World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:4, No:1, 2010

# A Simple *Affymetrix* Ratio-transformation Method Yields Comparable Expression Level Quantifications with cDNA Data

Chintanu K. Sarmah, Sandhya Samarasinghe, Don Kulasiri, and Daniel Catchpoole



***Abstract*—**Gene expression profiling is rapidly evolving into a powerful technique for investigating tumor malignancies. The researchers are overwhelmed with the microarray-based platforms and methods that confer them the freedom to conduct large-scale gene expression profiling measurements. Simultaneously, investigations into cross-platform integration methods have started gaining momentum due to their underlying potential to help comprehend a myriad of broad biological issues in tumor diagnosis, prognosis, and therapy. However, comparing results from different platforms remains to be a challenging task as various inherent technical differences exist between the microarray platforms. In this paper, we explain a simple ratio-transformation method, which can provide some common ground for cDNA and Affymetrix platform towards cross-platform integration. The method is based on the characteristic data attributes of Affymetrix- and cDNA- platform. In the work, we considered seven childhood leukemia patients and their gene expression levels in either platform. With a dataset of 822 differentially expressed genes from both these platforms, we carried out a specific *ratio-treatment* to Affymetrix data, which subsequently showed an improvement in the relationship with the cDNA data.

***Keywords*—**Gene expression profiling, microarray, cDNA, Affymetrix, childhood leukaemia.

## I. Introduction

In 1995, two seminal publications, [1] and [2], by the lead investigator, Patric O. Brown of the *Howard Hughes Medical Institute* and his collaborators launched the era of gene-expression microarray analysis, and revolutionized the field of molecular biology. The technique of microarrays, which started off with simultaneous gene expression analysis of 45 genes within one experiment, provides high throughput capability of simultaneously interrogating the RNA expression of the whole genomes. Microarray technology has gradually become an indispensable tool for monitoring genome wide expression levels of gene. From the Patric Brown's lab, the technology has evolved representing both a technological and conceptual advancement of the field, and has gone worldwide, where many laboratories are now making their own arrays, in addition to the availability of commercial vendors like *Affymetrix* (Santa Clara, CA), *Agilent* (Palo Alto, CA), Rosetta[1]. With the increasing number as well as the availability of gene expression studies of various organisms, there has been a pressing need to develop approaches for integrating results across multiple studies. There are different practical advantages in such studies.

In a cross-study analysis, the data, relevant results and statistics of several studies are combined. Cross-study analysis has the potential to strengthen and extend the results gathered from the individual studies. This can turn an investigation to have higher accuracy and consistency, and thereby, helping in robust information mining. Besides, output of such a study can provide a broader picture of gene-expression as the final 'integrated'-result emerges based on a set of individual studies. Cross study analysis can also compensate for the possible data-errors of the individual study. The cost of such a study is possible to keep low by using the existing studies, as otherwise the set up of each microarray investigation is not inexpensive. It can also amplify the sample-size.

Despite having various advantages, while attempting to actualize integration of microarray studies, there are much higher challenges and difficulties as genetic expressions of different studies are neither readily comparable nor can directly be combined. There are several published works on cross-study analysis, where the observation on accuracy, reliability and reproducibility of microarray platforms clearly ranges from relatively discouraging [3]-[4] through cautious optimism [5]-[6] to impressive [7]-[8].

In this paper, we explain a ratio-transformation method for Affymetrix data which may potentially lead in the direction of integration of Affymetrix and cDNA (also called, spotted microarrays) microarray studies. Here, we have examined closely the data structures of the two diverse platforms, cDNA and Affymetrix, towards combing their gene expression data based on the commonalities within the basic data-attributes. Our example involves determining differentially expressed genes, and a simple ratio-based approach while considering seven childhood leukemia patients from either platform.

Chintanu Kumar Sarmah is a PhD student at Lincoln University, Canterbury, New Zealand. (phone: + 64-3-3218377; fax: +64-3-3253845; e-mail: sarmahc3@ Lincoln.ac.nz).

Sandhya Samarasinghe is an Associate Professor from Centre for Advanced Computational Solutions (C-fACS) at Lincoln University (e-mail: sandhya@lincoln.ac.nz).

Don Kulasiri is a Professor of Computational Modelling and Simulation at Centre for Advanced Computational Solutions (C-fACS) at Lincoln University (e-mail: Don.Kulasiri@lincoln.ac.nz).

Daniel Catchpoole is the Head of Tumour Bank, The Children's Hospital at Westmead, Australia (e-mail: DanielC@chw.edu.au).

[1] http://www.rii.com/

## II. Method and Results

### A. Data Collection

Affymetrix *GeneChip*® and GenePix® cDNA data were obtained from Tumour Bank, The Children's Hospital at Westmead, Australia. The data belonged to childhood leukemia patients; and seven of these were analyzed both in Affymetrix (*HGU-133A chip*) and in cDNA platforms.

### B. Quality check of the raw data

Microarray experiments measuring genetic expression levels are conducted through an elaborated procedure, and are subject to many potential variations. This makes it critical to carry out adequate assessment to make sure that the data is of good quality, and is consistent and comparable for further analysis. Accordingly, we carried out an extensive quality assessment using open-source statistical software, R [9] and *Bioconductor* [10] towards confirming it.

### C. Data Normalization

The purpose of normalization is to minimize systematic variations in the measured gene expression levels.

Following several available literature including, [11] and [12], web-information[#] from *National Cancer Institute* (a component of the *U.S. National Institute of Health*) and a detailed review on comparison of normalization methods [13], *quantile normalization* is said to provide the best overall performance. For more about this method, readers may refer to [13]-[14]. Quantile normalization method is a robust, one of the most widely as well as routinely used methods in the analysis of microarray experiments. Therefore, we picked quantile normalization to use in our Affymetrix (*HGU-133A*) chips. The *Robust Multichip Average* (RMA) algorithm [15] for Affymetrix arrays use quantile normalization; and, as it is apparently the best method available at present[#], this method was used for processing our Affymetrix chips. RMA is largely the work of Terry Speed's group at University of California at Berkeley, and is an expression measure consisting of three particular preprocessing steps : convolution background correction, quantile normalization, and a summarization method based on a multi-array model fit robustly using the median polish algorithm [16].

In case of cDNA data, prior to normalization process, an adaptive background correction, *Normexp+offset* was used, as recommended by [17]. It is an usual assumption in background correction of cDNA arrays that given the observed foreground intensities, $R_f$ and $G_f$, background correction for two-colour microarray data allows the true signal to be estimated by subtracting the foreground and background values, such that $R = R_f - R_b$ and $G = G_f - G_b$, and the corrected intensities are then used to form the log-ratio, $M = \log_2 (R/G)$, and average log intensity, $A = ½ \log_2 (RG)$, for each spot. The *normexp+offset* method of background correction is based on the normal and exponential convolution model previously used to background correct Affymetrix data as a part of the RMA algorithm, as given in [15] and [18]. In this method, a convolution of normal and exponential distributions is fitted to the foreground intensities using the background intensities as a covariate, and the expected signal given the observed foreground becomes the corrected intensity. The corrected intensities, thus obtained, are positive, but may be close to zero. Therefore, a small positive offset is added to effectively move the corrected intensities away from zero. This should also reduce the variation of the low intensity M-values since $\log_2$ [(R+offset)/(G+offset)] will be close to 0 for R and G, both small relative to the offset. Based on the findings of [17], an offset value of 50 was used here for background correction.

As illustrated by [19], there is a range of normalization methods for spotted microarrays, and these methods may be broadly classified into *within-array* normalization and *between-array* normalization methods. The former group includes those methods that normalize the M-values for each array separately, while the latter normalizes the intensities or log-ratios to be comparable across arrays.

Several of within-array normalization methods were independently applied to the spotted arrays after background-correction to assess which normalization method would work relatively well. The method, *printtiploess* [20]-[21] was found to give better results while comparing to a few other commonly used methods, namely *median*, *loess*, *robustspline*. *Printtiploess* is also regarded as an effective method because of its ability to adjust for systematic differences between different print-tips [22]-[23] and it assumes that the ratios from each print-tip to have the same distributions.

The next question we addressed was whether between-array normalization would be required. We observed that the within-array normalized arrays were showing different spreads of M-values rather than having similar spread and found that from the list of between-array normalization methods including *scale*, *quantile* and vsn, the scale normalization method, proposed by [24] and [20] and further explained by [19], rendered the best result producing similar spread of the M-values across the cDNA arrays. The basic idea of this normalization is to simply scale the log-ratios to have the same median-abolute-deviation (MAD) across arrays.

Finally, we carried out a post-normalization quality assessment for both Affymetrix and cDNA data to make sure that the normalized data would be devoid of any systematic bias and other anomalies.

### D. Finding Differentially Expressed Genes

Besides the leukemic data, we also obtained Affymetrix data of 10-healthy children. We used these two sets of data, to find out DE genes.

For finding the DE genes, we first removed the control probes and then, filtered out the low quality data based on the fact that the Affymetrix probe sets scoring *absent* or *marginal* can be considered suspect, whereas *present* scores are good

[#] http://discover.nci.nih.gov/index.jsp

indicators of a signal reliably above background noise. To get rid of data with low content of information, we used the *relative standard deviation*, also known as the *coefficient of variability*, CV that would filter out the least variable genes, defined by the 90$^{th}$ percentile of the distribution of CV-values. Fig. 1 shows the chosen cut-off that picked the highest ranked 10% of CV-values.

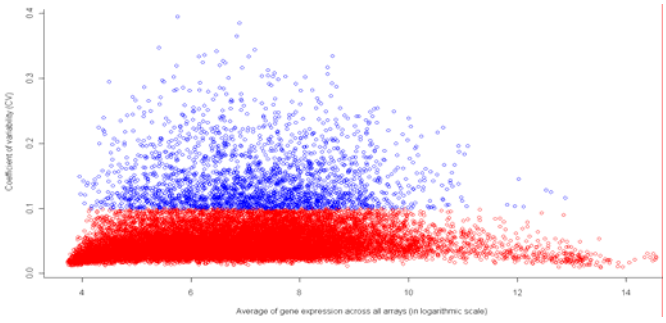

Fig. 1 CV as a function of average gene expression across all arrays (in logarithmic scale)

With a shortened list of genes, including only the most reliable and highly variable Affymetrix data, we specified the analysis design such that the samples belonging to the two experimental conditions (healthy and leukemic) were assigned to their respective group. A simple linear model was then fit to the data, as explained in [25]. To increase statistical power, and simultaneously reduce the risk of false positives, an empirical Bayes method [26] was used. This would improve on the accuracy of estimating variability for individual genes through shrinking of the standard deviation by including genes expressed at similar levels.

Finally, we adjusted the obtained p-values to account for the *multiple testing* (or, *multiple comparisons*) *problem*. As described by [27], multiple testing problems bring in error in inferences when one considers a set of statistical inferences simultaneously; and, such loss of statistical power in inference imposed by the multiple testing is common during simultaneous analysis of thousands of genes. Out of the methods on offer, including [28], [29], [30], [31] and [32], to prevent this from happening, we used the method of *Benjamini and Hochberg* [29] for adjusting p-values for multiple comparisons. The method controls the *false discovery rate*, the expected proportion of false discoveries (i.e., the false positives, or, type I errors) amongst the rejected hypotheses in multiple comparisons. The false discovery rate is a relaxed condition; and the method, [29] is a better compromise between sensitivity and specificity. We arbitrarily set the FDR control to a conservative value of 0.05.

Fig. 2 gives a histogram of the raw, unadjusted p-values and compares the distribution to that observed after adjustment to account for multiple testing correction. Information is also overlaid in that figure about how the distribution would be expected if there were no experiment effect (i.e., a uniform distribution), as well as a line indicating the cut-off for statistical significance, i.e., FDR control=0.05. Further, in Fig. 3, an MA-plot displays the log fold change between leukemic and normal samples as a function of the average expression level across all samples, where the two-fold limits are indicated by horizontal lines and statistically significant genes are the sharp dots outside the area covered by the horizontal lines.

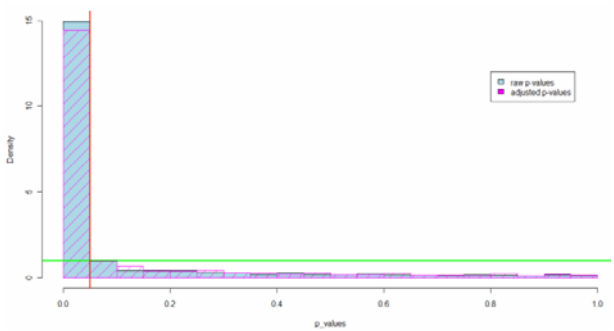

Fig. 2 p-value with theoretical uniform distribution (horizontal line), FDR cutoff arbitrarily set at 0.05 (vertical line)

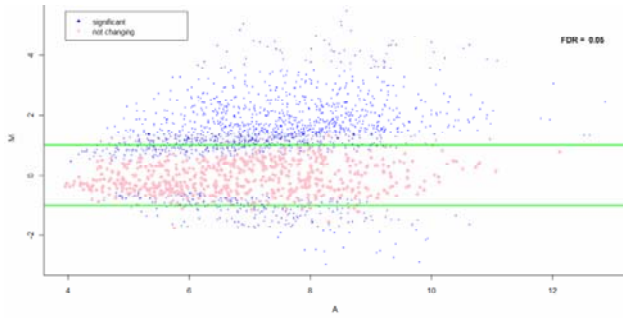

Fig. 3 MA plot comparing healthy and leukaemic samples. Significant genes are the sharp dots, and the horizontal lines are the 2-fold limits.

We used UniGene database [33] to annotate the genes. The overall statistical procedure on Affymetrix chips finally recognized a total of 822 genes as differentially expressed. Further, considering the fact that Affymetrix data contain relatively less noise than cDNA, we selected the same set of 822 genes from cDNA platform as well for our downstream analysis. We assumed here that as the arrays in both the platforms belonged to the same 7-childhood leukemia patients, therefore, the same set of genes would ideally be expressed differentially in either platform.

### *E. Ratio-transformation*

Our Affymetrix and cDNA data had no correlation between them as such. Using the DE genes, the correlation between the normalized data from either platform became 0.13, referring that there was absolutely no relation between them.

Fundamentally, Affymetrix and cDNA data have difference

in their data structure. The measurement used for relative expression level for a gene in cDNA platform is expressed in terms of *Expression Ratio*, which is denoted by:

$$T_k = \frac{R_k}{G_k} \quad (1)$$

where for each gene, $k$ on the array, $R_k$ and $G_k$ represents intensity metric for the tumor sample and the healthy sample, respectively. However, as a measure of expression of a gene in Affymetrix platform, the average difference between all the PM and MM probes of a gene is considered proportional to the actual expression level of the gene, as shown in Equation 2:

$$Average\ \ difference_{\ probe\ pair} \cong \frac{1}{n}\sum_i (PM_i - MM_i) \quad (2)$$

where, n represents the total number of probe pairs for the gene. $PM_i$ and $MM_i$ indicates the corresponding PM and MM probe intensities for the $i^{th}$ probe pair for the gene. It is, therefore, apparent that one of the basic differences between Affymetrix and cDNA lies in the nature of the retrieved data – while cDNA provides expression ratio, Affymetrix gives actual expression level of a gene. We acknowledged this difference, and considered the possibility that addressing this difference might lead towards giving an improved relationship between the two platforms.

Using the healthy Affymetrix arrays, and their average expression levels across all the chips for the 822 DE genes, we converted the Affymetrix data to *Affymetrix-ratio*. In this regard, if expression level of a gene, $x$ from one of the diseased Affymetrix chips is $D$, and each gene's
expression from the set of 10 healthy Affymetrix chips is $H$, then *Affymetrix-ratio* ($Affy_{ratio}$) can be denoted as in equation [3].

$$Affy_{ratio} = \frac{D_x}{\frac{\sum_{i=1}^{10} H_{i_x}}{10}} \quad (3)$$

Similar to cDNA where the expression level of a gene remains in the form of a tumor to healthy ratio, the transformation through equation [3] converts the Affymetrix expression data into the form of a tumor to healthy ratio.

Finally, it was found that this transformation improved the correlation between cDNA and $Affy_{ratio}$ for the 822 DE genes from 0.13 to 0.6. Fig. 4 shows the change in the Affymetrix data due to the ratio-transformation. Here, $Affy_{ratio}$ data can be seen to have roughly aligned at the same level as the cDNA data. It is possible to further align the $Affy_{ratio}$ data by converting them to their respective standard normal distributions, as shown in Fig. 5.

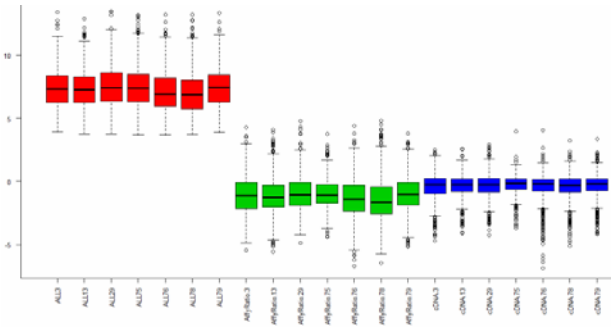

Fig. 4 Boxplots of the arrays (before and after ratio transformation)

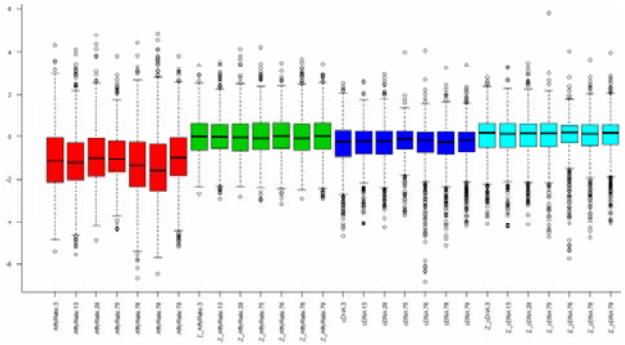

Fig. 5 Boxplots of AffyRatio, cDNA and their respective standard normal distribution

### F. Validation

Towards verifying whether the changes to the Affymetrix dataset after their ratio-transformation has brought in any unwanted change to the overall dataset, we carried out hierarchical clustering of both original as well as of the $Affy_{ratio}$ data separately using *Euclidean distance* and *Ward* agglomeration method. As observed in Fig. 6 and 7, both the clustering showed that the overall relationship was successfully preserved.

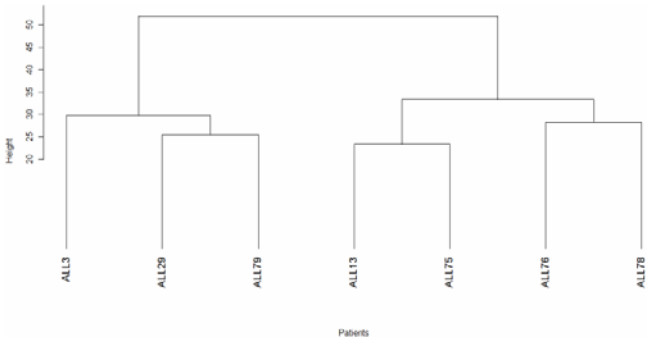


Fig. 6 Hierarchical Clustering of original Affymetrix data (with Euclidean distance and Ward Agglomeration method)

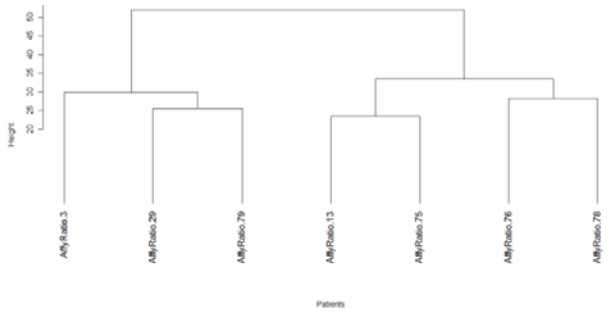


Fig. 7 Hierarchical Clustering of $Affy_{ratio}$ (with Euclidean distance and Ward Agglomeration method)

## III. Conclusion

The increasing number and publicly available microarray studies have provided the opportunity for cross platform studies. However, the cDNA and Affymetrix platform has considerably large disagreement which makes them difficult for their direct comparison. We have introduced here a simple and direct comparison of the data from the two platforms that brings the cDNA and Affymetrix data to a common and comparable level. It appears that further implementation of integrative bioinformatics applications and robust statistical techniques may render substantial improvement towards extrapolation of this idea in the direction of integrating data from cDNA and Affymetrix platform.

## References

[1] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270(5235):467–70.

[2] Smith V, Botstein D, Brown PO. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. Proc Natl Acad Sci USA 1995;92(14):6479-83.

[3] Tan PK, Downey TJ, Spitznagel EL, Jr., Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 2003;31(19):5676-84.

[4] Severgnini M, Bicciato S, Mangano E, Scarlatti F, Mezzelani A, Mattioli M, Ghidoni R, Peano C, Bonnal R, Viti F, Milanesi L, De Bellis G, Battaglia C. Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment. Anal Biochem 2006;353(1):43-56.

[5] Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W. Multiple-laboratory comparison of microarray platforms. Nat Methods 2005;2(5):345-50.

[6] Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. Nat Methods 2005;2(5):337-44.

[7] Canales RD, Luo Y, Willey JC, Austermiller B, Barbacioru CC, Boysen C, Hunkapiller K, Jensen RV, Knight CR, Lee KY, Ma Y, Maqsodi B, Papallo A, Peters EH, Poulter K, Ruppel PL, Samaha RR, Shi L, Yang W, Zhang L, Goodsaid FM. Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol 2006;24(9):1115-22.

[8] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Longueville Fd, Fischer GM, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Scherf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Tong W, William Slikker J. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology 2006;24(9):1151-61.

[9] Ihaka R, Gentleman R. R: a language for data analysis and graphics. J. Comput. Graph. Statist. 1996;5:299–314.

[10] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.

[11] Grewal A, Lambert P, Stockton J. Analysis of expression data: an overview. Curr Protoc Bioinformatics 2007;Chapter 7:Unit 7.1.

[12] Mar JC, Kimura Y, Schroder K, Irvine KM, Hayashizaki Y, Suzuki H, Hume D, Quackenbush J. Data-driven normalization strategies for high-throughput quantitative RT-PCR. Bmc Bioinformatics 2009;10:110.

[13] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185-93.

[14] Bomstad BM, Irizarry RA, Gautier L, Wu Z. Preprocessing high-density oligonucleotide arrays. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S Bioinformatics and computational biology solutions using R and Bioconductor. NY, USA: Springer, 2005. p. 13-32.

[15] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4(2):249-64.

[16] Emerson J, Hoaglin D. Analysis of two way tables by medians. In: Hoaglin DC, Mosteller F, Tukey JW Understanding robust and exploratory data analysis. NY, USA: Wiley-Interscience, 2000. p. 166-207.

[17] Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK. A comparison of background correction methods for two-colour microarrays. Bioinformatics 2007;23(20):2700–07.

[18] McGee M, Chen Z. Parameter estimation for the exponential-normal convolution model for background correction of affymetrix GeneChip data. Stat Appl Genet Mol Biol 2006;5(1):Article24.

[19] Smyth GK, Speed T. Normalization of cDNA microarray data. Methods 2003;31(4):265-73.

[20] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30(4):e15.

[21] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 1979;74(368):829-36.

[22] Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R. Evaluation of normalization methods for microarray data. BMC Bioinformatics 2003;4:33.

[23] Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 2001;29(12):2549-57.

[24] Yang YH, Dudoit S, Luu P, Speed TP. Normalization for cDNA microarray data. In: Bittner ML, Chen Y, Dorsel AN, Dougherty ER Microarrays: optical technologies and informatics (proceedings of SPIE). San Jose, California: SPIE-International Society for Optical Engineering, 2001. p. 141-52.

[25] Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer, 2005. p. 397-420.

[26] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3:Article3.

[27] Miller RGJ. Simultaneous statistical inference, 2 ed. Springer, 1981.

[28] Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 1988;75(2):383-86.

[29] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. 1995;57:289-300.

[30] Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 1979;6(2):65-70.

[31] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75(4):800-02.

[32] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 2001;29(4):1165-88.
[33] Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2000;28(1):10-4.