uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Huiling Xiong
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Biology)
GRADE / DEGREE

Department of Biology
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Novel Methods and Strategies for Microarray Data Analysis

TITRE DE LA THÈSE / TITLE OF THESIS

Xuhua Xia
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Jonny St-Amand

George Carmody

Stephane Aris-Brosou

Vance Trudeau

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

NOVEL METHODS AND STRATEGIES FOR MICROARRAY DATA ANALYSIS

Huiling Xiong

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
in partial fulfillment of the requirements for the
Ph.D. degree in the

Ottawa-Carleton Institute of Biology

Thèse soumise à la
Faculté des études supérieures et postdoctorales
Université d'Ottawa
en vue de l'obtention du doctorat ès sciences

L'Institut de biologie d'Ottawa-Carleton

© Huiling Xiong, Ottawa, Canada, 2008

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-48665-8
Our file Notre référence
ISBN: 978-0-494-48665-8

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

ABSTRACT

Microarray technology has been used as a routine high-throughput tool in biological research to characterize gene expression, and overwhelming volumes of data are generated in every microarray experiment as a consequence. However, there are many kinds of non-biological variations and systematic biases in microarray data which can confound the extraction of the true signals of gene expression. Thus comprehensive bioinformatic and statistical analyses are crucially required, typically including normalization, regulated gene identification, clustering and meta-analysis. The main purpose of my study is to develop robust analytical methods and programs for spotted cDNA-type microarray data. First, I established a novel normalization method based on the Generalized Procrustes Analysis (GPA) algorithm. I compared the GPA-based method with six other popular normalization methods, including Global, Lowess, Scale, Quantile, Variance Stabilization Normalization, and one boutique array-specific housekeeping gene method by using several different empirical criteria, and demonstrated that the GPA-based method was consistently better in reducing across-slide variability and removing systematic bias. In particular, being free from the biological assumptions that most genes (95%) are not differentially expressed on the array, the GPA method is therefore more robust, and appropriate for diverse types of array sets, including the boutique array where the majority of genes may be differentially expressed. Second, I utilized statistical analysis to assess the quality of a novel goldfish brain cDNA microarray, which provides statistical validation of microarray data result. Thirdly, I developed a new program suite as a user-friendly analytical pipeline integrating most popular analytical methods for microarray data analysis. Finally, I proposed a novel analytical strategy to extract season-related gene expression information from multiple microarray data sets by using comprehensive data transformation and normalization analysis, differential gene identification, and multivariate analysis.

RÉSUMÉ

La technologie des puces à ADN (microarray) est fréquemment utilisée en biologie pour caractériser l'expression de millier de gènes à la fois. Il est donc primordial de développer des méthodes statistiques et bioinformatiques capable de différentier le niveau d'expression des gènes des sources d'artefacts souvent causées par des erreurs systématiques et des variations non biologiques propres aux puces à ADN. Plusieurs méthodes ont déjà été développées pour réduire le bruit ce type d'expérience telle que des méthodes de normalisations, d'identifications des gènes diférentiellement exprimés, de regroupement et méta-analyses. Le but principal de mon projet est de développer des méthodes d'analyses robustes et des programmes pour les données de puce à ADN de type ADNc. Premièrement, j'ai établi une nouvelle méthode de normalisation sans postulat basé sur l'algorithme ''Generalized Procrusted Analysis'' (GPA). Utilisant plusieurs critères empiriques, j'ai comparé cette méthode avec six autres méthodes de normalisation populaires, qui incluent Global, Lowess, Scale, Quantile, VSN et une méthode boutique spécifique pour les puces de gènes de maintien. J'ai démontré que la méthode basée sur algorithme GPA est constamment supérieure pour réduire la variabilité entre puces pour ADN et pour retrancher leserreurs systématiques. Sans les conditions biologiques et statistiques qui sont inhérente aux autres méthodes de normalisation, la méthode GPA est plus robuste et appropriée pour les diverses catégories de puces à ADN, incluant la puce boutique où la majorité des gènes sont exprimés de manière différentielle. Deuxièmement, j'ai utilisé des analyses statistiques pour caractériser la qualité d'une nouvelle puce à ADN pour le cerveau de poissons rouges, ce qui valide les résultats des puces à ADN. Troisièmement, j'ai développé une nouvelle série de programmes intégrant les méthodes d'analyses les plus fréquemment utilisées pour l'analyse des puces à ADN. Je propose finalement une nouvelle stratégie d'analyse pour extraire l'information d'expression des gènes régulés par saison, pour plusieurs groupes de données de puces à ADN. Cette méthode utilise une transformation compréhensive des données et une analyse de normalisation, l'identification des gènes exprimés différentiellement et une analyse mutlivariée.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Dr. Xuhua Xia, for his expertise, support and constant encouragement during my graduate experience. It was a great pleasure for me to conduct this thesis under his supervision. I also wish to thank Dr. Vance Trudeau who initiated *the Goldfish Environmental Genomics Project* that enabled me to participate in the research field of microarray data analysis. I greatly appreciate his patience and constructive comments during my Ph.D. studies. Thanks are also due to the other members of my committee, Dr. Guy Drouin and George R. Carmody for their creative comments and guidance throughout my thesis research. Dr. Stéphane Aris-Brosou helped clarifying my thoughts through advice and discussion.

Colleagues in Dr Xia's and Dr. Trudeau's labs have contributed substantially to this work. I am thankful to Dapeng Zhang for his help, discussion and suggestions in several projects. Also I appreciate the help and the advice from Christopher J Martyniuk, Kate Crump and Jason T. Popesku at the beginning of my PhD project. I would like to thank Dr. Nicolas Rodrigue, Gareth Palidwor, Dr. Tom Hazle, Sam Khalouei, Xiaoquan Yao, Pinchao Ma, Ziyu Song, and Malisa Carullo for their help and discussions. It is only with their help could I publish in prestigious journals such as Physiological Genomics, BMC Bioinformatics, and Molecular and Cellular Endocrinology.

I must thank some of my fellow graduate students for their friendship and encouragements: Robert Carter, Robert T. Morris, Brady Tracey, Sophie Calixte, Dr. Jennifer Li-Pook-Than and Akiko Shoji. They each helped make my time in the PhD program fun and interesting.

I am also indebted to Geoge Rajoh who was so kind as to read, correct and edit the whole manuscript thoroughly. Particularly I am grateful to his family for their kindness and affection. Home staying with them really made me feel at home and made me concentrate on my research throughout my stay in Ottawa.

I in particular wish to thank my family for offering me unconditional love and support throughout my PhD program. This work is dedicated to them.

Finally, I wish to acknowledge research funding from NSERC and an OGSST scholarship from Ontario government.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction to the microarray technique and its data analysis

1.1 Introduction to DNA microarrays

DNA Microarray technology is now playing an increasingly important role in biomedical research. It allows relative measurements of transcriptional levels for thousands of genes simultaneously. The most important application of DNA microarray is to identify the differentially expressed genes to facilitate the identification of functionally important genes [1-3]. Identification of co-expressed genes has also paved the way to large-scale identification of co-regulated genes [4-7] and gene-regulation networks [8-10].

Currently, there are two widely used microarray technologies: high-density oligonucleotide arrays [11] and spotted cDNA microarrays [12], with the main difference between the two being the probe sets. Gene probes in oligonucleotide arrays consist typically of 16-20 oligonucleotides of 25-mers, whereas gene probes in a cDNA microarray are spotted cDNA segments, typically expressed sequence tags (ESTs) resulting from large-scale preliminary characterization of gene expression. This thesis focuses on the analysis of spotted cDNA microarrays. The main advantage of cDNA arrays is that uncharacterized cDNA sequences can be spotted, which enables the microarray technology to be applied to organisms with no or only limited genome sequence information. If such unknown cDNAs are found to be regulated, the researcher can focus sequencing efforts on those that are found to be regulated by treatments of interest.

1.2 AURATUS genomics project and a novel goldfish brain cDNA microarray

My PhD project is part of an AURATUS goldfish environmental genomics project, in which a novel goldfish brain cDNA microarray has been established and used to screen differential gene expression in goldfish brain in experiments involving treatments with diverse hormones and endocrine disrupting chemicals. Here, I use the goldfish to detail

cDNA microarray technology, with the analytical pipeline for cDNA microarray experiment and data analysis illustrated in Figure 1.1.

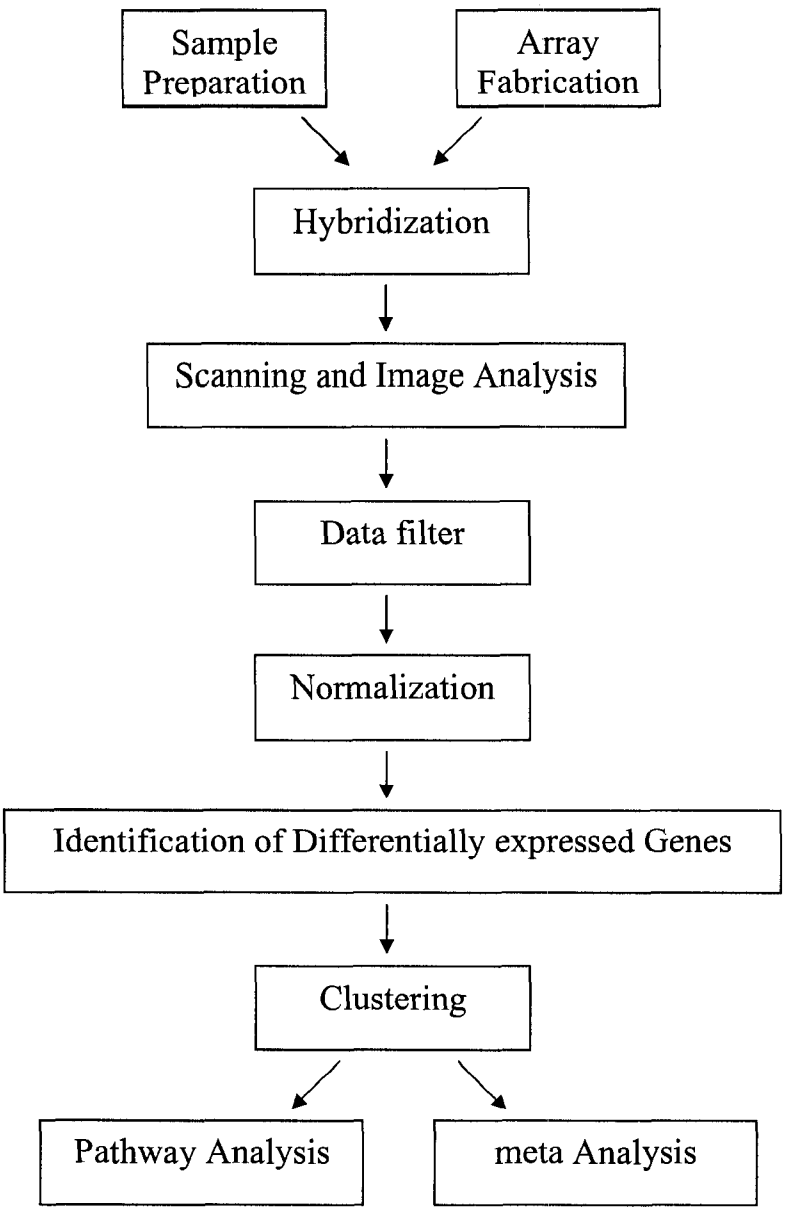

Figure 1.1. The flowchart of the procedure of microarray hybridization and the process of the data analysis.

This goldfish brain cDNA microarray contains about 9000 cDNA clones including 3000 goldfish brain cDNAs plus 6000 carp cDNAs [13, 14]. The inclusion of carp sequences to supplement our goldfish microarray is justified by the close phylogenetic relationship between the two species. Many genes are strongly conserved between goldfish and carp [15] and natural hybrids between the two species are frequently found [16]. There are also some special spots used as positive and negative controls, which are useful for evaluating microarray quality. The array is printed with 24 pins in a 2x12 format. The cDNAs are printed in duplicate side by side to give 48 grids.

The common microarray experimental protocol is composed of RNA sample extraction, labelling with fluorescent dyes, hybridization with probes, laser scanning, image processing and data analysis (Figure 1.1). Two different RNA samples, one from the experimental goldfish and the other from the control, are extracted from brain, and reverse transcribed into cDNAs. The resulting cDNA samples are labelled with cy5 (red) and cy3 (green) fluorescent dyes, respectively, and hybridized to the DNA probes on the array. The unhybridized cDNAs are washed off and the amount of chemically bound cDNA is quantified by the intensity of the fluorescence in each spot, as measured by the laser scanner. For each gene on the array, four fluorescence values are obtained: foreground value for the green channel, foreground value for the red channel, background fluorescence for the green channel, and background fluorescence for the red channel. These raw data represent the beginning of microarray data analysis pipeline.

1.3 Microarray data normalization

The cDNA microarray is a widely used high-throughput technique for gene expression profiling. However, much random variation and systematic bias exist in microarray data, which can confound the extraction of the true fluorescence intensity signals, and thus compromise downstream data analysis and interpretation of the experimental data. The sources of these biases include efficiency of dye incorporation, different processing procedures, concentration of DNA on arrays, difference in the amount of mRNA, variability in reverse transcription, uneven hybridization, different detection efficiencies by the scanner and experimenter bias [17]. Therefore, proper data normalization is required to remove these biases before accurate identification of differential expression [17-20]. Based on different

biological or statistical assumptions about data distribution or experimental design, various normalization methods have been proposed. The main objective of normalization is to ensure that measured intensities within and across slides are comparable. In this section, I will offer a general review of those diverse normalization methods. How to use these normalization methods in practical data analysis is detailed in chapter 4 by using a user-friendly program that integrates much of microarray data analysis.

1.3.1 Within-slide normalization

The objective of within-slide normalization in cDNA array data is to ensure that light-intensity data from the two channels (red and green) are comparable. The housekeeping gene method [21] is an early normalization method, which assumes that the expression levels of housekeeping genes remain constant even when the expression of many other genes is substantially changed. However, many so-called housekeeping genes have been reported to exhibit considerable variability under different experimental conditions, making them unsuitable and unrepresentative of the whole expression intensity range. The Global normalization approach [21] assumes that the center (mean or median) of the distribution of log ratio M values in each slide is zero. However, the Global normalization method does not consider intensity-dependent and spatially-dependent effects, which are usually major biases among the slides. In order to remove such biases, Yang et al. [17] proposed to use the local regression smoothing procedure (Lowess) that is applied to each slide separately to normalize the log ratio intensities. Lowess normalization has been one of the most popular methods but it relies on two important assumptions. Lowess assumes that most genes on the array are not differentially expressed across the experiments and also that the numbers of up- and down-regulated genes at each intensity level are roughly equal within each slide. Other methods including the semiparametric [19], neural network [22], and common array dye-swap methods [23] have been proposed to remove intensity-dependent biases.

1.3.2 Between-slide normalization

While the within-slide normalization methods can effectively remove the intensity-dependent or spatially-dependent biases within each slide, additional statistical methods are needed for normalization across multiple replicated slides [17]. Scale normalization [17] is

one popular approach for such across-slide normalization [20, 24], in which log ratio intensities are assumed to follow a normal distribution with expectation zero and homogeneity of variance across replicated arrays. Other effective across-slide normalization methods include Quantile [25] and Variance stabilization normalization (VSN) [26]. Quantile normalization was initially developed for the Affymetrix single channel chip [25], and then extended for two-colour cDNA microarrays in the Limma package of the bioconductor project (http://bioinf.wehi.edu.au/limma/). It relies on the assumption that the true probe intensities for each array in a set of replicated arrays are approximately equally distributed. The goal therefore is to adjust for the difference in distribution among multiple slides, and data points are shifted such that the sample densities of slides are identical. In contrast, the VSN method assumes that most of the genes on the arrays are not differentially expressed in a given experiment and utilizes the arcsine rather than log transformation to stabilize the variance so as to remove the dependence of the variance on the total intensity. This gives genes with higher intensities an equal chance of being ranked high as genes with lower intensity. VSN has been used for both the Affymetrix [27] and cDNA microarray platforms [28].

1.4 Other types of microarray data analysis: quality control, differential gene identification and cluster analysis

In addition to data normalization, a series of other statistical analysis procedures are also required to extract biological information from microarray data. These include quality control, differential gene identification, and cluster analysis. We will introduce several statistical methods for checking quality control of microarray data in Chapter 3. The methods for differential gene identification and cluster analysis will be detailed in Chapter 4.

1.5 Objective of my PhD study

The main objective of my PhD study is to establish new methods and software for cDNA microarray analysis and to assist the development of a new goldfish-carp cDNA microarray platform. Specifically, four main topics will be covered in my thesis:

a) To develop new General Procrustes Analysis (GPA) -based normalization method for microarray data analysis (chapter 2);

b) To assess the reliability of the goldfish-carp cDNA microarray platform (chapter 3);
c) To develop a user-friendly analytical program suite (chapter 4);
d) To extract biologically important information from multiple microarray data sets (chapter 5).

Chapter 2. Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data

(This chapter has been published in *BMC Bioinformatics* 2008, 9:25)

Summary

Normalization is essential in dual-labelled microarray data analysis to remove non-biological variations and systematic biases. Many normalization methods have been used to remove such biases within slides (Global, Lowess) and across slides (Scale, Quantile and VSN). However, all these popular approaches have critical assumptions about data distributions, which are often not valid in practice. In this study, we have developed a novel normalization method based on the Generalized Procrustes Analysis (GPA) algorithm. Both publicly available data and simulated data were used in comparing our GPA-based method with six other popular normalization methods, including Global, Lowess, Scale, Quantile, VSN, and one boutique array-specific housekeeping gene method. The assessment of these methods is based on three different empirical criteria: across-slide variability, the Kolmogorov-Smirnov (K-S) statistic and the mean square error (MSE). Compared with other methods, the GPA-based method performs effectively and consistently better in reducing across-slide variability and removing systematic bias. In particular, the GPA method has no requirement that most genes (95%) are not differentially expressed. The GPA-based method is therefore appropriate for diverse types of array sets, including the boutique array where the majority of genes may be differentially expressed.

2.1 Introduction

Microarray data normalization is a crucial step in microarray data analysis to remove these non-random variations and systematic biases [17-20]. While many different normalization methods are available and diverse strategies are implicated, most of them require certain critical biological or statistical assumptions about data distribution. For example, one usual assumption underlying the Global, Scale, Lowess and VSN methods is that the array contains a large enough assortment of random genes that are not differentially expressed across the experiment. The assumption on data distribution inherent in these

methods is not always valid in practice. For example, in custom-made boutique arrays most of genes are expected to be differentially expressed [29, 30]. Therefore, most of the above normalization methods are inappropriate. Although some novel methods have been proposed for such boutique arrays, including the housekeeping gene [29], Zipf's law [31] and the mixture model based methods [32], they still have their own assumptions. The commonly used housekeeping gene method assumes that a set of prior housekeeping genes exists in the microarray in similar expression patterns, and could be utilized for Lowess normalization. However, the hybridization signals from these proposed housekeeping genes may not span the entire fluorescence range produced by boutique arrays. Zipf's law method [31] assumes that the microarray data set exhibits an observed power-law distribution with an exponent close to -1, whereas the mixture model based method [32] assumes that the log ratio of intensity in each channel follows a Gamma distribution. Therefore, although many available normalization methods are effective for removing some types of biases among arrays, their own assumptions about data distribution may limit their application, or even introduce new biases. There is a clear need for developing versatile and robust methods for normalization of microarray data.

Here we present evidence to show that the Generalized Procrustes Analysis [33-35] or GPA, is a powerful statistical method for normalization of microarray data. The GPA method differs from other methodologies in that it requires few assumptions about data distributions, which delineates its main advantage over the other methods. Procrustes analysis is one of the least-squares methods for translation, rotation, scaling and aligning matrices of corresponding coordinates to maximize their agreement [33-35]. The transformation parameters are computed in a direct and efficient manner based on a selected set of corresponding point coordinates. Microarray data are matrices of color intensities. Replicated slides are matrices of similar configurations and are therefore amenable to Procrustes analysis. On the basis of publicly available and simulated data, our GPA-based normalization strategy is systemically compared with six other popular normalization methods including two within-slide methods (Global and Lowess), three across-slide methods (Scale, Quantile and VSN), and one boutique array specific housekeeping gene method. The assessment of these methods is based on three empirical criteria: variability

among replicated slides [20, 36], the Kolmogorov-Smirnov statistic [31, 37] and the mean square error [25].

2.2 Materials and methods

2.2.1 Experimental data

2.2.1.1 Swirl zebrafish data set

The swirl zebrafish data set is available at http://www.bioconductor.org. The data set has been previously utilized in several studies on normalization procedures [22, 38]. The main goal of this experiment was to identify genes with altered expression in the Swirl mutant compared to the wild type zebrafish. There are 8448 genes in each slide and all experiments were replicated four times including two dye-swaps.

2.2.1.2 HCT116 data set

The microarray data set of HCT116 cancer cell line was retrieved from the study of Zhou et al [39] which identified the dose- and time-dependent changes of gene expression in HCT116 cell lines after treatment of the topoisomerase inhibitor 1 camptothecin compound (CPT). There are 10 slides in total and each slide contains 2208 cDNA clones. This set of data was used previously in one study to compare the effectiveness of different normalization methods [40].

2.2.1.3 Mouse apoptosis boutique microarray

The custom boutique array data set was retrieved from http://www.mucosa.de/zipfs/zipfs_normalization.html. The microarray experiment aimed to identify differences in apoptotic mechanisms between two different mouse cell lines. There are 5 replicate slides and each slide contains 1024 spots. The genes selected for this array are involved in apoptosis; therefore it is expected that a high proportion of genes are highly differentially expressed. The data set has been used previously for Zipf's law method in boutique microarray data normalization [31].

2.2.2 Data simulation study

2.2.2.1 Simulation based on Balagurunathan's method

The generation of our simulation is based on a parameterized random signal model [41], which has been utilized in several papers [42-48]. More precisely, our simulation approach works as follows:

1. Simulate true expression intensity level I from an exponential distribution with a mean of 3000, i.e. $I_g \sim \exp(\lambda = 1/3000)$, for $g=1,\ldots$, the number of genes.

2. Simulate the base intensities for the red and green channels, R_k and G_k, respectively. Let $R_k \sim N(I_k, \alpha_r I_k)$ and $G_k \sim N(I_k, \alpha_g I_k)$ where α_r and α_g are the coefficients of variation for the intensity in the red and green channel respectively.

3. Simulate an experiment with differentially expressed genes. Randomly choose ρ*100% genes which are outlier genes, then convert these intensities to $R_k^{'}=R_k\sqrt{t_k}, G_k^{'}=G_k/\sqrt{t_k}$, for $k=1,\ldots$, the number of the genes with $t_k=10^{\pm b_k}$ where b_k follows a beta distribution, $b_k \sim B(1.7,4.8)$ and where +/- sign means over- or under- expressed with the desired probability which you want. Specially, $\rho=0$ when the experiment is self-self hybridized.

4. Simulate the dye-biased characteristics. Dye bias is often caused by differential labelling and detection efficiencies between the fluorescent dyes used [49]. The dyes commonly used for microarray experiments show nonlinear response characteristics, and different dyes give different responses. The fluorescent intensity is often not linearly related to the expression level. This effect is modified by the nonlinear function

$$f(x)=a_3[a_0+x(1-e^{-x/a_1})^{a_2}], a_3 \geq 1 \qquad (1)$$

where the four parameters a_0, a_1, a_2, a_3 in the formula can characterize the shape of the MA plot. Our fluorescent intensities for the two channels were transformed as $R_k^{''} = f(R_k^{'})$ and $G_k^{''} = f(G_k^{'})$ respectively.

5. Simulate background noise. The background noise for red and green channels is determined by a normal distribution, whose parameters are randomly chosen to describe the process, i.e. $I_{rb} \sim N(I_b, \alpha_{rb} I_b)$ and $I_{gb} \sim N(I_b, \alpha_{gb} I_b)$ where α_{rb} and α_{gb} are the coefficients of

variation of the background noise for red and green channels, respectively. A signal to noise ratio is then the true mean of the signal over the true mean of the background noise (I_b). Our simulated intensities are corrected by subtracting the background noise.

6. Simulate the red and green foreground noise of the spots that follows with normal distribution $I_{rf} \sim N(\mu_{R_k^{"}}, \sigma^2_{R_k^{"}})$ and $I_{gf} \sim N(\mu_{G_k^{"}}, \sigma^2_{G_k^{"}})$ respectively. Then the signal intensity of each spot is

$$SR_k = R_k^{"} + I_{rf}. SG_k = G_k^{"} + I_{gf}$$

$$with\ \mu_{R_{k\,i}^{"}} = R_k^{"} a_{m_1}; a_{m_1} \sim U[f_{a_1}, f_{b_1}]$$

$$\mu_{G_{k\,i}^{"}} = G_k^{"} a_{m_2}; a_{m_2} \sim U[f_{a_2}, f_{b_2}]$$

$$\sigma^2_{R_k^{"}} = a_{s_1} \mu_{R_k^{"}}; a_{s_1} \sim U[f_{c_1}, f_{d_1}]$$

$$\sigma^2_{G_k^{"}} = a_{s_2} \mu_{G_k^{"}}; a_{s_2} \sim U[f_{c_2}, f_{d_2}] \qquad (2)$$

where f_{a_1}, f_{b_1}, f_{a_2}, f_{b_2}, f_{c_1}, f_{c_2}, f_{d_1}, f_{d_2} are given parameters.

Based on the simulation procedure described above, a series of data sets were produced to gain a better insight into the effects of normalization. All of them include 1000 genes in 10 replicate slides. For each realization, we replicated the data set 100 times. Specific characteristics of these data sets are given below.

1. Simulation data 1 is a 1000×10 log ratio intensity of expression matrix. Based on the assumption that most genes are not differentially expressed, our data was simulated to hold 3, 5, 10, and 30 percent levels of differentially expressed genes, at the same time the proportion of up-regulated genes equals that of down-regulated genes without dye bias.

2. Simulation data 2 is the same as simulation data 1, but with dye bias which generates the banana shaped MA plot which is commonly observed in real cDNA microarray data sets.

3. Simulation data 3 is a 1000×10 log ratio intensity of expression matrix which is simulated as the boutique arrays in which 60 percent of genes are differently expressed. The ratios of the up-regulated genes to the down-regulated genes are 5:5, 7:3, and 9:1.

2.2.2.2 Simulation based on the SIMAGE method

In the SIMAGE method [50], simulated microarray data are divided into gene expression and biases from several sources including a raw background gradient signal, a channel effect, a spot pin effect, a nonlinear effect, a quantization and saturation effect, and random error due to unknown factors. All these effects are specified in 29 parameters in SIMAGE, which were roughly estimated from real microarray data. Based on these estimated parameters, the SIMAGE method simulated microarray data mimicking the real experimental data as close as possible. The detailed input parameters in this study are listed in Table 1.1. All microarray dataset included 1000 genes in 50 replicated slides. For normal microarray data, three differential levels (5, 10, 30 %) were considered with same ratio (1:1) of up-regulated gene to down-regulated gene. For boutique array, 60% genes were differentially expressed and three ratios (5:5, 7:3 and 9:1) of up-regulated to down-regulated genes were considered.

Table 1.1. Input parameters used in SIMAGE simulation.

Parameter Description	Value
Array number of grid rows	9
Array number of grid columns	4
Number of spots in a grid row	18
Number of spots in a grid column	18
Number of spot pins	12
Number of technical replicates	1
Number of genes (0 = max)	1000
Number of slides	50
Perform dye swaps	no
Gene expression filter	yes
Reset gene filter for each slide	no
Mean signal	11.492
Change in log2ratio due to upregulation	0.832
Change in log2ratio due to downregulation	0.605
Variance of gene expression	5, 10, 30 (1:1);
% of differential genes (up: down=1:1)	60(5:5, 7:3, 9:1);
	60(5:5, 7:3, 9:1);
	100(9:1)
Correlation between channels	0.981
Dye filter	yes
Reset dye filter for each slide	yes
Channel (dye) variation	0.51
Gene x Dye	0
Error filter	yes
Reset error filter for each slide	yes
Random noise standard deviation	0.219
Tail behaviour in the MA plot	0.09
Non-linearity filter	yes
Reset non-linearity filter for each slide	yes
Non-linearity parameter curvature	0.025
Non-linearity parameter tilt	0.777
Non-linearity from scanner filter	yes
Reset non-linearity scanner filter for each slide	yes
Scanning device bias	0.295
spotpin deviation filter	yes
Reset spotpin filter for each slide	no
spotpin variation	0.36
Background filter	yes
Reset background filter for each slide	yes
Number of background densities	5
Mean SD per background density	0.3
Maximum of the background signal (%) relative to the non-background	100
SD of the random noise for the background signals	0.1
Background gradient filter	yes
Reset gradient filter for each slide	yes
Maximum slope of the linear tilt	700
Missing values filter	yes
Reset missing spots filter for each slide	yes
Number of hairs	10
Maximum length of hair	20
Number of discs	6
Average radius disc	10
Number of missing spots	0

2.2.3 Generalized Procrustes Analysis (GPA)

The Procrustes analysis method is available as the package vegan in the R project http://www.r-project.org. We consider the log-transformed data, log ratio intensity $M = \log_2(R/G)$ and the mean log intensity $A = \log_2 \sqrt{RG}$ where R and G are the red and green signal intensities respectively. We organized our log-transformed data in N replicated arrays as N series of matrices with g rows and two columns where g is the number of gene probes on the slide and the two columns are the log ratio intensity and mean log intensity on a scale (M,A). For each cDNA spot j in the *i-th* slide, it corresponds to a vector $\vec{w}_{ji} = (M_{ji}, A_{ji})$ (j=1,2,...,g; i=1,2,...N). M_{ji} and A_{ji} are the measurement of M and A for gene j in replication i. Our expression level matrix of the *i-th* slide becomes the matrix $S_i = (\vec{w}_{1i}, \vec{w}_{2i}, ..., \vec{w}_{gi})^T$, where i=1, 2, ..., N. Firstly, a reference array S_0 is generated through computing the median intensity of each gene over all slides. Then each slide S_i (i = 1,2,..., N) is translated so that its centroid is at the centroid point of the reference array S_0. Let S_i be rotated and scaled such that the residual discrepancy between S_i and S_0 is minimized. We wish to find the orthogonal matrix H_i with ($H_i H_i^T = I$) and scale factor c_i so as to minimize the sum of squared distances between the corresponding points in S_i and S_0, i.e.

$$M^2 = trace(c_i S_i H_i - S_0)(c_i S_i H_i - S_0)^T$$

$$= c_i^2 trace(S_i S_i^T) - 2c_i \cdot trace(L) + trace(S_0 S_0^T) \quad (1)$$

where $L = S_i H_i^T S_0^T$

A perfect match gives $M^2 = 0$. To minimize M^2, we need

$$c_i = trace(L) / trace(S_0 S_0^T) \quad (2)$$

$H_i = VU^T$, where V and U are products of the singular value decomposition of $S_i S_0^T = ULV^T$ [51].

Therefore, the output data after our Procrustes normalization method on S_i is $S_i^* = c_i S_i H_i$.

The GPA method needs approximately a few seconds on a regular PC platform to normalize the data. Since the GPA algorithm computes the matrix from all the spots, it requires that all the spots have values. Here we used K-nearest neighbor averaging scheme [52] to impute missing values.

2.2.4 Other normalization methods

Six other normalization methods were compared against our GPA-based method. These include Global [21], Lowess [17], Scale [17], Quantile [25], VSN [26], and the housekeeping gene method for boutique array [29]. The detailed descriptions of those methods are found in Chapter 4.

2.2.5 Evaluation methods

2.2.5.1 Replicated Variability

The criterion of replicated variability is based on the rationale that the expression level of a gene should ideally remain the same across multiple replicate slides. For N replicated slides, the variability of N values for each gene can therefore be used to compare normalization methods [20, 53]. The standard deviation for gene g, *written as* σ_g, can be estimated as

$$\hat{\sigma}_g = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(M_{gi}^{*} - \overline{M}_{g\cdot}^{*})^2} \qquad (3)$$

where N is the number of replicates in the data set, M_{gi}^{*} is the normalized $log_2(R/G)$ value for gene g in slide i, while $\overline{M}_{g\cdot}^{*}$ is the average log ratio intensity over the slides for gene g. A smaller $\hat{\sigma}_g$ is indicative of a more effective normalization procedure. The mean of such $\hat{\sigma}_g$ estimates over all genes is a global measure of the performance of the normalization methods, with smaller mean $\hat{\sigma}_g$ indicative of a better performance of the normalization method.

2.2.5.2 Kolmogorov-Smirnov (K-S) test

The Kolmogorov-Smirnov (K-S) test is a goodness-of-fit test of two continuous distributions. It is used as a criterion for assessing normalization methods and is based on the

rationale that an effective normalization procedure should result in two similar (ideally identical) distributions with a small, ideally zero-valued, K-S statistic [31, 37]. In contrast, two very different distributions will generate a large K-S statistic.

2.2.5.3 Mean square error (MSE)

MSE is used for simulation studies where the true answer is known. For this reason, it is the ideal criterion for evaluating alternative normalization methods. Denote M_{ji} as the true expression log-ratio for the *j-th* gene of *i-th* replicate, $\hat{M}_{ji}$ as the estimated value for M_{ji}, and $\overline{M}_{j.}$ as the mean of $\hat{M}_{ji}$ ($i = 1, 2, ...N, j = 1, 2, \ldots,$ the number of genes) in N replicates. MSE for the *j-th* gene is defined as

$$\begin{aligned} MSE_j &= \frac{1}{N}\sum_{i=1}^{N}\left(M_{ji} - \hat{M}_{ji}\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{M}_{ji} - \overline{M}_{j.}\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left(M_{ji} - \overline{M}_{j.}\right)^2 \\ &= Var\left(\hat{M}_{ji}\right) + bias^2\left(M_{ji}\right) \end{aligned} \tag{4}$$

2.3 Results

2.3.1 Generalized Procrustes Analysis (GPA) in normalization of microarray data

GPA is a standard multivariate statistical method widely applied in shape analysis to find the optimal superimposition of two or multiple configurations [33-35]. The algorithm involves three transformations: translation, rotation and scaling. Translation is a movement in which the centroid of each configuration is shifted to the common origin by subtracting centroid coordinate. Rotation is a fixed displacement of all points by a constant angle, keeping the distance of each point from the centroid unchanged. Scaling is a stretching or shrinking of all points by a constant amount in a straight line from the point to the centroid of the configuration. The optimal transformation is defined as one with the smallest sum of the squared distances among corresponding points in the configurations. In our study, the GPA method is used to minimize the deviation of signal intensities among microarray slides. A detailed geometric transformation of microarray MA-plots in GPA normalization is

illustrated in Figure 2.1 with one set of simulated microarray data. Figure 2.1(a) shows the transformations for one slide and Figure 2.1(b) shows the superimposition among multiple slides (4 here) after each GPA transformation procedure. Two features of GPA normalization are that it does not change the relative position of points (genes) within each MA-plot and the transformations (translation, rotation and scaling) are based on a global optimization instead of local optimization.

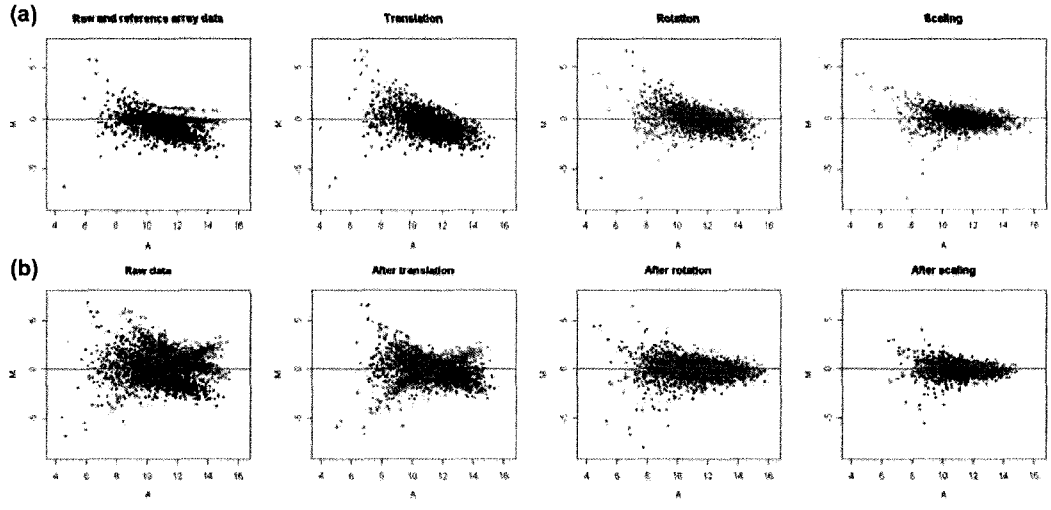

Figure 2.1. A geometric transformation of microarray MA-plots in GPA normalization. (a) shows how the MA-plot for one slide transformed during each GPA transformation procedure. The blue points represent raw data; pink points represent reference slide; red, green and purple points represent data points after translation, rotation and scaling, respectively; (b) shows how the MA-plots for four slides represented by four colours (blue, red, pink, green) transformed after each GPA transformation procedure. The SIMAGE method was used to simulate the microarray data set used here, which includes 50 slides with 10% differentially expressed genes and ratio of up-regulated to down-regulated genes is 1:1.

2.3.2 Strategy for comparison of normalization methods

Based on the different data sets, we compared our GPA-based method with six other normalization methods, including Global [21], Lowess [17], Scale [17], Quantile [25], VSN [26], and the housekeeping gene method for boutique array [29]. The comparison includes three levels. We first evaluated the performance of these different normalization methods in removing biases individually, no matter what strategies are behind them. Thereafter, we systematically compared several pairs of within-slide and across-slide normalization methods to (1) evaluate the potential of combinations of different methods in normalization procedures, and (2) assess the ability of decreasing across-slide bias of the GPA method compared to other across-slide normalization methods based on the same within-slide normalization background. This bears significance in practice since it is often necessary to combine different normalization strategies to decrease variation and potential bias within each slide and across slides. Third, for the boutique data in which common normalization methods are not suitable, we evaluated the GPA method with the housekeeping gene normalization method without other methods involved.

We first evaluated the performance of these normalization methods on two real microarray data sets through comparing data variability and similarity of data distribution among replicated slides. The rationale behind these two criteria is that an effective normalization method should result in lower replicate variability, and more similar (ideally identical) data distributions for replicated slides. Then we applied two simulation methods to simulate several types of microarray data. Mean square error (MSE) was used to evaluate different normalization methods through calculating the true difference between simulated data and normalized data. For the boutique array type, both real data and simulated data were utilized and subjected to these three empirical criteria. In addition, the differential effects of these normalization methods on MA-plots for both real data and simulated data are also compared.

2.3.3 Comparison based on the criterion of replicated variability

Figure 2.2 shows the plots of the variance estimates for the (a) swirl zebrafish and (b) HCT116 cancer data sets. Each bar represents the mean value of replicated variability $\hat{\sigma}_g$ for all genes. For both data sets, all normalization methods decrease variability

of the raw data. However, the GPA method alone yields lower variability than the Lowess, Quantile, Global, and Scale methods do. A Wilcoxon test indicates that the differences are significant ($p<0.01$). Here, the VSN method performs better than the GPA method, which is expected because VSN method specifically aims to stabilize the variance across the replicated arrays.

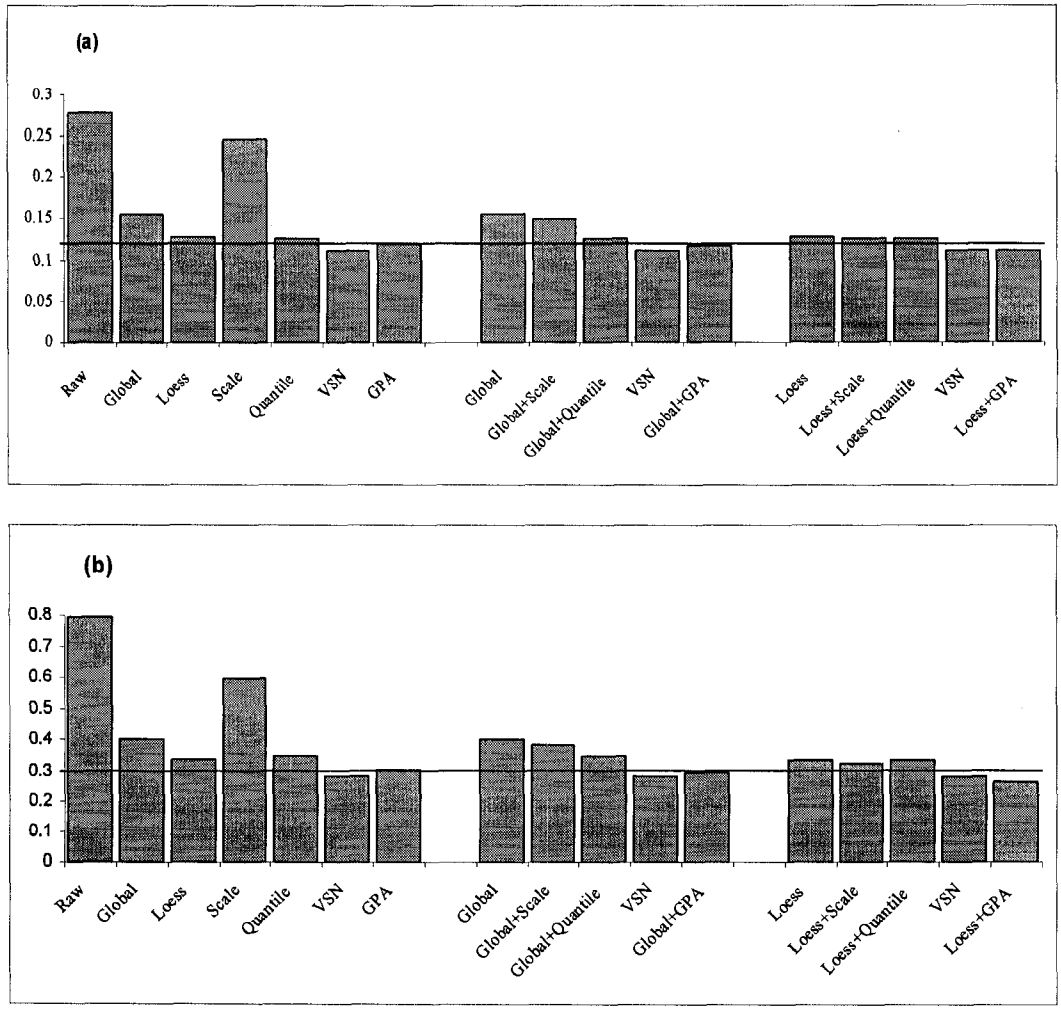

Figure 2.2. Mean of replicate variability for the (a) swirl zebrafish data set and (b) HCT116 data set. Larger value indicates a higher variability across slides. The reference line indicates the variability value for the GPA method.

When we compare the across-slide normalization methods based on the same within-slide normalization background (after Global or Lowess normalization), we can see (1) different combinations of within- and across-slide normalization methods can further reduce the variability values (Figure 2.2), and (2) the GPA method can result in a greater decrease than Scale and Quantile do ($p<0.01$). It is particularly evident in the combination of Lowess and GPA methods compared with other dual normalizations including Global-Scale, Global-Quantile, as well as Lowess-Scale, Lowess-Quantile pairs. The performance of Lowess-GPA pair (0.11 in swirl data and 0.2671 in HCT116 cancer data) is better than that of Lowess and GPA alone ($p<0.01$). It is also somewhat better than that of the VSN method (0.1102 in swirl data and 0.2793 in HCT116 cancer data) ($p<0.01$). In conclusion, our GPA method provides greater reduction of replicate error individually and in combination with other methods such as Lowess.

2.3.4 Comparison using the Kolmogorov-Smirnov (K-S) test

The K-S test is utilized to measure the similarity of the data distributions among replicated slides after diverse normalization procedures. Figure 2.3 (a) shows that for the swirl data set the K-S statistic for the GPA normalization method is much lower than that for the Global, Lowess, Scale and VSN methods alone. However, as expected, the K-S statistic was lowest with Quantile, which forces the empirical distributions in different slides to be identical. This also reveals that the Quantile method is an aggressive normalization process as originally noted by [25]. Except for Quantile normalization, the different combinations of within- and across-slide normalization methods showed Scale, VSN, GPA can decrease the discrepancy of the data distribution after within-slide normalization (Global or Lowess), but GPA effect is particularly evident (Figure 2.3(a)). The combination of Lowess and GPA methods produces a lower K-S value than other methods or method pairs. A similar result can be observed with the HCT116 cancer data set (Figure 2.3(b)) except that VSN is slightly better than GPA alone, but is somewhat outperformed by the Lowess-GPA pair (0.1003 vs. 0.097). For the HCT116 cancer data, although variability was decreased after Lowess or Global normalizations (Figure 2.3(b)), and an ideal MA-plot was produced after Lowess normalization; the discrepancies of data distribution among slides were increased. This is an example that the contributions of within-slide based normalization (Lowess) and across-slide

based normalization differ in various data types. Overall, the GPA method results in a more similarly distributed data than the other four normalization methods.

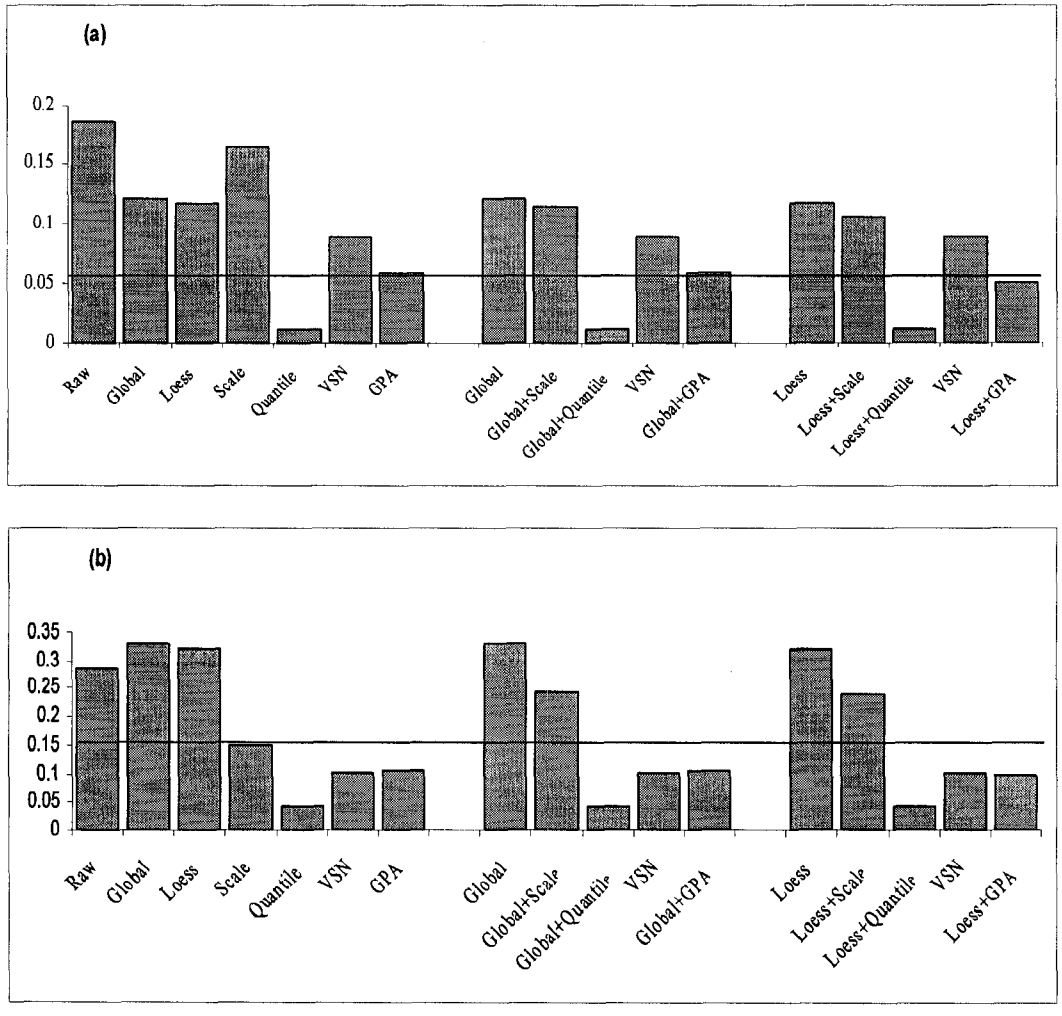

Figure 2.3. Mean of K-S statistic between pairs of slides for the (a) swirl zebrafish data set and (b) HCT116 data set. The reference line indicates the K-S value for the GPA method.

2.3.5 Comparison using the mean square error (MSE) criterion

We used simulation studies to further evaluate our GPA method. The advantage of using a simulated data set is that the true intensities are known, so we can assess the accuracy and precision of normalized data more systematically. Mean square error (MSE) is a widely used comparison criterion, which evaluates the true difference between simulated and normalized data [25, 42]. Note that MSE is decomposed into variance and squared bias. The variance component (ν) is an index of precision and the bias component (β) is an index of accuracy. Obtaining data with satisfactory precision and accuracy has been one of the biggest challenges in the application of microarrays. For this reason, MSE is an excellent criterion for assessing normalization methods [20, 25], with smaller ν and β values indicating better normalization.

We used two different methods to simulate microarray data: one is a parameterized random signal model from the study of Balagurunathan et al [41], which is flexible and has been widely utilized; the other is the recently published SIMAGE method [50], which simulates microarray data based on the estimated parameters from real microarray data. Based on the first method, two types of data were simulated, one without dye bias and one with the "banana-shaped" dye bias. Four different levels (3, 5, 10 and 30%) of differentially expressed genes were considered. The ratio of the up-regulated genes to the down-regulated genes was 1:1. Table 2.1 shows the MSE values for the data set with 5% differential genes following different normalization methods. In both cases, the median of variance (ν) is several times lower with the GPA normalization individually than with the other normalization methods. The median of bias (β) with GPA normalization is also significantly lower than with the other methods ($p<0.01$ by Wilcoxon test). This trend is observed for other data sets with 3, 10, 30% differentially expressed genes (not shown). Based on the SIMAGE method, three sets of microarray data were simulated with 5, 10, 30 % differential expressed genes (up: down=1:1). Table 2.2 shows that our GPA method produces the lowest ν and β values compared with all other methods in three data sets ($p<0.01$ by Wilcoxon test). These results indicate that the GPA method performs effectively and consistently better in reducing across-slide variability and removing systematic bias.

Table 2.1. Comparison among different normalization methods based on simulated normal microarray data with the Balagurunathan's method. The data sets are simulated without or with dye bias and include 1000 genes in 10 slides with 5% differentially expressed genes. The ratio of up-regulated to down-regulated genes is 1:1.

Method	**Without dye bias**		**With dye bias**	
	$\nu^{(1)}$	$\beta^{(1)}$	ν	β
Raw	0.1182	0.004627	0.1231	0.004586
Global	0.1179	0.004732	0.1223	0.00473
Lowess	0.1143	0.005368	0.1182	0.004976
Scale	0.1113	0.004666	0.1201	0.004515
Quantile	0.1149	0.005226	0.1212	0.004716
VSN	0.06294	0.002581	0.06314	0.002541
GPA	0.02122	0.00117	0.02069	0.00119
Global +Scale	0.1117	0.004579	0.1191	0.004595
Global +Quantile	0.1149	0.005226	0.1212	0.004716
Global+GPA	0.02122	0.001174	0.02068	0.001156
Lowess+Scale	0.1088	0.005161	0.1153	0.005067
Lowess+Quantile	0.1133	0.005085	0.1198	0.004812
Lowess+GPA	0.02215	0.001279	0.02205	0.00134

(1) ν and β: the median of variance and bias, respectively, of MSE.

Table 2.2. Comparison among different normalization methods based on simulated normal microarray data with the SIMAGE method. The data sets include 1000 genes in 50 slides with 5, 10, 30% differentially expressed genes and the ratio of up-regulated to down-regulated genes is 1:1.

Method	5%[1]		10%[1]		30%[1]	
	$\nu^{(2)}$	$\beta^{(2)}$	ν	β	ν	β
Raw	1.677	0.06525	1.678	0.08937	1.773	0.1615
Global	0.6751	0.06319	0.4611	0.08158	0.5158	0.1697
Loess	0.2089	0.05962	0.1863	0.07056	0.1891	0.1642
Scale	1.273	0.06859	1.145	0.09515	1.368	0.1921
Quantile	0.2546	0.06579	0.2085	0.07702	0.2167	0.1825
VSN	0.2331	0.05441	0.177	0.06695	0.2014	0.1573
GPA	0.08605	0.04569	0.09839	0.06639	0.1063	0.1328
Global+Scale	0.5449	0.06481	0.3928	0.08096	0.4448	0.1851
Global+Quantile	0.08674	0.04529	0.2085	0.07702	0.2167	0.1825
Global+GPA	0.2546	0.06579	0.09967	0.06784	0.107	0.13
Loess+Scale	0.1846	0.06	0.171	0.06968	0.1788	0.161
Loess+Quantile	0.2055	0.06124	0.1852	0.07132	0.1895	0.1638
Loess+GPA	0.1324	0.0536	0.1221	0.06078	0.1291	0.1468

(1) Percentage of differentially expressed genes
(2) ν and β: the median of variance and bias, respectively, of MSE

2.3.6 The MA-plots for both real data and simulated data after different normalization methods

MA-plots were used to give a general description about the effect of different normalizations on the raw microarray data. Based on different assumptions, these normalization methods result in various geometrical effects on the raw Swirl zebrafish data (Figure 2.4). The Lowess method produces an ideal lowess line along the mean of log intensity. For other methods including GPA without such Lowess type assumption, they can't make the Lowess line around zero along log ratio intensity in each plot. The Global method shifts the central of log ratio intensity to zero. The Scale method changes the scaling of data along the log ratio intensity directions for every slide. The MA-plots for Quantile and VSN are apparently similar to Lowess. For MA-plot after GPA normalization, the shifting and rotation of the data in each slide can be observed. Also a more obvious scaling can be observed in MA-plot of HCT116 cancer data (Figure not shown). Combinations of different methods produce combinatorial effect for plots. A similar result can be seen for simulation data (Figure 2.5).

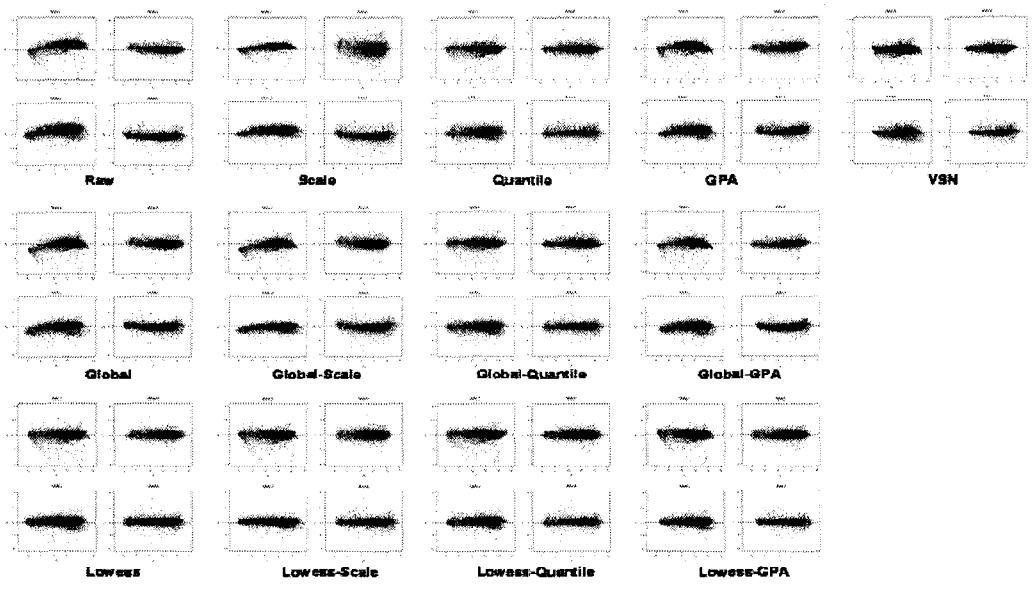

Figure 2.4. The MA-plots for real swirl zebrafish data after different normalization methods.

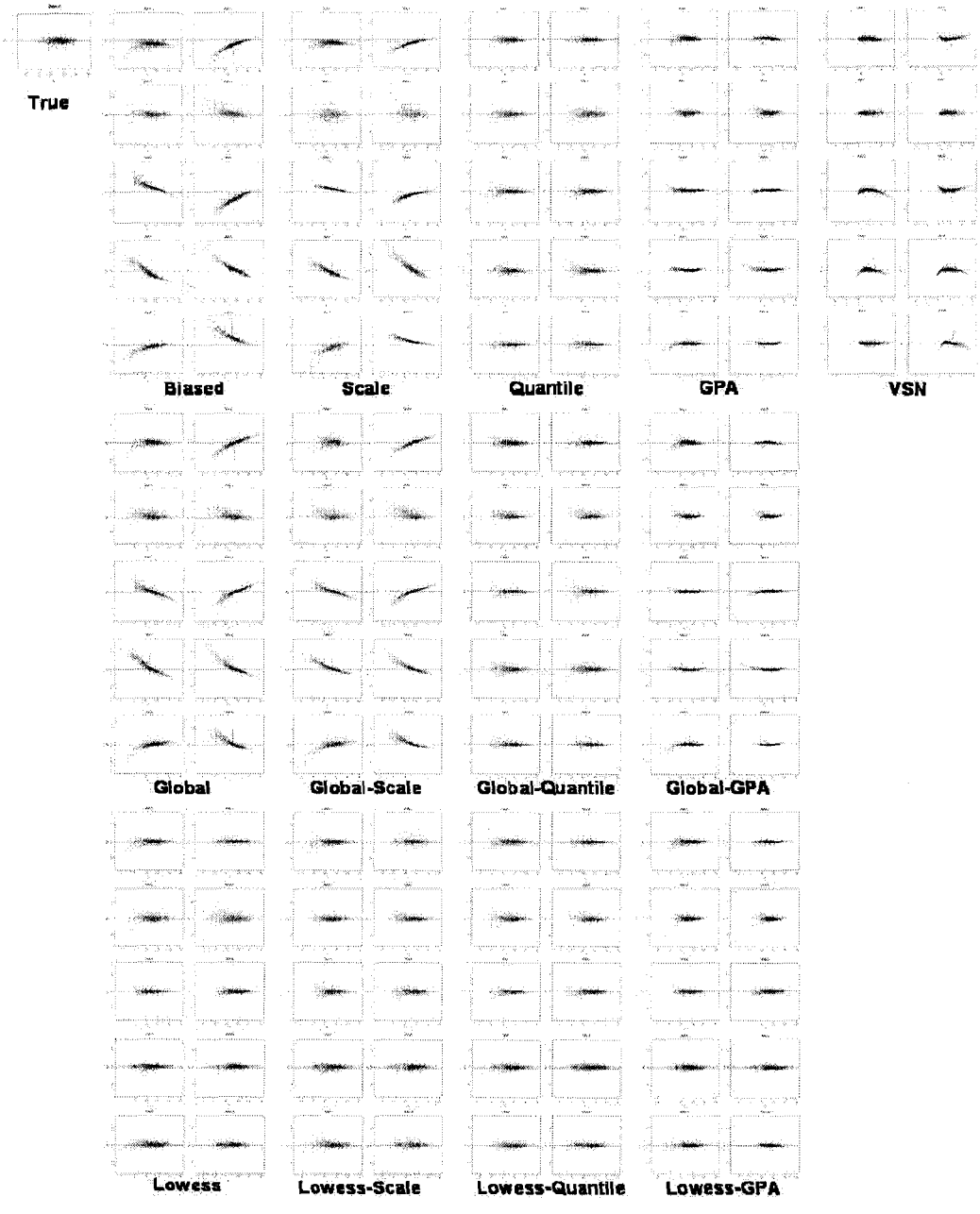

Figure 2.5. The MA-plots for SIMAGE simulated microarray data after different normalization methods. This data set includes 1000 genes in 50 slides. 5 percent genes are differentially expressed. The ratio of up-regulated to down-regulated genes is 1:1. Here we showed MA-plots for first 10 slides in this data set.

2.3.7 Application of the GPA method for boutique arrays

Being relatively assumption-free, the GPA method may be particularly useful in analyzing boutique arrays where the majority of genes may be differentially expressed. Here we show that the GPA method is superior to the popular housekeeping gene normalization approach. We used one real apoptosis microarray data set to evaluate these two methods. The hypothesis behind these microarray data is that the expression of most genes is changing significantly, so most normalization methods are not appropriate. Table 2.3 shows that our GPA normalization can yield lower replicated variability and more similar data distribution (K-S statistic) compared to housekeeping gene normalization. This is further supported in the simulated boutique arrays (Table 2.4). The simulation data set used here is an extreme example for boutique arrays. Approximately 60 percent of the genes are differentially expressed and three different values (5:5, 7:3 and 9:1) are applied for the ratio of up-regulated to down-regulated genes. The medians of ν and β are smaller with GPA normalization than with housekeeping gene normalization for diverse cases. These results demonstrate that the GPA method performs better, with consistently smaller ν and β value, than the housekeeping gene method.

Table 2.3. Variance and K-S values for mouse apoptosis boutique array after GPA and housekeeping gene normalizations.

Method	Mean of Variance	Mean of K-S Value
Raw	0.8760382	0.700391
Housekeeping gene	0.3053415	0.620606
GPA	0.175378	0.268652

Table 2.4. Comparison between the GPA and housekeeping gene normalizations based on simulated boutique array data with the SIMAGE method. The data sets include 1000 genes in 50 slides with 60% differentially expressed genes and the ratios of up-regulated to down-regulated genes are 5:5, 7:3, and 9:1, respectively.

Method	5:5[1]		7:3[1]		9:1[1]	
	$\nu^{(2)}$	$\beta^{(2)}$	ν	β	ν	β
Raw	1.467	0.4234	1.178	0.4687	1.532	0.8552
Housekeeping gene	0.2055	0.4642	0.1842	0.478	0.176	0.9346
GPA	0.111	0.3402	0.1059	0.3296	0.1194	0.5331

(1) Ratio of up-regulated to down-regulated genes

(2) ν and β: the median of variance and bias, respectively, of MSE

In order to further support the ability of GPA on boutique arrays and illustrate its freedom of general assumption that most genes are not differentially expressed, we simulate another extreme example of boutique arrays with 90% up-regulated genes at 10 fold and 10% down-regulated genes at 2 fold. In this case, the housekeeping gene normalization method cannot work since there are no assumed prior housekeeping genes in the experiment, whereas the GPA method can solve this problem. A geometric transformation of such extreme boutique arrays after GPA normalization procedures is supplied as Figure 2.6.

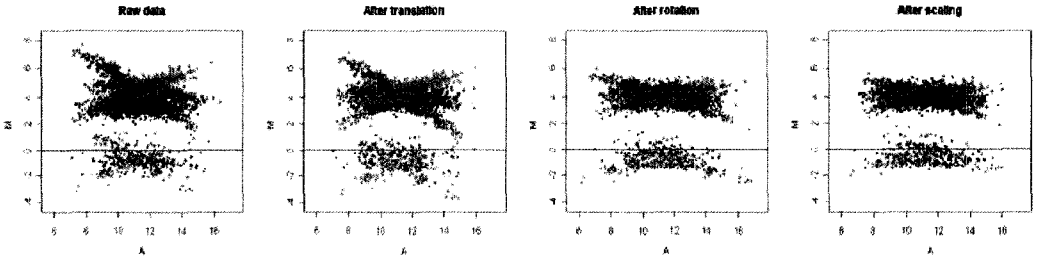

Figure 2.6. A geometric transformation of microarray MA-plots in GPA normalization on the extreme boutique arrays. The SIMAGE method was used to simulate the boutique array data set, which includes 50 slides with 90% up-regulated genes at 10 fold and 10% down-regulated genes at 2 fold. Four slides represented by four colours (blue, red, pink, green) were randomly selected to show their MA-plots after each GPA transformation procedure.

2.3.8 The effect of choice of a reference array on GPA normalizations

The GPA normalization employs a reference array in its first step, which is also utilized in other different normalization methods including Qspline [20, 53], Iset [25], Zipf [31], and Quantile [25]. In this paper the reference slide was established from median values across all slides since several papers illustrated that median or mean value-based reference array is more robust against random variation in the data [25, 54]. In our study we also investigated the effect of alternative methods of choosing reference slide on the performance of GPA normalizations. Figure 2.7 shows the replicated variability and K-S value for swirl zebrafish data after GPA normalizations based on different reference slides. We can see that for K-S statistic, median and mean value-based reference slides provide lower values than other individual reference slides in GPA normalizations. For replicated variability, median and mean value-based reference slides showed better results than other three individual slides except slide 2. Furthermore, although the GPA normalizations differ based on different reference slides, the overall better performance compared to other methods can be still observed here (Figure 2.8). Similar results can also be obtained from HCT116 cancer data (data not shown) and simulated studies. For different types of simulated microarray data (5, 10, 30, 60% differential levels), GPA normalizations with median and mean value-based reference arrays always exhibit a stable and relatively better performance than ones with

individual reference arrays. Figure 2.9 shows the MSE result for simulated data with 5% differential expressed genes after GPA normalizations. Overall, although GPA performance varied when different individual arrays were used as reference array, its use of median and mean value-based reference array can provide relatively stable and better results.

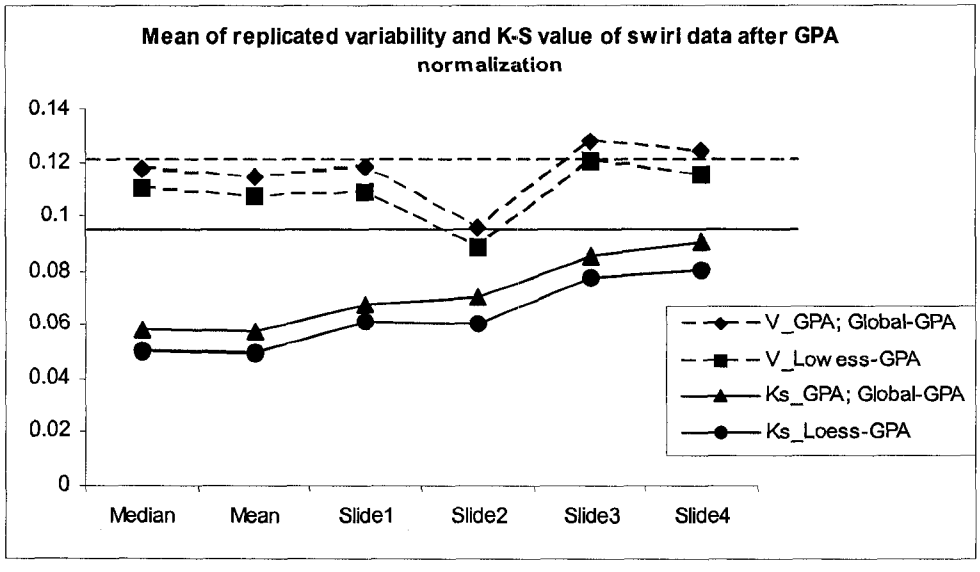

Figure 2.7. Replicate variability and K-S statistic for the swirl zebrafish data set after GPA normalizations with different reference arrays. The upper dashed and lower straight lines indicate variability and K-S value for Lowess method, respectively.

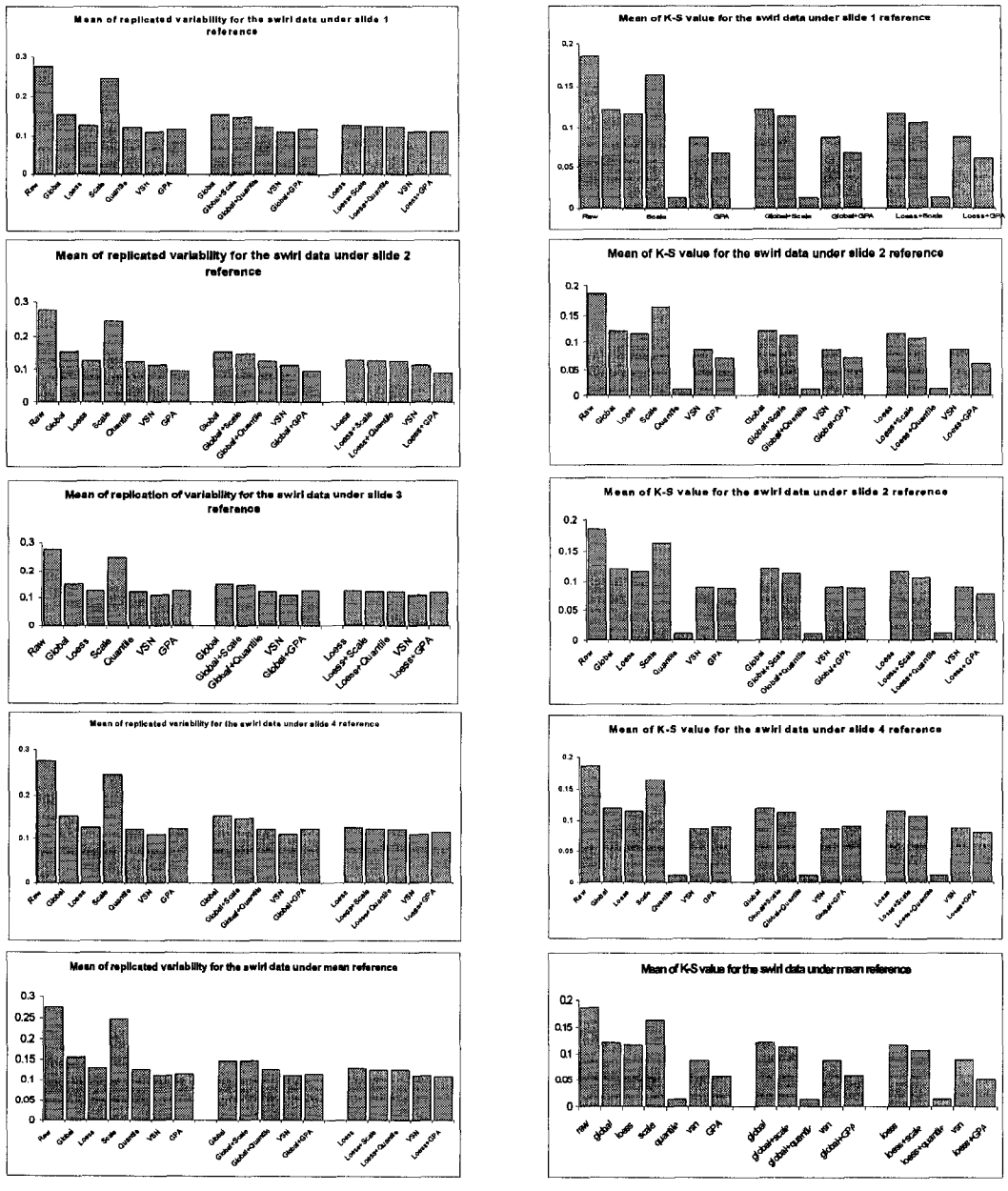

Figure 2.8. Replicate variability and K-S value for swirl zebrafish data after GPA normalizations based on different reference slides, and other normalization methods. Although the GPA normalizations differ based on different reference slides, the overall better performance compared to other methods can be still observed here.

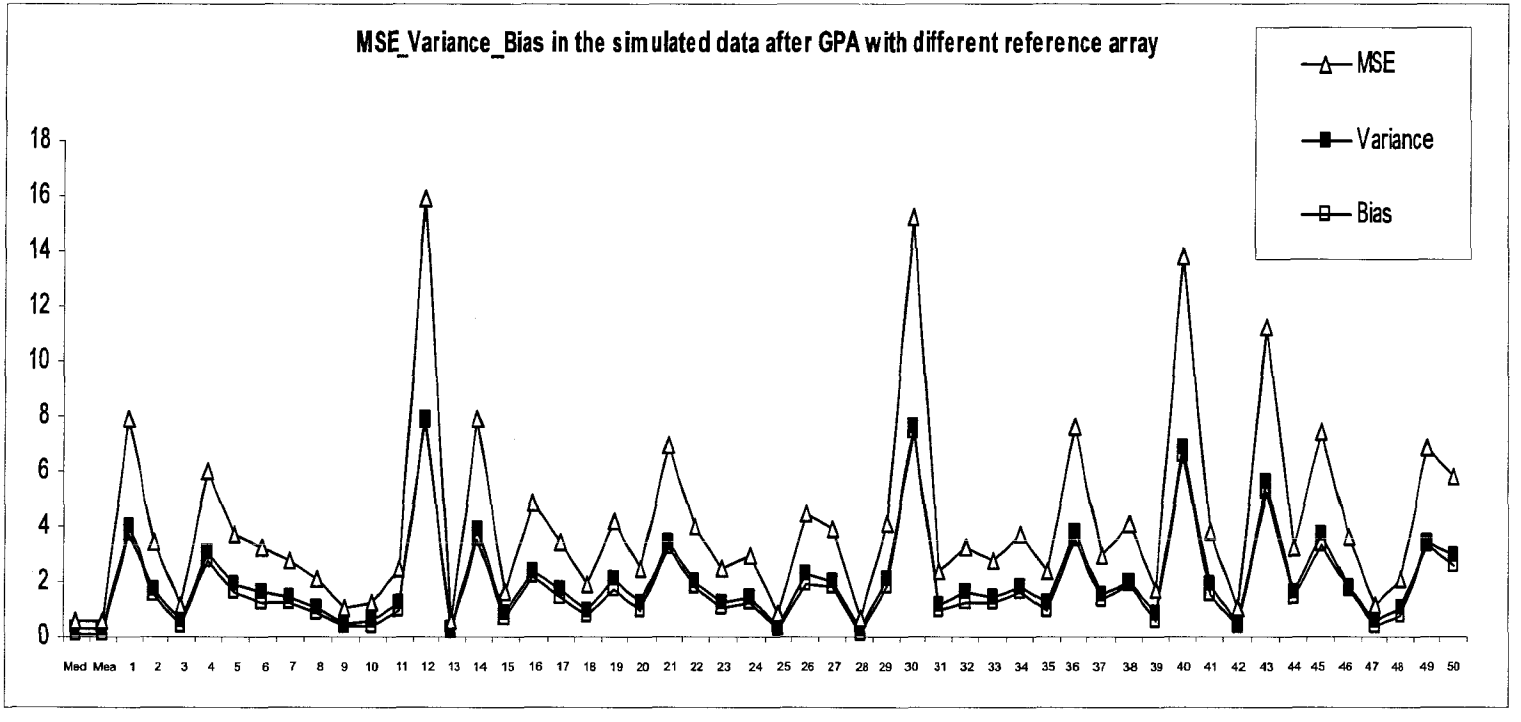

Figure 2.9. The MSE results for simulated data after GPA normalizations with different reference arrays. The data are simulated by SIMAGE method, include 1000 genes in 50 slides with 5% differentially expressed genes and ratio of up-regulated to down-regulated genes is 1:1. The first two tic marks read as Med and Mea, for median reference slide and mean reference slide respectively.

2.4 Discussion and Conclusion

Data normalization is an essential step for the spotted cDNA array and exerts important effects on the subsequent data analysis leading to identification of differentially expressed genes, as well as clustering and pathway analyses. Many normalization methods have been proposed. However, almost all the available normalization methods are based on biological or statistical assumptions about data distribution, which are often not valid in practice. For example, the usual assumption that the array contains a large enough assortment of random genes, most of which are not differentially expressed across the experiments, is inherent in many existing methods including the Global, Scale, Lowess and VSN methods considered here. The assumption is particularly problematic for custom-made boutique arrays where most genes are expected to be differentially expressed, or are directly implicated in the biological process being studied. Although some novel methods have been proposed for such boutique arrays, including the housekeeping gene method [29], Zipf's law method [31] and the mixture model based method [32], these methods have also their own assumptions about data distribution and suffer from similar problems. Therefore normalization methods without such assumptions must be developed. We were motivated to address this question largely because we have developed a custom cDNA array [13]. We demonstrate the potential of a GPA-based method for normalizing microarray data. With this approach, there is no need to assume that expression data follow any particular distribution, for example, that the distributions of up- and down-regulated genes are symmetric.

Procrustes analysis is a powerful least-squares approach by translating, rotating and isotropic scaling to achieve a better fit among matrices of similar configurations. It has been widely applied in many fields such as statistical analysis of shapes [55], analytical chemistry [56], photogrammetry [57] and protein structural alignment [58]. Its strength in dealing with similarity among shapes or configurations led us to consider its potential for microarray data normalization analysis in balancing signal intensity level across different slides. The GPA method only requires replicate slides in the experiment, but conducting replicate experiments is a very popular experiment design since it can greatly ascertain experimental errors and reduce noise bias in the measurement and greatly helps us to analyze the variability across slides.

We utilized both the real and simulated data to compare the GPA-based approach with six other normalization methods including Global, Lowess, Scale, Quantile, VSN, and one boutique array specific housekeeping gene method. For both real data sets, the GPA method showed a significant better performance in decreasing replicated variability and increasing similarity of data distribution than other methods. Global and Scale methods performed worse in both criteria. Both VSN and Quantile performed best only in one case of variance or K-S statistic because their assumptions specifically favour one of them. Although the Lowess can decrease variability, it showed no advantage in decreasing discrepancy of data distribution. A better performance of GPA than other across-slide based methods can still be observed on the same within-slide normalization background. The Lowess-GPA pair performed better than Lowess and GPA alone, and other dual normalizations, implying a better combination potential. Furthermore, we utilized two different parameterized models to simulate several types of microarray data, which were used to accurately evaluate normalization effects through assessing true difference between known true data and normalized data. The results indicate that our GPA method outperformed other methods in reducing across-slide variability and removing systematic bias in microarray data. Overall the GPA normalization can effectively decrease the replicated variability, discrepancy of data distribution among slides and retrieve underlying biological information. It not only can be used individually to balance the bias across slides, but also in practice can be combined with other methods such as Lowess to produce better results. Moreover, the combination of GPA with Lowess can reduce discrepancies in data distribution as a result of Lowess. This is logical because Lowess and GPA have different within-slide based and across-slide based strategies. Furthermore, the application of GPA normalization in analyzing the boutique array was demonstrated in both the real and simulated boutique arrays. The GPA method performs consistently better than the popular boutique array-specific housekeeping gene normalization method.

In conclusion, we have shown that GPA is a promising normalization method for microarray data analysis. In particular, GPA method is free of assumptions about data distribution inherent in other existing approaches. This makes GPA versatile and robust for diverse types of array sets, especially for custom-made boutique arrays where the majority of genes may be differentially expressed.

Chapter 3. Assessing the reliability of goldfish brain cDNA microarray platform

Summary

The goldfish-carp cDNA microarray is a novel gene expression profiling platform designed to identify EDC (endocrine disrupting chemical)-related gene profiles in the goldfish brain. There is a need to assess data quality and sources of errors in microarray experiments prior to any subsequent analysis. Therefore I carried out a series of statistical analyses to examine the inter-array and intra-array reproducibility. The statistical criteria to be measured include array quality filter test, statistical repeatability coefficient, linear correlation coefficient, and coefficient of variation. These investigations demonstrate a high degree of reproducibility both within and among arrays implying that this platform is capable of reliably and reproducibly detecting differences in gene expression patterns between two distinct biological conditions.

3.1 Materials and methods

3.1.1 Data source

The goldfish-carp cDNA microarray V1.0 has grid of 12 rows and 4 grid columns. In each grid, there are 20 rows and 20 columns. The total number of spots in the array is 19,200. We used one set of published data [13] to demonstrate the functionality of goldfish brain cDNA microarray. All derived raw data files are available at the GEO repository (http://www.nebi.nlm.nih.gov/geo/). Platform and series accession number are GPL3735 and GSE7025, respectively. The sample RNAs of both treatment and control groups (Table 3.1) were hybridized on four replicated slides with a balanced dye-swap design. Image files were obtained with ScanArray image analysis software.

Table 3.1. Slide details.

SlideNumber	FileName	Ch1(Cy5)	Ch2(Cy3)
3	slide3.txt	Treatment	Control
15	slide15.txt	Control	Treatment
31	slide31.txt	Treatment	control
41	slide41.txt	Control	Treatment

We extracted raw data including spot and background intensities of both channels, and technical flags from ScanArray output files. On one array, there are 9600 replicated clones including gene clones, positive controls (Alien PCR products, alien mRNA spikes and human actin PCR product), negative controls (alien PCR product, poly(dA) 40-60, salmon sperm DNA and human COT-1 DNA), and some blank spots serving as negative controls. The expression levels of about 8000 cDNA spots were assessed after removing control spots. For each cDNA spot, the local background was subtracted from the respective intensity values.

3.1.2 Data filtering

Spots manually flagged due to poor hybridization and spots in which the estimated fluorescence intensity was below or equal to the estimated background signal intensity in either channel were removed before further analysis. Table 3.2 shows the summary descriptive statistics of our data sets and provides the consistency of the resolved detection signals. Usually, slides with less than 2% of saturated features flagged are considered acceptable [59].

Table 3.2. Summary descriptive statistics for our data sets. The percentage of poor spots is shown. Flags represent the bad spots. Spot < background means the foreground intensity of the spot is lower than background intensity, spot < 2*background means the foreground intensity of the spot is lower than two times background intensity, while spot < 100 + background means the intensity of the spot is lower than background intensity plus 100 value.

	Total number of spots	% flags	%spot <background	%spot <2*background	%spot <100+background
Slide3	19200	3.05	2.84	2.98	27.03
Slide15	19200	0.27	0.66	0.95	18.52
Slide31	19200	0.05	0.82	1.06	17.42
Slide41	19200	0.07	1.8	1.94	18.43

3.1.3 Data transformation and data normalization

We consider the log-transformed data, log ratio intensity $M = log_2(R/G)$ and the mean log intensity $A = \log_2 \sqrt{RG}$ where R and G are the red and green background corrected signal intensities respectively. For n genes and N_{array} arrays, we organized our log-transformed data as Y_{ik}^{l}, the measured expression level of gene i (i = 1, 2,...the total of the number of genes n) and replicate k (k = 1, 2,... m_i) in array l (l = 1, 2,..., N_{array}). The first l = 1 and the last $l = N_{array}$ arrays are obtained under the replicated conditions.

The corrected intensity data were log_2-transformed. The ratio of net fluorescence from the Cy5-specific channel to that from Cy3-specific channel was calculated for each spot. This represents the ratio of the cDNA in the treated vs. untreated goldfish brains. Four independent experiments were performed to reduce variation related to labelling and hybridization efficiencies among the experiments.

We applied the Lowess normalization method [17] to remove intensity-dependent dye differences between channels within each slide.

3.1.4 Methods for assessing repeatability

3.1.4.1 Array Quality Filter test (AQF)

The AQF test is used to filter slides with low quality [60] for experiments which have positive control clones. This test assumes that the replicated log ratio intensities of positive control genes are normally distributed with mean 0. If the log ratio intensity with one positive control spot is in the 70-percent confidence interval, we flag it as "bad". We filter the arrays which have >50% of the positive controls flagged "bad".

3.1.4.2 Mean absolute pairwise deviation

The mean absolute deviation is used to measure the similarity of replicates in each array. The mean absolute pairwise deviation for a spot is defined as the average absolute differences between measurements from paired spots (repeatedly spotted clones). For one array, the mean absolute pairwise deviation is defined as the average taken over all clones.

3.1.4.3 Repeatability coefficient

The repeatability coefficient is an alternative criterion to test correlation values and is an indicator of the internal quality of a single array [61]. For each gene, we define the coefficient of repeatability as 1.96 x standard deviation (SD) of the replicated log ratio intensities, i.e. the 95% confidence intervals of such values. For one array, the repeatability coefficient is defined as the average of those repeatability coefficients over the whole spotted clones.

3.1.4.4 Correlation coefficient

The correlation coefficient is a common approach for assessing the concordance of replications. Both intra-array and inter-array replicates are checked. For all the spotted clones, the average correlation coefficient over the whole spotted clones is an indicator of the repeatability of our whole data.

3.1.4.5 Coefficient of variation

For one gene, the coefficient of variation (CV) is calculated by the standard deviation divided by the mean of the replicated log ratio intensities among inter-array replicates. For all arrays, CV is defined as the average of the CVs over the whole spotted clones.

3.2 Results and discussion

3.2.1 General data quality

Microarray data quality can be assessed from box-plots, density plots and image plots which can be generated automatically. This step identified outlier array. An example image plot which represents four arrays with consistent green foreground intensities is shown in Figure 3.1.

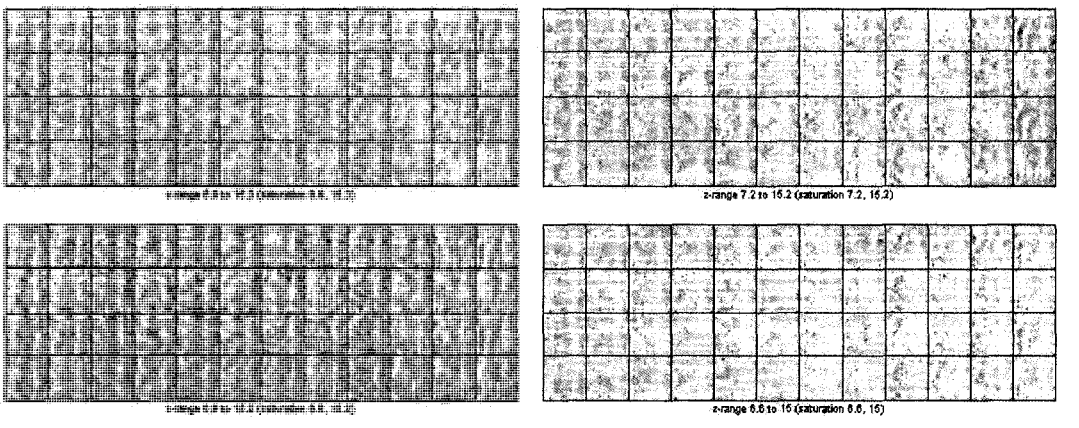

Figure 3.1. Image plots of the green foreground intensity for four arrays.

We next check the box plots for the data after Lowess normalization. The box-plots (Figure 3.2) are centered at zero and have fairly similar spreads, which implies that the data is consistent between slides.

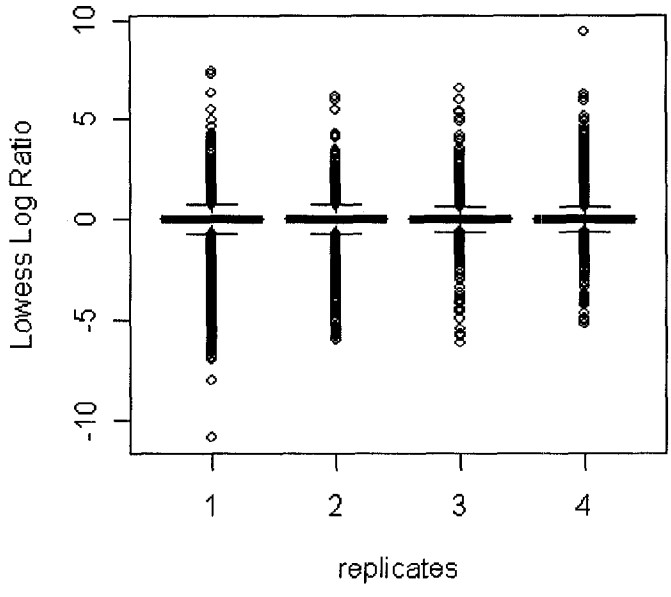

Figure 3.2. Box-plot displaying the log ratio for different microarray replicates after Lowess normalization. This shows the interquantile range, with the 75th percentile at the top and the 25th percentile at the bottom. The line in the middle of the box represents the 50th percentile, or median.

When we applied the normalized data to an AQF test, the values of four replicates were 0.37, 0.36, 0.29, and 0.29 respectively, with all under the threshold of 0.5. Also we generally assess the inter-array reproducibility by comparing the detected spots between replicates. Table 3.3 shows that there is high concordance percentage (96%, 99%, 99%, and 99%) of the spots with background-subtracted fluorescence levels higher than 2-fold background levels in these four slides.

Table 3.3. Comparison of spots with fluorescence levels above threshold in different microarray replicates. Values correspond to the percentage of spots selected in the replicate indicated on the left that are also selected in the replicate indicated on top.

Number of replicate	Number of replicate			
	Slide3	Slide15	Slide31	Slide41
Slide3		99	99	99
Slide15	97		99	99
Slide31	96	99		99
Slide41	97	99	99	

We next determined dynamic range and sensitivity levels using series of PCR controls, which are composed of 10 sets of PCR products (PCR1, PCR2, ...PCR10). Figure 3.3 shows that the intensities of these control PCR spots covers the whole range of intensities. And all of them lie close to the horizontal line corresponding to zero, which indicates they are not differentially expressed in the array.

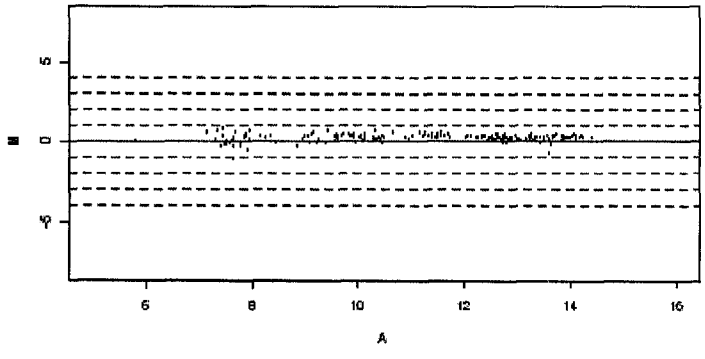

Figure 3.3. MA-plot for PCR control spots.

3.2.2 Statistical criteria for assessing intra-array and inter-array repeatability

We assessed intra-array repeatability using mean absolute pairwise deviation and repeatability coefficient. The intra-assay analysis compared the $\log_2$ (Cy3/Cy5) value of each spotted cDNA clone to its printed duplicate within a single array. As shown in Table 3.4, the values for these two criteria are between 0.2 and 0.4. This value range is consistent with those previously reported for replicate sample analysis [61] and indicates that the repeatability for intra-array is acceptable.

Table 3.4. Summary table of two statistical criteria for measuring intra-array duplication

	Mean absolute pairwise deviation	Repeatability coefficient
Slide3	0.297 (0.157)	0.369 (0.037)
Slide 15	0.296 (0.115)	0.398 (0.02)
Slide 31	0.253 (0.082)	0.341 (0.02)
Slide 41	0.262 (0.165)	0.348 (0.040)

We further verified the intra-array replication by applying the correlation coefficient and the coefficient of variance (CV). Figure 3.4 (A) and (C) show the results of Cy3 (Cy5) from two slides. The regression line, which is created from the data comparing replicated spots in one array, remains similar to the line of symmetry. Most duplicated measurements were located within the 99% confidence interval. The CV values for two comparisons are 10.7% and 12.8%; both are less than the 20% cutoff. One previous study [62] has shown that CVs between technical microarray replicates usually range between 10-20%. In addition, the correlation coefficients (0.95 and 0.96) show good agreement between the intensities of the replicates spots. Overall, the intra-array Cy3 (Cy5) measurements are highly reproducible between the duplicated spots on our cDNA microarray.

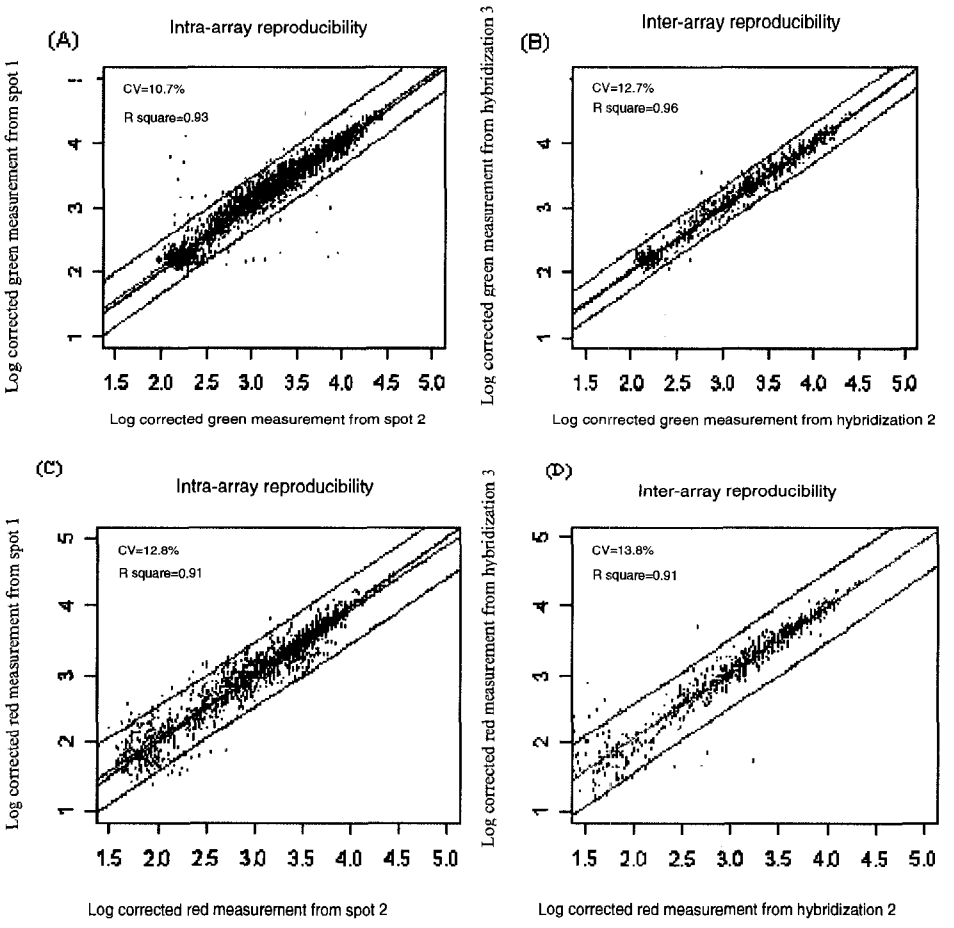

Figure 3.4. Statistical analysis of array variation. (A), (C) are for intra-array variation. The data is the log intensity of green (red) spots in the third slide. The vertical axis and horizontal axis are corresponding duplicated values from spots. The median coefficients of variation (CV) are 10.7% and 12.8%. The correlation coefficients(r) are 0.96 and 0.95. They indicate a high degree of agreement between duplicates. (B), (D) are for inter-array variation. The data is the log green (red) intensity from spots between the first and second hybridization. The horizon axis represents values from the second hybridization. The vertical axis represents the values from the first hybridization. The mean coefficients of variation (CV) are 12.7% and 13.8%. The correlation coefficients(r) are 0.98 and 0.95. They indicate a high degree of association between different experiments. Middle lines in (A-D) means the regression lines. The blue lines represent the predicted 99% confidence intervals of the data.

The inter-array replication was also checked by correlation coefficient and coefficient of variance. We plotted the measurement of log average Cy3 (Cy5) measurement between two cDNA microarray slides, as shown in (B) and (D). Again, the regression line of this comparison was identical to the line of symmetry and most measurements fall within the 99% confidence interval. The correlation coefficients are 0.98 and 0.95. Both CVs (12.7% and 13.8%) were less than 20%. The results for the other pairwise hybridized slides were similar (not shown here).

3.3 General conclusion

Microarray quality greatly affects data analysis. Determining the quality of microarray prior to further analysis is crucial for reliable data mining. In this chapter, I examined four biological repetitions of a single experiment to check the reliability of goldfish microarray system. Not only general data distribution profiles were provided here, but also statistical and systematic errors were quantified by using a series of statistical methods. Box plot, PCR control analysis, and AQF test all showed that goldfish brain microarray exhibits an overall technically good quality.

The intra-array and inter-array replication were extensively examined based on replicated hybridizations. All the statistical parameters showed that there was a high degree of consistency on each slide and between replicates on different slides. It implies that the goldfish-carp microarray platform is a reliable tool for profiling differences in gene expression between samples.

Chapter 4. A pipeline for cDNA microarray data analysis

Summary

Extracting meaningful biological information from microarray data is a major challenge. Various methods and algorithms have been created to facilitate the analysis of DNA microarray data, with a typical pipeline including data pre-processing, data normalization, differential gene identification, multivarite statistical analysis and gene ontology annotation. Unfortunately, bioinformatic tools for these different steps are often poorly documented and difficult to use, requiring a professional bioinformatician well versed in statistical analysis. Therefore, there is a crucial need of an integrative and user-friendly program suite to ease microarray data analysis in molecular biology laboratories with limited bioinformatic support.

Here I present a new R-based program suite, called GoldR, which has integrated the most popular microarray data analysis methods including data filtering, imputation for missing values, comprehensive normalization analysis, differential gene identification, and clustering analysis. This suite has been tested and refined over the last few years by analyzing the goldfish-carp cDNA microarray data. This chapter includes detailed instructions for using this program as well as the conceptual background for each data-processing step in the pipeline.

4.1 General introduction to GoldR program suite

This program suite is coded in the R environment and works on a Windows platform. The analysis procedure can be divided into several stages: uploading files and data visualization, quality control, data normalization, imputation of missing values, differential gene identification, and cluster analysis. Among these stages, the quality control stage has been described in chapter 3. A visual screenshot of GoldR opening menu is shown in Figure 4.1.

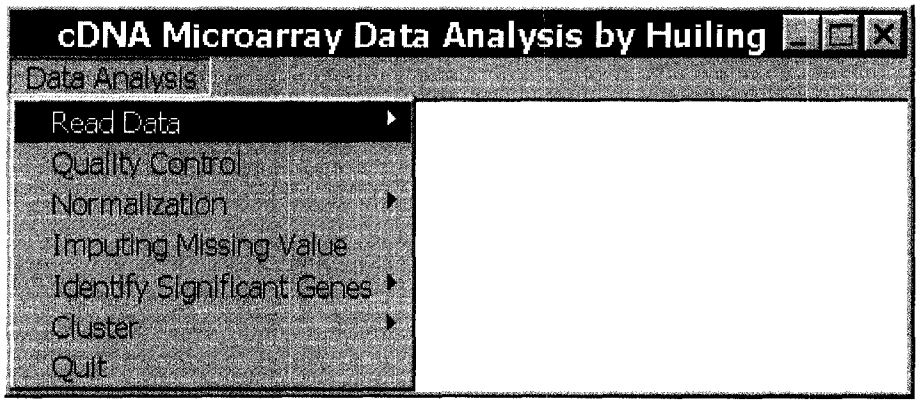

Figure 4.1. A visual screenshot of GoldR analysis suite showing six components of the pipeline including reading data, quality control, normalization, imputing missing value, identifying significant genes and cluster.

4.2 File input and data visualization.

GoldR takes three kinds of files as input: 1) light intensity files: text files exported by an image analysis program in specified format; 2) sample target files: a tab-delimited text file listing the targets hybridized to each channel on each array; 3) a control gene list file: a text file containing slide coordinates of control genes.

In this step, some visualization figures will be generated to give an overall view of hybridization patterns of microarray slides. For example, a spatial spot image allows us to detect scratches and artefacts on the each array. Also we can assess the necessity of background correction and evaluate spatial homogeneity of the data with these visual inspections. As shown in Figure 4.2, the water ink and local irregularities on the arrays are apparent.

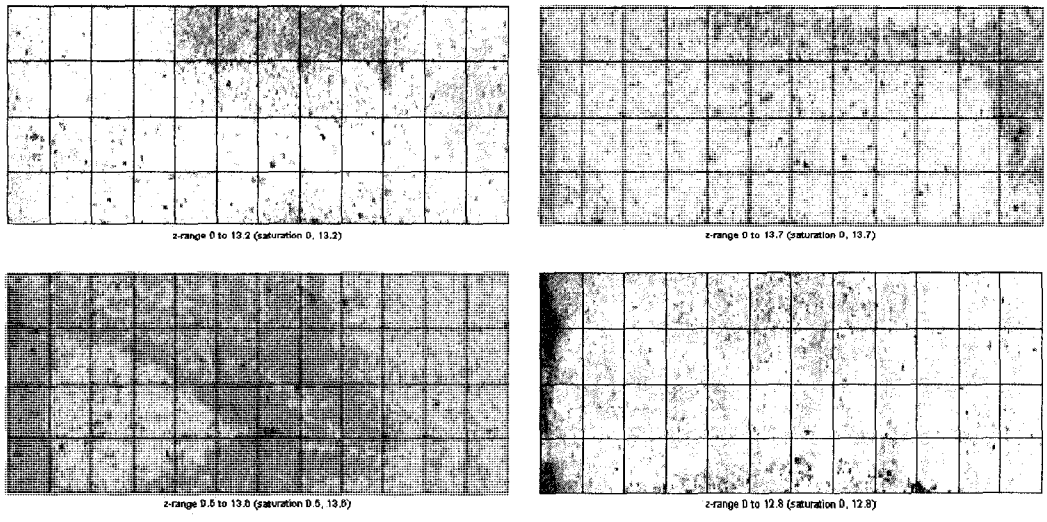

Figure 4.2. Image plots of the red background intensity for four arrays.

4.3 Data filtering

cDNA microarray data often contain outlier spots due to the scratches or dust on the surface or imperfections during array production or hybridization. Such spots should be removed at the very beginning of any analysis. First, the spots manually flagged due to poor hybridization and the spots for which the estimated fluorescence intensity was below or equal to the estimated background signal intensity in either channel will be removed. Second, for the replicate experiments, the genes whose intensities cannot be detected in more than 20% replicates will be removed.

4.4 Data normalization

The cDNA microarray, as with all experimental technologies, has its own limitations. cDNA microarray measurements are quite noisy and may also be plagued by systematic biases. Many steps in the experimental process can introduce substantial variability in the quality of the data: mRNA sample preparation, reverse transcription, possibly amplification

and labelling, PCR probe preparation, slide preparation, hybridization, laser scanning and image processing. In order to measure gene expression accurately, it is important for us to take into account these random and systematic variations. Therefore, normalization is an essential preprocessing procedure to make sure that (1) data from the experimental treatment and the control are comparable, (2) data from different arrays are comparable and (3) variation across replicate arrays is minimized as much as possible.

A number of publicly available normalization approaches have been developed, however there is no gold standard which can be applied to all cDNA microarray data. This program suite combines several of the most popular normalization methods. The strengths of various methodologies will be comprehensively discussed below.

4.4.1 MA-plot in normalization

Before introducing normalization methods, we describe the MA-plot, with $M = \log_2(R/G)$ for all spots on the array plotted as a function of $A = log_2(R*G)$. This MA-plot can help reveal the intensity-specific artifacts in the $log_2 (R/G)$ measurements. Figure 4.3 is the visualization of the MA-plot of raw data in one array. Note that, if there is no intensity-dependent dye bias and no differential gene expression (i.e., no difference between R and G globally), then the points are expected to distribute equally above and below the zero horizontal line. This is not the case for data plotted in Figure 4.3, which suggests the necessity of within-slide normalization.

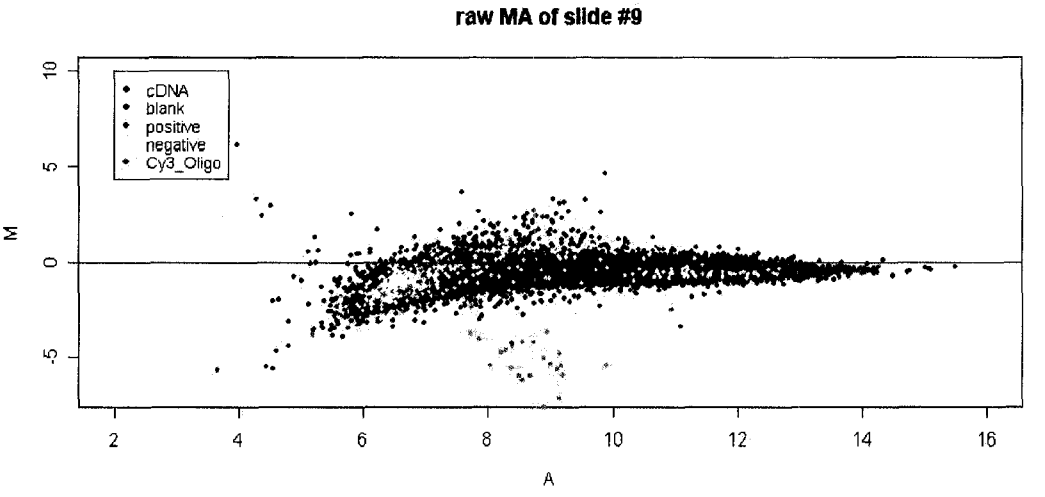

Figure 4.3. MA-plot of raw data for one slide

4.4.2 Global normalization

The Global normalization method assumes that most genes in a microarray experiment do not change their expression between control and treatment, and that only a small fraction of genes are differentially expressed, i.e. the center of the distribution of log ratio in each slide is expected to be zero.

We set $M_i = \log_2(R_i / G_i) = \mu_i + c + \varepsilon_i$, i= 1, 2,..., the number of genes, where M_i is the difference in log expression values between red and green intensities for *i-th* gene. G_i and R_i are the measured green and red intensities, μ_i is the true log ratio and ε_i is the error term which is independent and identically distributed with expectation zero and finite variance. Based on this approach, the shift factor c can be estimated by the median or the mean of the log ratios denoted by $\hat{c}$.

After global normalization, $\hat{c}$ is subtracted from all log intensity ratios, i.e. $\hat{M}_i = \log(R_i / G_i) = \mu_i - \hat{c}$, where i= 1, 2,..., the number of genes on the array.

The MA-plot is then shifted to the line M equal to zero by subtracting M_i by the mean or median. This can be illustrated by comparing Figure 4.4 with Figure 4.3. The data set shifts all data points in the up by the same amount; only the intercept of the x-axis changed.

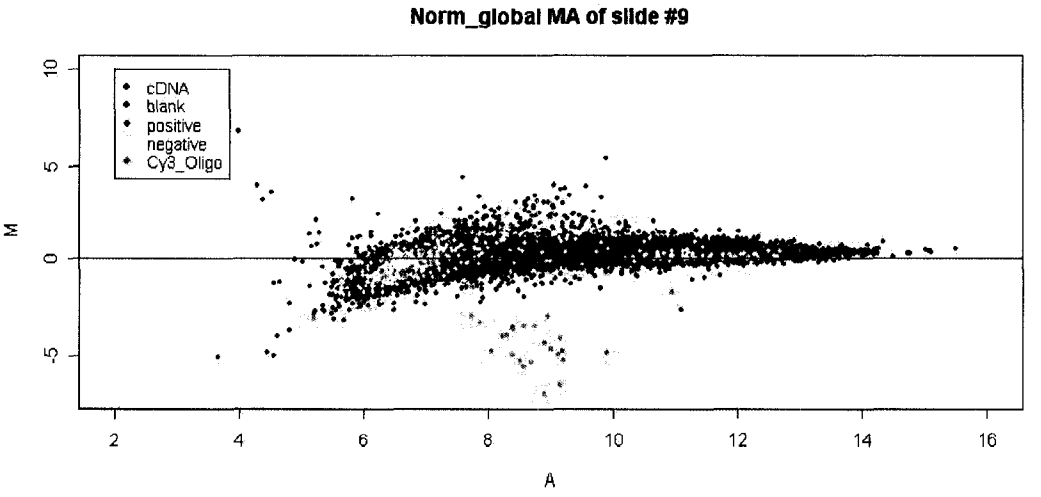

Figure 4.4. MA-plot of global normalized data for one slide

Global normalization is not usually recommended since it is a very simple adjustment of the median intensity value within each array and does not take account of the intensity-dependent bias of the MA-plot. In other words, the result from global normalization does not achieve the expected distribution with the expected M_i values distributed equally above and below the zero line over the entire intensity range. For this reason, other normalization methods have been developed.

4.4.3 Lowess normalization

Yang et al [17] proposed Lowess normalization to remove nonlinear intensity-dependent bias. In the literature, the terms Lowess and Loess are often used as synonyms, but they are in fact different. Lowess is an implementation of local regression involving two variables [63], Loess [64] is the implementation of a more advanced local regression method supporting not only univariate and multivariate predictor variables, but also local linear and local quadratic fits [65]. For microarray data, we have only two variables, M and A, so the local regression is done with Lowess or similar computational tools.

The Lowess algorithm assumes that up-regulated and down-regulated genes are balanced over the entire intensity range, e.g., 5 genes up-regulated by two-fold in the neighbourhood of intensity A_i are balanced by 5 genes down-regulated by two-fold. Such an assumption is generally invalid when only a small fraction of genes are differentially

expressed so that almost all points are expected to be equally distributed above and below the expected zero line over the entire intensity range.

After locally regressing M on A to obtain the predicted M_i value (M_i^p), we obtain the normalized value (M_i^n) as ($M_i - M_i^p$). This removes the strong intensity-dependent effects (Figure 4.5, where M represents M_i^n).

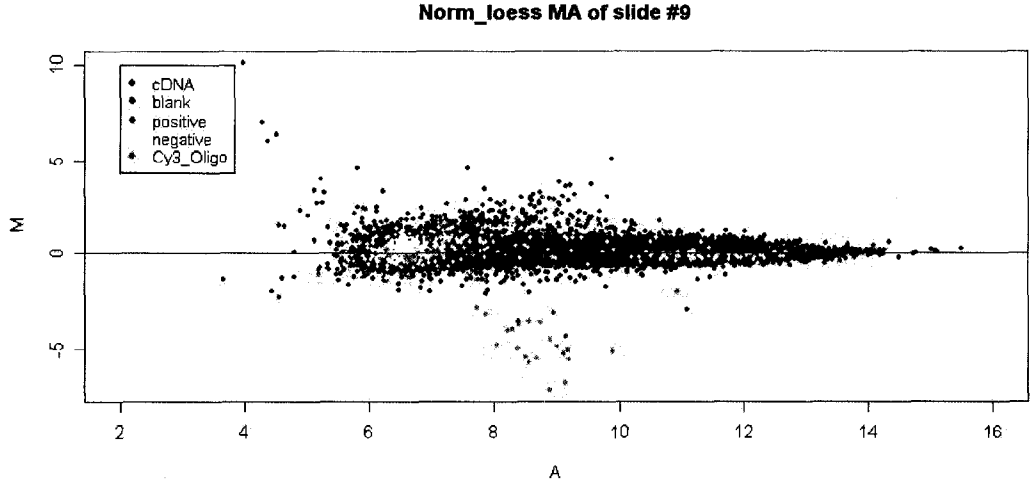

Figure 4.5. MA-plot of Lowess normalized data for one slide.

4.4.4 Print-tip normalization

Print-tip normalization removes systematic bias among print-tips. When control genes (i.e., genes whose expression is expected to be the same over all print-tips) are available, the print-tips can be normalized to have the mean of $M_i = M$, where i stands for print-tip i. When there are no control genes available, print-tip normalization assumes that average gene expression is the same across all print-tips, which may be a risky assumption. The effect of such normalization is illustrated in Figure 4.6.

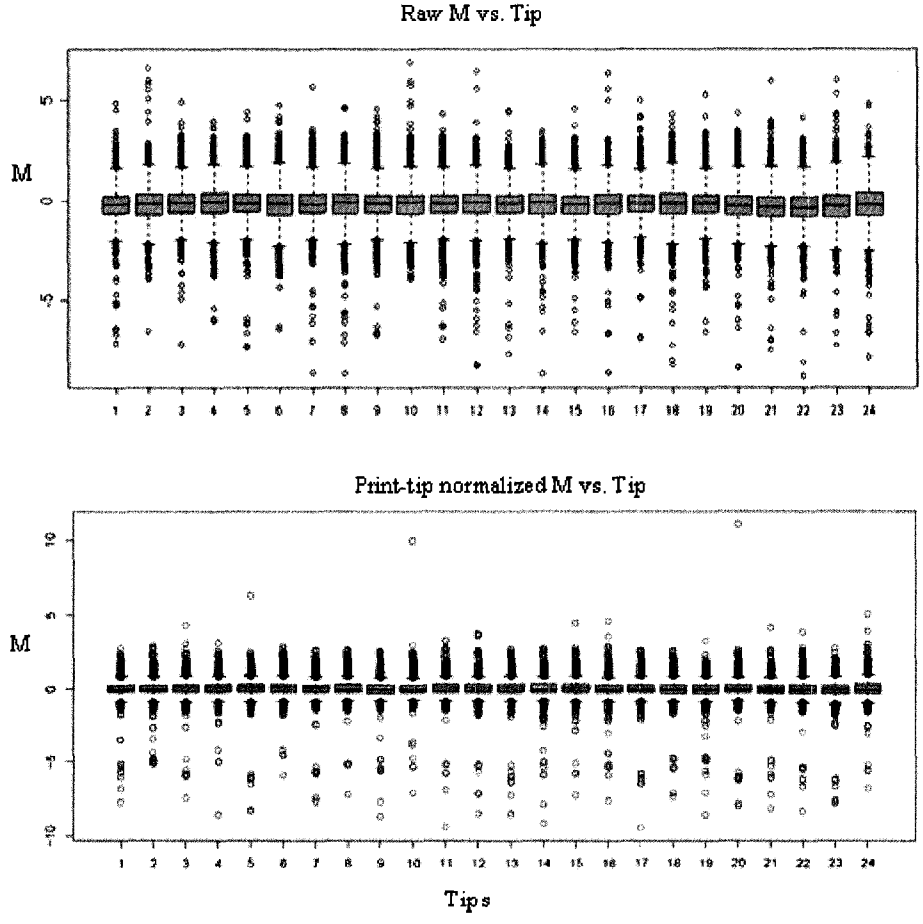

Figure 4.6. Boxplots of M values before and after print-tip normalization by tips.

4.4.5 Dye swap

Dye Swap is a widely used technical replication in two-color spotted array analysis [66]. For this strategy, the same labelling protocol is utilized throughout the experiment, whereas the dye is exchanged between the treatment and the reference samples (the reference is either a control or a pool of controls). This approach allows one to identify dye bias, if present. For example, suppose we have two samples, C and D. In the first hybridization, we label cDNA sample C with red dye and D with green dye and reverse the dye labelling in the second, so that the ratios for our measurements can be defined respectively as

$$M_i^1 = \log_2(R_i^1 / G_i^1) = \log_2(C_i^1 / D_i^1)$$

$$M_i^2 = \log_2(R_i^2 / G_i^2) = \log_2(D_i^2 / C_i^2)$$

$$A_i^1 = \frac{1}{2}\log_2(R_i^1 \times G_i^1) = \frac{1}{2}\log_2(C_i^1 \times D_i^1)$$

$$A_i^2 = \frac{1}{2}\log_2(R_i^2 \times G_i^2) = \frac{1}{2}\log_2(D_i^2 \times C_i^2)$$

As we are making two comparisons between identical samples, we expect

$$\frac{C_i^1}{D_i^1} \times \frac{D_i^2}{C_i^2} = 1 ,$$

which leads to the normalized data $M_i = \frac{M_i^1 - M_i^2}{2}$, $A_i = \frac{A_i^1 + A_i^2}{2}$.

This method effectively removes dye bias. However, this benefit is achieved at a high costs: two dye-swap arrays yield only one normalized data a set.

4.4.6 Scale normalization

Scale normalization [24] is typically used to normalize microarray data over replicated arrays. It assumes the range of the log ratio is the same among the slides (which is true for replicated slides). The transformed log ratios M_i of the *s-th* replicated slides are assumed to follow a normal distribution with expectation zero and variance $a_s^2\sigma^2$, where σ^2 denotes the variance of the true log ratios and a_s^2 denotes the scale factor for the *s-th* slide and s=1,2,... N_{array}, the total number of arrays.

Given the constraint $\sum_{s=1}^{m} \log_2 a_s^2 = 0$, the scale factors are estimated by $\hat{a}_s = MAD_s / (\prod_{j=1}^{N_{array}} MAD_j)^{1/N_{array}}$, with the median absolute deviation *MAD* defined by

$$MAD_s = median_{i=1,..,n}\{|M_i - median_{k=1,..,n}(M_k)|\}$$

Scale normalization between slides is then performed by dividing the log ratio M_i of each slide by the corresponding scaling factor $\hat{a}_s$. The homogeneity in log ratios between slides will be achieved. Figure 4.7 shows the boxplot of log ratio intensities before and after scale normalization among slides.

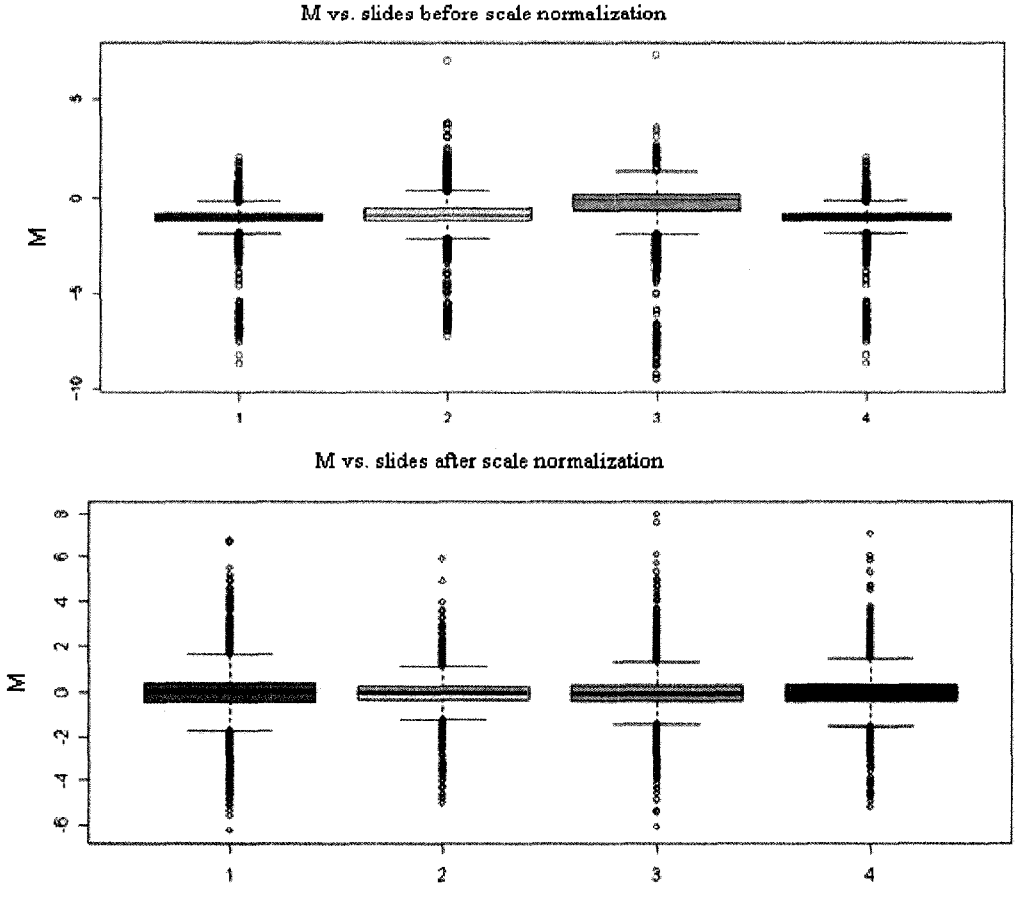

Figure 4.7. Log ratio intensity boxplots before and after scale normalization between slides.

4.4.7 Quantile normalization

Quantile normalization method is widely applied for single channel arrays [25]. It can also be used in between-slide normalization for cDNA microarray. The goal of Quantile normalization is to make the distribution of log ratio intensities the same for multiple arrays.

The rationale behind the method is an n-dimensional Q-Q plot shows that the distribution of multiple data vectors is the same if we project the points of n-dimensional Quantile plot onto the diagonal. The algorithm of Quantile normalization is explained as follows.

Designate Y as the matrix of log intensity data, with columns being the microarrays, and rows being genes.

1. Sort $Y \rightarrow Y_{sort}$ by array.

2. Calculate the mean of every row across all arrays and replace all values of each row by this mean, i.e. $Y_{sort} \rightarrow Y'_{sort}$.

3. Rearrange each array of Y'_{sort} to the same order as $Y \rightarrow Y_{norm}$.

After this transformation, all arrays have the same distribution of log ratio intensities as shown in Figure 4.8.

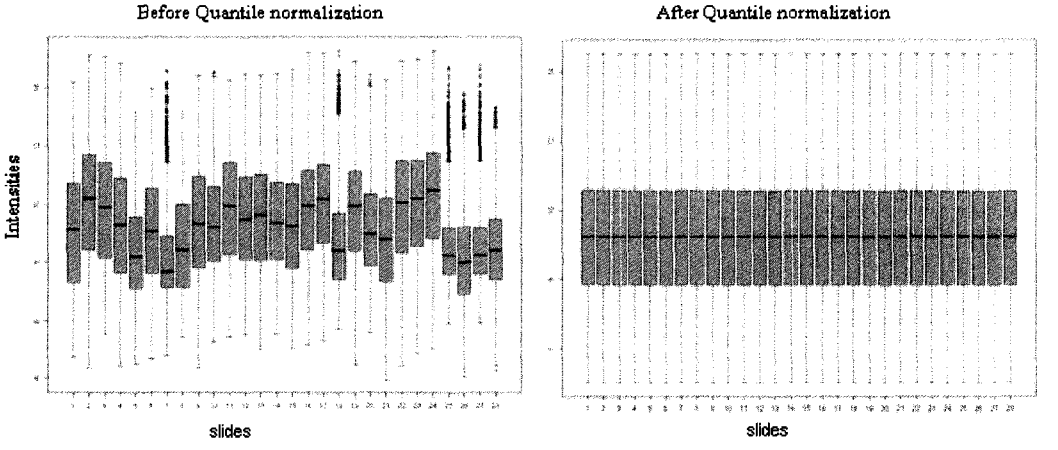

Figure 4.8. Boxplots of intensities before and after Quantile normalization

4.4.8 Variance stabilization normalization (VSN)

VSN [26] is a direct intensity-based normalization method. Its assumption is that most of genes are not differentially expressed in a given experiment. This method helps to remove systematic dye-bias and make the mean and the variance of the transformed data approximately independent of the mean intensity. It is given by $Y'_{ij} = \arcsin e(Y_{ij})$, where the intensity Y_{ij} is the value of gene i and replicate j for each channel.

4.4.9 Generalized Procrustes Analysis (GPA)

Generalized Procrustes analysis is a least-squares approach which uses translation, rotation and isotropic scaling to achieve a better fit among matrices of similar configurations. The recently developed across-slide normalization method based on GPA [67] is relatively independent of the assumptions inherent in other across-slide normalization methods. This makes it versatile and robust for diverse types of array sets, especially for custom-made boutique arrays where the majority of genes may be differentially expressed.

4.5 Estimation of missing values

Some statistical methods can handle data sets containing missing values such as those of multivariate data. However, many methods assume that the data set is complete. Thus, before using such methods, we need to estimate the missing values, a statistical protocol often referred to as imputation.

There are many methods for imputing missing values [68]. The simplest is to replace the missing value of a gene in one array by the value of the same gene in a replicated array. When there are N replicated arrays, then the missing value for a gene in array 1 can be replaced by the mean of the other N-1 values in the replicated arrays. This is often referred to as the row-average method. Such a treatment suffers from the problem that the array with the missing gene value may have much higher or lower average intensity than other replicated arrays. More advanced methods typically involve regression approaches. Of the three methods evaluated in the microarray context, i.e., a singular value decomposition (SVD) based method, weighted K-nearest neighbors (KNN), and the row average, the KNN regression method was found to outperform the other methods in terms of robustness and accuracy [68].

4.6 Identifying differentially expressed genes

Most cDNA microarray experiments aim to identify differentially expressed genes under particular experimental conditions involving a control and a treatment. The problem is that the number of tests (genes) to be investigated is always far greater than the limited number of arrays, leading to practically uncontrollable experimental error rate when the number of comparisons is very large. Here I review four widely used methods that alleviate this problem.

4.6.1 False Discovery Rate (FDR)

After the final normalized dataset has been generated, we can blend fold changes with statistical tests to evaluate the differential expression. This statistical validation is essential because the simple-minded fold change is very unreliable and may have variable significance depending on expression level. Normally, we can use some parametric or nonparametric tests such as t-test or rank test. The volcano plot (Figure 4.9) shows the association between the *p*-values and the $\log_2$ of the fold change. In addition, the false positives will increase sharply with the number of hypotheses. Therefore, multiple testing needs to be adjusted for assessing the statistical significance of findings. The false discovery rate (FDR) [69] is a strategy to provide the most accurate correction for multiple experiment .

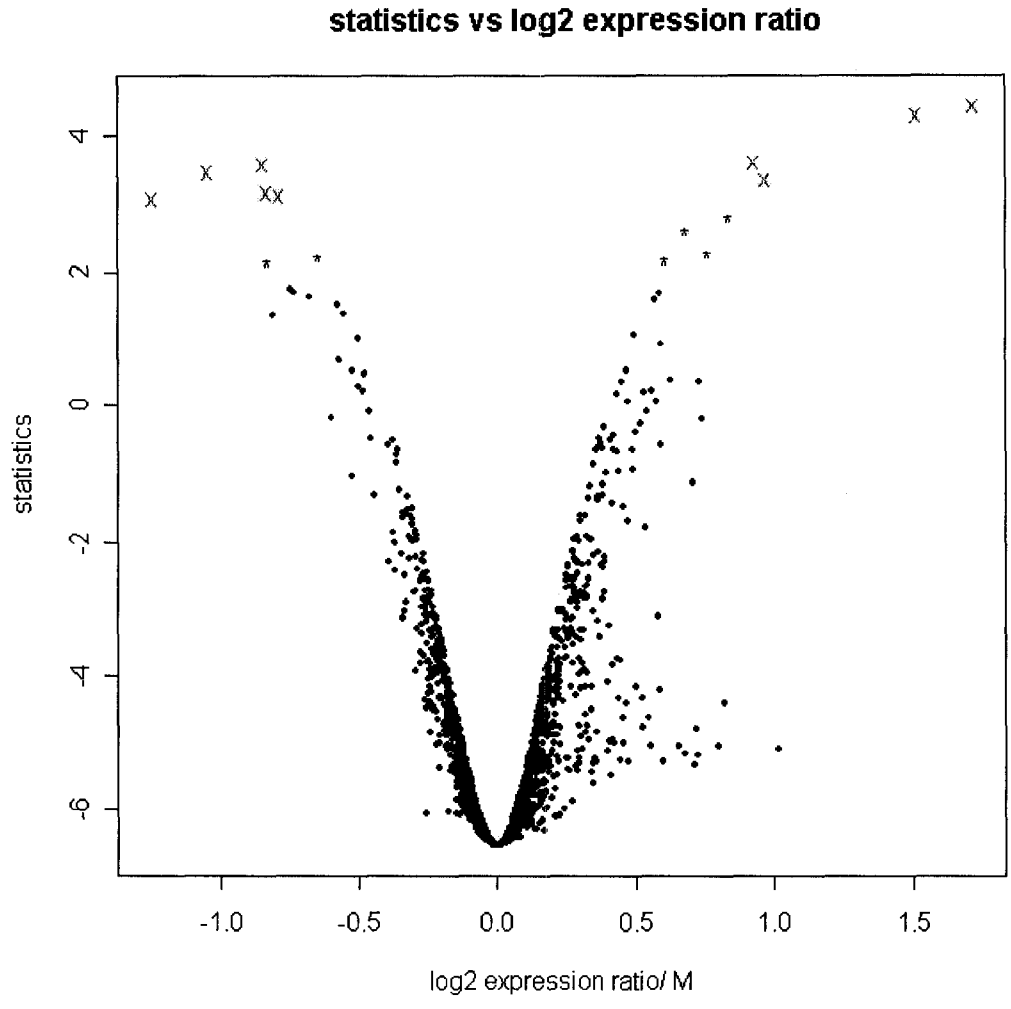

Figure 4.9. Volcano plot showing the association between the *p*-values and the $\log_2$ of the fold change.

FDR estimates the expected proportion of false positives in the resulting gene list through a permutation scheme based on the chosen cutoff-value for the test statistic. i.e. it is defined as $FDR = (p * n)/i$, where p is the mean of p-values of the gene among the permutations, n is the total number of genes and i is the number of genes with p value at or better than p.

4.6.2 Empirical Bayes statistical analysis

The Empirical Bayes statistical analysis [70] is an approach to shrink the estimated sample variances towards a pooled estimate. Thereby it makes more stable inference especially when the number of arrays is small. It creates a B-statistic which estimates the posterior log-odds that a gene is significantly expressed. The result will show more significance if the B value is higher. The B-statistic is defined by

$$B = \frac{\bar{x}_{Treatment}(i) - \bar{x}_{Control}(i)}{\sqrt{(a+s^2)/n}}$$

where the penalty a is estimated from the mean and standard deviation of the sample variances s^2.

4.6.3 Significance Analysis of Microarrays (SAM)

SAM, developed by Tusher and his coworkers [71], is a modified t-test statistic exploring sample-label permutations to evaluate statistical significance. This method adds a constant value s_0 to the standard deviation and is calculated from the distribution of gene-specific standard errors. It is designed to control the FDR and to eliminate small variances. This minimizes the dependence of the t-test variance on standard deviation levels. Another benefit of SAM is that it does not make any strong parametric assumptions and only involves only order statistics rather than the complex estimation procedures. The SAM statistic S for gene i is defined as:

$$S = \frac{\bar{x}_{Treatment}(i) - \bar{x}_{Control}(i)}{s(i) + s_0},$$

where $\bar{x}_{Treatment}(i)$ and $\bar{x}_{Control}(i)$ are the average levels of expression for gene i in treatment and control experiments, $s(i)$ is the standard deviation of repeated expression measurements and s_0 is the fudge factor.

4.6.4 Cyber-T

Cyber-T [72] assumes that genes of similar expression levels have similar measurement errors. It models the standard deviation as a function of signal intensity by using a Bayesian probabilistic framework. The background variance for each gene is

estimated by using the number of neighbouring genes of similar expression level to balance the experimental fluctuations within a limited numbers of replicates. The Cyber-T statistic is an equivalent regularized Bayesian t-test and is defined as the following:

$$CyberT = \frac{\bar{x}_{Treatment}(i) - \bar{x}_{Control}(i)}{\sqrt{\frac{v_0 s(i)^2 + (n-1)s_0^2}{v_0 + n - 2}}}$$

It also has the ability to estimate experiment wide false positives and negatives based on p-value distributions.

4.7 Cluster analysis

One of the main objectives in gene expression studies is to identify co-regulated genes or visualizing similarities in gene expression using cluster analysis. During this process, the genes will be grouped into several distinct clusters according to their expression patterns over the probed biological conditions. The hypothesis about the co-expressed genes is that they may be regulated by the same set of transcriptional factors or they may be involved in the same biological process. The common clustering algorithms include hierarchical clustering [4] , self-organizing maps [73], k-means, principal component analysis [74], and support vector machines [75]. All of them calculate the pair-wise similarity or dissimilarity of gene expression profiles to classify individual genes into gene clusters. Similarity distance measurements include simple Pearson correlation, the jack-knife correlation, mutual information or Spearman's rank correlation coefficients. For dissimilarity measures, the Euclidean distance, Manhattan metric, chord distance and geodesic distance have been proposed. Euclidean distance or Pearson's correlation is the mostly usual estimation of the distance between expression profiles.

It should be noted that due to the underlying assumptions in each clustering algorithm and necessity to adjust various parameters, the normalization methods used to decrease the bias within or across slides, and the used similarity measure, the outcome of the cluster analysis may differ substantially. Thus, it is imperative to apply different clustering algorithms and parameter values on the same dataset and illuminate different relationships between the data. The comparison of gene expression clusters obtained by using several

methods can provide the researcher with additional information [76]. In the following parts, three popular clustering algorithms including hierarchical clustering, self-organizing maps, and k-means will be introduced. More detailed discussion may refer to the book of Bioinformatics and the Cell [77].

4.7.1 Hierarchical Clustering

Hierarchical Clustering [4] is the most commonly used method for exploring the inherent class characteristics between gene expression profiles. It may be visualized as a tree-like dendrogram in which each cluster is nested in a parent cluster. The leaves of the tree are the genes which represent the smallest clusters measured by the distance between two gene expressions such as Euclidean distance or Pearson correlation. At each subsequent node of the tree, the two nearest clusters are grouped together to form a bigger cluster. This procedure is iterated until the root note of the tree is obtained a single cluster containing all genes. Besides the distance measure, we also need to decide the linkage rule to determine if two clusters are sufficiently similar to be linked together. The type of linkage includes single linkage, complete linkage and average linkage. The most attractive aspect of hierarchical clustering is that we only need to specify the distance measurement and the linkage rule. It allows us to investigate the similarities of either neighbouring genes or experimental samples.

4.7.2 Self-organizing maps (SOM)

An unsupervised learning algorithms, SOM [73] takes data that do not have prior group affiliation. It takes a training data set and goes through a training process to obtain a SOM of nodes (or artificial neurons) which can then be used for classification. Since SOM is also one of the artificial neural network (ANN) algorithms, it is necessarily associated with concepts such as nodes (neurons) and learning rate [77].

SOM are maps of a small, usually two-dimensional space in which each point represents a cluster. During the clustering algorithm, a mapping function is automatically built to assign the genes into one of the points of the map in which the points located closely are also similar in the original m-dimensional space.

4.7.3 K-mean

K-mean partitions an m-dimensional space of genes into k groups in order to minimize the variance of the data within each region. Firstly, it randomly selects k data points as the centroids of k different clusters and assigns each of genes to the cluster with the closest centre. Secondly the centroid of each cluster is computed and the cluster centers are updated. The genes are reassigned to the nearest clusters. The procedure is iterated until assignment of data points is unchanged and it converges to a stable solution.

One drawback of K-mean algorithm is that the number of clusters (k) must be specified at the beginning; however the true k is unknown to the researcher. If the guessed k value is too large, then co-expressed genes may be split into different clusters. If the value is too small, then unrelated genes may be forced into the same cluster. This problem is shared with self-organization maps. A practical solution may be to begin with hierarchical clustering to get a first impression on the number of patterns hidden in the dataset and then use this information to set the number of clusters required for k-means and self-organizing map techniques.

4.8 Conclusions and Discussion

This chapter described basic analysis subunits in the GoldR program suite and the pipeline instructions for using those programs. We have shown that these programs are of high quality, capable of reproducibly detecting gene-expression differences between two different samples, and are useful for identifying new biological associations. Also this new program suite is user-friendly and allows a biologist to perform all the functions without comprehensive statistical and computer knowledge.

This program takes only a couple of minutes to complete a computation. The output files include tab-delimited text files and graphic plots that can be saved in the user's directory. Based on the options of analysis, the text file may contain the gene information, M, moderated-T statistics such as B statistic, Cyber-T statistic, S statistic and the corresponding p values, FDR. The table of genes can be either unranked or ranked by M, p value or statistics. Graphic plots include array image plot, volcano plot, MA-plot and boxplot for each array (before and after within-array, between-array normalization), print-tip Lowess plots, heatmap plots and other clustering plots.

Chapter 5. Extracting seasonal gene expression information from multiple goldfish brain microarray data sets

Summary

In this chapter, a series of experimental data from the *AURATUS goldfish environmental genomics project* was used to extract seasonal gene expression information using a novel analytical strategy combining comprehensive data transformation and normalization, differential gene identification, and multivariate analysis. Multiple independent microarray data sets with mRNA source from both female goldfish hypothalamus (Hyp) and telencephalon (Tel) were studied. All these data sets are classified into three seasonal time points: May, August and December, which correspond to physiologically distinct periods, namely sexually mature prespawning, sexually regressed, and early gonadal redevelopment, respectively. A series of within-slide and between-slide normalizations were utilized to remove systemic and experimental bias. About 1000 genes in female Hyp were first identified to be significantly differentially expressed among different seasons by an optimal discovery procedure (ODP) method. Furthermore, four gene expression patterns were defined based on correlation relationship of differential genes, most of which fall into H-L-L and L-H-H (L, low expression and H, high expression) patterns. We also utilized multivariate analysis to examine the transcriptome similarity between Hyp and Tel in May and August, and found that the transcriptomes of Hyp and Tel in same season exhibit a global similarity and the differential gene patterns identified in Hyp were also evident in Tel. We hypothesized that these gene expression patterns may account for dynamic change in neuroendocrine function over the seasonal reproductive cycle. Our study provides a proof-of-concept for extracting seasonal characters from available microarray datasets.

5.1 Introduction

One character of fish is that their own internal physiological rhythm and reproductive development is closely associated with external seasonal factors like temperature and photoperiod [78]. Accordingly, fish reproductive development can be basically divided into

three seasonal stages: gonadal regression, recrudescence or redevelopment, maturation and spawning [79]. It is also well-known that fish reproductive development is closely regulated by neuroendocrine-pituitary-gonadal axis. The main pathway of this axis is initiated by the gonadotropin-releasing hormone (GnRH), the primary hypothalamic neurohormone, which stimulates the release of luteinizing hormone (LH) from the anterior pituitary. Then LH stimulates the synthesis of gonadal sex steroids, such as testosterone (T) and estrogen (E2), which exert a primary role in gonad development locally [78].

How the neuroendocrine system coincides with seasonal cycle is an open question. The previous work has shown that the expression of neuroendocrine genes modulating hypothalamo-pituitary-gonadal axis varies over the whole seasonal cycle[80-82]. However, no microarray study has been conducted to investigate the global transcriptome profile during a whole season. Microarray data sets from AURATUS project harbour seasonal information, thereby motivating the present analysis of such season-related gene expression characters. This type of information is beyond the original experimental designs, which aimed to understand the transcriptomic effects of goldfish brain by different hormones and endocrine disrupting chemicals. This project is collaborated with Dr. Trudeau's Lab and some analysis results are presented here. Results that have been verified by real-time PCR (not shown). This PCR data is not part of this thesis, but will be included in the PhD thesis of my collaborator, Dapeng Zhang.

5.2 Materials and methods

5.2.1 Experiments and microarray datasets

All microarray data used in this study are from the *Goldfish Environmental Genomics* project. Microarray data sets with mRNA source from female fish neuroendocrine brain hypothalamus and telencephalon have been collected to represent four seasonal time points. The detailed information about the experimental design, brain tissue, slide number, treatment seasonal time, and other information are listed in Table 5.1.

Table 5.1. Detailed information about the experimental design, brain tissue, slide number, treatment seasonal time, and other information. Chemical is the drug used in individual experiment. Purchase and treatment time indicate the exact time when fish came in and when the fish was treated.

Month	Sex	Brain	Slide number	Chemical	Purchase time	Treatment time	Gonad state
May	Female	Tel	28_48_43_34	MPTP-MPT	28-Apr-04	22-May-04	Prespawning
May	Female	Hyp	4_17_52_21	MPTP-MPT	28-Apr-04	23-May-04	Prespawning
May	Female	Hyp	9_19_33_44	D1R agonist, SKF 38393	22-Mar-06	May 10, 2006	Prespawning
May	Female	Tel	7_13_30_47	D1R agonist, SKF 38393	22-Mar-06	May 10, 2006	Prespawning
May	Female	Tel	8_16_24_40	D2R agonist, quinpirole	22-Mar-06	May 10, 2006	Prespawning
May	Female	Hyp	5_11_14_26	D2R agonist, quinpirole	22-Mar-06	May 10, 2006	Prespawning
August	Female	Tel	16_22_33_50	GABAA agonist, muscimol	9-Jun-04	late Aug. 2004	Intact, sexually regressed
August	Female	Hyp	2_31_46_48	GABAA agonist, muscimol	9-Jun-04	late Aug. 2004	Intact, sexually regressed
August	Female	Hyp	5_29_34_47	GABAB agonist, Baclofen	9-Jun-04	early Sep, 2004	Intact, sexually regressed
August	Female	Tel	8_21_41_44	GABAB agonist, Baclofen	9-Jun-04	early Sep, 2004	Intact, sexually regressed
December	Female	Hyp	19_30_36_39	Fluoxetine	6-Oct-04	12-Dec-04	Early redevelopment

5.2.2 Data normalization analysis

For each array, spots that had been manually flagged due to poor hybridization and spots in which the estimated fluorescence intensity was below or equal to the estimated background signal intensity in either channel were removed before further analysis. Lowess normalization [17] with span 0.4 was then used to decrease the intensity dependent biases within slide, and Scale normalization [17] was used to decrease biases across slides. Thereafter, control sample intensities were recovered and extracted to build a new series of expression data for all the slides. Following this, Quantile normalization [25] was used for across-slide normalization to minimize biases among different experiments. All these normalization analyses were conducted using GoldR program suite (Chapter 4).

5.2.3 Identification of differentially expressed genes

An optimal discovery procedure (ODP) method which is implemented in the Extraction of Differential Gene Expression (EDGE) program [69] was used to assess the significance of differential expression of the genes between different seasonal time points. The false discovery rate for selected genes was monitored and controlled by calculating the q value. Genes with q value <0.0001 were considered as significantly differentially expressed genes between comparison groups. The selected genes were further clustered and visualized using MultiExperiment Viewer (MeV) [83] with Pearson correlation as a distance function.

5.2.4 Multivariate data analysis

Two multivariate methods, principal component analysis (PCA) and hierarchical clustering analysis (HCA), were utilized to classify the transcriptome similarity relationship between slides that represent different seasonal time points. Briefly PCA combines two, or more, correlated factors (i.e. transcripts) into one new variable, a principal component (PC) [74, 84, 85]. In PCA the dimensionality of the dataset is reduced by replacing the original variables with a smaller number of newly formed variables that are linear combinations of the original variables and that explain the majority of the information (variability) from the experiment.

A hierarchical clustering tree [86] generates a binary dendrogram representing the association structure of pairs of arrays. The association between two arrays was measured in terms of their correlations. The Pearson correlation of the standardized intensity measurements over the differential genes was calculated for all pairwise combinations of the slides. The algorithm identifies the gene pair with the smallest distance and groups them with a link, where distance is defined to be one-minus the correlation. The algorithm proceeds in a recursive manner to build the tree structure step by step.

The MeV program [83] was utilized for both PCA and HCA analyse and visualization of the results.

5.2.5 Gene Ontology (GO) analysis

The resulting differentially expressed genes were submitted to the blast2go [87] web server (http://www.blast2go.de/). Enrichment of these GO functional categories was determined using the Gossip/Fisher Exact test package, through a comparison of the categories found in our list of genes compared to the list of all genes found on the goldfish cDNA microarray. FDR controlled *p*-values (FDR < 0.05) were used for the assessment of differentially significance.

5.3 Result

5.3.1 Boxplots of all female goldfish neuroendocrine brain slides during normalization procedures

In order to remove systematic and experimental biases among different slides used in this study, we conducted a series of normalization procedures including both between-slide and within-slide normalizations. For each experiment, we subjected the slides to Lowess and Scale normalization. We then extracted the raw control sample intensity, which was used to represent transcriptomic status of female brain in each season. We then further decreased the bias between different experiments through using Quantile normalization. We analyzed the slides with an mRNA source from female hypothalamus (Hyp). Figure 5.1 shows each boxplot during the normalization procedures. The microarray data sets derive from three seasonal time points (May, August, and December), which correspond to the distinct physiological periods of sexually mature prespawning, sexually regressed, and early redevelopment of goldfish ovary, respectively.

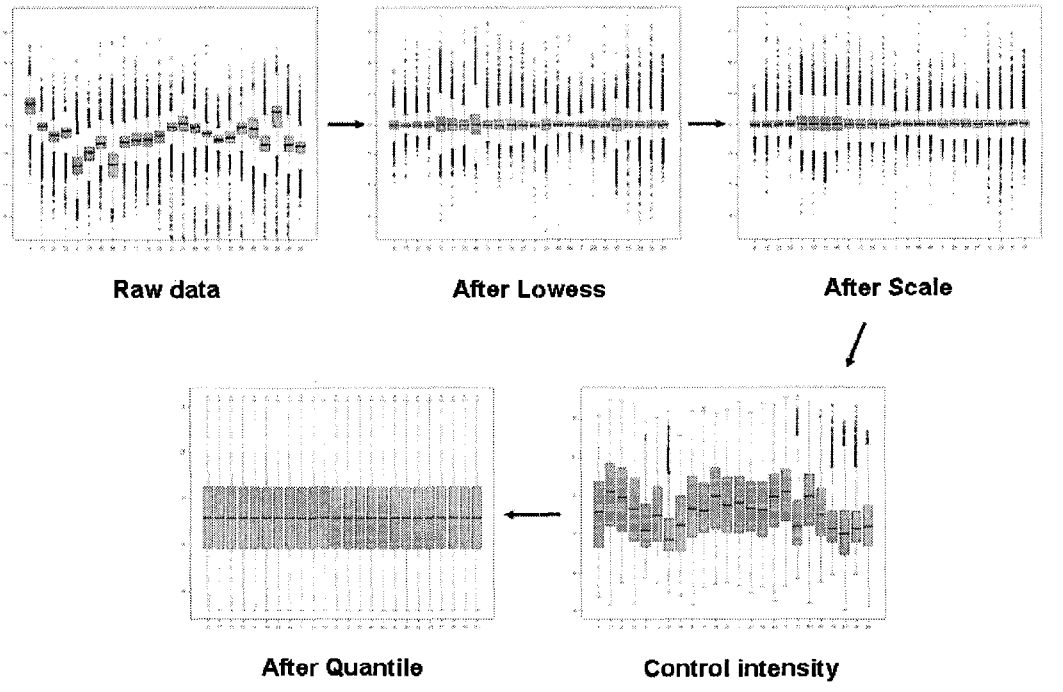

Figure 5.1. Boxplots of M values during the normalization procedures.

5.3.2 Identification of differentially expressed genes in female hypothalamus along the reproductive seasonal cycle

We tried to identify the differentially expressed genes between these three seasonal time points. The optimal discovery procedure (ODP) method [69] implemented in the EDGE program [88] was utilized for this purpose. A total of 998 differentially expressed genes with high statistical significance (q value < 0.0001) was identified. All the genes were further subjected to hierarchical cluster analysis (HCA) using Pearson correlation as the distance function. In the HCA, not only the relationships between different slides can be classified, but also the genes with similar expression patterns can be grouped by visual inspection of the hierarchical clusters results. The prominent co-expressed gene groups are believed to be specifically regulated in the different seasons considered here. As shown in Figure 5.2, the identified differential expressed genes can be grouped into four gene expression clusters comprising H-L-L, H-H-L, L-H-H and L-L-H (L, low expression and H, high expression)

patterns along the seasonal cycle. Most of the identified genes (800 genes) belong to H-L- L and L-H- H expression clusters.

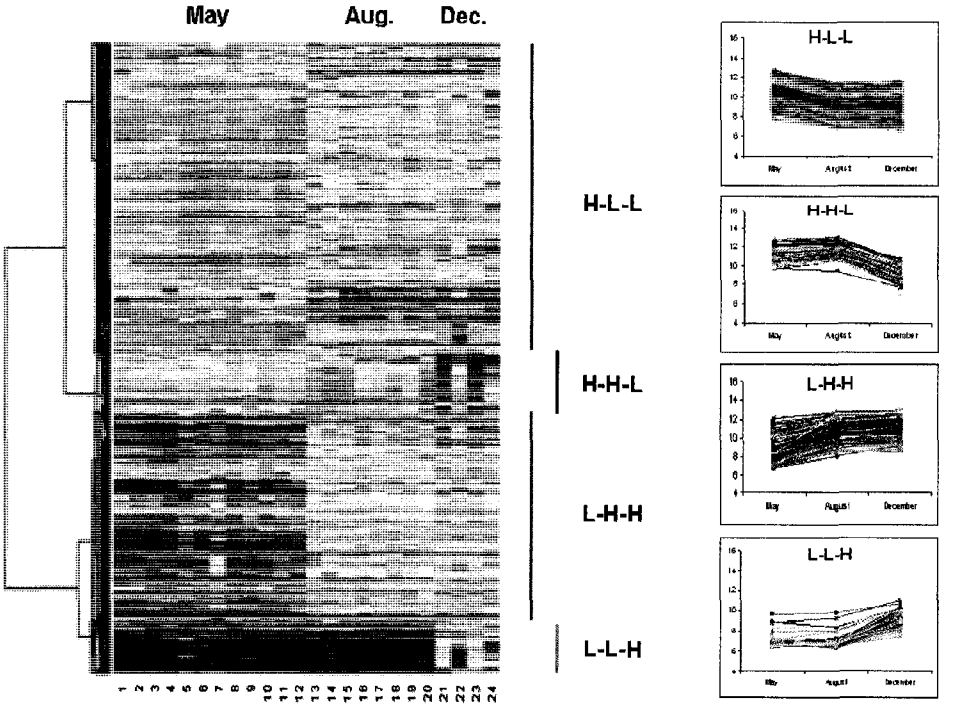

Figure 5.2. Hierarchical clustering of the expression profiles of the significantly differentially expressed genes along the reproductive seasonal cycle. The colour showed log ratio intensity of the gene expression. Red showed high expression and blue showed low expression.

5.3.3 Multivariate analysis showed that the female telencephalon exhibits a similar transcriptomic profile as the female hypothalamus

We next examined the global transcriptome similarity between Hyp and Tel during the seasonal cycle. Both Hyp and Tel are important brain tissues involved in neuroendocrine control of growth and reproduction [78, 79]. Two expression datasets for female Tel in May and August were available for this study. After the previously applied data normalizations,

the principal components analysis (PCA) was performed for Hyp and Tel data. PCA is an exploratory multivariate statistical technique for simplifying complex data sets [74, 84, 85]. It is frequently used to classify samples with a similar expression profile, e.g. related to a specific treatment or phenotype. The results of the PCA analysis are visualized in a two-dimensional plot (Figure 5.3). This sub-space, determined by PCA, captures the highest amount of the total variability. In our line and column centered data matrix, the first (PC1) and second (PC2) principal components captured 60.72% and 12.67% of the total variability, respectively. The points in the plots represent the global transcriptomes of the different slides. The transcriptomes that cluster together are overall more similar, while more dissimilar transcriptomes are further apart. Unexpectedly, the output of PCA analysis showed that the transcriptomes of Hyp and Tel in both May and August have significant overlap. This indicates that Hyp and Tel have the highly similar gene expression profiles in the same season, at least in sexually mature animals.

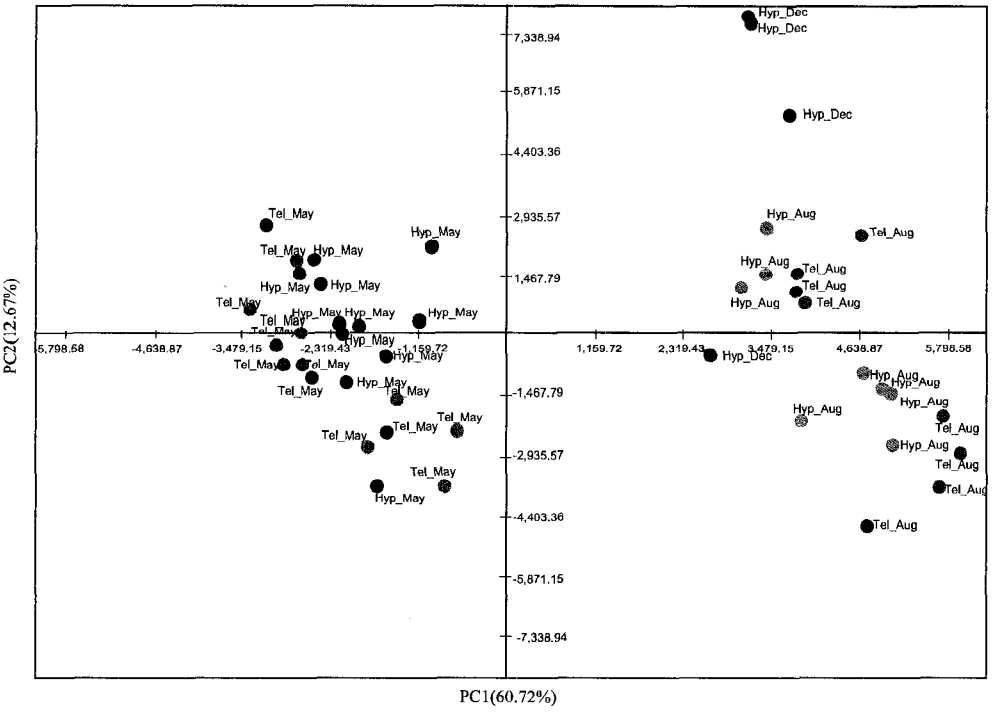

Figure 5.3. Two-dimensional PCA plot for multiple microarray slides from three seasonal time points. Red points indicate transcriptome status of Hyp brain samples in December; Blue points for Hyp brains in May; Cyan points for Hyp brains in August; Green points for Tel brains in May and pink points for Tel brains in August.

We next investigated whether the expression patterns of differentially expressed genes identified in Hyp are similar in Tel. We plotted all the differential genes in same order in Figure 5.4 as Figure 5.2 for Tel. We found those genes also exhibit clear differential expression patterns in Tel between May and August. This trend has also been observed by clustering global transcriptomes of both Hyp and Tel (not shown). Therefore, we conclude that the same gene group is differentially expressed over the seasons and showed similar expression patterns in Hyp and Tel brains.

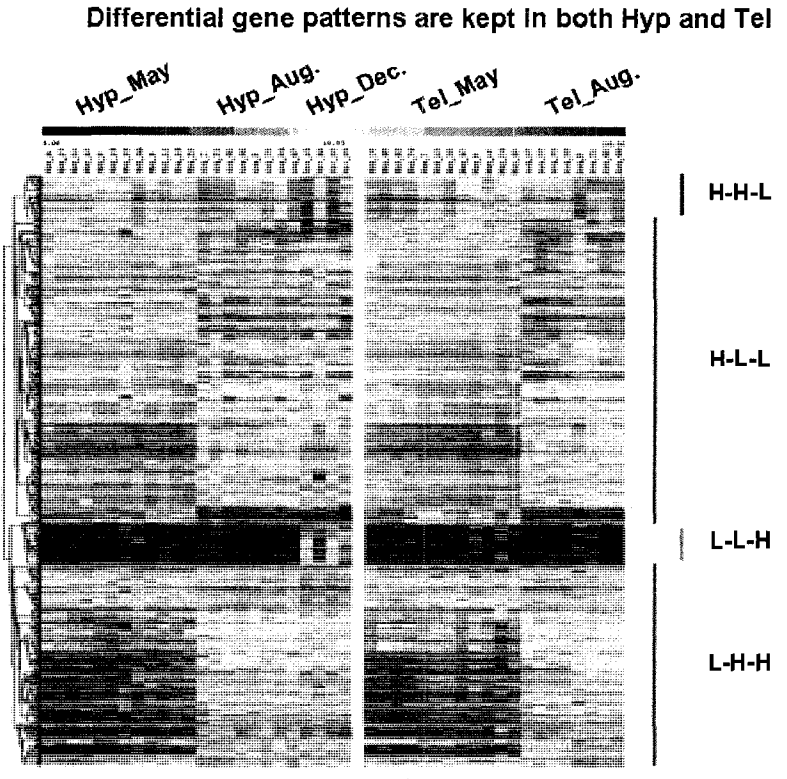

Figure 5.4. Different gene patterns are similar in both Hyp and Tel.

5.3.4 Gene Ontology analysis of differentially expressed genes

Since most the same genes were differentially expressed between seasons with similar expression patterns in both Hyp and Tel brain tissues, we hypothesize that these genes may be responsive to dynamic changes of the neuroendocrine system in the brain over the reproductive seasonal cycle. Gene Ontology (GO) analysis was further utilized to explore the potential functional significance among these genes. Importantly, a series of GO categories implicated in the neuroendocrine system are significantly over-presented for the genes of H-L-L pattern (Figure 5.5). For example, the identified molecular functions include anion channel activity, hormone activity, GABA receptor activity and chloride channel activity. Accordingly, the biological processes include chloride transport, GABA transport, glutamine metabolic process, G-protein coupled receptor signalling pathway, synaptic transmission, etc. This bears biological implication since the gene expression changes of the H-L-L pattern corresponds to the transition from prespawning to sexually regressed female goldfish, among which the significant change of gene expression is long believed.

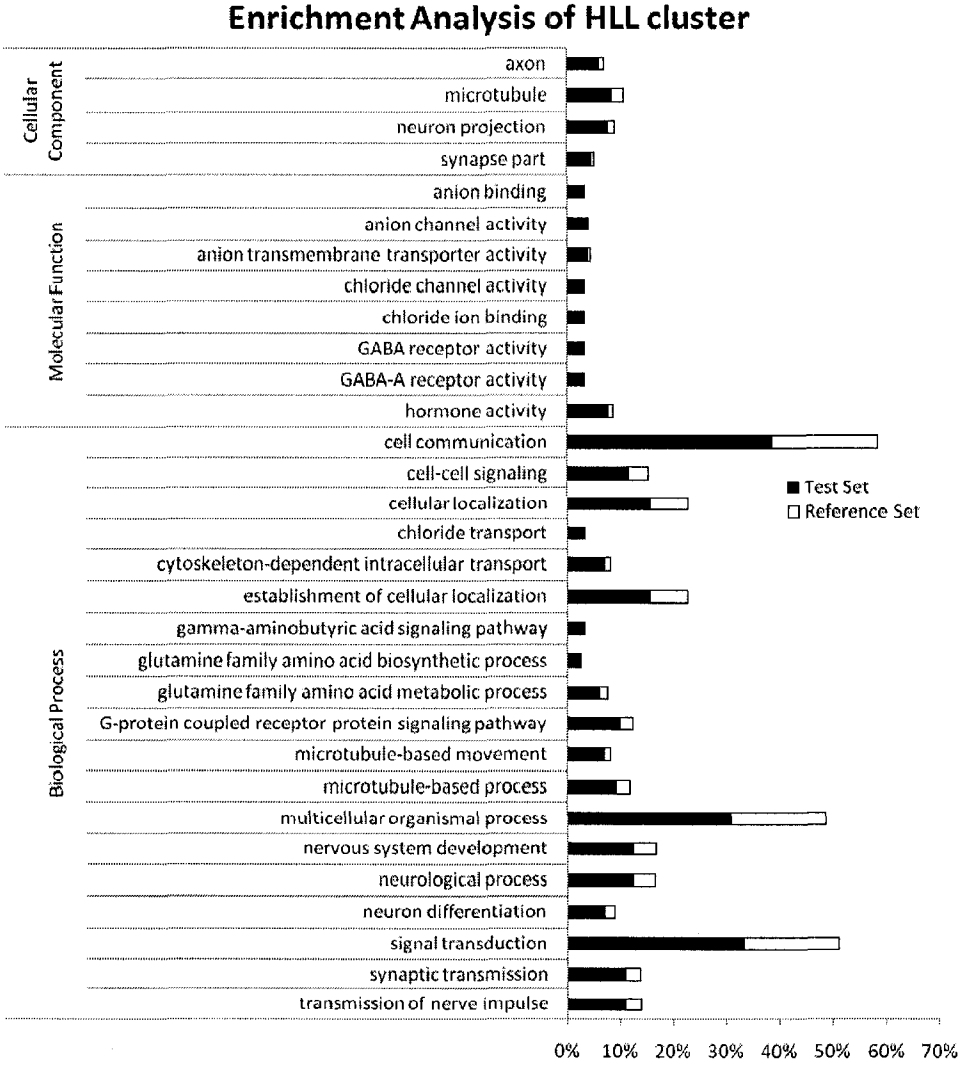

Figure 5.5. Gene Ontolgy (GO) term enrichment analysis for the genes in HLL pattern.

5.4 Discussion and conclusion

The purpose of this study was to develop a new analytical strategy for extracting seasonal gene expression pattern information from existing microarray data. We utilized a series of normalization procedures to remove systematic and experimental bias. The overall similarity among the slides from same season in PCA results illustrated that this strategy is effective and the resulting data can reflect the seasonal information. Although the number of

the slides used for the seasonal time point December is only 4 in the current study, the analyzing strategy we have established allows a further iterated investigation with an increasingly rich goldfish microarray database [13, 89, 90]

In the present study, we have provided for the first time biological systems-level investigations of transcriptomic change in female fish brain along the reproductive seasonal cycle and identified several clear gene expression patterns. Although documented evidence has assigned seasonal expression profiles for a few genes in the neuroendocrine system, no such global gene expression analysis has previously been done so far. Interestingly, the overall transcriptomes for all examined seasons did not change dramatically (not shown), indicating that the fish brain transcriptome is generally stable during the seasonal reproductive cycle. However, a set of genes has been identified as differentially expressed between seasons and most of those genes can be clustered into two simple gene expression patterns (H-L-L and L-H-H). The change in expression of those genes occurs between May and August, which correspond to fish reproductive status of prespawning and sexually regressed fish, respectively. We further examined the transcriptome similarity between female Hyp and Tel and found that in the same season Hyp and Tel exhibit similar transcriptomic profiles. Also the same set of genes identified in Hyp is differentially expressed between May and August in Tel. This suggests that this set of differentially expressed genes accounts for the seasonal functional change of neuroendocrine brain (both Hyp and Tel). Indeed, GO analysis has showed that a series of GO functional categories implicated in neuroendocrine system are significantly over-presented for the genes of H-L-L pattern, including chloride transport, GABA transport, glutamine metabolic process, and synaptic transmission. The expression patterns for several involved genes, enzymes or hormones in these processes have been experimentally illustrated. For instance, gonadotropin GTH II, a follicle stimulating hormone, has been shown with a seasonal expression pattern similar to HLL [91]. The seasonal expression profile for GABA related genes have also been observed in goldfish [92]. Moreover, several candidate genes identified in this study including isotocin, ependymin II, GABA-A gamma2, calmodulin 1b, and aromatase b were further examined using real-time PCR assays based on a set of seasonal fish brain samples (not shown). The result shows an overall same pattern with this analysis, which strongly supports our hypothesis.

Overall, we established an analytical strategy for extracting seasonal gene expression pattern information from existing goldfish microarray data sets. We found both Hyp and Tel transcriptomes exhibit an overall similarity in same season and a set of genes showed differentially expressed between seasons. Two simple gene expression patterns can be assigned to most of those genes and correspond to the transition of fish reproductive status. We hypothesized that the change of the dynamic expression patterns may account for the neuroendocrine control of reproductive process during the seasonal cycle.

Chapter 6. General conclusions and future work

Final remarks

In this thesis, I have discussed the major aspects of my research in the area of microarray data analysis. Since microarray techniques consist of a series of different components such as array platforms, experimental design, data analysis methods and tools, and Meta information inference, several new developments have been made for each of these individual steps in this study. The main themes can be summarized as new methods, new platforms, new tools and new analytical strategies. In this final chapter, I very briefly discuss the main contributions, focusing on important findings. Additional and more detailed discussions are available in the last sections of the Chapters 2 to 5, respectively.

6.1 New method

Normalization is essential in dual-labelled microarray data analysis to remove non-biological variations and systematic biases. Although many normalization methods have been proposed, most of them have critical assumptions about data distributions. In my study, I have developed a novel normalization method based on the Generalized Procrustes Analysis (GPA) algorithm. Compared to other popular established normalization methods including Global, Lowess, Scale, Quantile, VSN, and a boutique array-specific housekeeping gene method, the GPA-based method has consistently better performance in reducing across-slide variability and removing systematic bias. In particular, the GPA method is more robust and appropriate for diverse microarray types including the boutique array since there is no inherent statistical and biological assumption.

6.2 New platform

During my doctoral study, I closely cooperated with Dr. Trudeau Lab in establishing and analyzing a goldfish-carp brain cDNA microarray. The goldfish-carp cDNA microarray is a novel gene expression profiling platform designed to identify endocrine disrupting chemical, EDC-related gene profiles in the goldfish brain. Assessing data quality and sources of errors in microarray experiments is therefore necessary prior to any subsequent analysis. Here a series of statistical analyses including array quality filter test, statistical repeatability coefficient, linear correlation coefficient, and coefficient of vitiate have been used to

examine the inter-array and intra-array reproducibility of goldfish-carp microarrays. High degree of reproducibility both within and among arrays has been observed and indicates that this new platform is capable of reliably and reproducibly detecting differences in gene expression patterns between two distinct biological samples.

6.3 New program suite

Various methods and algorithms have been created to facilitate the analysis of DNA microarray data, with a typical pipeline including data pre-processing, data normalization, differential gene identification, multivariable statistical analysis and gene ontology annotation. Understanding or using these bioinformatic methods for these diverse steps usually requires a wide knowledge base covering biology, mathematics, statistics and computer science, which is very challenging for biologists. Therefore, there is a crucial need for an integrative and user-friendly program suite to ease microarray data analysis in molecular biology laboratories with limited bioinformatic support. Here a new R-based program suite, called GoldR, has been developed. It integrated the most popular microarray data analysis methods including data filtering, imputation for missing values, comprehensive normalization analysis, differential gene identification, and clustering analysis. This GoldR suite has been tested and refined over the last few years by analyzing the goldfish-carp cDNA microarray data.

6.4 New analytical strategy

Based on microarray data sets from the AURATUS genomics project, I proposed a new analytical strategy to extract seasonally-related gene expression information, which combines a series of comprehensive data transformation and normalization analysis, differential gene identification, and multivariate analysis. About 1000 genes were identified to be differentially expressed in the female goldfish hypothalamus (Hyp) and telencephalon (Tel) between May, August and December. Further gene clusters were defined and gene ontology analysis showed that most genes related to neuroendocrine functions are changed between May and August, indicating a potential seasonal trend within neuroendocrine systems. Our study provides a proof-of-concept for extracting seasonal characters from available microarray datasets.

6.5 Future works

Several projects can be extended based on my thesis work. Firstly, I would like to further investigate the relationship between normalizations and identification of differentially expressed genes. This topic is interesting, but few studies have been conducted. Secondly, I will try to develop more functions in my GoldR program suite. Especially, some well established microarray data analysis methods will be integrated. Thirdly, with the increasing amount of microarray data available in public databases, meta-mining of biological information or pathway/network construction is a new direction of interest. Most studies right now suffer their few considerations on systemic bias in microarray platform. I think introducing some statistical criteria will help retrieving more reliable information. One possible project is to develop new method for *Meta* analysis of time-point experiments. Time-course experiment is a very popularly used experimental design in microarray application. Many data in microarray database are with this trait. However there are few methods, to my knowledge, that can effectively extract the dynamic information from time-points experiments. Overall, microarray data analysis is a rapidly developing field. New mathematical or statistical algorithms, methods or strategies will introduce new powers in this field.

Appendix

List of published papers

1. **Xiong H**, Zhang D, Martyniuk CJ, Trudeau VL, Xia X: Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data, BMC Bioinformatics 2008, 9:25.
2. Marlatt VL, Martyniuk C, Zhang D, **Xiong H**, Xia X, Moon T, Trudeau V, Tissue-specific auto-regulation of estrogen receptor subtypes and profiling of estrogen-responsive genes in the neuroendocrine axis of male goldfish (*Carassius auratus*) exposed to 17β-estradiol. Molecular and Cellular Endocrinology, 2008, 283(1-2): 38-48.
3. Khalouei S, Yao X, Mennigen J, Carullo M, Ma P, Song Z, **Xiong H**, Xia X: Bioinformatic Approach to Identify Penultimate Amino Acids Efficient for N-Terminal Methionine Excision. Bioinformatics and Biomedical Engineering, ICBBE 2007.
4. Martyniuk CJ, **Xiong H**, Werry K, Chui S, Sardana R, Gerrie E, Trudeau VL. A goldfish brain enriched cDNA array to study endocrine disruption in the neuroendocrine brain. Physiological Genomics, 2006, 27(3):328-36.
5. Mennigen JA, Martyniuk CJ, Crump K, **Xiong H**, Zhao E, Popesku J, Anisman H, Cossins AR, Xia X, Trudeau VL. The effects of fluoxetine on the reproductive axis of female goldfish (Carassius auratus). Physiol Genomics. 2008, in press. PMID: 18765858.
6. Popesku JT, Martyniuk CJ, Mennigen J, **Xiong H**, Zhang D, Cossins AR, Xia X, Trudeau V. The goldfish (*Carassius auratus*) as a model for neuroendocrine signaling. Molecular and Cellular Endocrinology, 2008, 293(1-2):43-56.

REFERENCE

1. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer**. *Cancer research* 2002, **62**(15):4427-4433.
2. Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors**. *Nature genetics* 2003, **33**(1):49-54.
3. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S *et al*: **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(14):8418-8423.
4. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(25):14863-14868.
5. Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae**. *Nature genetics* 2001, **29**(4):482-486.
6. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FC: **Protein interaction verification and functional annotation by integrated analysis of genome-scale data**. *Molecular cell* 2002, **9**(5):1133-1143.
7. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets**. *Genome research* 2004, **14**(6):1085-1094.
8. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW: **oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes**. *Nucleic acids research* 2005, **33**(10):3154-3164.
9. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, Chinnaiyan AM: **Mining for regulatory programs in the cancer transcriptome**. *Nature genetics* 2005, **37**(6):579-583.
10. Chen KC, Wang TY, Tseng HH, Huang CY, Kao CY: **A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae**. *Bioinformatics (Oxford, England)* 2005, **21**(12):2883-2890.
11. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays**. *Nature genetics* 1999, **21**(1 Suppl):20-24.
12. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science (New York, NY* 1995, **270**(5235):467-470.
13. Martyniuk CJ, Xiong H, Crump K, Chiu S, Sardana R, Nadler A, Gerrie ER, Xia X, Trudeau VL: **Gene expression profiling in the neuroendocrine brain of male goldfish (Carassius auratus) exposed to 17alpha-ethinylestradiol**. *Physiological genomics* 2006, **27**(3):328-336.
14. Williams DR, Li W, Hughes MA, Gonzalez SF, Vernon C, Vidal MC, Jeney Z, Jeney G, Dixon P, McAndrew B *et al*: **Genomic resources and microarrays for the common carp Cyprinus carpio L.** *Journal of Fish Biology* 2008, **72**:2095-2117.

15. Neafsey DE, Hartl DL: **Convergent loss of an anciently duplicated, functionally divergent RH2 opsin gene in the fugu and Tetraodon pufferfish lineages.** *Gene* 2005, **350**(2):161-171.
16. Smith PJ, McVeagh SM: **Genetic analyses of carp, goldfish, and carp–goldfish hybrids in New Zealand.** In: *DOC Research & Development Series 219.* Wellington: Department of Conservation; 2005: 20.
17. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic acids research* 2002, **30**(4):e15.
18. Chou JW, Paules RS, Bushel PR: **Systematic variation normalization in microarray data to get gene expression comparison unbiased.** *J Bioinform Comput Biol* 2005, **3**(2):225-241.
19. Eckel JE, Gennings C, Therneau TM, Burgoon LD, Boverhof DR, Zacharewski TR: **Normalization of two-channel microarray experiments: a semiparametric approach.** *Bioinformatics* 2005, **21**(7):1078-1083.
20. Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R: **Evaluation of normalization methods for microarray data.** *BMC Bioinformatics* 2003, **4**:33.
21. Zien A, Aigner T, Zimmer R, Lengauer T: **Centralization: a new method for the normalization of gene expression data.** *Bioinformatics (Oxford, England)* 2001, **17 Suppl 1**:S323-331.
22. Tarca AL, Cooke JE, Mackay J: **A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data.** *Bioinformatics (Oxford, England)* 2005, **21**(11):2674-2683.
23. Dabney AR, Storey JD: **A new approach to intensity-dependent normalization of two-channel microarrays.** *Biostatistics (Oxford, England)* 2007, **8**(1):128-139.
24. Smyth GK, Speed T: **Normalization of cDNA microarray data.** *Methods* 2003, **31**(4):265-273.
25. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics (Oxford, England)* 2003, **19**(2):185-193.
26. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics (Oxford, England)* 2002, **18 Suppl 1**:S96-104.
27. Hummel M, Bentink S, Berger H, Klapper W, Wessendorf S, Barth TF, Bernd HW, Cogliatti SB, Dierlamm J, Feller AC *et al*: **A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling.** *N Engl J Med* 2006, **354**(23):2419-2430.
28. Gurok U, Steinhoff C, Lipkowitz B, Ropers HH, Scharff C, Nuber UA: **Gene expression changes in the course of neural progenitor cell differentiation.** *J Neurosci* 2004, **24**(26):5982-6002.
29. Wilson DL, Buckley MJ, Helliwell CA, Wilson IW: **New normalization methods for cDNA microarray data.** *Bioinformatics* 2003, **19**(11):1325-1332.
30. Yauk CL, Williams A, Boucher S, Berndt LM, Zhou G, Zheng JL, Rowan-Carroll A, Dong H, Lambert IB, Douglas GR *et al*: **Novel design and controls for focused DNA microarrays: applications in quality assurance/control and normalization for the Health Canada ToxArray.** *BMC genomics* 2006, **7**:266.

31. Lu T, Costello CM, Croucher PJ, Hasler R, Deuschl G, Schreiber S: **Can Zipf's law be adapted to normalize microarrays?** *BMC Bioinformatics* 2005, **6**:37.
32. Zhao Y, Li MC, Simon R: **An adaptive method for cDNA microarray normalization.** *BMC Bioinformatics* 2005, **6**:28.
33. Goodall C: **Procrustes methods in the statistical analysis of shape.** *J R Stat Soc 53* 1991:pp. 285–339.
34. Gower J: **Generalized Procrustes Analysis.** *Psychometrika* 1975, **40**(1):33-51.
35. Ten Berge J: **Orthogonal Procrustes rotation for two or more matrices.** *Psychometrika* 1977, **42**:267-276.
36. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S: **A new non-linear normalization method for reducing variability in DNA microarray experiments.** *Genome biology* 2002, **3**(9):research0048.
37. Metchev S, Grindlay J: **A two-dimensional Kolmogorov-Smirnov test for crowded field source detection.** *ROSAT sources in NGC 6397Monthly Notices of the Royal Astronomical Society* 2002, **335**:73-83.
38. Wettenhall JM, Smyth GK: **limmaGUI: a graphical user interface for linear modeling of microarray data.** *Bioinformatics (Oxford, England)* 2004, **20**(18):3705-3706.
39. Zhou Y, Gwadry FG, Reinhold WC, Miller LD, Smith LH, Scherf U, Liu ET, Kohn KW, Pommier Y, Weinstein JN: **Transcriptional regulation of mitotic genes by camptothecin-induced DNA damage: microarray analysis of dose- and time-dependent effects.** *Cancer Res* 2002, **62**(6):1688-1695.
40. Yoon D, Yi SG, Kim JH, Park T: **Two-stage normalization using background intensities in cDNA microarray data.** *BMC Bioinformatics* 2004, **5**:97.
41. Balagurunathan Y, Dougherty ER, Chen Y, Bittner ML, Trent JM: **Simulation of cDNA microarrays via a parameterized random signal model.** *Journal of biomedical optics* 2002, **7**(3):507-523.
42. Wang D, Huang J, Xie H, Manzella L, Soares MB: **A robust two-way semi-linear model for normalization of cDNA microarray data.** *BMC bioinformatics* 2005, **6**:14.
43. Zhou X, Wang X, Dougherty ER: **Binarization of microarray data on the basis of a mixture model.** *Molecular cancer therapeutics* 2003, **2**(7):679-684.
44. Balagurunathan Y, Wang N, Dougherty ER, Nguyen D, Chen Y, Bittner ML, Trent J, Carroll R: **Noise factor analysis for cDNA microarrays.** *Journal of biomedical optics* 2004, **9**(4):663-678.
45. Hua JaB, Yoganand and Chen, Yidong and others: **Normalization Benefits Microarray-Based Classification.** *EURASIP Journal on Bioinformatics and Systems Biology* 2006, **2006**:Article ID 43056, 43013 pages.
46. Demirkaya O, Asyali MH, Shoukri MM: **Segmentation of cDNA microarray spots using markov random field modeling.** *Bioinformatics (Oxford, England)* 2005, **21**(13):2994-3000.
47. Fujita A, Sato JR, Rodrigues Lde O, Ferreira CE, Sogayar MC: **Evaluating different methods of microarray data normalization.** *BMC bioinformatics* 2006, **7**:469.
48. Nykter M, Aho T, Ahdesmaki M, Ruusuvuori P, Lehmussola A, Yli-Harja O: **Simulation of microarray data with realistic characteristics.** *BMC bioinformatics* 2006, **7**:349.

49. Uchida S, Nishida Y, Satou K, Muta S, Tashiro K, Kuhara S: **Detection and normalization of biases present in spotted cDNA microarray data: a composite method addressing dye, intensity-dependent, spatially-dependent, and print-order biases.** *DNA Res* 2005, **12**(1):1-7.
50. Albers CJ, Jansen RC, Kok J, Kuipers OP, van Hijum SA: **SIMAGE: simulation of DNA-microarray gene expression data.** *BMC bioinformatics* 2006, **7**:205.
51. Krzanowski WJ: **Principles of Multivariate Analysis: A User's Perspective.** Oxford: Clarendon Press; 1988.
52. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics (Oxford, England)* 2001, **17**(6):520-525.
53. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S: **A new non-linear normalization method for reducing variability in DNA microarray experiments.** *Genome Biol* 2002, **3**(9):research0048.
54. Wu W, Dave N, Tseng GC, Richards T, Xing EP, Kaminski N: **Comparison of normalization methods for CodeLink Bioarray data.** *BMC bioinformatics* 2005, **6**:309.
55. Goodall C: **Procrustes methods in the statistical analysis of shape.** *Journal of the Royal Statistical Society Series B* 1991, **53**(2):285-239.
56. Krzanowski WJ: **Procrustes rotation in analytical chemistry, a tutorial.** *Chemometrics & Intelligent Laboratory Systems* 2004, **72**:123-132.
57. Akça MD, Institut für Geodäsie und Photogrammetrie (Zürich): **Generalized Procrustes analysis and its applications in photogrammetry.** Zürich: ETH Swiss Federal Institute of Technology Zurich Institute of Geodesy and Photogrammetry; 2003.
58. Theobald DL, Wuttke DS: **Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(49):18521-18527.
59. Burgoon LD, Eckel-Passow JE, Gennings C, Boverhof DR, Burt JW, Fong CJ, Zacharewski TR: **Protocols for the assurance of microarray data quality and process control.** *Nucleic acids research* 2005, **33**(19):e172.
60. Sauer U, Preininger C, Hany-Schmatzberger R: **Quick and simple: quality control of microarray data.** *Bioinformatics (Oxford, England)* 2005, **21**(8):1572-1578.
61. Jenssen TK, Langaas M, Kuo WP, Smith-Sorensen B, Myklebost O, Hovig E: **Analysis of repeatability in spotted cDNA microarrays.** *Nucleic acids research* 2002, **30**(14):3235-3244.
62. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M *et al*: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic acids research* 2001, **29**(8):E41-41.
63. Cleveland WS: **Robust Locally Weighted Fitting and Smoothing Scatterplots.** *Journal of the American Statistical Association* 1979, **74**:829-836.
64. Cleveland WS, Grosse E: **Computational Methods for Local Fitting.** *Statistics and Computing* 1991, 1:47-62.

65. Cleveland WS, Devlin SJ: **Locally-Weighted Fitting: An Approach to Fitting Analysis by Local Fitting.** *Journal of the American Statistical Association* 1988, **83**:596-610.
66. Dobbin K, Shih JH, Simon R: **Statistical design of reverse dye microarrays.** *Bioinformatics (Oxford, England)* 2003, **19**(7):803-810.
67. Xiong H, Zhang D, Martyniuk CJ, Trudeau VL, Xia X: **Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data.** *BMC Bioinformatics* 2008, **9**(1):25.
68. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17**(6):520-525.
69. Storey JD, Dai JY, Leek JT: **The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments.** *Biostatistics (Oxford, England)* 2007, **8**(2):414-432.
70. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Statistical applications in genetics and molecular biology* 2004, **3**:Article3.
71. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(9):5116-5121.
72. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes.** *Bioinformatics (Oxford, England)* 2001, **17**(6):509-519.
73. Kohonen T: **Self-Organizing Maps**, vol. 30. Berlin:Springer-Verlag: Springer Series in Information Sciences; 1995.
74. Basilevsky A: **Statistical Factor Analysis and Related Methods: Theory and Applications** 1994.
75. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics (Oxford, England)* 2000, **16**(10):906-914.
76. Sturn A, Quackenbush J, Trajanoski Z: **Genesis: cluster analysis of microarray data.** *Bioinformatics (Oxford, England)* 2002, **18**(1):207-208.
77. Xia X: **Bioinformatics and the cell : modern computational approaches in genomics, proteomics, and transcriptomics.** New York: Springer; 2007.
78. Trudeau VL: **Neuroendocrine regulation of gonadotrophin II release and gonadal growth in the goldfish, Carassius auratus.** *Reviews of reproduction* 1997, **2**(1):55-68.
79. Blazquez M, Bosma PT, Fraser EJ, Van Look KJ, Trudeau VL: **Fish as models for the neuroendocrine regulation of reproduction and growth.** *Comparative biochemistry and physiology* 1998, **119**(3):345-364.
80. Samia M, Lariviere KE, Rochon MH, Hibbert BM, Basak A, Trudeau VL: **Seasonal cyclicity of secretogranin-II expression and its modulation by sex steroids and GnRH in the female goldfish pituitary.** *General and comparative endocrinology* 2004, **139**(3):198-205.

81. Peyon P, Saied H, Lin X, Peter RE: **Postprandial, seasonal and sexual variations in cholecystokinin gene expression in goldfish brain.** *Brain research* 1999, **74**(1-2):190-196.
82. Peyon P, Saied H, Lin X, Peter RE: **Preprotachykinin gene expression in goldfish brain: sexual, seasonal, and postprandial variations.** *Peptides* 2000, **21**(2):225-231.
83. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M *et al*: **TM4: a free, open-source system for microarray data management and analysis.** *BioTechniques* 2003, **34**(2):374-378.
84. Pearson K: **On lines and planes of closest fit to systems of points in space.** *Philosophical Magazine* 1901, **2**(6):449-572.
85. Sherlock G: **Analysis of large-scale gene expression data.** *Briefings in bioinformatics* 2001, **2**(4):350-362.
86. Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes.** *Genome research* 1999, **9**(11):1106-1115.
87. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics (Oxford, England)* 2005, **21**(18):3674-3676.
88. Leek JT, Monsen E, Dabney AR, Storey JD: **EDGE: extraction and analysis of differential gene expression.** *Bioinformatics (Oxford, England)* 2006, **22**(4):507-508.
89. Popesku JT, Martyniuk CJ, Mennigen J, Xiong H, Zhang D, Xia X, Cossins AR, Trudeau. VL: **The goldfish (Carassius auratus) as a model for neuroendocrine signaling.** *Accepted in Molecular and Cellular Endocrinology* 2008.
90. Marlatt VL, Martyniuk CJ, Zhang D, Xiong H, Watt J, Xia X, Moon T, Trudeau VL: **Auto-regulation of estrogen receptor subtypes and gene expression profiling of 17beta-estradiol action in the neuroendocrine axis of male goldfish.** *Mol Cell Endocrinol* 2008, **283**(1-2):38-48.
91. Sohn YC, Yoshiura Y, Kobayashi M, Aida K: **Seasonal changes in mRNA levels of gonadotropin and thyrotropin subunits in the goldfish, Carassius auratus.** *General and comparative endocrinology* 1999, **113**(3):436-444.
92. Lariviere K, Samia M, Lister A, Van Der Kraak G, Trudeau VL: **Sex steroid regulation of brain glutamic acid decarboxylase (GAD) mRNA is season-dependent and sexually dimorphic in the goldfish Carassius auratus.** *Brain research* 2005, **141**(1):1-9.