A Composite Three-stage Normalization Method
for cDNA Microarray Data

by

Ming-Jie Ma

Thesis Advisor: Jinsoo Hwang

A THESIS
Submitted to the faculty of
INHA UNIVERSITY
in partial fulfillment of the requirements
for the degree of
MASTER OF SCIENCE

The Graduate School of
Statistics
February 2008
A Composite Three-stage Normalization Method
for cDNA Microarray Data

by

Ming-Jie Ma

A thesis submitted to the faculty of inha University in partial fulfillment of the requirements for the degree of master of science in the Graduate School of Statistics

Incheon, Korea
February 2008

Approved by Prof. Jinsoo Hwang
Submitted to the faculty of

Major Advisor
This certifies that the thesis of Mingjie Ma is approved.

Chair

Referee

Referee

The Graduate School of Statistics
Inha University
February 2008
CONTENTS

CONTENTS ........................................................................................................... i

LIST OF FIGURES ............................................................................................ ii

ABSTRACT ......................................................................................................... iii

목론 요약 ........................................................................................................ iv

1. Introduction .................................................................................................... 1

II. Review of normalization methods ................................................................. 3

1. Normalization illustration ............................................................................. 3

2. Current normalization methods ..................................................................... 5

III. Main Results ............................................................................................... 9

1. Proposed method .......................................................................................... 9

2. Simulation ...................................................................................................... 11

IV. Conclusion .................................................................................................. 31

Reference ......................................................................................................... 33
LIST OF FIGURES

Figure 1-1 Typical microarray experiment .................................................. 1
Figure 1-2 DNA microarray experiment mimetic diagram ...................... 2
Figure 2-1 The general shape of microarray array ................................... 9
Figure 2-2 RSC data gene expression profile ........................................ 11
Figure 2-3 The 36 slides of RSC data ..................................................... 12
Figure 2-4 The sequence of 36 slides .................................................... 13
Figure 3-5 G vs R plot ........................................................................ 13
Figure 3-6 logG vs logR plot ................................................................. 15
Figure 3-7 MA-plot illustration ............................................................... 16
Figure 3-8 logG vs logR plot and M vs A plot of the first slide .............. 17
Figure 3-9 M vs A plot ......................................................................... 17
Figure 3-10 Original MA-plot vs after first stage normalization MA-plot 19
Figure 3-11 The MA-plot of the first slide use RSC data ....................... 20
Figure 3-12 The M vs A plot of RSC data ............................................. 22
Figure 3-13 Comparison four figures .................................................... 23
Figure 3-14 Extraction the first slide from figure 3-13 ......................... 24
Figure 3-15 Box plot of 36 slides use original RSC data ..................... 25
Figure 3-16 MAD normalized box plot ............................................... 27
Figure 3-17 Print-tip, MAD normalized box plot .................................. 28
Figure 3-18 Shift-log, MAD normalized box plot ................................... 29
Figure 3-19 Print-tip, Shift-log, MAD normalized box plot .................. 30
Figure 4-1 Comparison normalized box plots (1)-(5) ......................... 31
ABSTRACT

In microarray experiment, we can find there are many undesirable systematic variations. The focus of this paper is to propose a robust normalization method, called a composite three-stage normalization method for cDNA microarray data in order to reduce these systematic variation among or within arrays. The proposed new normalization method is a composite of current normalization methods. It consists of three stages. The first stage is print-tip loess normalization, it provides a well-tested general purpose normalization method which has given good results on a wide range of arrays. The second stage is shift-log method that can balance the dye bias. In the third stage, we will make the M value normalized using the median absolute deviance value, we can call it mad method for short. According to the result, we can see the proposed method is a robust normalization method for cDNA microarray data. After normalization, the variance between slides is very small, the result is just what we have expected. This composite three-stage normalization method is much better than only making one normalization for cDNA microarray data. And through comparing with other simple normalization methods, we showed that our method performs better than the other methods.

Keywords: cDNA microarray, gene expression, normalization, dye bias, variation, MA-plot, three-stage, print-tip, shift-log, MAD.
국문 요약

마이크로어레이 실험에서 혼합 불필요한 체계적인 편의가 존재한다. 본 논문에서는 슬라이드 간 혹은 슬라이드 내의 체계적인 편의를 감소시키는 cDNA 마이크로어레이 데이터에 대한 강한 심단계표준화방법을 제안한다. 제안한 새로운 표준화방법은 기존의 표준화 방법의 종합으로써 세 단계로 나누어진다. 첫 번째 단계에서 넓은 영역의 어레이에 대해 좋은 결과를 얻을 수 있는 Print-tip 표준화방법을 적용하고, 두 번째 단계는 Shift-log표준화방법을 적용하여 염색체 평형을 이루게 하며, 세 번째 단계는 절대편차증강값(rad)을 이용하여 M값의 평형을 얻는다.

제안한 표준화 방법을 실험데이터에 적용하였을 때 슬라이드간의 편차가 아주 작아짐으로써 좋은 결과를 얻을 수 있었다. 이 실험 결과로부터 본 논문에서 제안한 심단계표준화방법은 cDNA 마이크로어레이 데이터에 대해 아주 강한 편차를 얻을 수 있었고, 기존 단일 표준화방법에 비해 훨씬 좋은 성능을 나타낼 수 있었다.

Keywords: cDNA microarray, gene expression, normalization, dye bias, variation, MA-plot, three-stage, print-tip, shift-log, MAD.
Chapter 1 Introduction

There are many undesirable systematic variations in the microarray experiments. Normalization is the process of removing such variations that affects the measured gene expression levels. Normalization plays an important role in the earlier stage of microarray data analysis. The subsequent analysis results are highly dependent on normalization.

Before we discuss the normalization methods of cDNA microarray data, at first, we suppose that an experiment has been conducted using a series of two-color cDNA microarrays.

Figure 1-1 can help us to realize a cDNA microarray experiment.

![Figure 1-1 Typical microarray experiment](image)

Each microarray has been hybridized with RNA from two sources labeled with different fluorescent dyes, then mixed and hybridized with the arrayed probes. The two color channels will be referred to by convention as red (Cy5) and green (Cy3). We suppose that the arrays have been scanned to produce images and that these images have been further processed by an image analysis program to produce measured red and green foreground and
background intensities for each spot on each array. Before the gene expression profiles of the RNA samples can be analyzed and interpreted, the red and green intensities must be normalized relative to one another so that the red/green ratios are as far as possible an unbiased representation of relative expression.

We can describe the process of DNA microarray experiment as the mimetic diagram as following:

![DNA microarray experiment mimetic diagram](image)

Figure 1-2 DNA microarray experiment mimetic diagram

In Figure 1-2, the penultimate stage is "Data preprocessing and normalization". We will discuss this stage and propose a new normalization method for cDNA microarray experiment in this paper.
Chapter 2 Review of normalization methods

2.1 Normalization illustration

This section we will make illustration for cDNA microarray normalization. We will talk about the concept of normalization, the purpose of normalization and the sources of non-biological variation.

2.1.1 What is Normalization?

Normalization means to adjust microarray data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between the printed probes.

In order to accurately and precisely measure gene expression changes, it is important to take into account the random(experimental) and systematic variations that occur in every microarray experiment.

Normalization of microarray data is any procedure to reduce or adjust for such systematic variation among or within arrays.

Normalization methods are often applied prior to the application of statistical analysis methods, which are usually designed to detect differential expression. Also, normalization includes procedures that adjust for known effects as part of the statistical analysis.

2.1.2 Purpose of Normalization

The aim of any normalization method is to remove the biases within each microarray involved in an experiment, so that the microarray data can be used for downstream analyses, such as detection of differentially
expressed genes or multivariate analyses.

The need for normalization can be seen most clearly in self-self experiments, in which two identical mRNA samples are labeled with different dyes and hybridized to the same slide.

### 2.1.3 Sources of Non-Biological Variation

The sources of noise can be divided into instrument noise, which results either from substrate, coating, and sample or by nonspecific hybridization to the probes on the microarray surface. Such as following:

1. **dye bias**: there are some differences in heat and light sensitivity
2. **differences in the amount of labeled cDNA hybridized to each channel in a microarray experiment**
3. **different amount of mRNA**
4. **variation across replicate slides**
5. **variation across hybridization conditions**
6. **variation in scanning conditions**
7. **variation among technicians doing the lab work**
2.2 Current normalization methods

2.2.1 Within-slide normalization

In this case, the normalization is done separately for each slide, using only the red and green intensities for this slide. Several approaches are described in the following:

1. Lowess normalization method

Lowess normalization assumes that the dye bias appears to be dependent on spot intensity. The adjusted ratio is computed by:

\[
\log(R/G) = \log(R/G) - c(A)
\]

(2.1)

where \(c(A)\) is the Lowess fit to the \(\log(R/G)\) vs \(\log(\sqrt{R \times G})\) plot.

If green has been chosen as the treatment dye and red as the control dye, then \(R\) and \(G\) are reversed in the above formula. Treatment and control dyes are designated when the data is imported into GeneLinker.

Lowess regression, or locally weighted least squares regression, is a technique for fitting a smoothing curve to a dataset. The degree of smoothing is determined by the window width parameter. A larger window width results in a smoother curve, a smaller window results in more local variation.

2. Global normalization method

Global methods assume that the red and green intensities are related by a constant factor. That is, \(R = k \cdot G\), and in practice, the center of the
distribution of log-ratios is shifted to zero:

$$\log_e R/G \rightarrow \log_e R/G - c = \log_e R'(kG)$$  \hspace{1cm} (2-2)

A common choice for the location parameter $c = \log_e k$ is the median or mean of the log-intensity ratios for a particular gene set. Global normalization methods are mentioned as pre-processing steps in a number of papers on the identification of differentially expressed genes in single-slide cDNA microarray experiments. Chen et al. (2000) assume that $R = k \cdot G$ and propose an interactive method for estimating the constant normalization factor $k$ and cut-offs for the red and green intensity ratio $R/G$. In some software packages (e.g. GenePix), a constant normalization factor is estimated such that the arithmetic mean of the intensity ratios of all the genes on a given microarray is one. Global normalization methods are still the most widely used methods in spite of the evidence of spatial or intensity dependent dye biases in numerous experiments.

3. Intensity dependent normalization

In many cases, the dye bias appears to be dependent on spot intensity, as revealed by plots of log-ratio $M$ vs overall spot intensity $A$. An intensity or $A$-dependent dye normalization method may thus be preferable to global methods.

We use the robust scatter-plot smoother lowess from the statistical software package R to perform a local $A$-dependent normalization:

$$\log_e R'/G \rightarrow \log_e R'/G - c(A) = \log_e R'(k(A)G)$$  \hspace{1cm} (2-3)

where $c(A)$ is the lowess fit to the $M$ vs $A$ plot. The lowess() function is a
scatter-plot smoother which performs robust locally linear fits. In particular, the loweress() function will not be affected by a small percentage of differentially expressed genes which will appear as outliers in the M vs A plot. The user-defined parameter \( f \) is the fraction of the data used for smoothing at each point; the larger the \( f \) value, the smoother the fit. We typically use \( f = 20\% \).

Sapir and Churchill (2000) suggest using the orthogonal residuals from the robust regression of logR vs logG as normalized log-ratios. Since an M vs A plot amounts to a 45° counterclockwise rotation of the (logG, logR) coordinate system (up to multiplicative constants), their method is similar to fitting a robust regression line through the M vs A plot instead of a loweress curve. One can view this linear normalization as a more constrained version of intensity dependent normalization. Kepler propose a more general intensity dependent normalization approach which uses a different local regression method instead of the loweress() function.

### 2.2.2 Multiple slide normalization

After within-slide normalization, all normalization log-ratios will be centered around zero, regardless of the normalization method. Multiple slide normalization methods, which aim to allow experiment to experiment comparisons, may also need to be adjusted for scale when the different slides have substantially different spreads in their log-ratios. Failing to perform a scale normalization could lead to one or more slides having undue weight when averaging log-ratios across experiments.

The main reason that can lead to error in single-slide are Cy3 and Cy5 intensities. When there are several slides (replicated experiment), the differences between these slides can become the reason of variation for the microarray experiment. We can use linear models for normalization.
between-slide.

The popular approach is to use the analysis of variance (ANOVA) model on logarithmic intensity scale that accounts for variations at each stage (dye, spot, slide, etc.) and their interaction terms at the same time. The model introduced in (Kerr et al., 2000) is as follows:

$$\log_2(y_{ijk}) = \mu + A_i + D_j + V_k + G_y + (AG)_{iy} + (VG)_{ky} + e_{ijk}$$  \hspace{1cm} (2.4)$$

Here, $\mu$ is the overall average signal, $A_i$ represents the effect of the $i^{th}$ array, $D_j$ represents the effect of the $j^{th}$ dye, $V_k$ represents the effect of the $k^{th}$ variety, $G_y$ represents the effect of the $y^{th}$ gene, $(AG)_{iy}$ represents a combination of array $i$ and gene $y$, and $(VG)_{ky}$ represents the interaction between the $k^{th}$ variety and the $y^{th}$ gene. The error term $e_{ijk}$ is assumed to be independent and identically distributed with mean zero. After estimating the coefficients, the gene-(sample) variety term $(VG)_{ky}$ is the parameter we want to infer. Note that this approach also assumes linear normalization factors.

The within-slide scale normalization method may also be used for multiple slide scale adjustment. We are currently evaluating this approach with experiments where a scale normalization seems called for.
Chapter 3 Main Results

3.1 Proposed method

The aim of an effective normalization method for cDNA microarray data is to remove noise while retaining signal in the data. I will propose a new composite three-stage normalization method here. The composite three-stage normalization method is more effective than the other normalization methods.

The first stage and second stage normalization method will be done within-slide, and the third stage is between-slide normalization method.

We can call one slide as one array.

![Diagram](image)

Figure 3-1 The general shape of microarray array

We usually divide one slide to several subarrays like in Figure 3-1. And every subarray forms a different groups of pins. The pin groups is made up of many spots. Usually these spots genes have been expressed sequence tages already.

The variance between the print-tips and the dye bias are major sources of variation.

There are three stages during the normalization process, we will make
three times normalization work for cDNA microarray data. The first stage is
Print-tip loess normalization, it provides a well-tested general purpose
normalization method which has given good results on a wide range of
arrays. The method may be refined by using quality weights for good
results on a wide range of arrays. Then after the print-tip normalization
method, we use shift-log normalization method to treat the data that we
have gotten from the first stage. The second stage shift-log method can
balance the dye bias. Since the red intensities often tend to be lower than
the green intensities, so in the shift-log method we let \( R - s = G + s \), the
value of \( s \) can be calculated. In the third stage, we will make the M value
normalized by using the median absolute deviance value. We called it mad
method.
3.2 Simulation

In this paper, we used the Rat Stem Cell data to do the simulation for normalization. The RSC Data was downloaded from this homepage: "http://biostat.smu.ac.kr/public_dataset/rat_stem_cell.csv".

3.2.1 Rat Stem Cell Data (RSC Data)

The following Figure 3-2 is the profile of RSC data.

![Gene expression profile](image)

**Figure 3-2** RSC data gene expression profile

The neural stem cell of rat can be self-renewal. We extract the cortex cell of 14 days rat, and divide the data into two groups. One is ciliary neurotrophic factor (CNTF) data, and the other one is no CNTF data. Here we want to study the mitogen cell and novel cell intensity. We extract the mRNA from nerve tissue that we have gotten. Then we dye with Cy5 and
Cy3, Cy3 as the reference cDNA and hybridize them. In this experiment, the number of probes that we use is 3,840. The experiment time is separated to 6 groups: 12 hours, 1 day, 2 days, 3 days, 4 days, 5 days for each 6 time groups we repeat the experiment 3 times. So there are 2 groups in this experiment (CNTF and no CNTF), 6 time groups, and repeat 3 times. We can get 2×6×3=36 cDNA microarray slides totally in this experiment. We scan these to get the image file for image analysis.

The 36 slides of rat stem cell data can be shown as following:

![Figure 3-3 The 36 slides of RSC data](image-url)
The sequence of the 36 slides are described as following:

<table>
<thead>
<tr>
<th>Slide number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1     2     3     19    20    21</td>
</tr>
<tr>
<td>4     5     6     22    23    24</td>
</tr>
<tr>
<td>7     8     9     25    26    27</td>
</tr>
<tr>
<td>10    11    12    28    29    30</td>
</tr>
<tr>
<td>13    14    15    31    32    33</td>
</tr>
<tr>
<td>16    17    18    34    35    36</td>
</tr>
</tbody>
</table>

Figure 3.4 The sequence of 36 slides

3.2.2 G vs R plot of RSC Data

Figure 3.5 G vs R plot
Let the intensity of Cy3 and Cy5 in cDNA microarray data be G and R respectively. If there isn’t any variation exist, the red and green intensities should be equal. After normalization, the G vs R scatter plot should be seen like a direct proportion function straight line.

The Figure 3-5 shows the G vs R plot of RSC data.

### 3.2.2 Log transform for R & G

If we use logarithm for the ratio, then the difference of brightness value can be described use log value. Some researchers use base 2 in the logarithm because the intensities usually range from 0 to $2^5$, other researchers use base 10 for the ease of intuition. Using a different base, however, does not affect the conclusion of analysis.

Let $M$ represents log-ratio of two dyes, we can express $M$ as following:

$$M = \log \frac{Cy5}{Cy3} = \log \frac{R}{G} - \log R - \log G \quad (3.1)$$

We use log transform brightness value in stead of brightness value, there are several advantages. First, the multiplicative effects become additive effects in log transformation. The linear effect like this is very effective for statistical analysis. The second advantage, after log transform the distribution become relative symmetric, it will be very useful for data analysis.

Let's look at the plot that after log transform, the plot of logG and logR for RSC data can be shown as Figure 3-5. It was drawn by the original rat stem cell data.
3.2.4 M vs A plot

As we have discussed, we can see there are several advantages to do log transform. But we are worried about losing information if we use log-ratio. In order to retain signal in the data, we need a new variable. In Yang et al. (2001) proposed an intensity variable $A$, which is given as follows.

$$A = \log\sqrt{GR} = \frac{\log(G) + \log(R)}{2} = \frac{1}{2} (\log G + \log R)$$  \hspace{1cm} (3-2)

$A$ represents the averaged logarithmic intensity.

The MA-plot is actually a 45-degree rotation and rescaling of log-intensity plot of Cy5 and Cy3. Write $R$ and $G$ for the background-corrected red and green intensities for each spot. Normalization is usually applied to the
log-ratios of expression, which will be written as $M = \log_2 R - \log_2 G$. The log-intensity of each spot will be written as $A = (\log_2 R + \log_2 G)/2$, a measure of the overall brightness of the spot. (The letter M is a mnemonic for minus while A is a mnemonic for add.) It is convenient to use base-2 logarithms for M and A so that M is units of 2-fold change and A is in units of 2-fold increase in brightness. On this scale, $M=0$ represents equal expression, $M=1$ represents a 2-fold change between the RNA samples, $M=2$ represents a 4-fold change, and so on. We can illustrate the MA-plot as following:

![Figure 3-7 MA-plot illustration](image)

If there isn't any variation, the dye effect of Cy5 and Cy3 are equal, so the ratio of Cy5 and Cy3 are nearly 1. Because $M=\log(R/G)$, $R/G=1$, the M value should nearly 0.

Let us choose the first slide among the 36 slides of the original RSC data to illustrate the logR vs logG plot and the transformed M vs A plot.
Figure 3-8 logG vs logR plot and M vs A plot of the first slide

The MA-plot of 36 slides use RSC data, the original plot (before normalization) can be shown as following:

Figure 3-9 M vs A plot
3.25 Composite three-stage normalization method

(1) The first stage; make normalization for every pin group

The actual process of making a cDNA microarray involves the use of a robotic arm (called an arrayer) that presses a print-head gently onto the array surface, thereby transferring the probe (i.e. the polynucleotide). However, as with most machinery, printing parts might need to be changed on a regular basis. Furthermore, during the investigation itself, different print-tips might be used. One must therefore consider the possibility that such changes will introduce systematic variation into the data.

In addition to intensity-dependent variation in log ratios, spatial bias can also be a significant source of systematic error.

Most normalization methods do not correct for spatial effects produced by hybridization artifacts or print-tip or plate effects during the construction of the microarrays.

It is possible to correct for both print-tip and intensity-dependent bias by performing locally weighted regression and smoothing (LOWESS) fits to the data within print-tip groups.

Every grid in an array is printed using the same print-tip. Different experiments may be done using different printing set-ups. Some systematic differences may exist between the print-tips, such as slight differences in the length or in the opening of the tips, and deformation after many hours of printing. Alternatively, print-tip groups are proxies for spatial effects on the slide. Within-print-tip-group normalization is simply a (print-tip+A)-dependent normalization, that is:

$$\log_{2} R/G \rightarrow \frac{\log_{2} R}{G} - c_{t}(A) = \frac{\log_{2} R}{(k(A)G)}$$  \hspace{1cm} (3-3)
where $q_i(A)$ is the lowess fit to the $M$ vs $A$ plot for the $i$th grid only, $i=1,...,I$ and $I$ represents the number of print-tips. The formula above shows that the $M$ value minus $q_i(A)$ in order to make the normalized $M$ value nearly zero. And we can write the difference value as log-ratio, $\log_2 R/(q_i(A)C)$.

We use the RSC data to test the first stage normalization method and the result is shown as following:

![Image of MA-plots](image)

**Figure 3-10** Original MA-plot vs after first stage normalization MA-plot

In Figure 3-10, the left MA-plot (1) is the 36 slides of RSC cDNA microarray data before normalization, i.e. the original MA-plot. The right MA-plot (2) is the 36 slides of RSC data after first stage print-tip normalization.

In order to compare them more clearly, we choose the first slide from original MA-plot and after first stage normalization MA-plot individually.
In Figure 3-11, the left plot (1) is the first slide of original cDNA microarray data, the right plot (2) is the first slide of RSC data M vs A plot after the first stage normalization which is made between the pin groups.

We can find the first stage normalization is effective, the M vs A plot above the first slide above we can see clearly that the plot after first stage normalization is much better than the original MA-plot. In order to get the more precise result, we proceed to do the second normalization work.

(2) The second stage: make normalization within-slide for red and green dyes

The differences in channel offsets is a major systematic effects in cDNA microarray. In this paper we discuss about the three-stage normalization method, we have discussed the first stage about the adjust for pin groups. In the second normalization stage, we will do something to balance the channel offsets, i.e. adjust the red and green fluorescent dyes in one slide. So the second stage is also within-slide normalization.
The shift-log normalization method proposed by Kerr et al. (2002) corrects for differences in the channel offsets, adjusts log-ratios by adding a constant to the dyes. The shift-log transformations for each array could remove the systematic dye effects. It normalizes log-ratios using a translation constant. We let the constant as a letter \( s \). Because we want to make the dyes normalization in one slide, if the red color fluorescent dye and the green color fluorescent dye are balanced, the cDNA microarray data is normalized for this experiment. But actually, since the red intensities often tend to be lower than the green intensities, so in the shift-log method we adjust \( R \) and \( G \), where the value of \( s \) can be calculated later. We can adjust the dyes and make a linear function as following:

\[
\log(R - s) = \log(G + s) \tag{3-4}
\]

The constant \( s \) is calculated as:

\[
s = \text{median} \left( \frac{\log(R_i - s)}{\log(G_i + s)} \right) \tag{3-5}
\]

In this formula, \( \text{mad} \) it means the median absolute deviation of brightness value. The number of gene is \( p \). The estimate \( s \) is robust, so it can’t be influenced by outlier strongly.

We can see the effect of the second stage normalization to make balance for dyes through the MA-plot as following. Let’s compare the original data \( M \) vs \( A \) plot and the MA-plot only after shift-log normalization.
Figure 3-12 The M vs A plot of RSC data

In Figure 3-12, the left plot (1) is the original data 36 slides, the right plot (2) is the MA-plot normalized by shift-log normalization method by using the original RSC data.

We can't see there are any big differences between the two plots. The left 36 slides are similar to the right 36 slides. So we can get the conclusion that if we just do the shift-log normalization for the original data, there isn't big effect on it. The normalization result of only doing shift-log is worse than only doing the first stage normalization print-tip method.
(1) The original data MA-plot
(2) Only print-tip normalization
(3) Only shift-log normalization
(4) First print-tip then shift-log, two stages

Figure 3-13 Comparison of four figures

So we wonder if the result will be better if we do the shift-log normalization approach after the first stage print-tip normalization method.

The Figure 3-13 shows that if we adjust the pin groups within slide first and then adjust the dye balance, we will get a very nice result. The two-stage normalization method is much better than the one stage.
individually.

In order to observe Figure 3-13 more clearly, we choose the first slide of the four figures above respectively as following:

(1) The original data MA-plot  (2) Only print-tip normalization

(3) Only shift-log normalization  (4) First print-tip then shift-log, two stages

Figure 3-14 Extraction the first slide from figure 3-13

In terms of observing the four MA-plots of RSC data in Figure 3-14, we chose the ranges of M are all from -5 to 5. After we did the two stages normalization for the data, the all spots in the first slide that we have chosen are all centered around zero. It illustrates the normalization method
is very effective.

(3) The third stage; make normalization between-slide use the mad scale

We have done the first stage and second stage within-slide normalization for the data.

After the two stages normalization we will do the third stage normalization. The third stage normalization will be done between-slide.

Since the RSC data that we referred in this paper is an iterative experiment data. After we did the normalization within-slide, we should consider the variance between slides.

Let us look at the box plot of the 36 slides. The box plot is drawn using the original RSC data.

![Box plot of 36 slides using original RSC data](image)

Figure 3-15 Box plot of 36 slides use original RSC data

We can see from Figure 3-15 that the difference of variance exist between slides. The variance among slide is big. The range of log-ratio M of 36 slides are from less than -10 to more than 10. So we need to do
normalization for variance between slides.

In order to do normalization for variance between slides, we should calculate the variance for slides.

We suppose the number of gene is \( p \), the number of slide is \( n \), the formula to calculate the variance for \( i \)th slide is given by

\[
\hat{s}_i^2 = \frac{1}{p-1} \sum_{j=1}^{p} M_{ij}^2 \quad i = 1, \ldots, n
\]  

(3-6)

If we want to compare the variances between the slides, we define a estimated constant \( \hat{q} \),

\[
\hat{q}_i = \frac{\hat{s}_i^2}{\sqrt{\prod_{k=1}^{n} \hat{s}_k^2}} \quad i = 1, \ldots, n
\]  

(3-7)

The estimate value \( \hat{s}_i^2 \), which we have calculated above, is the variance of \( i \)th slide. If the value of \( \hat{q}_i = 1 \), then we can see the variance of \( i \)th slide is the average value among all of the slides. If the value of \( \hat{q}_i < 1 \), it illustrates the variance of \( i \)th slide is smaller compared than the other slides. The experiment of \( i \)th slide is very exact. If the value of \( \hat{q}_i > 1 \), it shows the variance of \( i \)th slide is bigger than the other slides. But if the value of \( M_{ij} \) is far from median, for the data whose \( M_{ij} \) value is either too big or too small, then the value of \( \hat{q}_i \) will be very sensitive. So we should find a robust estimate instead of \( \hat{q}_i \).

The median absolute deviation (MAD) are insensitive to outliers. Hence, a normalization scheme using MAD would be robust and the estimate is given by
\[ \tilde{a}_i = \frac{\text{mad}_i}{\sqrt{\prod_{k=1}^{n} \text{mad}_k}} \quad i = 1, \ldots, n \] (3-8)

So we make normalization for variance between-slide use the scale MAD. We can see the normalization result through the following box plots.

At first, we just use the original RSC data make normalization for variance between slides by scale MAD. The box plot is shown as following:

![Box plot image]

**Figure 3-16 MAD normalized box plot**

Figure 3-16 shows the box plot of 36 slides normalized by scale MAD by using the original RSC data. Comparing the box plot above with the Figure 3-15 that hasn't been normalized for variance between slides. We can see there are not big differences between the two figures. The variance between slides can't be reduced enough. The range of log-ratio M has been changed a little. So we can see if we only make variance normalization between-slide by using scale MAD, there is a certain effect for the original data.

Next, we use the RSC data that has been normalized by print-tip method within-slide in the first stage to make normalization for variance between
slides by scale MAD. The box plot is shown as following.

![Box plot diagram](image)

**Figure 3-17** Print-tip MAD normalized box plot

Compare the box plot Figure 3-17 with Figure 3-16 which hasn't been normalized for variance between slides. We can see the median of the box plots of every slide are all centered around zero. The variances between slides have been reduced well, and the range of log-ratio M has been changed. The range of M became from -5 to 10.

Then, we use the RSC data that has been normalized only by the shift-log method within-slide in the second stage to make normalization for variance between slides by scale MAD. The box plot is shown as following:
Figure 3-18 is made by the RSC data that have been normalized by using the shift-log within-slide normalization method in the second stage.

This normalization result in variance between-slide is similar with the above when we only make mad normalization of the original RSC data. Comparing with Figure 3-15, we can see the median of log-ratio M isn't centered nearly zero. The variance between slides won't be reduce enough, the range of log-ratio M has been changed a little.

Since the results above we have gotten after between-slide normalization for variance are not very ideal, we proposed the three-stage normalization for cDNA microarray data.
Figure 3-19 is made by the RSC data that have been normalized by the print-tip and shift-log within-slide normalization methods in the first and second normalization stages. Then at the third stage we did mad normalization for variance method for the whole slides.

From Figure 3-19, we can see that the composite three-stage normalization method is the best one. The Figure 3-19 shows the median of log-ratio $M$ is centered very nearly zero. The variance between slides can be reduce well, the range of log-ratio $M$ has been changed a lot. The range of $M$ became from -2 to 1. In the original box plot, the range of log-ratio $M$ is from -10 to 10. So the three-stage method can reduce the variance as well.
Chapter 4 Conclusion

In terms of comparison the composite three-stage normalization method that we have proposed in this paper is the best method. First we normalize by print-tip & shift-log to balance the variation within slide, and finally we normalize the variance between slides by MAD. After we normalized by using the composite three-stage normalization method, the systemic variance in the experiment can be reduced well.

We can compare the box plots to see the result more clearly.
Figure 4.1 shows us that the whole composite three-stage normalization method is the best one.

The composite three-stage normalization method that we propose here, is the most effective normalization method for cDNA microarray data. So I think this composite three-stage normalization method can be used as a new method in the future.
Reference


Acknowledgement

As a foreigner, I finished my master course in statistics department of Inha University in Korea. During the two years and a half, so many kindly persons helped me. I will never forget the helps and I will say thanks to them here.

At first, I'll always be deeply indebted to my advisor, professor 홍일주, since I came to Korea, professor 홍 took care of me not only on the study but also on my living, he treats me like his daughter. If there wasn't the guidance of my professor, I would not finish this paper. So I really appreciate my advisor professor.

During the master course, I have learnt a lot from other professors in our department. I will also say thanks to grandfatherly professor 전홍석, kindly professor 이재호, smart professor 박진호, capable professor 박회진 and tender professor 김진경. All of the professors shew solicitude for me, I will remember that forever.

In particular, I would like to thank my grandparents and my parents who are in my hometown. They always encouraged me to study hard and to face any difficulties. The supports that come from my relatives helped me to finish my study during these years.

At the same time, I will say thanks to the graduate students in our statistics department. From the first day I came to Korea, 지연언니 treated me like her younger sister and always helped me whenever I was in trouble. I will say sincere thanks to her. I really want to say thanks to funny and kindly 중일, kindly 준성, capable 전수언니, intelligent 진희, pretty 미진언니, good friend 지영, cute 미화언니, clever 승희, and handsome 형진, 용재, 혁준, 민석 who are all the seniors in my department. All of the seniors helped me so much. 혁준, 민선, 브람, who studied with me in the whole two years, helped me how to do the homework and explained any problems, which I couldn't understand to me. I will never forget these good friends of mine. And I will say thanks to my juniors 최경, 하나 and 계성, 민우선생님 and 바야르다, are also be thanked a lot who started their Ph.D studies this semester for their warm
encouragement to me.

I will deeply say thanks to my good Korean friend 수정 and her family. 수정 looks me just like her older sister and 수정’s parents are my Korean father and mother. They always say that I am the Chinese daughter of theirs and they take care of me so patiently. I really felt happy because of their warm hearts. I will never forget the grace of 수정’s parents.

At last, I would also like to thank my boyfriend 학남 and his parents as well his aunts. Without 학남, I couldn’t finish this paper. He has helped me so much that I can’t express my thanks in words. I will thank and love him in my whole life.

In a word, I say thanks here to any persons who love me and always help me. I have benefited considerably from the advice, suggestions, and critical reading of many people. I will try my best in my future study to make my life more beautifully.

Ming-He, Ma
February, 2008