Comparison of Normalization Methods in Microarray Analysis

Comparison of Normalization Methods in Microarray Analysis

By

Rong Yang, B.S.

A Project

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2006) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Comparison of Normalization Methods in Microarray Analysis

AUTHOR: Rong Yang, B.S.

(Peking University, P. R. China)

SUPERVISOR: Dr. Angelo Canty

NUMBER OF PAGES: x, 67

Abstract

DNA microarrays can measure the gene expression of thousands of genes at a time to identify differentially expressed genes. The Affymetrix GeneChip system is a platform for the high-density oligonucleotide microarray to measure gene expression using hundreds of thousands of 25-mer oligonucleotide probes.

To deal with Affymetrix microarray data, there are three stages of preprocessing to produce gene expression measurements/values. These are background correction, normalization and summarization. At each stage, numerous methods have been developed.

Our study is based on Affymetrix MG_U74Av2 chip with 12488 probe sets. Two strains of mice called NOR and NOR.NOD_Idd4/11 mouse are hybridized for the experiment. We apply a number of commonly used and state-of-art normalization methods to the data set, thus compute the expression measurements for different methods. The major methods we discuss include Robust Multi-chip Average (RMA), MAS 5.0, GCRMA, PLIER and dChip.

Comparisons in terms of correlation coefficient, pairwise expression measures plot, fold change and Significance Analysis of Microarray (SAM) are conducted.

Acknowledgements

I would like to thank my supervisor, Dr. Angelo Canty, for his encouragement, patience, and guidance throughout the completion of this project.

I would like to dedicate this project to my husband, Jayden, my daughter, Sarah and my parents, Mei Liang and Chengyun Yang, who have shown their love, support and encouragement throughout my schooling. Thank you for everything you have sacrificed, all the advice you have given, and the unconditional love.

I also express my special thank to Dr. Jayne Danska and Dr. Zhenya Ivakine at The Hospital for Sick Children in Toronto for the data.

Contents

1 Introduction **1**

1.1 Background in DNA and Microarrays 1

1.2 Affymetrix Genechip Microarray 5

1.3 Applications of DNA Microarray Technology 7

1.4 Description of Data 8

2 Preprocessing Stages and Methods **9**

2.1 Background Correction 10

2.1.1 RMA Convolution 10

2.2 Normalization 11

2.2.1 Quantile Normalization 12

2.2.2 Invariant Set Normalization 12

2.3 Summarization 13

2.3.1 Average Difference Summarization 13

2.3.2 Median Polish Summarization 14

2.4 Commonly Used Methods for the Preprocessing 14

2.4.1 RMA 14

2.4.2 MAS 5.0 15

2.4.3 GCRMA 16

2.4.4 PLIER 18

2.4.5 DChip (Li and Wong) 19

3 Comparisons Based on Expression Measurements 20

3.1 Correlation Coefficient of Expression Measurements 20

3.2 Pairwise Plots of Expression Measurements 25

3.2.1 MAS 5.0 Present/Absent Calls 25

3.2.2 RMA vs. GCRMA based on the affinity 30

4 Comparisons based on fold-change 35

4.1 Fold Change 35

4.2 Housekeeping Genes 36

4.3 Quantitative (Real-Time) PCR 36

4.4 Results 40

5 Comparisons based on SAM 45

5.1 Introduction 45

5.2 Model and main idea about SAM 47

5.3 Comparison results between RMA, MAS5, GCRMA and PLIER 49

5.4 Comparison results between RMA and dChip 52

6 Conclusion and Discussion **57**

A Table of the Definition of Methods **60**

List of Tables

3.1 *Reduced preprocessing methods list* 22

3.2 *Pairwise correlation coefficients between different preprocessing methods* 23

3.3 *Correlation coefficients comparison for the Li and Wong* 23

3.4 *Methods Description for comparing with dChip* 24

4.1 *Part of normalized gene data by take Actin housekeeping gene and one of genes of interest called C1qb.* 39

4.2 *Fold change comparisons between microarray and Real-Time PCR (A)* . 42

4.3 *Fold change comparisons between microarray and Real-Time PCR (B)* 43

5.1 *Ranking correlations between RMA, MAS5, GCRMA and PLIER.* . . . 49

5.2 *q-value correlations between RMA, MAS5, GCRMA and PLIER.* 49

5.3 *d-statistic correlations between RMA, MAS5, GCRMA and PLIER* . . 50

5.4 *Ranking correlations between RMA and dChip* 53

5.5 *q-value correlations between RMA and dChip* 53

5.6 *d-statistic correlations between RMA and dChip* 53

A.1 *Possible preprocessing methods list (part A)* 61

A.2 *Possible preprocessing methods list (part B)* 62

A.3 *Possible preprocessing methods list (part C)* 63

List of Figures

1.1 *A schematic of the role of RNA in gene expression and protein production. Graphics from http://www.accessexcellence.org* 2

1.2 *Perfect Match and Mismatch Probes. Figure came from the slides of Dr. Roger Bumgarner at University of Washington* 6

3.1 *Expression measurements comparison between RMA and MAS 5.0. (A) Original expression measures for RMA vs. RMA; (B) RMA vs. GCRMA expression measurements for those genes that present at once ("good genes" subset); (C) RMA vs. GCRMA expression measurements for which the detection algorithm declares them present.* 28

3.2 *Expression measurements comparison between GCRMA and RMA, and GCRMA and MAS 5.0.* 29

3.3 *Expressions of RMA vs. GCRMA based on the affinity. Red=low, blue=medium and yellow=high.* 31

3.4 *Expressions of RMA vs. GCRMA based on the different levels of affinity. (A) RMA vs. RMA for low affinity; (B) RMA vs. RMA for medium affinity; (C) RMA vs. RMA for high affinity; (D) RMA vs. RMA for low and medium affinity;* 32

3.5 *Histogram of expressions of RMA and GCRMA. (A) and (E) are original histogram of expression values for RMA and GCRMA respectively. (B) and (F) are expression histograms of high affinity of RMA and GCRMA; (C) and (G) are expression histograms of medium affinity of RMA and GCRMA; (D) and (H) are expression histograms of low affinity of RMA and GCRMA;* 34

5.1 *q-value pairwise correlation coefficient plot from RMA, MAS 5.0, GCRMA and PLIER. From left to right and from top to bottom, the corresponding orders of the graphs are RMA, MAS 5.0, GCRMA and PLIER. Var 1, var 2, var 3 and var 4 refer to RMA, MAS 5.0, GCRMA and PLIER respectively.* 50

5.2 *d-statistic pairwise correlation coefficient plot from RMA, MAS5, GCRMA and PLIER. From left to right and from top to bottom, the corresponding orders of the graphs are RMA, MAS5, GCRMA and PLIER. Var 1, var 2, var 3 and var 4 refer to RMA, MAS 5.0, GCRMA and PLIER respectively.* 51

5.3 *q-values pairwise plot from RMA, dChip PM only and dChip PM-MM model. From left to right and from top to bottom, they follow the same order of RMA, dChip PM only and dChip PM-MM. Var 1, var 2 and var 3 refer to RMA, dChip PM and dChip PM-MM respectively.* 54

5.4 *d-statistic pairwise plot from RMA, dChip PM only, dChip PM-MM model. From left to right and from top to bottom, they follow the same order of RMA, dChip PM only and dChip PM-MM. Var 1, var 2 and var 3 refer to RMA, dChip PM and dChip PM-MM respectively.* 55

Chapter 1

Introduction

This chapter gives a brief introduction to the required background in genetics and microarrays. Section 1.2 provides information about Affymetrix GeneChip technology. Section 1.3 describes the data set we use.

1.1 Background in DNA and Microarrays

Proteins are the structural components of cells and tissues and perform many key functions of biological systems. The production of proteins is controlled by genes, which are coded in deoxyribonucleic acid (DNA), common to all cells in one being, and mostly static over one's lifetime. Protein production from genes involves two principal stages, known as transcription and translation, as illustrated in the schematic of Figure 1.1 (Parmigiani *et al.*, 2003).

A deoxyribonucleic acid (DNA) molecule is a double-stranded polymer composed of nucleotides. A nucleotide consists of a phosphate group, a sugar and one of four

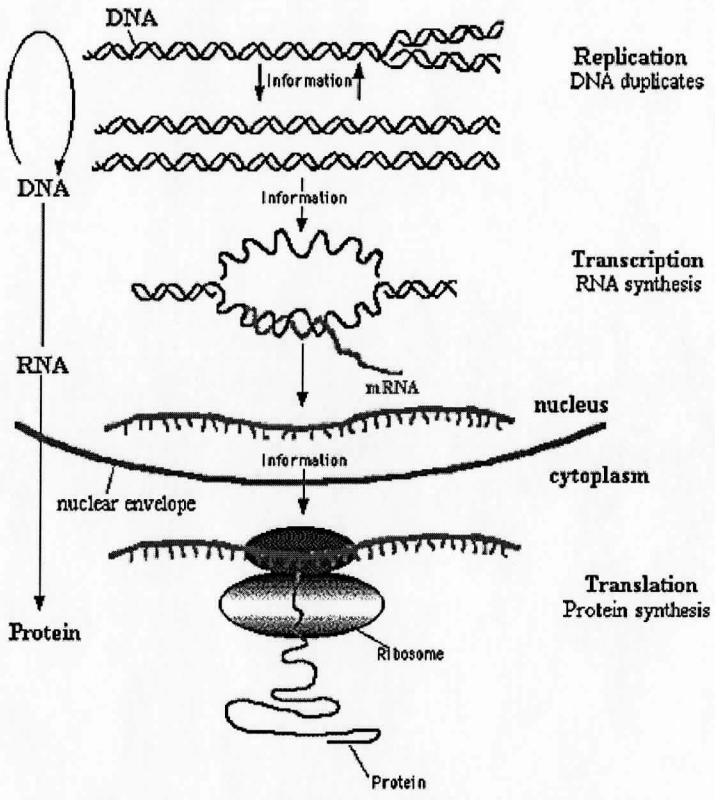

Figure 1.1: *A schematic of the role of RNA in gene expression and protein production. Graphics from http://www.accessexcellence.org*

nitrogen bases, which are purines composed of adenine (A) and guanine (G) and pyrimidines composed of cytosine (C) and thymine (T). Two strands are joined together by binding of the complementary bases to form a double helix structure. The binding rules are that A binds with T and G binds with C. These rules are also called Watson-Crick base pairings rules. As the complementary property, the nucleotide sequence of one strand determines the sequence of another strand. There are millions of nucleotides on DNA strands in length.

The two strands can be separated by heating. During transcription, single strands of mRNA are formed as spliced copies of the DNA segment coding a gene. After transcription, the mRNA is used as a template to assemble a chain of amino acids to form a protein. Gene expression investigations study the amount of transcribed mRNA in a biological system. Gene expression is a multi-step process that begins with transcription. In genetics, transcription is the first process in gene expression. In transcription, DNA is copied to RNA by an enzyme called RNA polymerase (RNAP). Transcription to yield an mRNA is the first step of protein biosynthesis.

Several techniques are available for measuring gene expression, including serial analysis of gene expression (SAGE), cDNA library sequencing, differential display, cDNA subtraction, multiplex quantitative RT-PCR, and gene expression microarrays.

DNA microarray is a new technology which differs from traditional methods in molecular biology that generally work on a "one gene in one experiment" basis, which means that the throughput is very limited and the "whole picture" of gene function is hard to obtain. Microarrays allow us to examine the whole genome on a single chip so that researchers can look at the interactions among thousands of genes simultaneously.

There are several microarray technologies. Currently, two approaches are prevalent: cDNA arrays and oligonucleotide arrays. Although they both exploit hybridization, they differ in how DNA sequences are laid on the array and in the length of these sequences.

In spotted DNA arrays, mRNA from two different biological samples is reverse-transcribed into cDNA, labeled with dyes of different colors, and hybridized to DNA sequences, each of which is spotted on a small region, or spot, on a glass slide. After hybridization, a laser scanner measure dye fluorescence of each color at a fine grid of pixels. Higher fluorescence indicates higher amounts of hybridized cDNA, which in turn indicates higher gene expression in the sample. A spot typically consists of a number of pixels. Image analysis algorithms either assign pixels to a spot or not and produce summaries of fluorescence at each spot as well as summaries of fluorescence in the surrounding unspotted areas (background).

For each location on the array, a typical output consists of at least four quantities, one of each color for both are spot and the background. Sometimes these are accompanied by measures of quality of the spot, to flag technical problems, or by measures of pixels intensity variability. The use of two channels allows for measurement of relative gene expression across two sources of cDNA, controlling for the amount of spotted DNA, which can be variable, as well as other experimental variation. This had led to emphasis on ratios of intensities at each spot. Although this ratio is critical, there is relevant information in all four of the quantities above.

The second commmon approach involves the use of high-density oligonucleotide arrays. This is an area of active technological development. The most widely used oligonucleotide array type is the Affymetrix GeneChip (for brevity Affy)(Parmigiani *et*

al., 2003). Our research focuses on the analysis of data from the Affymetrix technology.

1.2 Affymetrix Genechip Microarray

In this section, we introduce the terminology used to describe Affymetrix GeneChips.

Affymetrix GeneChip microarrays have become a crucial component of gene expression and genotype research for many laboratories. Affymetrix uses equipment similar to that which is used for making silicon chips for computers, and thus allows mass production of very large chips at reasonable cost. Where computer chips are made by creating masks that control a photolithographic process for removal or deposition of silicon material on the chip surface, Affymetrix uses masks to control synthesis of oligonucleotides on the surface of a chip, where an oligonucleotide is a molecule usually composed of 25 or fewer nucleotides, used as a DNA synthesis primer and usually called oligo. The masks controls the synthesis of several hundred thousand squares, each containing many copies of an oligo. So the result is several hundred thousand different oligos, each of them present in millions of copies (Knudsen, 2004).

In order to produce a GeneChip array, we need to know the sequence of the target organism. When given a known sequence, a number of 25 base sequences complementary to the sequence for target genes are chosen. These sequences are known as probes. The collection of probes is called a probe set. Affymetrix microarrays use a probe set consisting of 11-20 probe pairs to represent a gene. Sometimes there is more than one probe set that correspond to the same gene, but each uses a different part of the sequence. Each Affymetrix probe pair consists of a perfect match (PM) and a mismatch (MM) oligonucleotide. Perfect Match (PM) is defined as a probe that exactly comple-

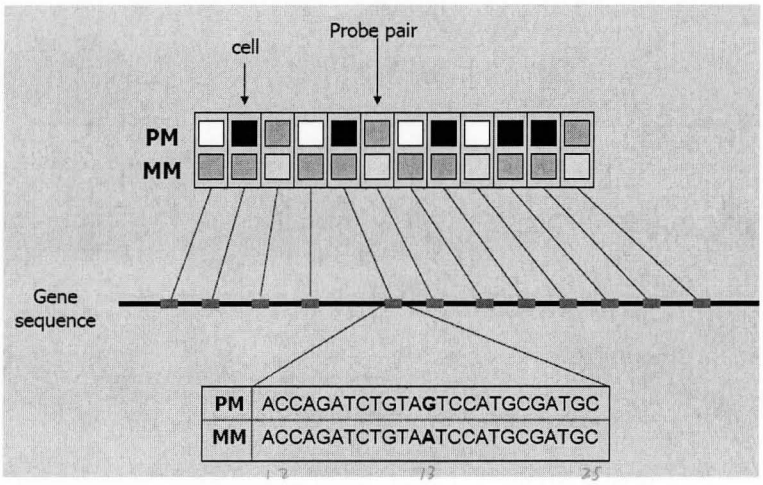

Figure 1.2: *Perfect Match and Mismatch Probes. Figure came from the slides of Dr. Roger Bumgarner at University of Washington*

mentary to the sequence of interest in a probe set. PM and Mismatch (MM) probes are identical in sequence except for the middle (13th) base position. MM probes are used to detect nonspecific binding, where nonspecific binding is characterized by the reversal of the central Watson-Crick pairing for each PM/MM probe pair, whereas specific binding refers to the combination of a WC and a self-complementary (SC) pairing in PM and MM probes, respectively (Binder and Preibisch , 2005). Examples of PM and MM probes are given in Figure 1.2.

Affymetrix provides various chip types, for example, the human HGU95 and mouse MGU74 Chip are two widely used. One can hybridize various arrays of a specific chip type. For each of them, millions molecules of a particular probe are attached to a 400 μm^2 area on the chip. After processing the raw image produced by the Affymetrix scanner, each probe is represented by about 100 pixels at a specific location of the image. At the final stage, the image processing-software stores the location and two summary statistics, a mean and standard deviation, for each probe in a file denoted

with the extension CEL (Parmigiani *et al.*, 2003).

After generating the image data to a CEL file, one needs to process the data to produce expression values. The PM and MM probe intensities for each probe set are combined together to produce a summary value. Originally, this was done using the average difference (AvDiff) algorithm (Affymetrix, 1999). But several drawbacks do exist. Afterwards, MAS 5.0 (Affymetrix, 2001), Model Based Expression Index (MBEI) (Li and Wong, 2001a) and the Robust Multi-chip Average (RMA) (Irizarry *et al.*, 2003a, 2003b) and Probe Logarithmic Intensity Error (PLIER) (Affymetrix, 2005) were developed to compute expression measurements. More details about those methods will be introduced in Chapter 2 and followed by comparisons of these methods in Chapter 3, 4 and 5.

1.3 Applications of DNA Microarray Technology

Microarrays can be applied to the problems of gene discovery, the diagnosis of diseases, pharmacogenomics and toxicogenomics among others. Gene discovery is the process of finding genes that are differentially expressed between tissues from different conditions. When given expression profiles for a diseased and non-diseased tissues, a new sample can be diagnosed by measuring its expression profile and comparing it with the reference profiles. Pharmacogenomics is the process of discovering how a therapeutic response from a drug affects the expression profile of a patient (Regalado, 1999). More specifically, pharmacogenomics seeks to answer such questions as: Why does a drug work better in some patients and not others? Why is a drug toxic for some people? Toxicogenomics is the study of how exposure to toxicants affects the genetic profiles

of the exposed tissues. See Nuwaysir *et al.* (1999) and Bolstad (2004).

1.4 Description of Data

A number of genetic regions which are called loci play a important role in Type 1 Diabetes susceptibility. We have two parental strains of mice Non-Obese Resistant (NOR) and Non-Obese Diabetic (NOD). These are identical by descent in 88% of the genome but NOD mice get Type 1 Diabetes at much higher rates than NOR mice in which $82 - 85\%$ compared to $3 - 5\%$ by age 6 months.

NOR.NOD_Idd4, can be constructed through selective multi-generational in-breeding or other genetically engineered of these mice. NOR.NOD_Idd4 congenic strain is identical to the parental NOR except in region IDD4 which it inherits from the NOD mice. In our case, the Double Congenic NOR.NOD_Idd4/11 strain is identical to the parental NOR except in regions Idd4 and Idd11 which it inherits from the NOD mice.

The experiment is set up by Affymetrix MGU74Av2 chip with 12488 probe sets. DNA is taken from activated bone-marrow derived macrophages of two strains of mice that are NOR mouse and NOR.NOD_Idd4/11 mouse.

The data is processed on two different days to obtain 9 arrays. On day 1, 5 replicates that include 2 NOR strains and 3 NOR.NOD_Idd4/11 strains are obtained. On day 2, 4 replicates that include 2 NOR strains and 2 NOR.NOD_Idd4/11 strains are obtained.

Chapter 2

Preprocessing Stages and Methods

The preprocessing procedure of a microarray data analysis starts with CEL files as described in section 1.2. A three-stage procedure, background correction, normalization and summarization, is used to produce an expression measurement. At each stage of the procedure, numerous methods have been proposed for GeneChip arrays. The result of preprocessing is the expression measurements/ expression values for each chip.

In this chapter, we discuss the purpose of each stage and the different methods used for each of three stages. In the last section of this chapter, we introduce several commonly used methods for computing the expression measurements in the recent research. The examination of the behavior of different methods, in the next chapter, is mainly based on the content in this chapter.

2.1 Background Correction

The term background correction is also referred to as signal adjustment. The scanning of arrays results in optical and background noise affecting pixel intensities. On some array images, a slight signal is seen in the area that is in between spots. Therefore, background information is difficult to obtain. Numerous background correction methods are proposed. For example, the image processing software will produce an absolute expression X and a background measurement B for each spot or cell, where X is the result of signal and additional background noise, then it is a biased estimate of the true hybridization we are going to measure. One of the way to obtain an unbiased expressions is to subtract the background, i.e. consider $X - B$. The other way is just to use X to estimate the expression level. See details in Parmigiani *et al.* (2003). Commonly used methods include the RMA and MAS algorithms. The RMA convolution for the background correction will be introduced in this section. The MAS 5.0 algorithm will be discussed in Section 2.4.

2.1.1 RMA Convolution

Irizarry *et al.* (2003a) find that the $PM - MM$ transformation results in expression estimates with exaggerated variance. They propose a background adjustment step that ignores the MM intensities. This approach sacrifices some accuracy for large gains in precision. The resulting algorithm, the robust multi-array analysis (RMA), has become a popular alternative to the MAS algorithm provided by Affymetrix.

The RMA convolution model is suggested by looking at plots of the empirical distribution of probe intensities. The method is to model the observed PM probes as

the sum of a signal and a background component. In particular, the model is that we observe $S = X + Y$, where X is signal and Y is background. Assume that X is distributed $\exp(\alpha)$ and that Y is distributed $N(\mu, \sigma^2)$, with X and Y independent. In order to avoid any possibility of negative values, assume that $Y \geq 0$. Thus, Y is normally distributed with truncation at 0. Given we have S the observed intensity, this then leads to an adjustment.

$$E(X|S = s) = a + b\frac{\phi(\frac{a}{b}) - \phi(\frac{s-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{s-a}{b}) - 1}$$

where $a = s - \mu - \sigma^2\alpha$ and $b = \sigma$. Note that ϕ and Φ are the standard normal distribution density and distribution functions respectively. See Bolstad (2004) for details about the derived quantity.

2.2 Normalization

In many applications, the goal in analyzing the gene expression is to learn how RNA populations differ in expression in response to genetic and environmental differences. As defined in Hartemink *et al.* (2001), the sources of variation can be classified as interesting variation and obscuring variation. Interesting variation refers to how cells variously express their different genes in response to the diverse genetic and environmental environments they encounter. Sources of obscuring variation include variation introduced during the process of sample preparation, during the manufacture of the array, during the hybridization of the sample on the array, and during the scanning and analysis of fluorescent intensity after hybridization. The obscuring sources of variation can have many different effects on data, unless arrays are appropriately nor-

malized. Various methods have been proposed for normalizing GeneChip arrays, such as constant normalization, contrasts normalization, invariant set normalization, loess normalization, qspline and quantile normalization. The quantile, and invariant set normalization methods will be discussed as follows.

2.2.1 Quantile Normalization

The quantile normalization method provides a fast method to normalize multiple chips within a set of chips, provided one is willing to assume a common distribution, i.e., to give each chip the same empirical distribution.

Let X be a matrix of probe intensities (probes by arrays). Given n arrays of length p, form X of dimension $p \times n$. Sort each column of X to give X_{sort}. Take the means across rows of X_{sort} and assign this mean to each element in the row to get X'_{sort}. Get $X_{normalized}$ by rearranging each column of X'_{sort} to have the same ordering as original X. The quantile normalization method is a specific case of the transformation $x'_i = F^{-1}(G(x_i))$, where we estimate G by the empirical distribution of each array and F using the empirical distribution of the averaged sample quantiles. More details can be found in Bolstad (2004).

2.2.2 Invariant Set Normalization

Array images usually have different overall image brightness, especially when they are generated at different times and places. The main idea of invariant set normalization is that, for a group of arrays, we need to normalize all arrays to a common baseline array having the median overall brightness, then using them to fit a non-linear relationship

between the "treatment" and "baseline" arrays. The non-linear relationship is used to carry out the normalization (Li and Wong, 2001a). Dchip software developed by Li and Wong (2001b) uses the invariant set normalization method as the normalization method.

2.3 Summarization

Summarization is the last step in the production of a gene expression measurement. As introduced in Chapter 1, each Affymetrix GeneChip microarrays probe set is composed of probes. Within a probe set, each probe accounts for a different part of the sequence for a particular gene. Summarization is the process that combine the multiple probe intensities for each probe set to generate an expression measurement. Commonly discussed summarization methods include Average Difference (Avgdiff) summarization, median polish summarization, MAS summarization, Li and Wong summarization and playerout summarization. Only those methods related to our method comparison will be discussed in this chapter.

2.3.1 Average Difference Summarization

AvDiff is the most commonly used method. For each probe set n on each array i, AvDiff is defined by

$$\text{AvDiff} = \frac{1}{\#A} \sum_{j \in A} (PM_j - MM_j)$$

with A the subset of probes for which $d_j = PM_j - MM_j$ are within 3 SDs away from the average of $d_{(2)}, \ldots, d_{(J-1)}$ with $d_{(j)}$ the j-th smallest difference. $\#A$ represents

the cardinality of A. Many of the other expression measures are versions of AvDiff with different ways of removing outliers and different ways of dealing with small values (Irizarry *et al.*, 2003b).

2.3.2 Median Polish Summarization

This is the summarization method used in the RMA expression summary. A multichip linear model is fitted to data from each probeset. In particular for a probeset k with $i = 1, \ldots, I_k$ probes and data from $j = 1, \ldots, J$ arrays we fit the following model $\log_2(PM_{ij}^{(k)}) = \alpha_i^{(k)} + \beta_j^{(k)} + \epsilon_{ij}^{(k)}$, where α_i is a probe effect and β_j is the $\log_2$ expression value (Bolstad, 2005). Median polish is a data analysis technique (more robust than ANOVA) for examining the significance of the various factors in a multifactor model (Tukey, 1977). The expression values we get using this summary measure will be in $\log_2$ scale.

2.4 Commonly Used Methods for the Preprocessing

2.4.1 RMA

RMA is a popular algorithm that was implemented in Bioconductor to calculate the expression measurements by Affymetrix. It integrate RMA convolution model as the background correction, quantile normalization and median polish summarization. The description of each stage's methods can be looked at section 2.1-2.3.

2.4.2 MAS 5.0

The Affymetirx MAS 5.0 algorithm (Affymetrix, 2002) uses a method to calculate signal value that comes from the combined, background-adjusted, PM and MM values of the probe set.

In order to remove background noise (background correction step), the chip is broken into a grid of 16 rectangular regions. For each region the lowest 2% of probe intensities are used to compute a background value for that grid. Each probe is then adjusted based upon a weighted average of the backgrounds for each of the regions. The weights are based on the distances between the location of the probe and the center of 16 different zones, denoted by d. A weighted sum is then calculated based on the reciprocal of a constant plus the square of the distances to all the zone centers. If the distance d between the chip coordinate (x, y) and the center of the k-th zone is d_k, a weighting factor can be calculated, which is d^2. A small factor, is added to d^2 to ensure that the value will never be zero.

In MAS 4.0, Affymetrix attempt to use the transformation $PM - MM$ to adjust for non-specific binding and background noise. But in general, $MM \geq PM$ for about 1/3 of the probes on any given array (Irizarry *et al.* , 2003a) which results in negative adjusted intensity values. Thus, when raw MM intensities are subtracted from the PM intensities it is possible to compute negative expression values. Additionally, since the use of logarithms has proven useful in microarray data analysis, the negative values cause problems in the analysis. To solve the problem of negative values using raw MM values, Affymetrix introduced the Ideal Mismatch (IM) (Affymetrix, 2002) in which the adjusted PM intensity was guaranteed to be positive.

To calculate a specific background ratio representative for the probe set, the one-step biweight algorithm (T_{bi}) is being used. The log ratio of PM to MM is simply an estimate of the difference of log intensities for a selected probe set. The biweight specific background (SB) for probe pair j in probe set i is:

$$SB_i = T_{bi}(\log_2(PM_{i,j}) - \log_2(MM_{i,j}) : j = 1, \ldots, n_i).$$

If SB_i is large, then the values from the probe set are generally reliable, and we can use SB_i to construct the ideal mismatch IM for a probe pair if needed. If SB_i is small, we smoothly degrade to use more of the PM value as the ideal mismatch. In general, when $MM < PM$, $ID = MM$; when MM$\leq$PM, IM is equal to an adjusted MM. For details see Affymetrix (2002).

Then compute the absolute expression value for probe set i as the one-step biweight estimate of the i n adjusted probe values:

$$SignalLogValue_i = T_{bi}(PV_{i,1}, \ldots, PV_{i,n_i}),$$

where define that the probe value PV for every probe pair j in probe set i, n is the number of probe pairs in the probe set and

$$PV_{i,j} = \log_2(V_{i,j}), j = 1, \ldots, n_i$$

and $V_{i,j} = \max(PM_{i,j} - IM_{i,j}, d)$, where d is defined before. For further details of MAS 5.0 see Affymetrix (2002).

2.4.3 GCRMA

In R, the function `gcrma()` converts background adjusted probe intensities to expression measures using the same normalization and summarization methods as `rma()` (Ro-

bust Multiarray Average). GCRMA adjusts for background intensities in Affymetrix array data which include optical noise and non-specific binding (NSB) using probe sequence information to estimate probe affinity to non-specific binding (NSB).

Naef and Magnasco (2003) introduced the idea of probe affinity. This refers to the fact that a G or C nucleotide leads to stronger hybridization because each G-C pair forms three hydrogen bonds whereas each A-T pair forms two and they propose a solution useful for predicting specific hybridization effects with base composition of the probes. Probe affinity is modeled as a sum of position-dependent base effects:

$$\alpha = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k=j}$$

with

$$\mu_{j,k} = \sum_{l=0}^{3} \beta_{j,l} k^l,$$

where $k = 1, \ldots, 25$ indicates the position along the probe, j indicates the base letter, b_k represents the base at position k, $1_{b_k=j}$ is an indicator function that is 1 when the k-th base is of type j and 0 otherwise, and $\mu_{j,k}$ represents the contribution to affinity of base j in position k. For fixed j, the effect $\mu_{j,k}$ is assumed to be a polynomial of degree 3. The model is fitted to log intensities from many arrays using least squares (Naef and Magnasco, 2003).

In GCRMA procedure, probe affinity model is fit to get affinity estimates to describe nonspecific binding (NSB) noise. The affinities predict NSB quite well, almost as well as the MM intensities. The advantage of the affinities over the MM is that they will not detect signal since they are pre-computed numbers (Wu *et al.*, 2004).

After the affinity estimation, background values will be estimated with either a maximum likelihood estimate or an Empirical Bayes estimate. Thus, use the same

normalization and summarization methods as RMA to get resulting GCRMA expression measurements.

2.4.4 PLIER

PLIER (Probe Logarithmic Intensity Error) is a new algorithm developed by Affymetrix. This method produces an improved signal (a summary value for a probe set) by accounting for experimentally observed patterns in probe behavior and handling error appropriately at low and high signal values. Resulting benefits include: Higher reproducibility of signal (lower coefficient of variation) without loss of accuracy; Higher sensitivity to changes in abundance for targets near background; Dynamic weighting of the most informative probes in an experiment to determine signal (Affymetrix, 2005).

Similar to other model-based approaches, PLIER accounts for the systematic differences in intensity between features by including parameters describing these differences. These parameters are termed "feature responses" (also called affinity in the literature) and one such parameter is included in the model for each feature (or pair of features, when subtracting Mismatch (MM) intensities). Feature responses represent the relative differences in intensity between features hybridizing to a common target (Affymetrix, 2005).

PLIER is an M-estimator model-based framework for finding expression estimates that is designed to handle near-background probe intensities well with minimal positive bias to the results. While the estimates from PLIER are by design not variance stabilized, PLIER shows good performance at detecting differential change, and can be variance stabilized by standard means. M-estimators form a very flexible framework

for analysis. It can handle PM-B, PM-MM, PM-only approaches in same framework and handle zero/near-zero concentration & affinities in model directly. See details about M-estimators in Huber (1981).

2.4.5 DChip (Li and Wong)

This is an implementation of the methods proposed in Li and Wong (2001a) and Li and Wong (2001b). DChip use invariant set normalization described in section 2.2.2. After normalization, the Li and Wong's model-based expression index (MBEI) is based upon fitting the multi-chip model to each probe set to compute the expression level of each gene in all samples. The statistical model is:

$$y_{ij} = \theta_i \phi_j + \epsilon_{ij}, \sum \phi_j^2 = J, \epsilon_{ij} \sim N(0, \sigma^2),$$

where y_{ij} can be PM_{ij} or the difference between $PM_{ij} - MM_{ij}$. The ϕ_j parameter is a probe-sensitivity index of probe j, θ_i is an expression index for the gene in the ith sample, J is the number of probe pairs in the probe set and PM_{ij} and MM_{ij} denote the PM and MM intensity values for the ith array and the jth probe pair for this gene. Fitting the model, they identify cross-hybridizing probes and arrays with image contamination at this probe set as well as single outliers, which are replaced by the fitted values. The estimated expression index $\hat{\theta}_i$ is a weighted average of PM-MM differences, $\hat{\theta}_i = (\sum_j p_{ij}\phi_j)/J$, with larger weights given to probes with larger ϕ (Li and Wong, 2001a).

Chapter 3

Comparisons Based on Expression Measurements

In this chapter, we discuss the comparison of expression measurements between different methods in terms of correlation coefficient and the pairwise plots. The explanation of the plots induce some special techniques during the preprocessing.

3.1 Correlation Coefficient of Expression Measurements

As described in Chapter 2, there are a number of methods at each preprocessing stage. We can combine different methods at different stages of preprocessing to calculate the expression measures for each probe set. It is important that not every preprocessing method can be combined together. In particular the RMA method background adjusts

only PM probe intensities and so should only be used in conjunction with the pmonly PM correction. Also remember that the mas and median polish summarization methods log2 transform the data, thus they should not be used in connection with any preprocessing steps that are likely to yield negative values like the subtractmm correction method. Therefore, all the expression measures of our comparison are log2 transformed. Furthermore, due to possible negative values produced by the subtractmm PM correction, we could not include subtractmm as PM correction method in the combinations. With those limitations on combinations of preprocessing methods, we finally obtain 53 different methods for our further discussion. See Appendix A for the definition of all the methods.

In Table A.1 at Appendix A , Methods 1 to 7 use the same background correction method, RMA, the same PM correction and median polish summarization, but 7 different normalization methods. The only step being varied is the normalization method. However, when we check the correlation coefficients for different normalization methods which corresponds to methods M1-M7, we notice that all the correlation coefficients are greater than 0.99. A similar situation happens to the methods M8-M14. If we define M8-M14, M15-M21, and so on, as a subgroup, within each subgroup, correlation coefficients between any two methods are all greater than 0.99, i.e. expression values between any two methods are highly correlated. Therefore, we can conclude that relative position of the gene across all the arrays are similar.

Due to the small influence of normalization among the preprocessing, we fix the normalization method as quantile and vary the other preprocessing steps to produce Table 3.1, then check the other factors' influence.

Table 3.2 shows the pairwise correlation coefficients. We can make some conclusions

Method	Bkg Correction	Normalization	PM Correction	Summarization
M1	rma	quantiles	pm only	median polish
M8	none	quantiles	pm only	median polish
M22	none	quantiles	pm only	avgdiff
M36	none	quantiles	pm only	mas
M15	none	quantiles	mas	median polish
M29	none	quantiles	mas	avgdiff
M43	none	quantiles	mas	mas
M52	mas	none	mas	mas

Table 3.1: *Reduced preprocessing methods list*

according to this comparison as follows:

- M15, M29 and M43 are similar except that they use different summarization methods, but they have pairwise correlation coefficients close to 1. M8, M22 and M36 are in a similar situation. We conclude that the different summarization methods results in relative expression measures similar.

- If we only vary pm correction and other steps are kept the same, we can see that correlations are all smaller than .90, such as the relationships between M8 and M15, M22 and M29, and M36 and M43. We conjecture that pm correction factor is the most important part among the preprocessing. On the other hand, if we fix pm correction, varying other step methods, the correlations are also no less than .98, such as M15 and M52, M29 and M52, and M43 and M52.

Now let us add the Li and Wong method to the comparison as shown in Table 3.3.

R	M8	M22	M36	M15	M29	M43	M52	GCRMA
M1	.9686	.9652	.9631	.8963	.8928	.8941	.887	.6256
M8		.9972	.9942	.8916	.8901	.8906	.8731	.6919
M22			.9959	.8899	.8899	.8903	.8728	.6901
M36				.8912	.8911	.893	.8751	.6903
M15					.9935	.9927	.9806	.7514
M29						.9975	.9852	.7521
M43							.9875	.7503
M52								.7225

Table 3.2: *Pairwise correlation coefficients between different preprocessing methods*

	M1	M4	M11	M18	M25
dChip-PM	0.9191	0.9194	0.8855	0.8479	0.8061
dChip-PM/MM	-0.0252	-0.0251	-0.0274	-0.0261	-0.0208

Table 3.3: *Correlation coefficients comparison for the Li and Wong*

In order to examine the conclusion we get above, we choose method M1, M14, M11, M18 and M25 to compare with dChip-MM and dChip-PM/MM. Unfortunately, we can't verify the conclusion as we did between RMA, MAS 5.0, GCRMA and PLIER. There even exists some negative correlations between dChip-PM/MM and the other methods in Table 3.3. But we can't find a proper way to explain at this point of time.

At the end of this chapter, we will compare our conclusions with those in the literature. Our conclusions are based on the comparison of pairwise correlation coefficients of the methods at expression measurement level. But in Irizarry & Wu (2005), they made the conclusion that background correction has the largest effect on performance.

	Bkg Correction	Normalization	PM Correction	Summarization
M1	RMA	Quantile	PM Only	Medianpolish
M4	RMA	Invariantset	PM Only	Medianpolish
M11	None	Invariantset	PM Only	Medianpolish
dChip-PM	None	Invariantset	PM Only	MBEI
M18	None	Invariantset	MAS	Medianpolish
M25	None	Invariantset	PM only	Avgdiff
dChip-PM/MM	None	Invariantset	Subtractmm	MBEI

Table 3.4: *Methods Description for comparing with dChip*

Their conclusion is set up by Affycomp II that is a graphic tool for evaluating and comparing of expression measures of the Affymetrix GeneChip. More details can be found in Cope *et al.* (2004).

In their research, the assessments evaluate performance in terms of bias (lack of accuracy) and variance (precision). The different comparison guideline leads to one of the reasons of different conclusions. On the other hand, the key of the Affycomp is that a benchmark data set comprises a dilution data set prepared by Gene Logic and a spike-in data set prepared by Affymetrix. Although a number of the conclusions are based on the spike-in and dilution data set, we have to say that these two data set are "over-training". The spike-in study by Affymetrix is a subset of the data used to develop and validate the MAS 5.0 algorithm. In particular, the conclusion about background correction made by Irizarry & Wu (2005) only use those two data sets and no other data sets are applied to their research and nobody could confirm their conclusion using the other data sets except spike-in and dilution. However, we use

the different data set which is a "real" one. This results in the second reason for the different conclusion.

3.2 Pairwise Plots of Expression Measurements

In this section, we focus on the comparison between RMA, MAS 5.0 and GCRMA by pairwise plots of expression values to find the relationships between 3 popular normalization methods.

When we plot the original pairwise expressions plot between RMA, MAS 5.0 and GCRMA, we can't find a clear pattern in the figure between any two methods (as shown in Figure 3.1(A) and the graphs in the first column of Figure 3.2). Therefore, we try to implement MAS 5.0 present/absent calls and affinity attribute technique to make more clear shape of those figures and explain the pairwise plots, as described in section 3.2.1 and section 3.2.2.

3.2.1 MAS 5.0 Present/Absent Calls

In Figure 3.1, the first plot is the original RMA against MAS 5.0 expression measurements performed on the same data set. We notice that there are some negative expression measures in MAS 5.0 method. It means that within all the arrays, the signals are very near to the background in some probe sets, i.e. the values of perfect match and mismatch are close for some genes in our data set. Now we use MAS 5.0 absent call algorithm to remove those probe sets and check what will happen.

MAS 5.0 assigns to each probe set an expression detection call. This detection call

is determined by a "detection p value" generating by the function, indicating whether a transcript of a particular gene is detected (present) or not (absent or marginal), i.e. when the expression level is below the threshold of detection, we called it absent, otherwise called present. For more details, see Affymetrix (2002). In this way, we may only look at genes whose transcripts are detectable and check the relationship with RMA or GCRMA.

For a given probe set, there are two main steps to calculate the "detection p-value":

1. Calculate the discrimination scores. For the i-th probe pair, the discrimination score is defined as:

$$R_i = \frac{PM_i - MM_i}{PM_i + MM_i}.$$

It measures the intensity difference of the probe pair $(PM_i - MM_i)$ relative to its overall hybridization intensity$(PM_i + MM_i)$. Let

$$r = \frac{PM + MM}{PM - MM}$$

and because of the relationship between the discrimination scores and the log2 ratio used in the Specific Background (SB) calculation, we have

$$\log_2 \frac{1+r}{1-r} = \log_2 \frac{1 + \frac{PM+MM}{PM-MM}}{1 - \frac{PM+MM}{PM-MM}} = \log_2 \frac{2 * PM}{2 * MM} = \log_2(PM) - \log_2(MM).$$

This tells us how different the PM and MM cells are.

Then we compare the discrimination score R_i with a user-definable the threshold τ, where τ is a small threshold between 0 and 1 and by default $\tau = 0.015$. We can increase or reduce the number of the detected calls by adjusting the value of τ. Increasing the threshold τ can reduce the number of false detected calls, but may also reduce the number of true detected calls (Affymetrix, 2002).

2. Use one-sided Wilcoxon rank test to calculate detection p-value. It assigns each probe pair a rank based on how far the probe pair discrimination score is from τ.

The hypothesis of this test is:

$$H_0 : \text{median}(R_i) - \tau = 0, \text{corresponding to absence of transcript}$$

$$H_1 : \text{median}(R_i) - \tau > 0, \text{corresponding to presence of transcript}$$

If $\text{median}(R_i) - \tau > 0$, we can reject the hypothesis that PM and MM are equally hybridizing to the sample. We can make a detection call based on the strength of this rejection. More details can be found in Affymetrix (2002).

By implementing MAS 5.0 present/absent calls, we remove the genes never present across all 9 arrays to get 6728 present genes out of 12488 genes, then generate a subset of the original data set. The genes in the subset present at least once in the 9 arrays. We call these 6728 genes "good genes" temporarily. We use this subset data set to plot RMA expressions vs. MAS 5.0 expressions as shown in Figure 3.1(B).

In the Figure 3.1(B), the expression measurement plot shows relative more clear shape than Figure 3.1(A) but still some negative expression values exist in MAS 5.0. The correlation coefficient improves from .7853 to .9135.

Since for each "good gene" in the subset data set, it might be present in array 1, for example, but absent at the other arrays. So for each gene in the subset, it might be present or absent across 9 arrays. Now we only plot the expressions for which the detection algorithm declares them present to get Figure 3.1(C). It appears linear shape, and now RMA and MAS 5.0 are correlated with $r = .9331$.

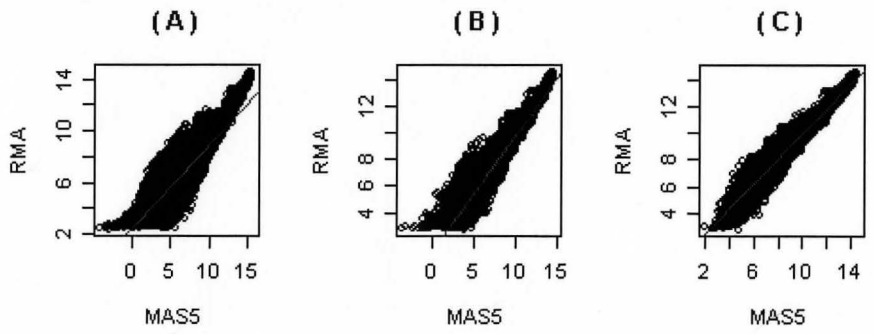

Figure 3.1: *Expression measurements comparison between RMA and MAS 5.0. (A) Original expression measures for RMA vs. RMA; (B) RMA vs. GCRMA expression measurements for those genes that present at once ("good genes" subset); (C) RMA vs. GCRMA expression measurements for which the detection algorithm declares them present.*

Therefore, as for the comparison between RMA and MAS 5.0, the correlation coefficient between them increases when we use MAS 5.0 detection calls algorithm. The correlation coefficients, from left to right, are .7853, .9135 and .9331, respectively.

Owing to the benefit we obtain by applying MAS 5.0 detection call algorithm, we use obtained subset to calculate the expressions for GCRMA and then plot the expression values of GCRMA vs. RMA and GCRMA vs. MAS 5.0 with those "good genes" to investigate the shape of graph.

The graph in column 1 of Figure 3.2 are the original plot RMA vs. GCRMA and MAS 5.0 vs. GCRMA. It is similar to the comparison in Figure 3.1 that the graph in column 2 is the comparison RMA against GCRMA and MAS 5.0 against GCRMA for "good genes" with all expression values whatever they are present, absent or marginal. In column 3, the graph is the comparison of expression values for "good genes" and

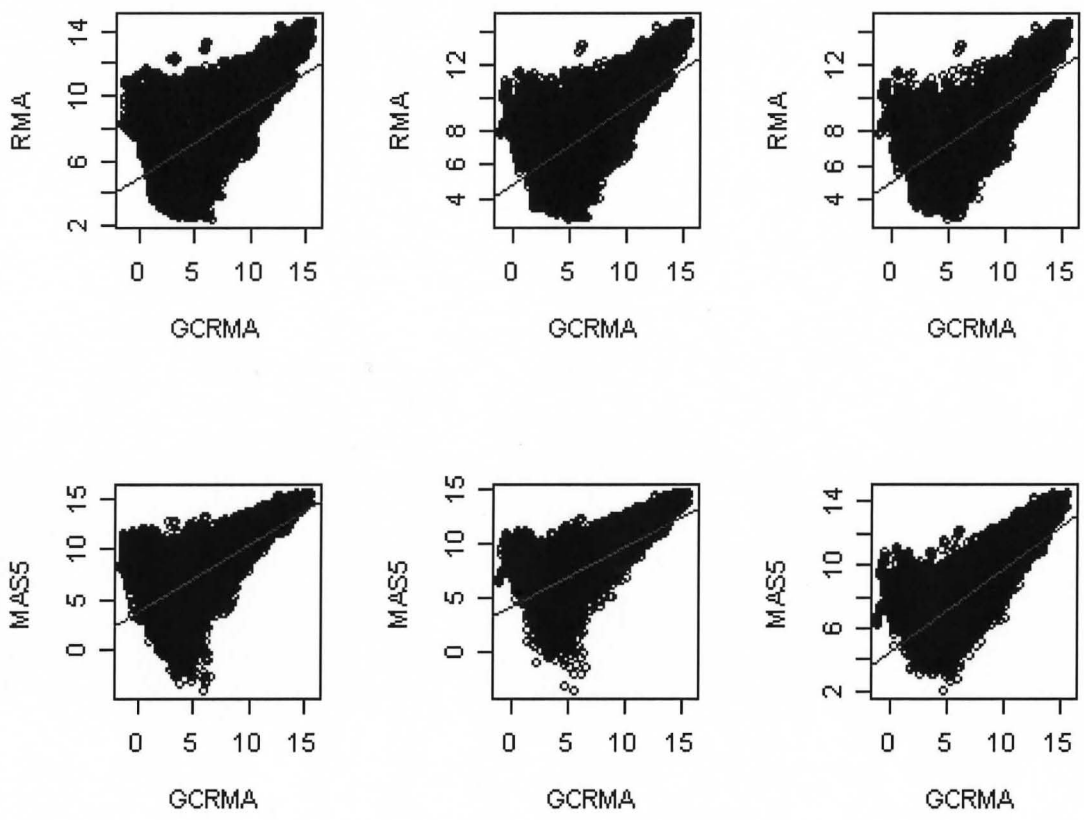

Figure 3.2: *Expression measurements comparison between GCRMA and RMA, and GCRMA and MAS 5.0.*

only when present.

From Figure 3.2, we see that MAS 5.0 Absolute Detection algorithm does not influence GCRMA too much, only removes some outliers and make the graphs look smoother. However, it is a fact that the use of subsetting by MAS 5.0 detection calls improves the correlation coefficient among the different methods. The correlation coefficients between RMA and GCRMA, according to the graph order, from left to right in Figure 3.2, vary from .3913, .5013 to .5114. Similarly, in the comparison of MAS 5.0 vs. GCRMA, the correlation coefficients change from .5213, .6116 to .6293. We also notice that MAS 5.0 shows higher correlation with GCRMA than does RMA.

3.2.2 RMA vs. GCRMA based on the affinity

In Section 3.2.1, we find that the correlation coefficient between RMA and GCRMA is as low as .3913, and the graph shows no clear pattern between them. The graph pattern has no improvement though we apply MAS 5.0 Detection algorithm to GCRMA. As for the GCRMA method, we conjecture that whether there is a method similar to MAS 5.0 Detection algorithm that we could explain the shape of the graph.

As described in Chapter 2, GCRMA uses probe sequence information to estimate probe affinity to non-specific binding (NSB). Therefore, we could use affinity information to separate the probe set with high, medium and low affinity and then look at the relationship.

The affinity of a probe is described as the sum of position-dependent base affinities. Each base at each position contributes to the total affinity of a probe in an additive

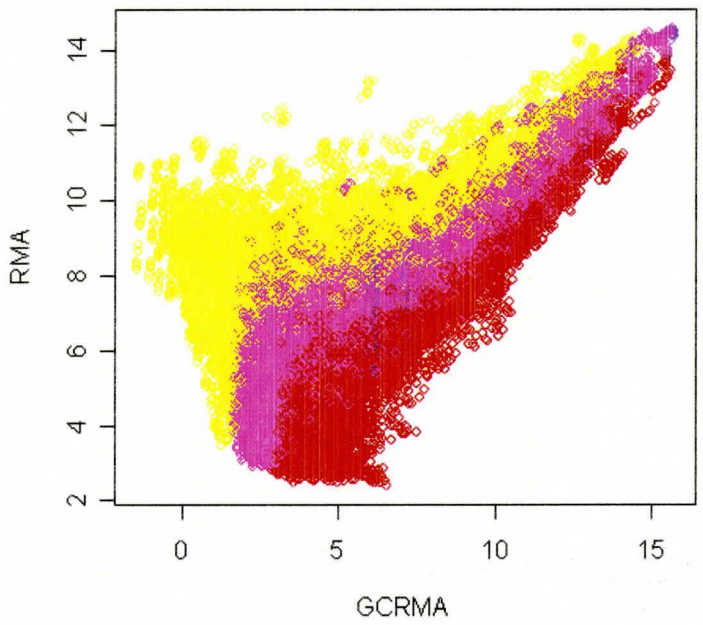

Figure 3.3: *Expressions of RMA vs. GCRMA based on the affinity. Red=low, blue=medium and yellow=high.*

fashion. For a given type of base, the positional effect is modeled as a spline function with 5 degrees of freedom. Note that one type of GeneChip array has unique affinity.

We make use of this affinity information to find the median of the PM affinities for each probe set, then discretize affinities into three groups (high, medium and low) based on the quantiles of the affinities. Define that $> 75th$ %-ile = high, $< 25th$ %-ile = low and everything else is medium. Then plot (Figure 3.3) GCRMA against RMA with different colors for these three groups.

By the analysis of the different three groups as Figure 3.4, we find that, for those genes with low affinity, the relationship between RMA and GCRMA seems to be linear

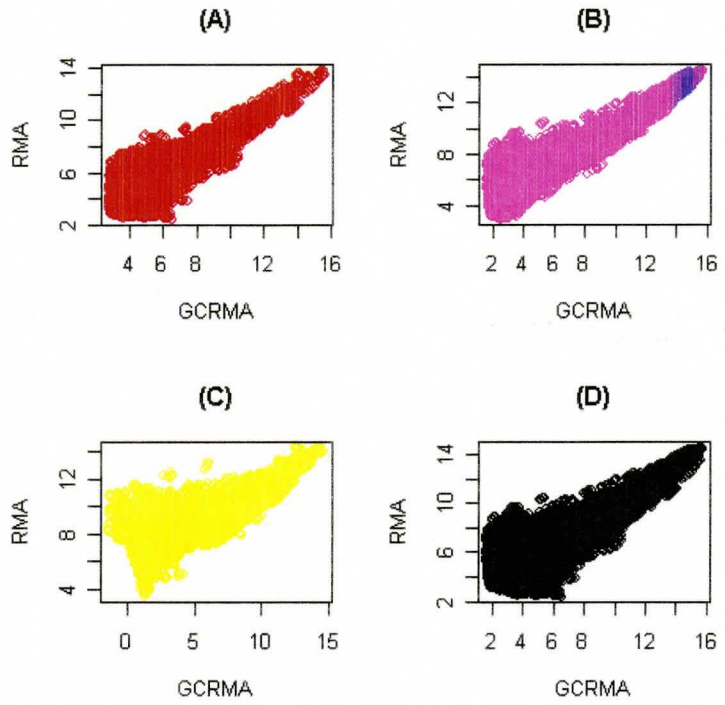

Figure 3.4: *Expressions of RMA vs. GCRMA based on the different levels of affinity. (A) RMA vs. RMA for low affinity; (B) RMA vs. RMA for medium affinity; (C) RMA vs. RMA for high affinity; (D) RMA vs. RMA for low and medium affinity;*

as shown in Figure 3.4 (A). But genes with medium or high affinity, especially in high affinity, show nonlinear relationship at all.

The correlation coefficients in Figure 3.4 from (A) to (D) are .7705, .7476, .5344 and .6152 respectively. Especially, Figure 3.4 (A) is the comparison between RMA and GCRMA for the low affinity, but the correlation at this situation is the highest one. On the other hand, Figure 3.4(C) is the comparison between RMA and GCRMA for the high affinity, but the correlation at this situation is the lowest one. We say that high affinity corresponds to lower correlations between RMA and GCRMA method, and low

affinity corresponds to higher correlation between RMA and GCRMA method. And RMA and GCRMA expression measures for high signal genes always appear linear pattern, but those low signal genes do not.

From the analysis of expression histograms of RMA and GCRMA, we find that the distributions of expressions between RMA and GCRMA normalization make big difference in which the expression of RMA tends to be normally distributed but GCRMA shows right skewed non-systematic distribution. In Figure 3.5(B), the expression of RMA with high affinity distributed approximately normal.

After comparison between RMA and GCRMA, we conclude that they do not show any linear relationship even by the classification of the different affinity information, or even though the algorithm to RMA and GCRMA are very similar. In addition, as reported in Section 3.2.1, the correlation coefficients between GCRMA and MAS 5.0 or GCRMA and RMA are very low, but why MAS 5.0 shows higher correlation with GCRMA than does RMA is an interesting question. It seems GCRMA has higher correlation with MAS 5.0 than with RMA. Therefore, the more study to explore the true relationship between RMA and GCRMA need to be done.

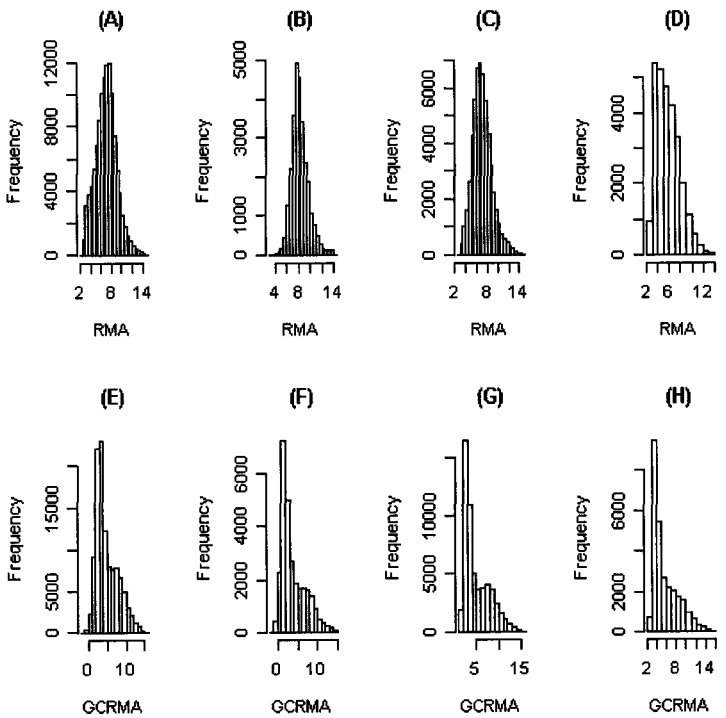

Figure 3.5: *Histogram of expressions of RMA and GCRMA. (A) and (E) are original histogram of expression values for RMA and GCRMA respectively. (B) and (F) are expression histograms of high affinity of RMA and GCRMA; (C) and (G) are expression histograms of medium affinity of RMA and GCRMA; (D) and (H) are expression histograms of low affinity of RMA and GCRMA;*

Chapter 4

Comparisons based on fold-change

In this chapter, we introduce RT-PCR, a common method for detecting differential expression and the fold change. In Section 4.3, we compare the fold change from the microarray data for different normalization methods with the quantitative RT-PCR fold change.

4.1 Fold Change

Fold change is one of approaches used for detecting differential expression in common practice. After having performed normalization, one could be able to compare the expression level of any gene in the sample to the expression level of the same gene in the control.

The simplest approach to calculate fold change is to divide the expression level of a gene in the sample by the expression level of the same gene in the control. Then you get the fold change, which is 1 for an unchanged expression, less than 1 for a

down-regulated gene, and larger than 1 for an up-regulated gene. The definition of fold change will not make any sense if the expression value in the sample or in the control is zero or negative (Knudsen, 2004).

For our data set, we define the fold-change as:

$$\text{fold change} = \frac{\text{mean of expression measures for NOR.NOD\_Idd4/11}}{\text{mean of expression measures for NOR}}.$$

4.2 Housekeeping Genes

A housekeeping gene is defined as a gene involved in basic functions needed for the sustenance of the cell, for instance, Actin and HPRT housekeeping genes in our study. Housekeeping genes are constitutively expressed (they are always turned ON). Housekeeping genes are used as internal standards in quantitative polymerase chain reaction since it is generally assumed that their expression is unaffected by experimental conditions (`http://www.medterms.com/script/main/art.asp?articlekey=24232`).

4.3 Quantitative (Real-Time) PCR

In biology, polymerase chain reaction (PCR) is a method that allows exponential amplification of short DNA sequences (usually 100 to 600 bases) within a longer double stranded DNA molecule. PCR entails the use of a pair of primers, each about 20 nucleotides in length, that are complementary to a defined sequence on each of the two strands of the DNA. These primers are extended by a DNA polymerase so that a copy is made of the designated sequence. After making this copy, the same primers can be used again, not only to make another copy of the input DNA strand but

also of the short copy made in the first round of synthesis. This leads to exponential amplification. After several (often about 40) rounds of amplification, the PCR product is analyzed on an agarose gel and is abundant enough to be detected with an ethidium bromide stain. For reasons that will be outlined below, this method of analysis is at best semi-quantitative and, in many cases, the amount of product is not related to the amount of input DNA making this type of PCR a qualitative tool for detecting the presence or absence of a particular DNA sequence. In order to measure messenger RNA (mRNA), the method was extended using reverse transcriptase to convert mRNA into complementary DNA (cDNA) which was then amplified by PCR and, again analyzed by agarose gel electrophoresis. In many cases this method has been used to measure the levels of a particular mRNA under different conditions but the method is actually even less quantitative than PCR of DNA because of the extra reverse transcriptase step. Reverse transcriptase-PCR analysis of mRNA is often referred to as "RT-PCR" which is unfortunate as it can be confused with "Real-Time PCR".

Real-Time PCR, also called quantitative (real-time) PCR, is a method of simultaneous DNA quantification and amplification. It is the real-time version of Q-PCR (Quantitative PCR). DNA is specifically amplified by polymerase chain reaction. After each round of amplification, the DNA is quantified. Common methods of quantification include the use of fluorescent dyes that intercalate with double-strand DNA and modified DNA oligonucleotides (called probes) that fluoresce when hybridized with a complementary DNA. Frequently, Real-Time PCR is combined with reverse transcription-polymerase chain reaction to quantify low abundance messenger RNA, enabling a researcher to quantify relative gene expression at a particular time, or in a particular cell or tissue type. More details about Real-Time PCR can be found at

http://pathmicro.med.sc.edu/pcr/realtime-home.htm.

In a Real-Time PCR experiment, some measurement of gene expression (copy number) is found for the gene of interest on each of n independent biological replicates (different animals) and within each biological replicate there are a number m of technical replicates primarily to control for minor differences in the experimental conditions. To account for natural variability in gene expression levels, n animals have the expression measure of a housekeeping gene found by the same process again with m technical replicates per animal. This setup is repeated independently for another sample of n animals from a second strain. Both the gene of interest and the housekeeping gene are identical in each of the two samples. The aim of the experiment is to test for differential gene expression between the two strains to confirm the results of a microarray experiment.

Traditionally, gene expression studies were done by one gene at a time using technologies such as RT-PCR. But the microarray technologies allows the simultaneous measurement of the expression level of thousands of genes. In addition, data from microarray experiments are both quantitative (expression level) and qualitative (the gene is expressed or not). However, a major drawback of using quantitative data is its accuracy and precision. Normalization is performed by scaling the gene expression levels of one chip to a control microarray, to a control target intensity, or to another color standard (in the case of printed microarrays). When the scaling factor departs from one, the magnitude of the correction may have an effect on the accuracy of the resulting data. That is the reason why a validation study using quantitative RT-PCR is needed.

We denote the expression measurements on the gene of interest in strain 1 by

Gene Name	Sample	Mean Qty	Std Dev	Actin Normalized Mean
Actin	NORF1A	5423329.5	505638.16	
	NORF2A	4867596.5	234921.3	
	NOR.NOD_Idd4/11F3A	8284620.5	273700.44	
	NOR.NOD_Idd4/11F4A	5913631.5	272737	
C1qb	NORF1A	45178.516	7162.103	8.33
	NORF2A	42346.875	4791.4136	8.70
	NOR.NOD_Idd4/11F3A	3964.7637	595.3106	0.48
	NOR.NOD_Idd4/11F4A	4314.191	720.0945	0.73

Table 4.1: *Part of normalized gene data by take Actin housekeeping gene and one of genes of interest called C1qb.*

$x_{1ij}, i = 1, \ldots, n, j = 1, \ldots, m$ and the housekeeping gene in strain 1 by $y_{1ij}, i = 1, \ldots, n, j = 1, \ldots, m$.

The software reports the mean and standard deviation of the m technical replicates for each of the four sets of data per animal. From the files of RT-PCR results supplied to me by Dr. Evakine who did the experiment, We can construct Table 4.1 using Actin as the housekeeping gene. And the values for NORF1A, NORF2A, NOR.NOD_Idd4/11F3A and NOR.NOD_Idd4/11F4A are derived from activated macrophages.

Since housekeeping gene had much higher levels of expression (10^6) than the genes of interest (10^4), the Actin normalized mean is obtained by

$$\text{Normalized mean} = \frac{1000 \times \text{mean of gene of interest}}{\text{mean of housekeeping gene}},$$

for example, for strain NORF1NA of gene C1qb, the Actin normalized mean is calculated by: $1000 \times 45178.516/5423329.5 \approx 8.33$.

In Table 4.1, there are two sample NORF1NA and NORF2NA to contribute the strain NOR, so the sample mean of strain NOR for Actin housekeeping gene is calculated by taking the average of two samples, i.e. $(8.33 + 8.70)/2 = 8.52$ for the computation of Real-Time PCR fold change. Similarly, the sample mean of strain NOR.NOD_Idd4/11 is $(0.48 + 0.73)/2 = 0.60$.

Our purpose, especially for this chapter, is to compare the fold change from Real-Time PCR to that one from microarray. Thus, we have to compute the Real-Time PCR fold change. Regarding to the definition of fold change in our study, we could calculate the Real-Time PCR fold change by the ratio of sample mean from NOR.NOD_Idd4/11NOR to sample mean from NOR , i.e. $0.60/8.52 = 0.07$ as shown in column 2 of Table 4.2. In a similar manner we got the Real-Time PCR fold change for 20 genes selected due to biological interest.

4.4 Results

Traditionally, gene expression studies were done by one gene at a time using technologies such as RT-PCR. But the microarray technologies allows the simultaneous measurement of the expression level of thousands of genes. That is the reason why a

validation study using quantitative RT-PCR need to be done.

In our Real-Time PCR table, 20 genes of interest by biologist are listed. We find the probe sets corresponding to each gene and the expression values for each methods to obtain the Table 4.2 and 4.3.

In Table 4.2 and Table 4.3, we compare the fold changes from the different methods to those of from two housekeeping genes which are called Actin and HPRT.

MAS5 shows the biggest correlation with RT-PCR (we use Actin housekeeping gene here) which is .7457. The correlations with RT-PCR for different methods, RMA, GCRMA and PLIER are .5576, .5323, .6210.

From Table 4.3 and 4.4, the following reality and conclusions can be made:

1. We notice that, for several genes, there are two or more probe set correspond to one gene. This is the fact as we described in Chapter 1.

2. It appears two "pending" genes called Ns-pending and Magmas-pending in Table 4.3 and 4.4. It means that these genes are unknown in biology, but it is known that they belong to "Ns" and "Magmas" family, respectively.

3. The normalization methods do make a difference during the analysis. It is hard to tell which method is better than others. However, comparing the fold change from different methods, the performance of GCRMA are close to Real-Time PCR fold change in almost half of all the genes (9 out of 20). We say that GCRMA has less variability than the other normalization methods.

4. According to the conclusion from Affymetrix, PLIER works better than MAS 5.0 where these two methods are developed by Affymetrix. However, from the

Gene Name	Actin	HPRT	Probe Set	RMA	GC-RMA	MAS5	PLIER	dchip PM	dchip PM/MM
C1qb	0.07	0.09	96020_at	0.22	0.15	0.2	0.23	0.23	0.22
			162276_i_at	0.94	1	0.89	1.13	1	1.06
C1qg	0.21	0.26	92223_at	0.45	0.34	0.37	0.47	0.48	0.48
Rbp1	0.29	0.4	104716_at	0.71	0.65	0.42	0.56	0.75	0.71
C1qa	0.39	0.5	98562_at	0.58	0.47	0.52	0.56	0.52	0.55
Arcn1	0.57	0.74	94512_f_at	1.01	1.02	0.95	1.03	1.02	1.13
			94513_r_at	1.02	1.02	1.43	1.05	0.99	1.07
			94514_s_at	0.97	0.89	0.98	1.06	1.01	0.95
Narg1	0.61	0.79	103791_at	0.91	0.99	0.95	0.93	0.92	0.83
			93246_at	1	1.03	1.04	1.07	1	1.03
			96152_at	1.01	0.97	0.97	1.06	0.98	0.98
Ns-Pending	0.58	0.79	98948_at	0.94	0.75	0.95	0.99	1.15	0.8
Apoe	1.6	2.09	161321_i_at	0.89	1.01	0.73	0.9	0.86	0.71
			95356_at	1.35	1.41	1.37	1.38	1.32	1.41
Ly75	0.47	0.61	103258_at	1.02	1.04	1.1	1.09	1.03	1.05
Ly6c	1.7	2.17	93077_s_at	1.23	1.58	1.29	1.5	1.2	1.33
Tnfrsf6	0.42	0.55	102921_s_at	0.9	0.83	0.83	0.91	0.9	0.87
Ifnar1	0.52	0.69	100483_at	1.1	1.24	1.12	1.18	1.09	1.1

Table 4.2: *Fold change comparisons between microarray and Real-Time PCR (A)*

Gene Name	Actin	HPRT Set	Probe	RMA	GC-RMA	MAS5	PLIER	dchip PM	dchip PM/MM
Inpp5b	0.91	1.22	94398_s_at	1.96	2.83	2.16	2.38	1.98	1.91
			94399_at	2.05	2.86	2.15	2.34	1.9	2.11
Apoc2	3.43	4.55	97887_at	1.29	1.45	2.02	1.62	1.17	1.32
Cd59a	0.43	0.58	101516_at	0.81	0.64	0.75	0.76	0.87	0.71
Rab3d	1.66	2.08	97415_at	1.3	1.8	1.4	1.53	1.24	1.39
Il6	0.6	0.76	102218_at	0.75	0.73	0.72	0.79	0.76	0.75
Magmas-Pending	0.52	0.66	160258_at	0.86	0.84	0.57	0.82	0.88	0.79
Mmp13	0.76	1.01	100484_at	1.29	1.38	1.22	1.34	1.34	1.33
Il1a	0.55	0.71	94755_at	0.85	0.82	0.84	0.89	0.85	0.85

Table 4.3: *Fold change comparisons between microarray and Real-Time PCR (B)*

result of our study which is based on the real data set, PLIER does not display better performance than MAS 5.0 does.

5. The performance of PLIER are very close to that of RMA, except for genes Inpp5b and Apoc2.

6. As for the comparison between two models of dChip, in the literature, it concludes that PM-MM model is better than PM only model for Li and Wong (MBEI) (Irizarry and Wu, 2005). This conclusion can be confirmed in our study as well, because the fold change for PM-MM model perform closer to housekeeping gene than that for PM model does, except on gene C1qa, Arcn1, Ly75, Ifnar1 and Apoc2.

Chapter 5

Comparisons based on SAM

In this chapter, we introduce the SAM methodology and SAM computation procedure. The comparisons of the ranking, q-value and d-statistic between different normalization methods are conducted at the end of this chapter.

5.1 Introduction

SAM (Significance Analysis of Microarrays) is a statistical technique for finding significant genes in a set of microarray experiments. The input to SAM is a matrix of gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment.

SAM computes a statistic d_i for each gene i, measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine if the expression of any genes are significantly related to the response. The cutoff for significance is determined by a tuning parameter Δ, chosen

by the user based on the false positive rate.

We need to identify differentially expressed genes from a set of microarray experiments. We must perform hypothesis tests on all genes simultaneously to determine whether each one is differentially expressed or not. Hence, the null hypothesis is that there is no change in expression levels between experimental conditions. The alternative hypothesis is that there is some change. We reject the null hypothesis if there is enough evidence in favor of the alternative. This amounts to rejecting the null hypothesis if its corresponding statistic falls into some predetermined rejection region. Hypothesis testing is also concerned with measuring the probability of rejecting the null hypothesis when it is really true (called a false positive) and the probability of rejecting the null hypothesis when the alternative hypothesis is really true (called power).

There are four important steps one must take in testing for differential gene expression. The first is that a statistic must be formed for each gene. The choice of this statistic is important in that one wants to make sure that no relevant information is lost with respect to the test of interest, yet all measurements on the gene are condensed into one number. The second step is to calculate the null distribution(s) for the statistics. One can assume that each gene has a different null distribution or one can calculate a null distribution for each gene. The third step is choosing the rejection regions. One can take a priori symmetric or one-sided rejection regions, or one can choose them adaptively. This involves comparing the original statistics to null versions of the statistics. The fourth step is to assess or control the number of false positives at the traditional 5% level. For more details, see Chu *et al.* (2003).

5.2 Model and main idea about SAM

In our experiment, there are two strains for the hybridization and 9 arrays were processed on two different days. Since the day effect is not biological variation and not of our interest, the only different experimental condition is the difference between strains. The model for fitting is:

$$Y = \beta_0 + \beta_1 \text{Day} + \beta_2 \text{Strain} + \epsilon, \tag{5.1}$$

where Day=which of 2 days the array was processed on, and Strain=strain of mouse on array.

We are interested in determining which genes show a statistically significant difference in gene expression between different strains. Therefore, the null hypothesis for each gene is that the data we observe have some common distributional parameter among the conditions. For each gene we form a statistic, d-statistic, that is a function of the data.

Suppose we separate our microarrays into two groups. The first group refers to microarrays from Day 1 and second group refers to those from Day 2. Thus, there are 5 samples in group 1 and 4 samples in group 2.

We use a modified t-statistic to calculate the "relative difference" in gene expression:

$$d_i = \frac{\hat{\beta}_2}{s_i + s_0}, i = 1, \ldots, 12488$$

where the quantity s_0 is called a fudge factor and is here to deal with cases for which the variability across arrays is very low and s_i is the standard error of $\hat{\beta}_2$.

Permutation technique making drawing of a very large number of samples possible using only the available data need to apply to our data set. Purpose of permutation is to establish significance in the case of without any assumptions. Regarding to our data set, totally $C_5^2 \times C_4^2 = 10 \times 6 = 60$ permutations are produced. Then recalculate the statistics for the permuted condition labels.

The plot of the average order statistics $\bar{d}_{(i)}$ against the observed $d_{(i)}$ can be drew.

For a value of Δ, one can draw two lines with slope 1 and intercepts $-\Delta$ and Δ. Then the points $t_1(\Delta)$ and $t_2(\Delta)$ can be found, where the plot first crosses these lines. The observations further from the center than these are declared significant. By the values of $t_1(\Delta)$ and $t_2(\Delta)$, we can find the numbers of significant genes in each permutation, and the average of these numbers is called the average number of falsely detected differences for the given value of Δ. Then the False Discovery Rate $FDR(\Delta)$ for the given value of Δ can be defined as

$$FDR(\Delta) = \frac{\text{the average number of falsely detected differences}}{\text{number detected in the original sample}} \times \hat{\pi}_0, \qquad (5.2)$$

where $\hat{\pi}_0$ is an estimate of π_0. π_0 is the proportion of true null (unaffected) genes in the original data set, and the algorithm for calculating $\hat{\pi}_0$ see Storey and Tibshirani (2003).

The q-value is similar to the p-value and is a measure of significant genes in terms of the false discovery rate. For each gene i, we can find the value of Δ_i, Δ_i=maximum Δ such that gene i is significant. Then the q-value is defined as

$$q_i = \min\{\text{FDR}(\Delta) : \Delta \leq \Delta_i\}$$

Further details can be found in Tusher *et al.* (2001).

	RMA	MAS5	GCRMA	PLIER
RMA	1	0.29	0.36	0.25
MAS5	0.29	1	0.20	0.11
GCRMA	0.36	0.20	1	0.43
PLIER	0.25	0.11	0.43	1

Table 5.1: *Ranking correlations between RMA, MAS5, GCRMA and PLIER.*

	RMA	MAS5	GCRMA	PLIER
RMA	1	0.34	0.37	0.17
MAS5	0.34	1	0.22	0.06
GCRMA	0.37	0.22	1	0.22
PLIER	0.17	0.06	0.22	1

Table 5.2: *q-value correlations between RMA, MAS5, GCRMA and PLIER.*

5.3 Comparison results between RMA, MAS5, GCRMA and PLIER

We apply the expression measurements as the input of the SAM procedure, in which the expression measurements are from the different normalization methods, RMA, GCRMA, MAS5, PLIER and several from dChip. In order to investigate the performance of different normalization methods, we compare the ranks, q-values and the d-statistics.

From the comparisons based on the rankings, q-values and d-statistics of the SAM output, we can conclude that RMA is more correlated with GCRMA than with MAS5 and Plier. The two methods from Affymetrix, MAS5 or Plier show low correlations.

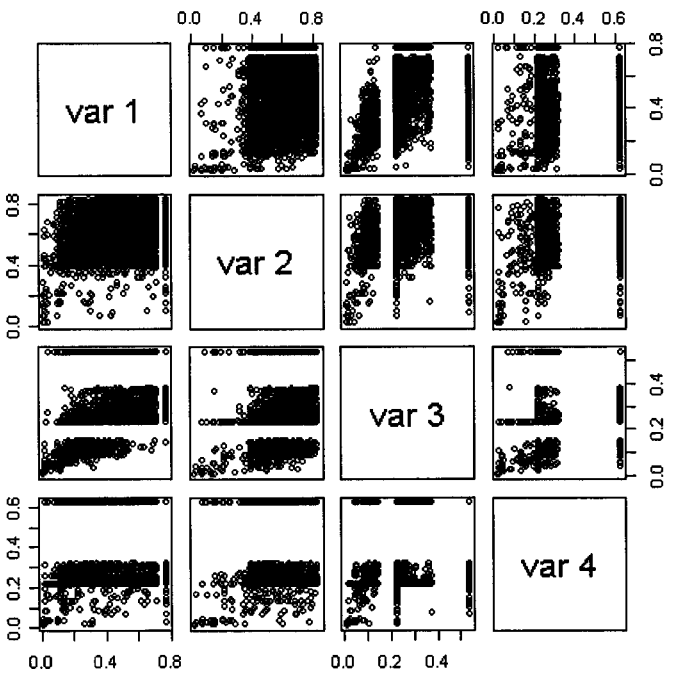

Figure 5.1: *q-value pairwise correlation coefficient plot from RMA, MAS 5.0, GCRMA and PLIER. From left to right and from top to bottom, the corresponding orders of the graphs are RMA, MAS 5.0, GCRMA and PLIER. Var 1, var 2, var 3 and var 4 refer to RMA, MAS 5.0, GCRMA and PLIER respectively.*

	RMA	MAS5	GCRMA	PLIER
RMA	1	0.63	0.79	0.74
MAS5	0.63	1	0.56	0.58
GCRMA	0.79	0.56	1	0.71
PLIER	0.74	0.58	0.71	1

Table 5.3: *d-statistic correlations between RMA, MAS5, GCRMA and PLIER*

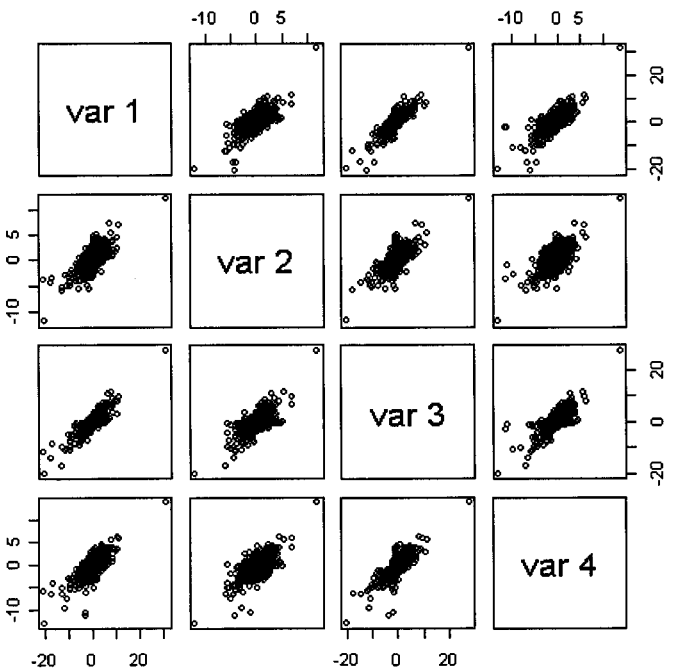

Figure 5.2: *d-statistic pairwise correlation coefficient plot from RMA, MAS5, GCRMA and PLIER. From left to right and from top to bottom, the corresponding orders of the graphs are RMA, MAS5, GCRMA and PLIER. Var 1, var 2, var 3 and var 4 refer to RMA, MAS 5.0, GCRMA and PLIER respectively.*

Surprisingly, the pairwise plots based on these three summaries indicate totally different patterns (as shown in Figure 5.1 and 5.2). In the q-value plot, there is always at least one gap. What the meaning is behind these gaps and why they appear requires further research.

In addition, we check the significant genes for different methods in which q-values are less than 0.05. As a result, we obtain 38 significant genes from GCRMA normalization method, 35 from RMA normalization method and 28 from PLIER normalization method, but only 4 for normalization method. It means that when we fix the q-value at 0.05, using GCRMA method could detect more significant genes than the other methods. Furthermore, the 4 significant genes from MAS 5.0 are also included in the list of significant genes from the other methods. This confirms the conclusion identically at Chapter 4 that GCRMA is a better method on performance of validation of significant genes than the other method and the performance of RMA and PLIER are very close.

5.4 Comparison results between RMA and dChip

The comparison between RMA and dChip is analogous to the one in Section 5.3. We examine the ranking correlation, d-statistic and q-statistic from dChip and compare with those from RMA.

Surprisingly, RMA and dChip show little relationship whatever in the comparison on the basis of ranking, d-statistic or q-value. In Figure 5.3, by looking at the scales of RMA and dChip we notice that the q-values from RMA cover from 0 to 1, but the q-values from dChip only in the range of $0 - 0.6$ approximately. From Figure 5.4, we

	RMA	PM only	PM-MM
RMA	1	0.10	0.07
PM only	0.10	1	0.28
PM-MM	0.07	0.28	1

Table 5.4: *Ranking correlations between RMA and dChip*

	RMA	PM only	PM-MM
RMA	1	0.09	0.07
PM only	0.09	1	0.23
PM-MM	0.07	0.23	1

Table 5.5: *q-value correlations between RMA and dChip*

	RMA	PM only	PM-MM
RMA	1	-0.14	-0.17
PM only	-0.14	1	0.66
PM-MM	-0.17	0.66	1

Table 5.6: *d-statistic correlations between RMA and dChip*

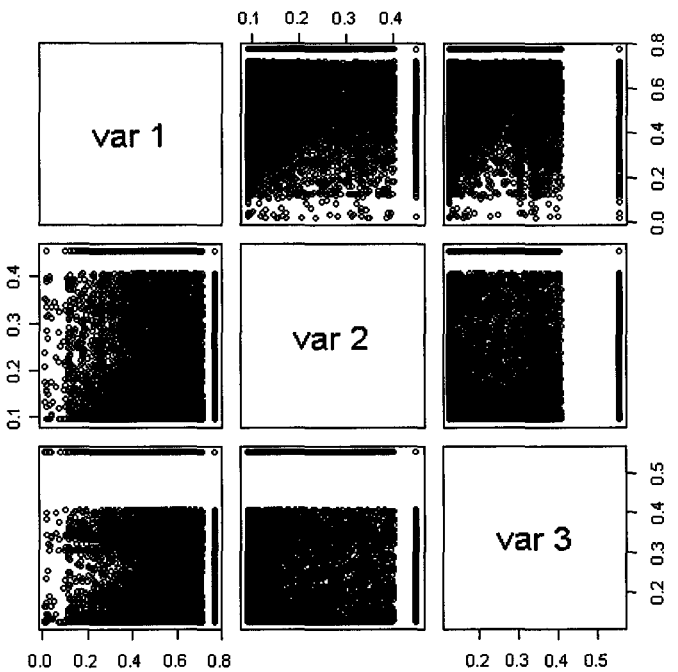

Figure 5.3: *q-values pairwise plot from RMA, dChip PM only and dChip PM-MM model. From left to right and from top to bottom, they follow the same order of RMA, dChip PM only and dChip PM-MM. Var 1, var 2 and var 3 refer to RMA, dChip PM and dChip PM-MM respectively.*

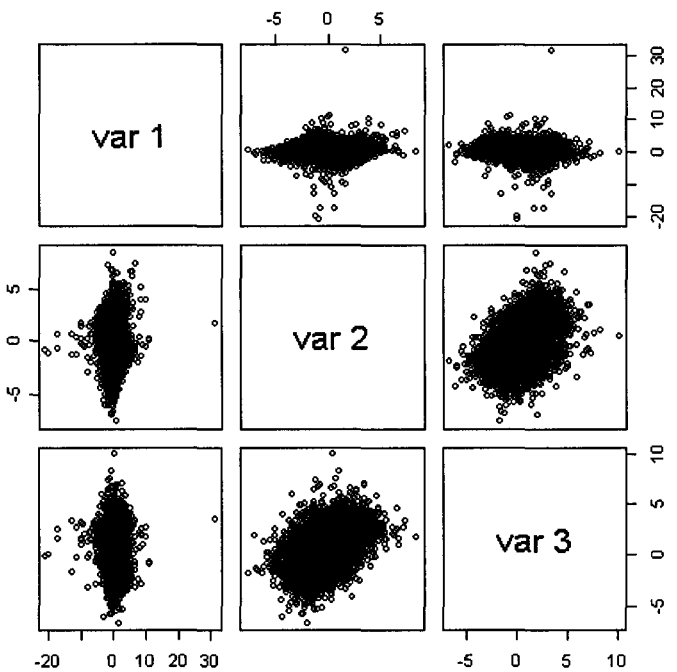

Figure 5.4: *d-statistic pairwise plot from RMA, dChip PM only, dChip PM-MM model. From left to right and from top to bottom, they follow the same order of RMA, dChip PM only and dChip PM-MM. Var 1, var 2 and var 3 refer to RMA, dChip PM and dChip PM-MM respectively.*

can see linear relationship between PM only model and PM-MM model of dChip.

Furthermore, there are a lot of negative correlation coefficients in the comparison of d-statistic as shown in Figure 5.4. The similar situation exists in the expression measures of dChip as well.

An other similar result to Section 5.3 is that the pairwise q-values graph (as shown in Figure 5.3) display some gaps and there is almost complete coverage of the plot area. Why the graph exist gaps and why q-value have no such values between the gaps will be focused in the further study.

Chapter 6

Conclusion and Discussion

Gene expression measures play an important role in the microarray data analysis. A better pre-processing and then expression measure leads the more efficient analysis to detect the differentially expressed genes.

After a number of normalization methods, RMA, MAS 5.0, GCRMA, PLIER dChip-PM and dChip-PM/MM, we obtain a set of expression measures for different methods.

The comparison in terms of pairwise correlation coefficient between different methods of 3-stage in preprocessing shows that the normalization method and summarization method results in the expression measures similar, but PM correction seems more important since it determines the model. Unfortunately, while comparing the different normalization and summarization methods, we did not check the scale from the pairwise plots from different methods. The similar comparison done by Irizarry and Wu (2005) based on bias and variance conclude that background correction has the largest effect. Further work could be applied to our real data set to compare the bias

and variance for different methods to verify the conclusion of Irizarry and Wu (2005). Also we could carry out an analysis similar to that done here using the spike-in and dilution data sets to see if we get a different conclusion for Iriarry and Wu (2005).

By the comparison in terms of pairwise plot, we examine some relationship between different methods. It is worth mentioning that RMA and MAS 5.0 appear to have a linear relationship, especially when applying the MAS 5.0 detection calls algorithm. Unfortunately, affinity information is not a way to explain the strange pattern between RMA and GCRMA. It seems that only at high signal ends, RMA and GCRMA show a little linear pattern. It probably have other better ideas to explore their relationship at expression level.

As for the comparison between microarray fold change and qRT-PCR fold change, GCRMA shows better performance on validation of significant genes than other methods. And the performance of RMA and PLIER are similar. In addition, the fold change of all the genes except gene Inpp56 and Apoc2 in Table 4.3 and 4.4 are significant.

The comparison on the basis of SAM confirm part of conclusion in the previous comparison, for instance, GCRMA is the best preprocessing method in use of validating the differentially expressed genes performance on validation of significant genes; RMA and PLIER perform similarly on the production of expression measures. But more research needs to be done in the pairwise relationship in terms of ranking correlation, d-statistic and q-statistic because of the strange graph pattern.

As for the two models of dChip, that is dChip PM only model and dChip PM-MM model, more study is needed to explore the relationship related to dChip with the other methods, where it is a little weak in this study. With the current findings, we

notice that dChip sometimes returns log-expression values of negative infinity which we have replaced with very small finite numbers. We also report that the relationships between dChip and the other show negative correlation.

Appendix A

Table of the Definition of Methods

Method	Bkg Correction	Normalization	PM Correction	Summarization
M1	RMA	Quantile	PM Only	Median Polish
M2		Constant		
M3		Contrasts		
M4		Invariantset		
M5		Loess		
M6		Qspline		
M7		VSN		
M8	None	Quantile	PM Only	Median Polish
M9		Constant		
M10		Contrasts		
M11		Invariantset		
M12		Loess		
M13		Qspline		
M14		VSN		

Table A.1: *Possible preprocessing methods list (part A)*

Method	Bkg Correction	Normalization	PM Correction	Summarization
M15	None	Quantile	Mas	Median Polish
M16		Constant		
M17		Contrasts		
M18		Invariantset		
M19		Loess		
M20		Qspline		
M21		VSN		
M22	None	Quantile	PM Only	Avgdiff
M23		Constant		
M24		Contrasts		
M25		Invariantset		
M26		Loess		
M27		Qspline		
M28		VSN		
M29	None	Quantile	Mas	Avgdiff
M30		Constant		
M31		Contrasts		
M32		Invariantset		
M33		Loess		
M34		Qspline		
M35		VSN		

Table A.2: *Possible preprocessing methods list (part B)*

Method	Bkg Correction	Normalization	PM Correction	Summarization
M36	None	Quantile	PM Only	MAS
M37		Constant		
M38		Contrasts		
M39		Invariantset		
M40		Loess		
M41		Qspline		
M42		VSN		
M43	None	Quantile	MAS	MAS
M44		Constant		
M45		Contrasts		
M46		Invariantset		
M47		Loess		
M48		Qspline		
M49		VSN		
M50	None	Invariantset	PM Only	Li and Wong
M51	None	Invariantset	Subtractmm	Li and Wong
M52	MAS	False	MAS	MAS
M53	Sequence Info.	Quantile		Median Polish

Table A.3: *Possible preprocessing methods list (part C)*

Bibliography

[1] Affymetrix (1999) Affymetrix GeneChip Analysis Suite User Guide.

[2] Affymetrix (2001) Affymetrix Microarray Suite User Guide, Version 5.0.

[3] Affymetrix (2002) Statistical Algorithms Description Document.

[4] Affymetrix (2005) Guide to Probe Logarithmic Intensity Error (PLIER) Estimation.

[5] Binder H. and Preibisch S. (2005) Specific and Nonspecific Hybridization of Oligonucleotide Probes on Microarrays, *Biophysical Journal*, 2005, 89: 337-352

[6] Blalock, E. (2003) *A Beginner's Guide to Microarray*, Kluwer Academic Publishers, USA.

[7] Bolstad, B. (2004) Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization, *Dissertation*, University of California, Berkeley.

[8] Bolstad, B. (2005) affy: Built-in Processing Methods, *Users' Guide Attached to Bioconductor.*

[9] Cope, L. M., Irizarry, R. A., Jaffee, H., Wu, Z. and Speed, T. P. (2004) A Benchmark for Affymetrix GeneChip Expression Measures. *Bioinformatics* 20: 323-331.

[10] Chu, G., Narasimhan, B., Tibshirani, R. and Tusher V. (2003) "Significance Analysis of Microarrays" Users guide and technical document.

[11] Hartemink, A. J., Gifford, D. K., Jaakola, T. S. and Young, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. *The International Society for Optical Engineering; International Biomedical Optics Symposium.*

[12] Huber, P. J. (1981), *Robust statistics*, John Wiley & Sons, Inc, New York, New York.

[13] Irizarry R. A., Bolstad B. M., Collin F., Cope L. M., Hobbs B. and Speed T. P. (2003a) Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research* 31(4):e15.

[14] Irizarry R. A., Hobbs B., Collin F., Beazer-Barclay YD, Antonellis K. J., Scherf U. and Speed T. P. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics* 4(2):249-64.

[15] Irizarry, R. A. and Wu, Z (2005) Comparison of Affymetrix GeneChip Expression Measures, *Bioinformatics*, Vol.1, no. 1, 2005, pp.1-7.

[16] Knudsen S. (2004) *Guide to Analysis of DNA Microarray Data*, (2nd ed.), John Wiley & Sons, New Jersey.

[17] Li C. and Wong W. H. (2001a), Model-based analysis of oligonucleotides arrays: model validation, design issues and standard error application, *Genome Biology*, 2(8): research0032.1-0032.11

[18] Li C. and Wong W. H. (2001b), Model-Based analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection, *Proceedings of the National Academy of Sciences of the United States of America*, USA **98**.

[19] Naef, F. and Magnasco, M. O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays, *Physical Review*, E 68, 011906.

[20] Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C., and Afshari, C. A. (1999) Microarrays and toxicology: the advent of toxicogenomics. *Molecular Carcinogenesis*, 24(3):153-159.

[21] Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (2003) The Analysis of Gene Expression Data: An Overview of Methods and Software in *The Analysis of Gene Expression Data: Methods and Software*, (Parmigiani, Garrett, Irizarry & Zeger, editors) New York: Springer.

[22] Regalado, A. (1999) Inventing the pharmacogenomics business. *American Journal of Health System Pharmacy*, 56(1):40-50.

[23] Storey J. D. and Tibshirani R. (2003a) SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays, in *The Analysis of Gene Expression Data: Methods and Software*, (Parmigiani, Garrett, Irizarry & Zeger, editors) New York: Springer.

[24] Storey J. D. and Tibshirani R. (2003b) Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences of the United States of America*, USA **100**, 9440-9445.

[25] Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachusetts.

[26] Tusher V. G., Tibshirani R. and Chu G. (2001) Significance Analysis of Microarrays Applied to Transcriptional Responses to Ionizing Radiation, *Proceedings of the National Academy of Sciences of the United States of America*, USA **98**, 5116-5121.

[27] Wu Z., Irizarry R. A. , Gentleman R., Martinez-Murillo F. and Spencer F., (2004) A Model- Based Background Adjustment for Oligonucleotide Expression Arrays, *Journal of The American Statistical Association* **99**, 909-917.