Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

A Novel Stepwise Normalization Method for Two-Channel cDNA Microarrays

Yuanyuan Xiao[1], C. Anthony Hunt[1], Mark R. Segal[2], Yee Hwa Yang[3]
[1]Department of Biopharmaceutical Sciences
[2]Division of Biostatistics, [3]Department of Medicine
University of California, San Francisco, CA, USA

Abstract—**Microarray experiments contain many sources of systematic errors. In order to extract biologically relevant information from microarray data, normalization needs to be applied to remove such variations. Although a number of normalization models have been proposed, it has not been well researched on how to select the most appropriate model with respect to the observed data. To tackle this problem, we propose in this paper a new stepwise within-slide normalization method, STEPNORM. It is a normalization framework that integrates various models of different complexities to sequentially detect and adjust systematic variations associated with spot intensities, print-tips, plates and two-dimensional spatial effects. We demonstrate the utility of STEPNORM on a set of well-studied cDNA microarray experiment.**

Keywords—cDNA **microarray, normalization**

I. Introduction

Microarray technology enables simultaneous monitoring of the expression of thousands of genes [1]. Like other measuring technologies, microarray data contain inherent systematic measurement errors arise from variations in labeling, hybridization, spotting or other non-biological sources [2]. Normalization procedures, which adjust microarray data to remove such systematic variations, are therefore important for subsequent analysis of either differential expression or gene expression profiling. In this paper, we describe a new systematic stepwise normalization procedure on two-channel cDNA arrays and illustrate its usage on a *swirl* zebrafish slide. The *swirl* experiment is comprised of four replicate hybridizations that contain 8,448 spots. It was carried out using zebrafish as a model organism to study the effect of a point mutation in the BMP2 gene that affects early development in vertebrates.

Two-channel microarrays measure relative abundance of expression of thousands of genes in two mRNA populations. This relative abundance is usually expressed as ratios, $M = \log_2(R/G)$, where R and G are the fluorescent intensity measurements of the red and green channels. The most pronounced systematic variation embodied in the ratios that does not contribute to differential expression between the two mRNA populations is the imbalance of the green and red dye incorporation. This imbalance is manifested as the dependence of ratios on primarily two factors, the fluorescent intensity (hereafter represented by the symbol A) and the spatial heterogeneity (hereafter represented by the symbol S).

The A bias can be best illustrated using a *MA-plot* [3], where M is plotted against A ($A = \log_2 \sqrt{RG}$). As the assumption is that the majority of genes are constantly expressed between the two mRNA samples, symmetrical distribution of points around the horizontal $M=0$ line is expected. Yet frequently we observe linear or nonlinear trend between M and A signaling the undesirable dependence of M on A. An example of nonlinear dependence between M and A is illustrated in Fig. 1a).

The S bias originates from different experimental conditions applied on spots from different areas on the slide. There are usually three major sources that contribute to this spatial variation. First, spots on the same slide are divided into different grids, which are printed by different print-tips from the printing robot; the inequality among M from different print-tips is well illustrated in Fig. 1a). Second, spots of different rows of the slide are often of different well plate sources; one can imagine that there may be effects associated with different well plates. Last, the physical condition of the slide itself could also differ region from region. Fig. 1b) reveals such artifacts on a *swirl* slide by color-coding the ranks of ratios. It shows that the un-normalized ratios are not uniform, and are higher (red) at the middle and lower (green) around the edges.

We proceed in the next section to introduce a new stepwise normalization method and then showcase its usage on a *swirl* slide in the Result section. The last section concludes with discussion and some observations.

II. Methodology

Having illustrated the existence of non-biological variations in microarray data, we review a number of popular normalization methods before proposing a new systematic approach to remove such variations.

Within-slide intensity bias There are currently three most applied models for the removal of the A bias from

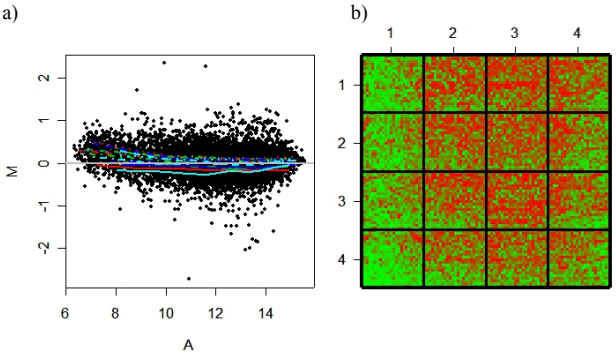


Fig. 1. Diagnostic plots for a *swirl* slide: (a) *MA-plot* with *loess* fits for individual print-tip groups. (b) Spatial plot of log ratios. The plot is divided into 16 grids representing the 16 different print-tips.

ratios. Suppose Normalized ratios are obtained as, $M' = M + c$. Models differ in determining the correction factor *c*. Global median shift, the most simplistic method, assumes that *c* is constant regardless of the intensities of the spots (*A*), and it merely shifts the median of the ratios to be zero. As a refinement to global normalization, robust linear regression (*rlm*) recognizes the need of *A*-dependent correction by fitting a linear regression model. The correction factor *c* in this model is therefore a linear function of *A*. This model is sufficient when the relationship between *M* and *A* is approximately linear, but fails to correct any nonlinear relationship between them. Nonlinear methods developed by Yang *et al.* [3] apply the robust scatter-plot smoother *loess* to perform a local intensity-dependent normalization; the correction factor *c* is therefore a nonlinear function of *A*.

Within-slide spatial bias: Print-tip and Plate Models, such as median shift, *rlm* and *loess*, if fitted within each print-tip (*PT*) or plate (*PL*), corrects the *PT* or *PL* bias. Median shift adjusts ratios within each *PL* (or *PT*) to be zero in an effort to correct for existing inequality between such spatial attributes, and it is a robust version of the ANOVA approach proposed by Sellers *et al.* [4]. Models *rlm* and *loess*, the latter being the most widely practiced normalization method, corrects for the *A* and *PT* (or *PL*) biases simultaneously.

Within-slide spatial bias: 2D Spatial Other than print-tip and plate effects, there could be other spatial attributes that contribute to the spatial heterogeneity (see Fig. 1b)). Sellers *et al.* [4] applied an ANOVA model to test the effects due to array rows and columns. They treat the row and column effects as categorical variables; hence the spatial heterogeneity is modeled as discreet and non-uniform changes. As a result the model fits a large number of parameters as the size of arrays commonly runs up to about a hundred rows and columns. An alternative approach proposed by Yang *et al.* [3] models the spatial heterogeneity as a smooth trend by treating rows and columns as continuous variables and fitting a two-dimensional *loess* curve. So doing requires much fewer parameters than the ANOVA model. Yet another way to model local spatial effects is proposed by Wilson *et al.* [5]. The spatial trend in this model is estimated by computing for each spot, the median log ratio over its spatial neighborhood (*3X3*). This model is able to correct any local spatial trend, for example, a small streak of artifacts, yet so doing costs a lot more degrees of freedom.

We have reviewed several models that could be applied for the elimination of non-biological biases in microarray data. These models differ in their assumptions and complexities. As biases are slide- and experiment-dependent, different slides may show different intensity and spatial trends. Using one model to correct all biases in a slide or using the same model for different slides exhibiting different biases might not be adequate. A more appropriate scheme that captures the particularities of each slide by assessing quantitatively the adequacy of each model with respect to the observed data is urgently called for. Precisely for this reason, we are proposing a new normalization framework that is stepwise and adaptive in nature. This new method is hereafter called STEPNORM.

Fig. 2 illustrates the procedures of STEPNORM using the example of the *swirl* experiment. It consists of four steps and in each step one bias is targeted for correction. The intensity *A* bias is usually the major source of variation and is therefore subjected to examination first. After the correction of the *A* bias, normalized log ratios are subjected to further normalizations based on the existence of spatial biases. As illustrated earlier, there are primarily three types of spatial biases: print-tip (*PT*), plate (*PL*) and two-dimensional effects (*2D*) and they will be tested sequentially. *PT* is subjected to testing first because the number of print-tips is usually smaller than plates and therefore costs fewer degrees of freedom; furthermore, various research [4] has shown the *PT* effect is usually more dominant than other spatial biases.

In each step of our new method, there are a number of competing models of different complexities. The solution to the problem of evaluating several candidate models is to select the model that provides an adequate description of the data while using a minimum number of parameters. Take the example of the first step -- removal of *A* bias, among the candidate models, median shift is the simplest, estimating only one parameter -- the median; however its effect is correspondingly very limited and probably only suits for data that do not show a significant amount of linear or nonlinear trend between *M* and *A*. On the other hand, the local fitting nature of the nonlinear local regression model *loess* is able to accommodate corrections for non-linearity, yet doing so requires fitting more parameters and runs the risk of over-fitting. Therefore, the challenge is to select the model that achieves the best balance between goodness of fit and simplicity. One of the most popular methods, taking both data fitting and model complexity into account, is the Bayesian Information Criterion (BIC), which is defined as, $BIC = -2\log(\hat{L}) + K\log(N)$, where $\hat{L}$ is the maximum likelihood of the normalization model, *K* the number of free parameters in the model and *N* the sample size. We integrate BIC into STEPNORM as the model selection criterion; the model with the lowest BIC value is considered to the preferred model in each step. Importantly, each step also includes testing a "null" model, which doesn't fit any parameters and represents the scenario that the systematic variation in this step is not statistically significant to warrant any correction.

III. Results

We illustrate in this section the application of STEPNORM on one of the *swirl* slides.

For the correction of the *A* bias, we have observed that almost all data have some sort of trend between *M* and *A,*

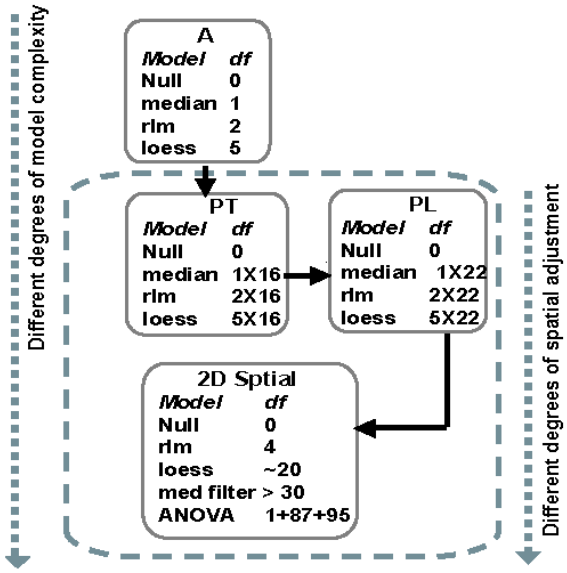


Fig.2 STEPNORM procedures for the *swirl* experiment. The *swirl* slides have 16 print-tips, 22 well plates, 88 rows and 96 columns.

though to various degrees. To find the most appropriate model for the swirl slide, we compare three models median shift, *rlm* and *loess*. From the inspection of the BIC values in Table 1, it is apparent that *loess* is the preferred model, possessing the lowest BIC value of all models. The span (*f*) parameter we used in *loess* (a function in the statistical software R) is: $f = 0.4$. The number of free parameters required in the fitting is estimated to be 4.95 from the output of this function, which could also be approximated as follows. This span parameter defines that about 40% of points are used for a local fitting in each moving window. Each fitting here is linear and therefore requires two degrees of freedom. Totally, about $\frac{1}{0.4} \times 2 = 5$ degrees of freedom are needed. The model *rlm* is a special case of *loess* when within each local area the fitting is linear and the span size is set to 1. We have observed good behaviors of *loess* for most of the slides that we have analyzed, the *swirl* slide being a typical example. The results in Table 1 illustrate that due to the typical high spot density nature of microarray data, a relative large span (0.4) is adequate to capture the nonlinear dependence of *M* on *A*, and it is also big enough to avoid the over-fitting concern.

We proceed next to the removal of the *PT* bias. Fig. 1a) reveals that before any normalization is carried out intensity trends within print-tips show nonlinear tendencies. Yet, such nonlinear trends disappeared after the first step *A*-bias correction. Indeed, *MA*-plot with *loess* fits within print-tips in Fig. 3c) shows largely vertical shifts and no evident curvature. Appropriately, Table 1 indicates that median shift is a better model than the more complex ones, such as *rlm* and *loess*. The same phenomenon also applies to the *PL* bias.

The last step in STEPNORM tests if there are remaining systematic variations associated with spot locations on the slide. Fig. 3g) highlights spots with the highest and lowest 15% pre-normalization ratios and reveals some spatial effects especially in the first and last column on the slide, where high (red spots) and low (green spots) ratios show noticeable separations. Table 1 indicates that *loess* is the preferred model to remove such spatial bias. Normalized ratios using the 2D-*loess* model is plotted in Fig. 3h) which shows an improved distribution of ratios on the slide.

We also applied STEPNORM on other slides in the *swirl* experiment and have obtained similar results (results not shown).

IV. Discussion

Efficient normalization is crucial for microarray research. It directly influences outcomes of downstream data analyses that could give rise to important biological implication and discoveries. In this paper we have presented a new normalization procedure STEPNORM, which integrates a number of published methodologies under the same framework and assesses their effectiveness via a quantitative criterion. Such a process is applied to each individual slide in an experiment so that data (slide) specificity could be achieved.

Unlike other normalization methods, STEPNORM could avoid data under-fitting or over-fitting as it implements both bias detection and removal in the same context. Intensity-dependent bias in ratios is usually the most common and dominant in microarray spot measurements. Very frequently such bias exhibits as a nonlinear trend between *M* and *A*; the curvature can be

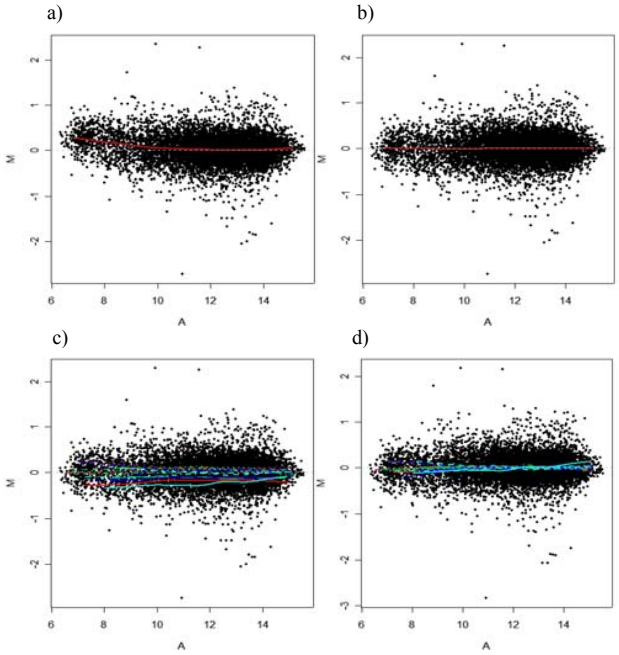
TABLE 1
STEPWISE NORMALIZATION (SWIRL DATA)

Bias	Models	*K*	*-2logL (X10^4)*	**BIC** ***(X10^4)***
A	Null	0	-1.978	-1.978
	median shift	1	-1.985	-1.984
	rlm	2	-2.002	-2.008
	loess	4.95	-2.016	**-2.011**
PT	Null	4.95	-2.016	-2.011
	median shift	4.95+16	-2.117	**-2.098**
	rlm	4.95+32	-2.122	-2.089
	loess	4.95+79.64	-2.128	-2.051
PL	Null	20.95	-2.116	-2.098
	median shift	20.95+22	-2.172	**-2.133**
	rlm	20.95+44	-2.178	-2.119
	loess	20.95+112	-2.098	-2.077
2D Spatial	Nulls	42.95	-2.172	-2.133
	rlm	42.95+4	-2.172	-2.130
	loess	42.95+13.6	-2.185	**-2.134**
	med filter	42.95+70	-2.208	-2.105
	ANOVA	42.95+183	-2.224	-2.020

estimated using a suitable robust scatter plot smoother, such as the *loess* procedure, which have shown good performance for the adjustment of the *A* bias for most slides we have analyzed using STEPNORM. However, we have also observed that the *A* bias is usually a whole-slide phenomenon and doesn't localize within spots related to a specific print-tip or plate. Therefore the current common practice that performs *loess* within each print-tip (LPT) to remove the *A* and *S* biases simultaneously appears to be over-fitting for most datasets. For slides like the s*wirl* experiment that have 16 print-tips, LPT estimates about $5 \times 16 = 80$ parameters when the span size is set at 0.4. On the other hand, the procedure preferred by STEPNORM applies *loess* for the removal of whole-slide *A* bias and then employs a simple median shift among ratios in different print-tips to remove the *PT* bias. So doing estimates only $5 + 16 = 21$ parameters.

The BIC criterion is no doubt an important component in the STEPNORM framework. It is chosen for model selection in STEPNORM for two reasons. First, it is quick to compute which makes it more appropriate than other computation-heavy criteria, such as cross-validation (CV), in the application of microarray datasets, which are usually large in size. Second, we have also observed that the outcome of applying BIC is largely compatible with that of applying CV (results not shown), which indicates the using BIC gives appropriate and reliable results and it is suitable in the application of microarray normalization.

The STEPNORM procedure currently addresses within-slide normalization issues, which is arguably the most important step in cDNA microarray normalization. We refer the reader to [6] for a detailed discussion on methods



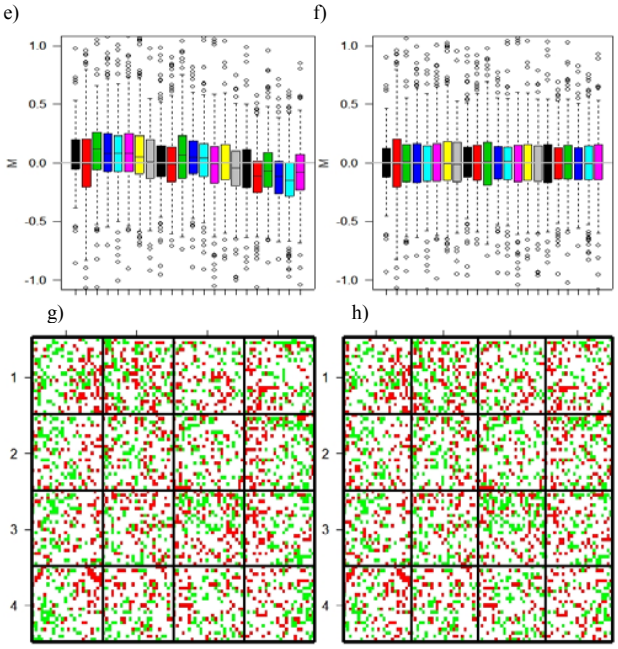


Fig. 3 Graphical display of bias before (left column) and after (right column) stepwise normalization for the removal of a, b) *A* bias; c, d) *PT* bias (with *loess* fits for individual print-tips); e, f) *PL* bias; g, h) 2D-spatial effects (highlighting the top 15% spots in both directions).

concerning between-slide normalizations. In addition, the methods in the STEPNORM procedure are implemented in an R package called STEPNORM, which is available for download form http://www.biostat.ucsf.edu/jean.

ACKNOWLEDGMENT

The authors thank the Ngai lab at the UC Berkeley for providing the *swirl* datasets and Ms. Pearl Johnson for administrative help.

REFERENCES

[1] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent, "Use of cDNA microarray to analyse gene expression patterns in human cancer," *Nature Genetics,* vol. 14, pp. 457-460, 1996.

[2] M. Schena, editor, *Microarray Biochip*. Eaton, 2000.

[3] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization of cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acid Research,* 30(4):e15, 2002.

[4] K. F. Sellers, J. Miecznikowski, and W. F. Eddy, "Removal of systematic variation in genetic microarray data," unpublished

[5] D. L. Wilson, M. J. Buckley, C. A. Helliwell, and I. W. Wilson, "New normalization methods for cDNA microarray data," *Bioinformatics*, vol. 19, pp. 1325-1332, 2003.

[6] Y. H. Yang and N. Thorne, "Normalization for two-color cDNA microarray data," IMS Lecture Notes, Monograph Series, vol. 40, pp. 403-418, 2003.