BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 21 2005, pages 4000–4006
doi:10.1093/bioinformatics/bti661

Gene expression

A two-stage normalization method for partially degraded mRNA microarray data

Li-yu D. Liu[1], Naisyin Wang[1,3,*], Joanne R. Lupton[2,3], Nancy D. Turner[2,3], Robert S. Chapkin[2,3] and Laurie A. Davidson[2,3]

[1]Department of Agronomy, Biostatistic Division, National Taiwan University, Taipei, Taiwan, [2]Department of Nutrition and Food Science and [3]Center for Environmental and Rural Health, Texas A&M University, College Station, TX, USA

Received on June 3, 2005; revised on August 25, 2005; accepted on September 1, 2005
Advance Access publication September 8, 2005

ABSTRACT

Motivation: The goal of the study is to obtain genetic information from exfoliated colonocytes in the fecal stream rather than directly from mucosa cells within the colon. The latter is obtained through invasive procedures. The difficulties encountered by this procedure are that certain probe information may be compromised due to partially degraded mRNA. Proper normalization is essential to obtaining useful information from these fecal array data.

Results: We propose a new two-stage semiparametric normalization method motivated by the features observed in fecal microarray data. A location–scale transformation and a robust inclusion step were used to roughly align arrays within the same treatment. A non-parametric estimated non-linear transformation was then used to remove the potential intensity-based biases. We compared the performance of the new method in analyzing a fecal microarray dataset with those achieved by two existing normalization approaches: global median transformation and quantile normalization. The new method favorably compared with the global median and quantile normalization methods.

Availability: The R codes implementing the two-stage method may be obtained from the corresponding author.

Contact: nwang@stat.tamu.edu

Supplementary information: Additional figures including scatter plots, MA plots and density plots of array differences may be found at http://stat.tamu.edu/~daisy/NSBRI/

1 INTRODUCTION

Colon cancer is the second leading cause of death from cancer in the United States. Early detection of colon cancer can result in a high cure rate. Unfortunately, the anxiety caused by invasive procedures such as colonoscopy could lead to resistance toward screening processes. The necessity of developing new non-invasive screening methods cannot be over-emphasized. Approximately one-sixth to one-third of normal adult colonic epithelial cells are shed daily. Sloughed colon cells as well as their genetic material can be collected from feces. The data studied in this paper were obtained from an on-going project, which aims to recover colonic gene expression information from exfoliated colonocytes in the fecal stream. It is important to determine whether valid information can be obtained from such data for a relatively large number of probes. The ultimate goal is to develop a non-invasive mRNA procedure for colon cancer screening.

A challenging issue on both the biological and bioinformatics fronts is that the RNA could be partially degraded. On the biological side, Schoor *et al*. (2003) suggested that partially degraded RNA samples can still lead to meaningful conclusions if they are handled appropriately. Davidson *et al*. (1995, 2003), and Kanaoka *et al*. (2004) have demonstrated that intact fecal eukaryotic mRNA can be successfully isolated. On the bioinformatics side, the need for sensible data processing/normalization techniques that accommodate the existence of partially degraded genetic material is essential. This paper investigates this issue.

Gene expression levels were collected using the CodeLink System from GE Healthcare. The foundation of a CodeLink array is a proprietary 3D aqueous gel matrix slide surface with 30-base oligonucleotide probes. Every array is inspected for spot morphology. If a spot does not meet quality control standards, scanning software automatically categorizes it as a specific sub-type corresponding to the problems encountered. Readings from the problematic spots are commonly eliminated from the final data analysis. Besides spots which are good (G) and spots with problems such as background contamination (C) or irregular shape (I), there are also spots marked as 'L' which are spots with readings 'near background'. The low measurements could reflect true low gene expression levels or they could be caused by degradation of the mRNA resulting in a low signal. When partial degradation is not an issue and the samples are collected from colonic mucosa cells (Davidson *et al*., 2004), there is a low proportion (5–8%) of L spots, so no special measures are needed to accommodate them.

Due to mRNA degradation in fecal samples, the proportion of L spots in fecal microarrays is markedly higher. The number of L probes varies from array to array. The chance of an L probe belonging to the class of degraded genes is also much higher. We note that direct applications of existing normalization methods to the fecal array data could lead to loss of information or potentially biased outcomes. Further explanation, including summary statistics about the data and illustrations of our findings using the existing methods will be provided in later sections. Since one cannot decide whether a low spot is caused by mRNA degradation, a reasonable strategy is to regularize the G probes and recover as much usable 'L' probe information as possible. As in all other microarray experiments, proper normalization of the data is essential to ensure that bias is eliminated in the final outcomes.

[*]To whom correspondence should be addressed.

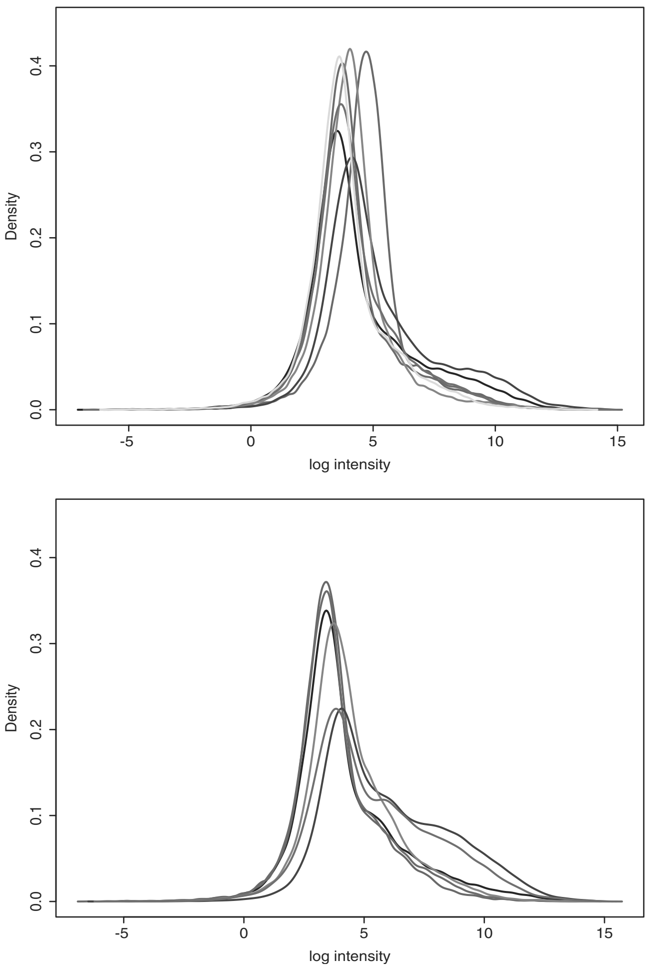

Fig. 1. Densities of raw log-2 intensity levels of G and L probes for each Diet 1 array (upper) and Diet 2 array (lower).

In this paper, we demonstrate the utility of a two-step normalization procedure applied to a CodeLink fecal microarray dataset. We compare the new method with two existing normalization methods. The assessment of all three methods is based on empirical outcomes that demonstrate their ability in removing bias without compromising underlying information.

2 DATA DESCRIPTION

The fecal bioarray data used to illustrate the procedures here were collected from rat fecal samples. These rats were randomly assigned to two diet groups: fish oil/pectin (D1) and corn oil/cellulose (D2) diets. Fecal samples were collected 14 weeks after the rats were exposed to carcinogen AOM and radiation. The goal of the experiment is to understand the effects of diet on genes being differentially expressed post carcinogen/radiation exposure (IRT). There are 7 and 6 bioarrays collected under IRT-D1 and IRT-D2, respectively. Figure 1 displays the densities of raw log-2 intensity levels of G and L probes for each array. The summaries of (1) total number of 'G' probes per array, (2) the proportion of morphologically disqualified spots (not included in Fig. 1) and (3) the ratios of the numbers of 'L' spots per array to the median of these numbers for all arrays are given in Table 1 (columns a–c), respectively. It is easy to see that there are a few bioarrays, e.g. arrays # 2 and 5 in D1, which tend to have higher proportions of 'L' spots compared with others. These arrays should tend to have more low measurements. Consequently, the densities of these arrays should tend to shift toward the left. Nonetheless, the phenomenon of certain arrays having much higher proportions of 'L' spots is not obvious when simply evaluating the density plots in Figure 1. Further, an array with a large number of L probes could have its distribution shifted to the right rather than to the left. We suspect that this is a consequence of an automatic calibration adjustment during the scanning process in order to better distinguish the large amount of low intensity entries. That is, what we have observed is most likely due to the automatic adjustment during the scanning process which is designed to regularize output. However, we were not able to obtain any information on how the scanning process is calibrated. Under the regular array scenario, it is expected that the distributions of measurements from different arrays should not differ much from each other. This is most likely the scenario under which the calibration process is designed. For the fecal array data, this automatic adjustment process is very likely to be one factor that actually introduces biases.

Table 1. The summaries of (a) total number of 'G' probes per array, (b) the proportion of morphologically disqualified spots and (c) ratios of the numbers of 'L' spots per array to the median of these numbers

Array ID	(a)		(b)		(c)	
	D1	D2	D1	D2	D1	D2
1	9635	9996	0.0125	0.0142	0.9816	0.9643
2	3940	7102	0.0624	0.0372	1.1468	1.0515
3	7164	8376	0.0148	0.0181	1.0802	1.0256
4	12453	17821	0.0232	0.0136	0.8503	0.6423
5	4967	8563	0.0220	0.0414	1.1609	0.9854
6	7965	14538	0.0486	0.0281	1.0000	0.7575
7	7070	—	0.0141	—	1.0852	—

2.1 Regular normalization procedures and some graphical display

It is well recognized now that to compare gene expression levels from two or more arrays, it is necessary to normalize the readings. Several different approaches have been used for normalization. Briefly, we summarize a non-exhaustive list here. The most commonly used normalization methods are the 'global' and 'local' normalization procedures. The former includes methods such as global median normalization (Zien *et al*., 2001; Quackenbush, 2002) which aim at standardizing the location and/or scale parameters of measurements from different arrays. The latter includes various non-parametric smoothing approaches (Schadt *et al*., 2001; Li and Wong, 2001) which aim at removing the intensity-based bias. Also see Quackenbush (2002) for an overall review of these two types of normalization methods. Besides these, the ANOVA model-based normalization methods of Kerr *et al*. (2000) and Kerr and Churchill (2001) account for multiple sources of variation via a statistical model rather than handling them during the pre-processing stage. Wolfinger *et al*. (2001) follow the same direction of consideration and go a step further by treating various

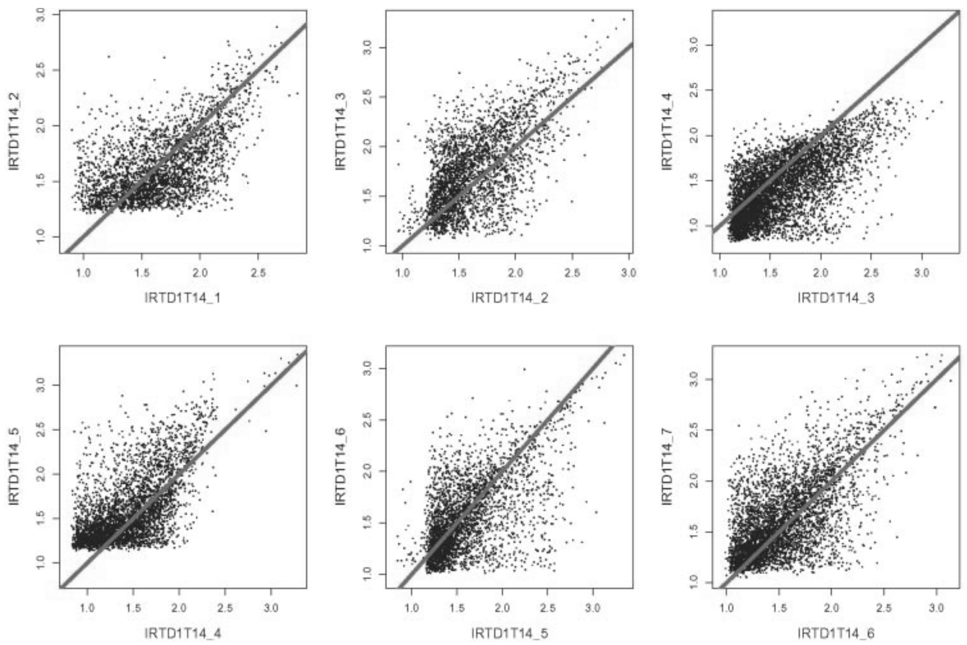

Fig. 2. Scatter plots between two consecutive Diet 1 arrays normalized by global median method.

coefficients as random. Several recent approaches such as Sidorov *et al*. (2002) and Bolstad *et al*. (2003) use parametric or non-parametric methods to transform the data so that the distribution of the transformed intensities is the same across a set of arrays. However, none of the above methods is readily usable for the partially degraded samples we have.

To illustrate the need to develop a new method for the fecal bioarray data and also to further display the data, we present some graphical outcomes obtained from two normalization procedures. They are (1) global median normalization and (2) quantile normalization (Bolstad *et al*., 2003). The global median method is perhaps the most commonly used method for CodeLink arrays. The quantile normalization method is among the most sophisticated normalization methods and is known to perform well when the measurements from different arrays share the same underlying distribution—an assumption which is most likely violated here simply due to different proportions of L probes from array to array. We provide a brief description of the two methods below and the readers are referred to Zien *et al*. (2001) and Bolstad *et al*. (2003) for details.

Global median normalization. This method is one of the standard procedures for normalization of single-channel arrays. The median log intensity of each array is subtracted from all of the log intensities in the same array. Let $\mathbf{m}_i$ be the median intensity of the *i*-th array. Then

$$\log \mathbf{x}_i' = \log \mathbf{x}_i - \log \mathbf{m}_i.$$

It simply shifts the distributions of arrays so that they are centered around zero.

Quantile normalization. This method is proposed by Bolstad *et al*. (2003). Yang and Throne (2003) adopted the same algorithm. The idea is to force arrays to have an identical distribution by letting

$$\mathbf{x}_i' = F^{-1}(G_i(\mathbf{x}_i)),$$

where G_i is the empirical distribution function for the *i*-th array and F is the empirical distribution function of the means of the quantiles over all arrays. The quantile normalization was calculated using software available in Bioconductor (http://www.bioconductor.org/).

In Figures 2 and 3, the scatter plots between 2 consecutive Diet 1 arrays normalized by global median and quantile methods are given, respectively. The 45° line was imposed. All morphologically qualified spots are included in the normalizing process. After global median normalization, non-linear trends are still observed in some of the scatter plots. The quantile method, on the other hand, removes the gene-to-gene correspondence between genes from two arrays within the same treatment group. Inferences on gene expression profiles become meaningless if no genetic information is left after normalization. Scatter plots and MA plots (Dudoit *et al*., 2002) with all possible pairwise combinations within each diet are provided in the Supplementary materials.

To check whether the normalization outcomes could be improved using G probes only, the scatter plots between two consecutive Diet 1 arrays using global median normalization is produced in Figure 4. Both Figures 2 and 4 indicate that there exist non-linear relationships between 2 arrays under the same treatment. The non-linear patterns in this scenario are often, though not always, similar to those of corresponding arrays when all morphologically disqualified spots were used. We also produced equivalent figures using quantile normalization with our own codes and the codes in bioconductor (results not shown). The former is similar to Figure 4 while the latter generates plots with very few points due to the fact that most probes have at least one observation not being 'G'. This simply suggests that a direct application of quantile normalization is not appropriate here.

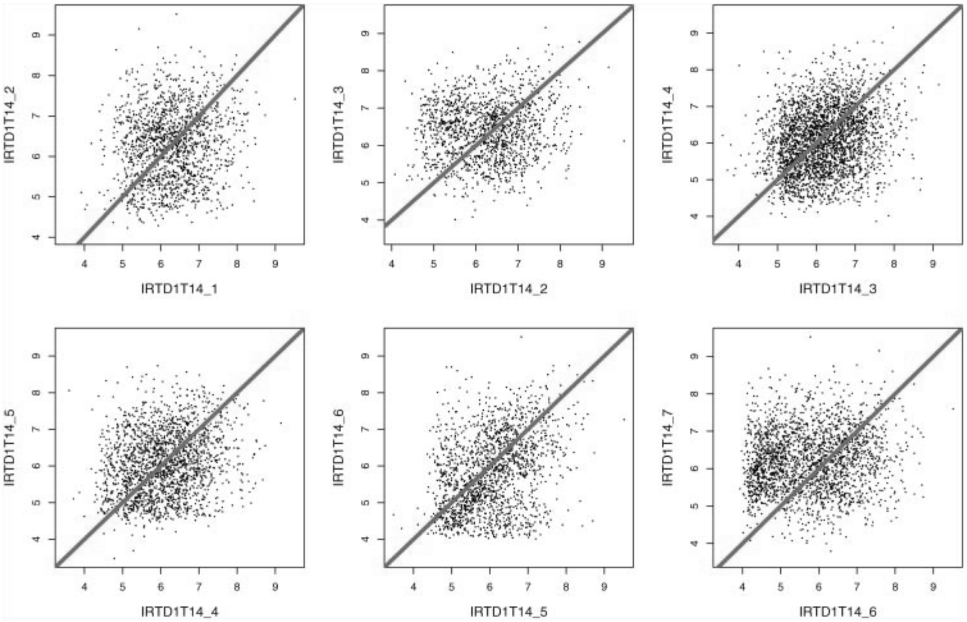

Fig. 3. Scatter plots between two consecutive Diet 1 arrays normalized by quantile method.

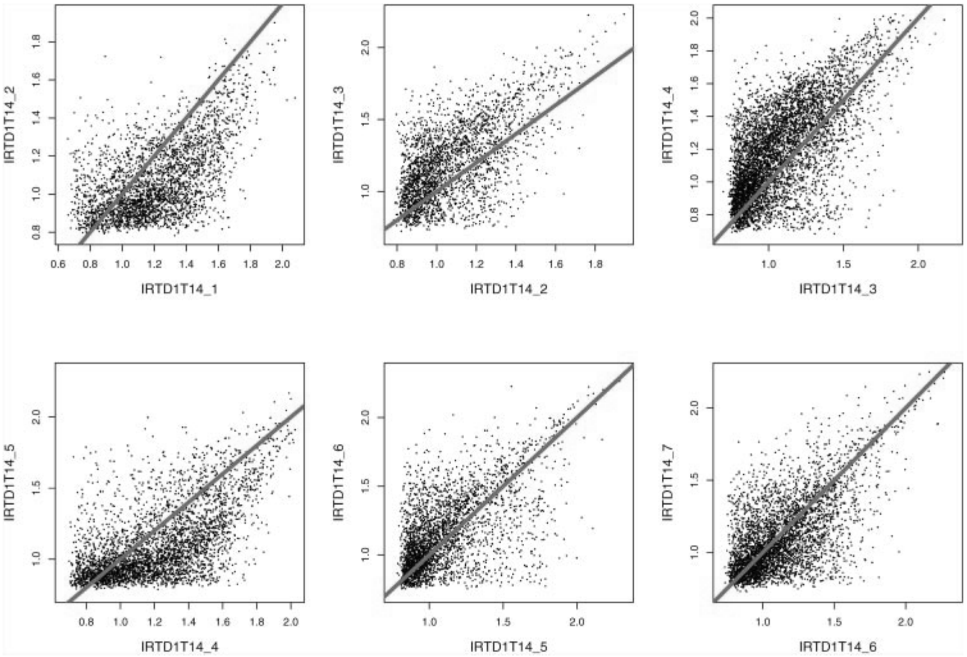

Fig. 4. Scatter plots between two consecutive Diet 1 arrays normalized by the global median normalization using G probes only.

3 TWO-STAGE METHOD

In this article, we propose a two-step normalization procedure motivated by features observed in CodeLink fecal microarray datasets. Before describing the steps within the normalization procedure, we have summarized the basic principles behind the proposed procedure.

- We assume that even with different proportions of L probes for different arrays, there exists a central 'peak' within the 'G' probes in each array such that the location and variation embedded in these central peaks enable us to perform a location–scale transformation to roughly align at least the central part of the majority of the arrays within the same treatment.

- Besides ‘aligning’ all arrays through location–scale transformations, we used local smoothing techniques to correct for intensity-based biases within each treatment group. This correction procedure mimics the idea behind non-linear transformation and corrects upward/downward biases of each array by comparison to the trend of its peer(s). Discussions on this type of methods for cDNA microarray data are given in Yang and Dudoit (2002). For CodeLink arrays, there are no G/R channels within the same array, as in two-channel cDNA arrays. Here, the peers are the arrays within the same treatment group. A simple robust inclusion step is introduced so that the majority of the ‘L’ probes were excluded during this process of identifying non-linear transformation to align signals in arrays within the same treatment group.

Suppose there are $i = 1, \ldots, I$ treatment groups, each with n_i arrays and the j-th arrays containing m_{ij} genes. In what follows, we describe the normalization procedure.

Step I. We apply a linear transformation to the ‘central’ $q\%$ of log-2 transformed gene expressions in each array to ‘best’ match the central $q\%$ of standard normal distribution. Here, the central location is defined to be the model of the central peak among ‘G’ probes. The procedure is carried out as follows:

(1) collecting the closest M items from both sides of the mode in the array to form a new set of intensities called $\mathbf{x}_c$; $2M = \min_{ij}(m_{ij} \times q/100)$;

(2) for the ij-th array, subtracting these central observations $\mathbf{x}_c$ by its mode b_{0ij} and then using the least square rule to find out the best scaling constant b_{1ij} that minimizes the sum of squared distances between the shifted percentiles and the corresponding central standard normal percentiles; and

(3) obtaining the new intensities by computing $x' = b_{1ij} \times (x - b_{0ij})$; here x is the log transformed gene expression value.

In practice, we let $q = 25$, that is, we use the central 25% of the data. We have let q vary up to the central 50% but observe no differences when performing gene by gene analysis at the latter stage. It is worth mentioning that we identified b_{0ij} and b_{1ij} using peaks among ‘G’ probes. Equivalent transformations can be roughly achieved using both ‘G’ and ‘L’ probes. This is not surprising. Certain systematic differences among measurements from one bioarray to another were formed most likely during the scanning and image processing stage—with the whole bioarray image being processed at the same time, regardless if a spot is ‘G’ or ‘L’.

(4) with our main objective being to regularize ‘G’ probes, we performed a very simple gene-by-gene step to include ‘L’ probes that behave similarly to other ‘G’ probes in the same gene, and excluded outlying ‘G’ probes. Basically, we did the following ‘robust inclusion step’:

- For a gene with less than three ‘G’ probes, we only included the ‘G’ probes.
- When there are three or more ‘G’ probes for a gene, we roughly estimate the central location of gene expression levels in this gene by the median of the ‘G’ probes and the standard deviation by the range divided by 4, following the empirical rule which states that central location ±2 SD covers the central 95% of the data. Any ‘L’ probes with values within the median ±3 SD are also included.

This simple step is motivated by the particular data structure we encounter here. It could certainly be applied in other situations for robustness consideration. Extreme outlying observations can be tentatively eliminated by the normalization procedure. The method is simple and consequently can be quickly calculated. Summary statistics from this step actually provide some additional interesting information which offers some insight into why a direct application of quantile transformation does not work well here. Specifically, the assumption that all ‘usable’ measurements from each array are from the same distribution could be violated, depending on which probes were used in the normalization steps. When the number of replicates within each treatment becomes large, further robustness considerations can be implemented. This simple procedure works rather well for our example. We also tried to use median absolute deviation (MAD) to estimate SD. However, since the number of arrays within the same treatment was not very large, MAD tends to be so conservative that many reasonable points were not used in the estimation.

Step II. With the central area roughly matched across all arrays, the second step in our method involved using the local smoothing method to correct the intensity-based biases associated with measurements in arrays within the same treatment group. A baseline array, which we chose to be the median of ‘qualified’ measurements of each gene, is specified for each treatment group. The objective is to correct for the shared upward or downward non-linear trend by genes from the same array particularly in the non-central area. This is carried out by identifying their trend against the baseline array. More precisely,

(1) taking the transformed intensities from the previous step for arrays within the same treatment group i and finding the medians of each gene to form the baseline array, $\mathbf{m}_i$;

(2) creating a local polynomial fit to the mean (or median), denoted by $\hat{f}_{ij}$, for the ij-th array, with the Xs being the baseline array values and the Ys being the transformed intensities in the ij-th array.

(3) updating the intensities in the ij-th array by subtracting the fitted values, $\hat{f}_{ij}$, and adding $\mathbf{m}_i$.

Note that since the central $q\%$ values of each array would already be well matched, the correction of the second step tends to occur in the area outside the central range. For example, if one array whose up-regulated genes consistently have higher values than those taken from the same genes in different arrays under the same treatment group, this phenomenon will be identified by an upward trend in $\hat{f}$ against the baseline median array. This intensity-based bias can therefore be removed through subtracting $\hat{f}$. We have tried both local polynomial fits to conditional mean and conditional median, where the latter estimation is carried out using local polynomial quantile regression (‘quantreg’ in R) by Koenker (2005).

The above transformation put the emphasis within treatment group alignment. When the mis-alignment among different treatment groups is a concern, a very simple modification using quantile transformation can be added. We can simply replace the $\mathbf{m}_i$ in Step 3 by $\mathbf{m}_i^*$ with $\mathbf{m}_i^*$ being the quantile transformed values of $\mathbf{m}_i$. For the majority of cases, if a gene has all ‘L’ spots in one treatment group, it tends to have all ‘L’ spots in all treatment groups. This is because bacteria may target the breakdown of certain nucleotide sequences. Consequently, for genes with multiple 'G' probes, we have $\mathbf{m}_i$ for almost all these genes in all groups. This simple step warrants that the probe medians from different treatment groups follow the same distribution. This is a much weaker assumption than requiring that the usable probe values in each bioarray follow the same distribution.

4 RESULTS

We applied our method to a CodeLink fecal microarray dataset. In the first step, we matched the 'central' 25% of the G probes to the equivalent central 25% standard normal distribution. Both the local polynomial mean and median estimates were carried out in the second step to adjust intensity-based biases among arrays within the same treatment. No noticeable differences between the two methods were observed for any arrays. Figure 5 shows one example of the correction by local smoothing methods. Transformed intensities after the first step are plotted against the 'baseline intensities', which are medians of qualified measurements of each gene. Even though the local polynomial estimated means/medians were calculated using genes after the robust-inclusion step, genes before the implementation of the robust-inclusion step were plotted in Figure 5. To reduce the size of the graph, only the highest 40% of L probes values were kept in the plot. No L probes in the lower 60% were included into the second step after the implementation of robust-inclusion step. Probes with low intensity values compared with the corresponding baseline intensities were visible in the lower half of the graph. The solid red line (local polynomial mean estimates) and dotted green line (local polynomial median estimates) were imposed. The final transformed observations were formed by subtracting the fitted values from the first stage transformed intensities and adding back the corresponding baseline intensities.

In Figure 6, gene-by-gene scatter plots between two consecutive Diet 1 arrays after two-stage normalization are given. The points are nicely scattered around the 45° lines. Unlike the results of global median transformation in Figure 2, the non-linear trends have been properly removed. Differing from observations after quantile transformations, Figure 6 shows that after two-stage transformation, if a gene is highly expressed in one array, it tends to have a larger chance to be highly expressed in another array within the same treatment group. Scatter plots and MA plots with all possible pairwise combinations are provided in the supplementary website. They collectively support the same conclusion.

5 DISCUSSION

We have proposed a two-stage normalization procedure for the processing of exfoliated colonocyte microarray data. One main feature of the data is the existence of a large number of outlying observations due to the fact that part of the mRNA has been degraded. Our method is built on simple non-parametric smoothing techniques with robustness consideration; consequently it can be applied to array data and the calculation can be completed in a reasonable amount of time. For the smoothing method, we have tried both local polynomial estimated mean and local polynomial estimated medians. They perform about the same since some precaution has been taken to exclude outlying observations before smoothing. The former can be computed faster than the latter. When in doubt, one can always compute both and compare the outcomes to ensure that the influence of outlying observations is properly controlled.

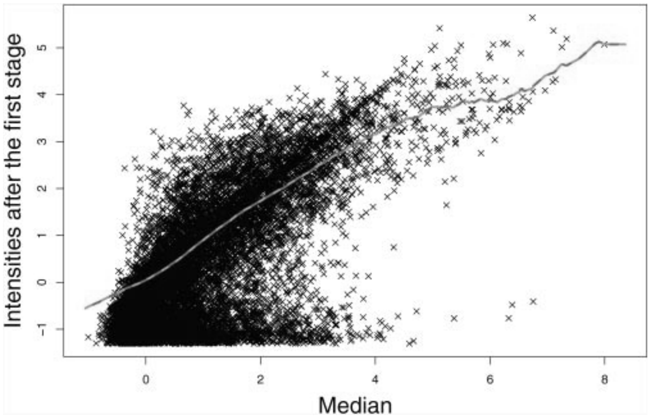

Fig. 5. Scatter plot of the intensities of the first array receiving Diet 1 versus corresponding medians after the first stage of normalization. The red line indicates the local polynomial mean estimates and the green line, the local polynomial median estimates at the second stage.

Currently, our emphasis on biological investigation focuses on evaluating the feasibility of using the newly developed biological and bioinformatics methods to properly extract information from the fecal microarray data. For this purpose, we compare the outcomes of testing the diet treatment effect using the responses transformed by the quantile-normalization method and by the new method, respectively. For both responses, we performed gene-by-gene mixed effect analyses, accounting for the dependence of replicative arrays produced with biological samples from the same subject. We then used the resulting diet effect p-values to calculate the false discovery rates, namely the q-values of Storey (2002). For the quantile-normalized responses, we found no gene that had a q-value <0.1 in the analysis. However, when we analyzed the responses normalized by the new method, there were 351 genes with q-values <0.1.

When we compared the outcomes with those from a previously conducted mucosal microarray experiment which shared a similar experimental setup as our current experiment, we also found certain evidence that comparable information can be obtained from the two experiments. For example, the diet treatment effects for the probe 'gap junction membrane channel protein beta 1' are significant in both experiments and the estimated diet 1 to diet 2 fold-changes for the mucosal and fecal samples are 0.7 and 0.72, respectively. For the fecal samples, the transformed data were rescaled, through the matching quantiles, back to the original scale in each treatment group before the fold-changes were calculated. We also found that the treatment effects for 'arginase 2' to be significant in both experiments and the estimated diet 1 to diet 2 fold-changes for the mucosal and fecal samples are 2.0 and 2.5, respectively. Obviously, with the gene-array platforms, the exact experimental setups and the sample sizes in the two experiments being different, we should not over-interpret the outcomes of such comparisons. Nonetheless, these findings provide support to the belief that one can obtain useful information from the partially degraded fecal microarray data.

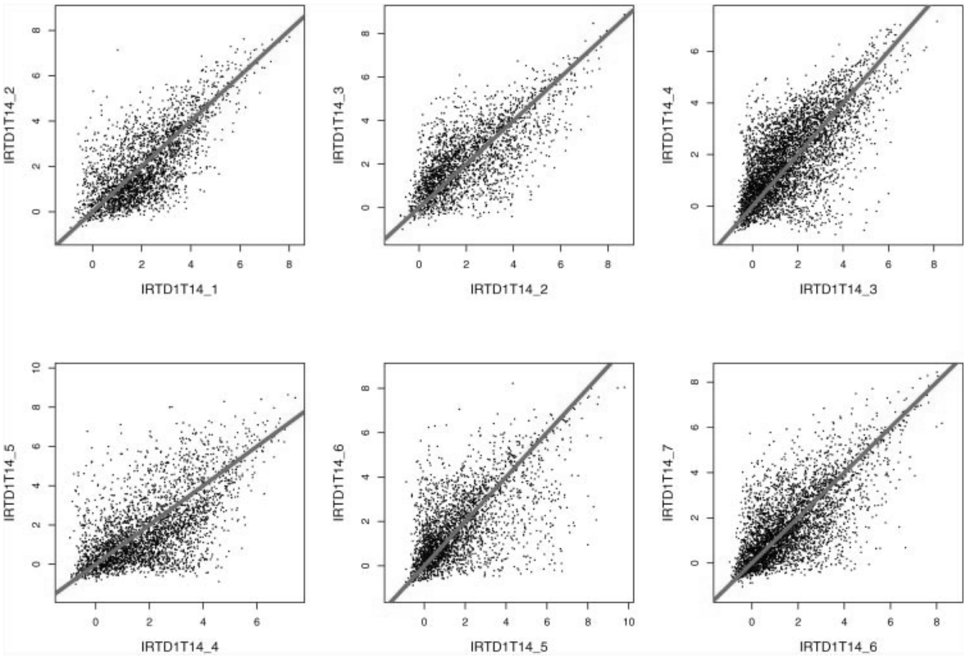

Fig. 6. Scatter plots between two consecutive Diet 1 arrays normalized by two-stage normalization.

ACKNOWLEDGEMENTS

This research is partially supported by NASA grant NAG 9-1523, NSBRI grant NASA NCC 9-58, NCI grants CA74552, CA59034, National Institute of Environmental Health Sciences grants P30-ES09106 and AICR 05A081.

Conflict of Interest: none declared.

REFERENCES

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Davidson,L.A. *et al.* (1995) Dietary fat and fiber alter rat colonic protein kinase C isozyme expression. *J. Nutr.*, **125**, 49–56.

Davidson,L.A. *et al.* (2003) Quantification of human intestinal gene expression profiles using exfoliated colonocytes: a pilot study. *Biomarkers*, **8**, 51–61.

Davidson,L.A. *et al.* (2004) Chemopreventive *n*-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. *Cancer Res.*, **64**, 6797–6804.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.

Kanaoka,S. *et al.* (2004) Potential usefulness of detecting cyclooxygenase 2 messenger RNA in feces for colorectal cancer screening. *Gastroenterology*, **127**, 422–427.

Kerr,M.K. and Churchill,G.A. (2001) Statistical design and the analysis of gene expression microarrays. *Genet. Res.*, **77**, 123–128.

Kerr,M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Koenker,R. (2005) *Quantile Regression.* Econometric Society Monograph Series, Cambridge University Press.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

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

Schadt,E.E. *et al.* (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem. Suppl.*, **37**, 120–125.

Schoor,O. *et al.* (2003) Moderate degration does not preclude microarray analysis of small amounts of RNA. *BioTechniques*, **35**, 1192–1201.

Sidorov,I.A. *et al.* (2002) Oligonucleotide microarray data distribution and normalization. *Inform. Sci.*, **146**, 65–71.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479–498.

Wolfinger,R.D. *et al.* (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–638.

Yang,Y.H. and Thorne,N.P. (2003) Normalization for two-color cDNA microarray data. *Science and Statistics: A Festschrift for Terry Speed, IMS Lecture Notes, Monograph Series*, **40**, 403–418.

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Zien,A. *et al.* (2001) Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, **17** (Suppl. 1), S323–S331.