Present (P, solid line), Marginal (M, dashed line) and Absent (A, dotted line) calls by the Affymetrix software make up 29%, 1.6%, and 69.4% of the data, respectively. The symmetry of the A call density around 0 indicates that the genes with A calls are likely to be noise.

Affymetrix data also provide another piece of useful information. Every expression level is annotated with P (Present), M (marginal), or A (absent) calls.

These confidence measures are derived based on the same PM (perfect match) and MM (mismatch) probe sets, usually 11-20 pairs, that are used to generate the expression level for each gene. To understand what these calls measure, we compute some basic statistics. We find that only 29.0% of the data received P calls; M calls and A calls take up 1.6% and 69.4%, respectively. To see how these calls are related to the actual expression values, we plot the density estimates of the expression values for each type of call in Fig. 2. (We computed the density estimates using a fast Fourier transform to convolve an empirical distribution with a Gaussian kernel.) We find that the values with the A calls are roughly symmetric around 0; in contrast, the M and P calls have distributions that cover substantially higher range.

This seems to indicate that those with A calls are likely to be noise. It then seems reasonable to filter out at least the 2,197 genes that have A calls in all 38 patients. In our computations, we eliminated those genes with A calls in 34 or more patients, based on the histogram showing the number of genes for each given number of A calls, leaving 3,609 out of 7,129 genes. 88% of the negative values are eliminated from the data and almost all of the genes eliminated have small magnitude below the nonlinearity region in which we are interested.

Algorithm

Our proposed solution is a rank-based method. First, we carry out a preliminary linear adjustment for overall brightness. Since we have often observed outliers of extremely high or extremely low (negative) values that distort the overall mean, we recommend the trimmed mean or the median for robustness. For trimmed mean, we have found that trimming 2.5% from each end of the data is sufficient. We then seek to identify some "non-differentially" expressed genes that will be used to correct for nonlinearity. Since any monotonic transformation will preserve the within-chip ranks of genes, we base our method on the ranks, and we look for genes whose ranks do not fluctuate much relative to other genes of similar signal intensity.

We use the following statistic for each gene k:

$$R_k = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} (x_{ki} - x_{kj})^2,$$

where x_{ki} is the rank of the expression value of the gene k in chip i. This is the squared summation of the changes in ranks for every pair of chips. The assumption is that if a gene is non-differentially expressed across the samples, their change in ranks should be relatively small. As we discuss further in the next section, we choose the genes whose statistic R_k is small compared to other genes of similar intensity.

The method we propose is an improvement over some methods introduced previously. In Tseng *et al.* (2001) and Li and Wong (2001), the within-chip rank for each gene was compared between a reference chip and the chip being normalized. If the difference between the ranks of a gene in the two chips are below some threshold, then it is assumed that those genes are non-differentially expressed. A locally weighted regression with smoothing is then performed using only these genes. One difficulty in the method, however, is that the threshold for ranks does not

depend on the intensity values. When the gene is expressed at a low level, a small perturbation can cause a large change in the ranks due to the presence of many similarly low-expressed genes. Iterating the same procedure on the set of non-differentially expressed genes, as suggested by Tseng *et al.* (2001), is one way of obtaining a more likely set of non-differentially expressed genes, but our method deals with this in a more direct way.

The more serious issue is that the nonlinear effect we observe is usually at the high intensity values, where we often have fewer data points. Due to this sparsity of points, it is difficult to determine which genes are likely to be non-differentially expressed by the changes in ranks in that range, and using a constant threshold is unreliable. Because of this problem, Tseng *et al.* (2001) suggest disregarding a certain number of points at the top or bottom, as specified by a parameter, and then using a linear regression based on the points in the middle to extend the normalization curve to this range. This solution, however, can give highly variable results depending on the parameter, and it is not clear how to choose the parameter.

The improvement we suggest in this paper is a more robust identification of the 'non-differentially expressed' genes. The fundamental difficulty in identifying non-differentially expressed genes is that often there is not enough information if the ranks on only two chips are used for each normalization.

Therefore, any simple method must resort to making a strong assumption, such as that the distribution of up-regulated and down-regulated genes is the same at all intensity ranges such as in Yang *et al.* (2001). We remedy this problem by using the ranks of each gene in *all* chips. Our method takes those genes whose sum of changes in ranks is small compared to others of similar median expression levels. In that process, we are still making the assumption that there are non-differentially expressed genes throughout the whole range of expression levels, but this seems to be the minimal assumption possible in the absence of some external measure such as spiked-in controls. Depending on the size or other characteristics of the data set, the statistic R_k can be modified. The current form, for instance, imposes a heavy penalty for deviance from the average rank; if the number of chips is very large and one does not wish to disqualify a gene based on a single chip, the statistic can be modified accordingly.

We note that in the presence of external information, such as class labels in the case of the leukemia data, there may be other ways of identifying differentially or non-differentially expressed genes. But our method does not require such information, which may not describe a correct phenotype classification even when it is present. In the Discussion section, we compare the result obtained using the class labels to that given by our method as a way of verifying our method.

Results

In Fig. 3, we plot R_k for all genes k=1,...,n as a function of each gene' s median intensity among all chips. We see that there is a sharp peak at a low intensity value close to zero, indicating that the ranks may change by a large factor at that range even when the changes in the expression levels are small. It is clear that the threshold for R_k should depend on the intensity to account for the density of

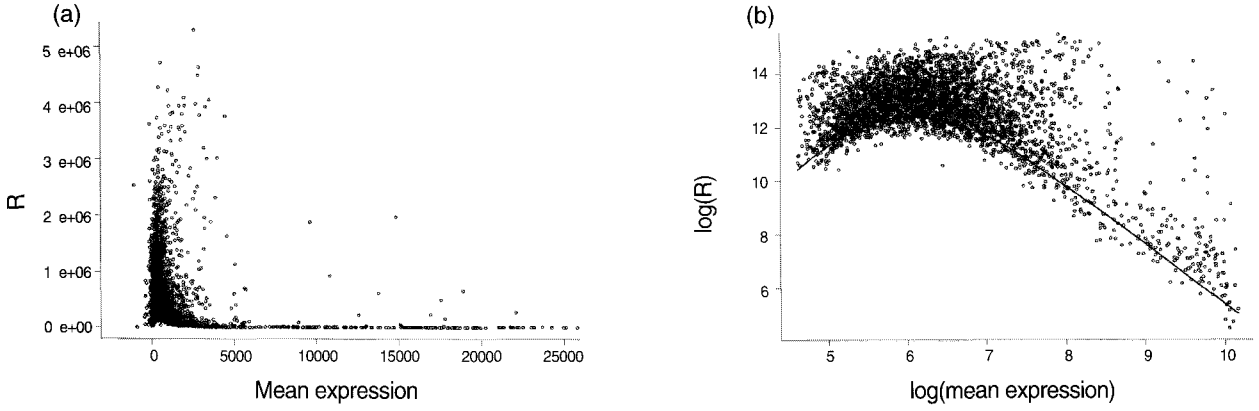

Fig. 3 (a) A plot of the mean expression level of each gene (across all samples) vs. the rank statistic, R_k. There is a sharp peak near zero because the high density of data points in that neighborhood results in large changes in ranks. To account for this effect, we need to find those genes that have high R_k relative to others at similar mean expression values. (b) We plot the same data in the log-log scale in order to fit a smooth curve. The cut-off line is a smoother (LOESS) computed on the bottom 5% running quantile. The points below this line are assumed to be non-differentially expressed.

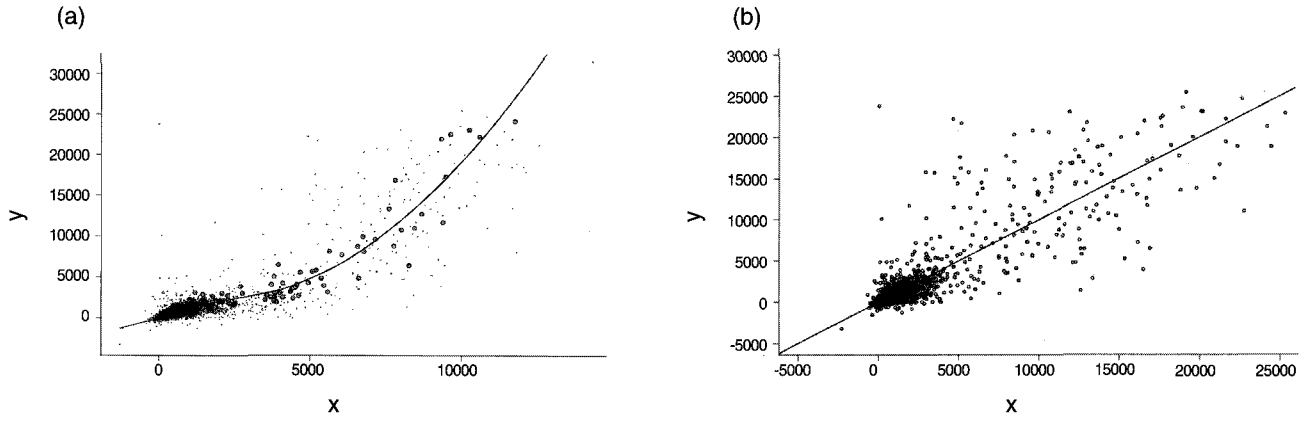

Fig. 4. (a) Before Normalization. Normalization curve for the two chips from Fig 1 (after linear normalization was applied to set the trimmed means the same for all chips). The circled points have been identified as the non-differentially expressed genes using the rank-based method using all 38 chips. A robust local linear regression is used to fit the curve through these points. (b) After normalization. The fitted curve is now the line with slope 1 and data points have been adjusted accordingly.

neighboring points.

In order to find those likely to be non-differentially expressed, we compute a running quantile of a given value across the points in Fig. 3. Then we compute a robust local regression to draw a smooth curve through the quantile points. As the quantiles at the ends of the distribution are biased, we extrapolate the curve in those cases. Because of the extremely sharp peak in the distribution, we actually convert both the x and y axes to a log scale; for negative numbers, we take the logarithm of their negatives and fit lines in such a way that the discontinuity near 0 can be ignored (computing the correct curve is not important for negative values.) In our example, we fit a smooth regression curve through the lower 5 or 10% quantiles at each mean intensity (in the log scale) and assume that the points below the curve are non-differentially expressed.

After identifying the genes, we can use them to compute the normalization curve. In Fig. 4, we show the chips shown in Fig. 1 before and after normalization. We identified 467 out of 7129 genes (6.5%) as most likely to be non-differentially expressed. These are circled on the left in Fig. 4. The solid curve is then fit through these genes using a locally weighted regression (Cleveland, 1979). That regression curve is then assumed to be the straight line that would have resulted if the experimental procedures had not introduced nonlinear effects. On the right, we transform that curve to be the straight line with slope 1 and adjust the other points accordingly. After the transformation, the picture is nearly symmetric.

Once we have obtained the subset of the non-differentially expressed genes, we can normalize all the chips against the same reference chip using the same

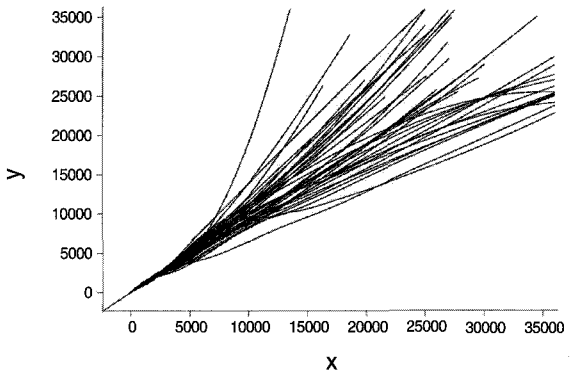

Fig. 5. Normalization curves computed for each of the chips vs. the reference chip. Most curves fall near $y=x$ but few have a noticeable nonlinear effect.

subset. In picking the reference chip, we should be careful not to pick the one with a severe nonlinearity problem. If such a chip were chosen as the reference sample, then all the curves would appear to have the nonlinearity problem even when that is not the case. A simple solution to this problem is to compute the linear correlation coefficient for all pairwise combinations and pick the chip with the largest sum of correlation coefficients with the rest of the samples. This procedure will identify a chip with little or no nonlinearity.

In Fig. 5, we plot the normalization curves of all chips (y-axis) against the reference chip (Chip 4). We see that most lines fall near the $y=x$ line, but a few appear to be problematic. Since the nonlinear effect we discuss here seems to be present in many data sets, we suggest that one compute all normalization curves as in Fig. 5 for a

given data set to determine if it is a problem.

Discussion

Our emphasis in this work has been to identify genes that are non-differentially expressed so that we can make more consistent comparisons across microarrays. But, we can also use Fig. 3 to find *differentially* expressed genes. It is already clear from Fig. 3 by visual inspection that some genes are outliers. We can, for example, mark those genes whose R_k statistic is in the top 1% compared to those with similar mean expression values, and examine them for biological significance.

In general, it is difficult to determine whether we have correctly identified certain genes as differentially or non-differentially expressed. For the type of data sets with class labels (ALL vs. AML), however, one can compare the result with that obtained from another method, such as the t-test (parametric) or the Wilcoxon (nonparametric) test. Strictly speaking, this is not a correct comparison, since either test presumes that the data set has been properly normalized. However, comparing the results this way gives some sense of the similarities between such a test and our method. In Fig. 6, we plot the running mean of the R_k statistics against the t-statistics. We see that there is a strong correlation (-.879 and .794, respectively, for negative and positive values of t-statistics), although the variance of the individual genes around the running means are high. This is evidence that the rank statistic is capturing an important feature of the data. We also note that our subsequent use of the R_k statistic is in the context of its neighbors in intensity values, not in place of the t-statistic.

This method can therefore be viewed as an exploratory tool in a limited way, as it does not require any phenotype information such as the disease labels when identifying either differentially or non-differentially expressed genes. In many cases, phenotype information may be available but inaccurate, or not available at all. It may not be clear how the information should be used even when available. For example, if there are multiple characteristics such as gender and age, we can discover interesting genes without having to divide them into all possible categories. The labels may also be continuous, making it difficult to put them into discrete categories.

In this paper, we have applied the normalization scheme at the level of the expression measures that have been derived from the probe-level data. Accurate and robust methods for reducing the probe-level information to the expression measures is under development by many researchers. The method described here can be applied at the probe level, at the expense of increased computational cost. While many studies using microarrays have resulted in significant progress in our understanding of various biological processes, it often has been the result of the prominent features in the data. As we look for less apparent features and construct a finer picture of the underlying mechanisms, having a "clean" data set will be increasingly important. For example, various clustering algorithms have been applied in the analysis in order to identify samples or experiments with similar characteristics or co-expressed genes.

Hierarchical clustering in particular has been a popular method. However, it is well-known that many clustering techniques such as hierarchical clustering are sensitive to small perturbations in the data set.

The validity of the results are therefore contingent on having good quality data. We have proposed one possible solution to address the nonlinearity between the transcript abundance and the signal intensity in oligonucleotide arrays. We identify genes based on their rank changes on

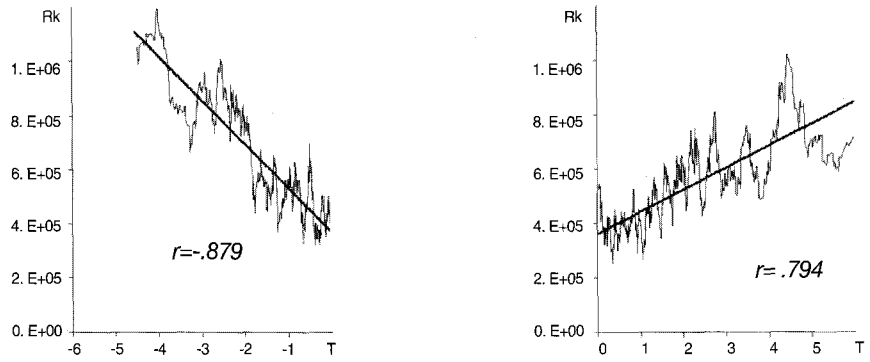

Fig. 6 t-statistic vs. R_k. When we know the division of the chips into two classes, we can see whether the t-statistics correspond with R_k. We plot the running mean of R_k as a function of t-statistic. Correlation coefficients are -.879 and .794 respectively for the negative and positive t-statistics.

all chips, rather than two at a time, and with the assumption that at least a very small proportion is non-differentially expressed at every intensity level. Until microarray technology can be improved, more such post-experiment corrections will need to be developed and applied.

Acknowledgments

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (03-PJ10-PG6-01GM01-0004).

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. G., Sabet, H., Tran, T., Yu, X. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503-511.

Cho, R. J., Campbell, M. J., Winzeler, E. A. *et al.* (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 2, 65-73.

Chudin, E., Walker, R., Kosaka, A., Wu, S. X., Rabert, D., Chang, T. K., and Kreder, D. E. (2001). The relationship between signal intensities and transcript concentration for affymetrix genechips. *Genome Biology* 3, research 0005.1-0005.10.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74, 829-836.

Collins, F. S. (1999). Microarrays and macroconsequences. *Nature Genetics* 21(Supp), 2.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-686.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. In *SPIE BiOS* 2001.

Hill, A. A., Brown, E. L., Whitley, M. Z., Tucker-Kellogg, G., Hunter, C. P., and Slonim, D. K. (2001). Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biology* 2, research 0055.1-research 0055.13.

Hoffmann, R., Seidl, T., and Dugas, M. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biology* 3, research 0033.1-0033.11.

Kepler, T. B., Crosby, L., and Morgan, K. T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology* 3, research 0037.1-0037.12.

Kroll, T. C. and Wolfl, S. (2002). Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. *Nucleic Acids Research* 30, e50-e55.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* 2, research 0032.1-0032.11.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo M. V., Chee, M. S., Mittmann, M., Want, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). DNA expression monitoring by hybridization of high density oligonucleotide arrays. *Nature Biotechnology* 14, 1675-1680.

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

Ramdas, L., Coombes, K. R., Baggerly, K., Abruzzo, L., Highsmith, W. E., Krogmann, T., Hamilton, S. R., and Zhang, W. (2001). Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology* 2, research 0047.1-0047.7.

Shmulevich, I. and Zhang, W. (2002). Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* 18, 555-565.

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

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielsen, H. B., Saxild, H. H., Nielsen, C., Brunak, S., and Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology* 3, research0048.1-0048.16.

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 Research* 30, e15.

Yang, Y. H., Dudoit, S., Luu, P., and Speed T. P. (2001). Normalization for cDNA microarray data. Technical Report 589, Statistics Dept, UC Berkeley.