UCSF
UC San Francisco Electronic Theses and Dissertations

Title

Statistical issues in the analysis of the DNA microarray data: Normalization and differential expression

Permalink

https://escholarship.org/uc/item/8526p1h9

Author

Xiao, Yuanyuan

Publication Date

2004

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Statistical Issues in the Analysis of the DNA Microarray Data: Normalization and Differential Expression

by

Yuanyuan Xiao

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Chemistry

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Date University Librarian

Degree Conferred:.....

Statistical Issues in the Analysis of the DNA Microarray Data: Normalization and Differential Expression

Copyright Spring 2004

by

Yuanyuan Xiao

Acknowledgments

Finally, I am sitting down to write my acknowledgments – the last and most personal chapter of my thesis, only now am I starting to feel and see the closure of my long, at times seemingly endless student life. Yes, I have been a student almost all my life, starting when I was 7 in a Northern city in China (and I still remembered the first Chinese character I was taught on my first day in school), and all the way until now, more than 20 years later and an Ocean apart. At this moment I am finally able to draw a conclusion to my life of learning in school, in a language that is not my mother tongue, but fittingly, it has become what I know best for my scientific endeavors.

I am not completely sure how I had chosen to be a scientist. Thinking back, it is probably not just that I "chose to" but rather that I was "chosen" to become a scientist. Having been born into a family of engineers and having always performed rather well in school, especially in the sciences, my scientific path has almost become a natural and inevitable fate. Fortunately, this choice is also one that I do not regret. Rather, I feel blessed of having been bestowed with the simple joy of finding things out and with the precious gift of being curious and inquisitive. But most of all, I feel blessed to have been surrounded by the many wonderful people that have helped, nurtured and supported my scientific pursuits along the way.

I am indebted to many people in my five years of PhD study at UCSF, or more generally, in my many years as a student. Without them, this thesis would not have come to fruition. First is my advisor Dr. C. Anthony Hunt, who has generously encouraged me to pursue my own research interests. I owe him lots of gratitude for his support, his advice and his confidence in me. My special thanks go out to Dr. Mark Segal, whose door is always open to me whenever I am in need of help, and without whom I would not have been able to complete this thesis. His vast statistical knowledge and his true personal kindness has made working with him a great joy. Thanks also to Dr. Jean Yang, who is not only a teacher, a colleague, but also a friend. Her enthusiasm in science is always contagious and she never fails to give me advice whenever I need it. Thanks to my other thesis committee members, Drs. Steve Shiboski and Hao Li, who took great effort in reading my long thesis

and providing me with valuable comments. Thanks to my friend Jenny, who has helped me tremendously during my difficult times in the first two years of my graduate study and who afterwards has remained a true and dear friend. Thanks to Yu, who is always willing to let me share her joy and her recipes, and I will always cherish our constant, sisterly and sometimes-hazardous-to-others chattings in the lab! Thanks to my parents, for instilling in me from early age an interest in science, and for tolerating me being an ocean away. Thanks to my sister, for being my best friend, and for rendering me the value of sisterhood. Last, but not the least, many many thanks to Armin and Yolanda, who are the sole reason that I look forward to coming home everyday, for teaching me appreciation and contendedness in life, and for accompanying me for my journey through life.

Abstract

Statistical Issues in the Analysis of the DNA Microarray Data: Normalization and Differential Expression

by

Yuanyuan Xiao

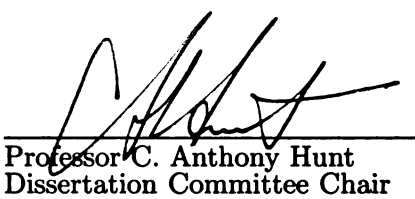
Doctor of Philosophy in Pharmaceutical Chemistry

University of California, San Francisco

Professor C. Anthony Hunt, Chair

DNA microarray technology, being one of the most powerful tools in functional genomics, profiles gene expression of any organism on a genomic scale and produces large and complex data. Analysis of such data has profound influence on the accurate extraction and deciphering of the underlying biological processes. This thesis addresses pertinent statistical issues from two aspects, the normalization of cDNA microarrays and the assessment of differential expression. The first half of this thesis focuses on normalization, which is the process that adjusts variations inherent in microarray technology rather than from biological differences. Although a number of normalization models have been proposed, it has not been well researched on how to select the most appropriate model with respect to the observed data. To this end, we propose a new within-slide normalization method, which integrates various normalization models of different complexities in the same framework and assesses their effectiveness via a quantitative criterion. Such a process is applied sequentially for the adjustments of the intensity and spatial biases. The latter half of this thesis examines the important issue of identifying differentially expressed genes amongst the many being investigated. We show that by using a distance synthesizing scheme that combines varying statistics, our proposed algorithm enjoys robust properties that are lacking in individual

statistics. Both algorithms in normalization and differential expression are tested using several microarray datasets that feature varying properties. Emphasis of the analysis of real microarray datasets is given to a novel splicing array experiment, for which we develop and illustrate appropriate methodologies in normalization and differential expression in the context of a complex experimental design.



Professor C. Anthony Hunt
Dissertation Committee Chair

Table of Contents

List of Tables x

List of Figures xii

1 Introduction 1

1.1 Background on DNA microarrays 3

1.1.1 Array construction 3

1.1.2 Target preparation and array hybridization 4

1.2 Array applications 4

1.2.1 Gene expression monitoring and comparison 4

1.2.2 Gene expression network analysis 6

1.2.3 Genetic screening 6

1.3 Thesis outline 7

2 Data 11

2.1 Spt splice experiment 11

2.1.1 Background on splicing 12

2.1.2 Background on the transcription elongation factors – Spt4 and Spt5 13

2.1.3 Spt splice microarray experiment 13

2.2 Quality control experiment (QC) 18

2.3 Affymetrix spike-in experiment 19

2.4 *Ro1* experiment 19

3 Normalization of Two-Color cDNA Arrays **22**

3.1 Introduction 22

3.2 Existing methodologies 26

3.2.1 Within-slide intensity bias 26

3.2.2 Within-slide spatial bias: Print-tip and Plate 29

3.2.3 Within-slide spatial bias: 2D Spatial 29

3.3 Stepwise normalization 30

3.3.1 Algorithm 31

3.3.2 Model selection criteria 32

3.3.3 Application of STEPNORM using BIC as the model selection criterion 35

3.3.4 Application of STEPNORM using CV as the model selection criterion 41

3.4 Comparison of normalization methods 43

3.5 Conclusion and discussion 52

4 Differential Expression **55**

4.1 Existing methods 56

4.1.1 Ranking statistics 56

4.1.2 Accessing significance (selecting cut-off) 59

4.2 Case study: Analysis of the Spt splice experiment 60

4.2.1 Data 60

4.2.2 Differential expression analysis 64

4.2.3 Comparisons of differential expression models 69

4.3 DEDS 70

4.3.1 Illustrative example 1: Affymetrix spike-in experiment 75

4.3.2 Illustrative example 2: *Ro1* experiment 78

4.3.3 Illustrative example 3: Spt splice experiment 80

4.4 Conclustion and discussion 88

4.4.1 Differential expression study of the Spt splice experiment 88

4.4.2 DEDS 88

A Appendix **102**

A.1 Differential expression models – Spt experiment 102

A.1.1 Specificities of the ANOVA mixed differential expression models . . 102

A.1.2 ANOVA table for Model I (SJ indices) 103

A.1.3 Variance components for Model I (SJ indices) 103

A.2 p-value based DEDS algorithm 104

List of Tables

2.1 Affymetrix latin square spike-in experiment. Note that transcript 407_at had concentrations mixed up, so it doesn't follow original latin square. 20

3.1 Results of application of STEPNORM on a QC slide. BIC is employed as the model selection criterion. Models chosen by STEPNORM along the normalization steps are in bold fonts. 36

3.2 Results of application of STEPNORM on a QC slide. CV is employed as the model selection criterion. "PE" in the cell indicates prediction errors estimated by CV and "se" indicates standard errors of the prediction error estimations. Models chosen by STEPNORM along the normalization steps are in bold fonts. 43

3.3 Description of normalization methods used in the comparison of bias and variance. Model complexity is approximated according to the splice data, which has 5760 data points and 16 print-tips. 47

4.1 A summary of the five competing differential expression models for the analysis of the Spt experiment. 65

4.2 ANOVA table for differential expression model I of the Spt experiment (SJ). 66

4.3 Estimation of random effects for differential expression model I. 66

4.4 Degrees of freedom table for the four ANOVA mixed models (SJ). 69

4.5 Numbers of differentially expressed genes by varying statistical measures in the *Ro1* experiment. Results of four measures for two sets of comparisons, C *vs.* T and C *vs.* Re, are presented. Differential expressed genes by the B statistics refer to genes that have positive posterior probabilities for differential expression. 80

4.6 A summary of the five competing differential expression models and the estimated number of differentially expressed genes from each model. 83

4.7 Number of differential expressed genes in SJ and IA indices of the Spt experiment 85

4.8 Distribtution of differential expressed genes in the Spt experiment. Genes are divided in ribosomal (RP) and non-ribosomal (non-RP) genes. 85

4.9 Properties of differential expressed genes with a positive fold change in the Spt experiment using the median summary. "start" is the nucleotide position in the open reading frame where intron begins; "mRNA/hr" is the number of times that a gene is transcribed per hour. Numbers in the parentheses for the non-RP genes of mutants are p-values quantifying the difference of a specific property between DE genes in a mutant and all non-RP intronic genes. For a specific property and a given mutant, the p-value is derived by calibrating the observed property against the expected under the referent null distribution, which is simulated by drawing 10,000 random samples of size n (n is different for different mutants; see Table 4.8 column 3) from the non-RP intronic gene population and taking medians of each sample. . . . 87

4.10 Properties of DE genes with a negative fold change using the median summary. 87

List of Figures

1.1 Scheme of the various steps involved in microarray experiment using: (A) spotted microarray and (B) high-density oligonucleotide microarrays. RNAs from test and reference samples are labeled with dyes and hybridized to microarray slides. The figure is modified from Dudda-Subramanya et al. (2003). 5

1.2 A summary of the major steps involved in a typical microarray project. . . 8

2.1 Splicing array probe design. There are three oligonucleotide probes for each intron-containing gene: intron(red), splice-junction(blue) and exon(green). In addition, there are also about 800 probes for intronless genes (yellow). The figure is modified from Clark et al. (2002). 14

2.2 Graphical representation of splice experimental designs. In this representation, *vertices* correspond to target mRNA samples and *edges* to hybridizations between two samples. By convention, we place the green-labeled sample at the tail and the red-labeled sample at the head of the arrow. 15

2.3 Graphical representation of the nested design of the splice experiment: the effect A is nested in S and S is in turn nested in M. Note that there are two samples (S) for each mutant, but only one sample for the wildtype. 16

2.4 Scatter plots of the logarithm intensities of splicing-related probes. Points are colored coded as indicated. 17

3.1 Graphical illustration of typical biases of expression ratios using a QC slide. a) MA plots of unnormalized ratios. The red line is *loess* (Cleveland (1979); Cleveland and Devlin (1988)) regression fit for ratios of the whole slide. The blue line displays the horizontal zero line. b) Boxplots of unnormalized ratios stratified by the 48 print-tips and 58 well-plates. 24

3.2 2D spatial plot of unnormalized ratios of a QC slide. a) Array representation of ranks of the unnormalized ratios. Spots are color coded according to their ranks; higher ratios are red and lower ratios are green. White stripes reside empty spots where no probes were deposited. b) A smooth version of a). The smoothing is carried out by *loess* 2D regression ($span = 0.2$; for details consult the Section 3.2.3.) 25

3.3 Stepwise normalization procedure using the example of the QC experiment. A QC slide has 48 print-tip-groups and 58 well plates. 32

3.4 Graphical display of biases in log-ratios of a QC slide before normalization. (a) *MA-plot* displaying the *loess* line for the whole slide (red line) and for each of the 48 print-tips (gray lines). (b) Boxplots of ratios stratified by print-tips (top) and well plates (bottom) are shown in the right panel. . . . 37

3.5 Graphical illustration of biases before (left column) and after (right column) stepwise normalization for the removal of the PT bias. (a, b)MA-plots with *loess* fit for the whole slide (red) and for each of the 48 print-tips (gray) ; (c, d) boxplots stratified by print-tip-groups and well-plate-groups. 38

3.6 Spatial plots a) before and b) after normalization by STEPNORM. Plotted are top 5% of extreme ratios of both directions. 39

3.7 Stepwise normalization procedure employing BIC as the model selection criterion on a QC slide. Colored bars represent different models within each step; the heights of the bars indicate the accuracy of fit ($-2log(\hat{L})$). Gray lines overlaying the bars are the corresponding BIC values for each model. The model chosen within each step is marked by a red star. 40

3.8 Stepwise normalization procedure on a QC slide. (a) CV is employed as the model selection criterion. Plotted are CV prediction errors with error bars for different models (colored spots) within each normalization step. Purple curve connects models selected by CV along the normalization steps. (b) BIC is employed as the model selection criterion. Plotted are BIC values for each model within each normalization step. Green curve connects models selected by BIC along the normalization steps. 44

3.9 Dot plot of MSE of the normalization methods being compared for the Spt experiment. Note to better illustrate differences between normalization methods, two of the *ceg1-250* arrays that have MSEs larger than 1.5 in the "none" method are not plotted. 48

3.10 Plot of variance and bias for the *spt4*Δ slides (left panel) and the wildtype self hybridization slides (right panel). 50

3.11 Plot of variance and bias for the stepwise normalization of a *spt4*Δ slide. Colored points represent the variance (black) or the bias (purple) of each model within a normalization step. Asterisks mark models selected by STEPNORM. 51

4.1 Boxplots of normalized ratios of splicing related probes stratified by mutants. Splice junction and exon probe ratios show a shift from the horizontal zero line in the negative direction, whereas intron probe ratios are centered at zero. 62

4.2 Scatter plots of normalized ratios of splicing related probes. Points are colored coded by their mutant identity. Gray horizontal and vertical reference lines indicate zero expression ratios. 63

4.3 Scatter plot matrix of differential expression models for SJ indices of the (a) *ceg1-250* and (b) *spt4dΔ* mutants. Plotted are the $-\log_{10}(p-values)$ of the corresponding models. Correlation coefficients between respective models are shown in the lower triangle of the matrix. 71

4.4 A graphical representation of the motivation behind DEDS. Plotted are the multivariate point cloud with a point corresponding to a gene's vector of two related differential expression measures, M1 and M2. E is the "extreme origin" in the direction of differential expression. Two scenarios are presented: when M1 and M2 show good (panel (a)) or bad (panel (b)) concordance. Red spots are genes that are ranked highly by both measures and blue spots are genes that are ranked highly by one measure but lowly but the other. . . . 73

4.5 Comparisons of different differential expression measures using ROC curve for the Affymetrix spike-in experiment. (a) Use RMA probe summary. (b) Use MAS 5.0 probe summary. 77

4.6 Scatter plot matrix of differential expression statistics for a two-sample comparison (12 *vs.* 12) study of the Affymetrix spike-in experiment. Plotted are the statistics of four measures: FC, t, penalized t and SAM. Correlation coefficients between respective measures are shown in the lower triangle of the matrix. Q-Q plots of the four measures are illustrated on the diagonal of the matrix. Genes, selected by DEDS summarizing the above four measures when controlling for FDR at 0.01, are highlighted in red. For better illustration, the most significant gene, whose statistics are much larger than the rest of genes, is not plotted. 79

4.7 (a) Histogram and (b) $Q-Q$-plot of p values from the mutant *ceg1-250* in the splice experiment. (c) Histogram and (d) $Q-Q$-plot of p values from the Affymetrix spike-in experiment where all but 14 genes remain unchanged. When no gene is expected to be differentially expression, the distribution of p values is uniform, shown as dashed line in the figure. 82

4.8 Three-dimensional scatter plots of $-log_{10}p$ values from differential expression (DE) models for the mutant *ceg1-250* (SJ indices) in the Spt experiment. Black spots are DE genes by DEDS and gray spots are non-DE genes by all measures including DEDS. Colored numbers represent non-DE genes by DEDS but are DE genes by one or more of the five models. Panel (a) provides coding for the colored numbers; 1 means DE and 0 means non-DE. 84

4.9 Volcano plot of fold change (FC) *vs.* $|t|$ for the Ro1 experiment (C versus T). 91

Chapter 1

Introduction

The first complete genome sequence of a free-living organism, *Haemophilus influenzae*, which consists of 1.8 million nucleotides and 1,749 genes, was completed in 1995 (Fleischmann et al. (1995)). It was followed by an explosion of determinations of complete genome sequences, including many genetics model organisms, for instance, *E. coli* (Blattner et al. (1997)), *S. cerevisiae* (Goffeau et al. (1996)), *C. elegans* (The *C. elegans* sequencing consortium (1998)) and *D. melanogaster* (Adams et al. (2000)). Most recently, scientists have completed the sequences of the laboratory rat (Rat genome sequencing consortium (2004)), marking the third successfully sequenced mammalian genome, after the human (Venter et al. (2001)) and the mouse (Mouse genome sequencing consortium (2002)). The resulted large volumes of genome-wide information have brought biology to an exciting post-genomic era, characterized by the challenging task of assigning biological meaning to the vast amount genomic data. To fulfill such a task, a new level of systematic approach is required to answer questions about when, where and how genes are expressed to create the complex and dynamic living systems. Tools in functional genomics in particular are making rapid progress to decipher genomic information by the analysis of gene expression profiles at the mRNA and protein levels and the analysis of polymorphisms and mutation patterns in the genome, and they do so by tackling all genes in a genome simultaneously in contrast to the classical approaches that did so one gene at a time (Kanehisa (2000)).

The DNA microarray technology, being one of the most powerful tools in functional genomics, profiles *gene expression* of any organism on a genomic scale using miniature biological assays (Lockhart et al. (1996); Cho et al. (2001); DeRisi et al. (1996)). *Gene expression* is the term used to describe the transcription of the information encoded in the DNA into mRNA, which is then translated into proteins that carry out most of the functions critical for a living cell. Gene expression profiles are major determinants of cellular phenotypes and functions. DNA microarrays examine gene expression profiles at the mRNA level. The principle of a microarray experiment exploits *hybridization specificity* for detection of the presence of individual target sequences in a complex transcription mixture. More specifically, mRNA from a given cell line or tissue is extracted, labeled and then hybridized in parallel to a large number of DNA sequences procured from individual genes and immobilized on a solid surface. Because of the highly parallel process of the hybridization between each RNA or DNA molecule in solution with a matching partner on the array, microarrays are capable of surveying the expression of hundreds or thousands of genes simultaneously, in sharp contrast to the much less powerful classical Northern-blotting analysis. During recent years, DNA microarray technology has been advancing rapidly. When the first biological application of DNA microarray was published in 1995 (Schena et al. (1995)), there were only 45 *Arabidopsis* genes on a slide. In sharp contrast, we are now able to quantify tens of thousands of transcripts simultaneously on a single slide.

Having presented a brief summarization of the microarray technology, we proceed next to Section 1.1 for a review of the technology underpinning the DNA microarray. A short introduction of the typical processes involved in a microarray experiment is provided. Section 1.2 examines the various questions in biology and medicine, especially in the post-genomic era, that microarray can be used to address. Section 1.3 concludes this chapter by discussing the challenges associated with the analysis of the complex data that a microarray experiment typically produces. An outline of this thesis, which explores several statistical issues in microarray data analysis, is provided at the end of this chapter.

1.1 Background on DNA microarrays

The most commonly used systems today can be divided into two groups, according to the array construction: spotted microarrays (Brown and Botstein (1999) and high density oligonucleotide arrays (Lockhart et al. (1996)).

1.1.1 Array construction

For oligonucleotide arrays, the arrayed materials, which have been generally termed "probes", are synthesized *in situ* as short 20-25mers either by photolithography onto silicon wafers or using ink-jet technology. For example, the Affymetrix GeneChip^{TM} typically employs multiple probe sequences, represented by 16-20 probe pairs, to assay expression of a given gene. Each probe pair consists of a perfect match (PM) probe and a mismatch (MM) probe; PM and MM probes differ only at the middle base, which strongly disrupt hybridization and therefore measures non-specific binding.

Spotted arrays, on the other hand, most commonly use cDNA probes, which are products of polymerase chain reaction (PCR) generated from cDNA libraries, but recently have seen more and more usage of presynthesized oligonucleotides (Barczak et al. (2003)), due to the falling of the prices of commercially synthesized oligonucleotides. These probes are then deposited onto a solid surface in defined locations by a robotic arrayer. The arrayer has a grid of pins, or print-tips, which pick up a set of probes from a 96- or 384-well plate and print them on the slide.

Spotted arrays offer a great degree of flexibility in probe selections, especially for the construction of customized arrays to the likings of the investigator, for instance arrays that assay only selected pathways. The usage of cDNAs as arrayed elements also provides the opportunity to examine unsequenced clones from cDNA libraries.

1.1.2 Target preparation and array hybridization

Another difference between the high-density, *in-situ* synthesized oligonucleotide microarrays and the spotted arrays lies in target preparation (Figure 1.1). Spotted microarrays uses a two-color scheme and it is standard practice to compare the gene expression of *two* biological samples on *one* chip. mRNAs from the two cell samples to be compared are extracted, amplified and labeled with two different dyes, typically the fluorescent dye Cy5 (red) and Cy3 (green). The labeled targets are then competitively hybridized to the slide and scanned at two different wavelengths to detect the relative transcript abundance for each spot in both channels. The high-density oligonucleotide arrays, on the other hand, employ a one-color scheme and require at least two arrays for a comparison of expression levels between two samples.

1.2 Array applications

Microarray based transcriptional profiling allows manifold applications.

1.2.1 Gene expression monitoring and comparison

So far, much of the interest in microarrays has been directed towards identifying genes, the regulated expression of which could help explain particular biological phenomena. More specifically, a typical question of microarray data analysis is to identify differentially expressed genes between different biological conditions or treatments. That is, genes whose expression levels are associated with a response variable of interest. The response variables could be either polytomous (e.g. treatment/control status, tumor class) or continuous (e.g. time, blood pressure), they could consist of censored survival times or multiple clinical phenotypes. Assessment of differential expression typically involves associating each gene with a statistical score indicative of the strength of the evidence of differential expression. We provide a detailed description of relevant statistics typically applied for the detection of differential expression in Section 4.1.

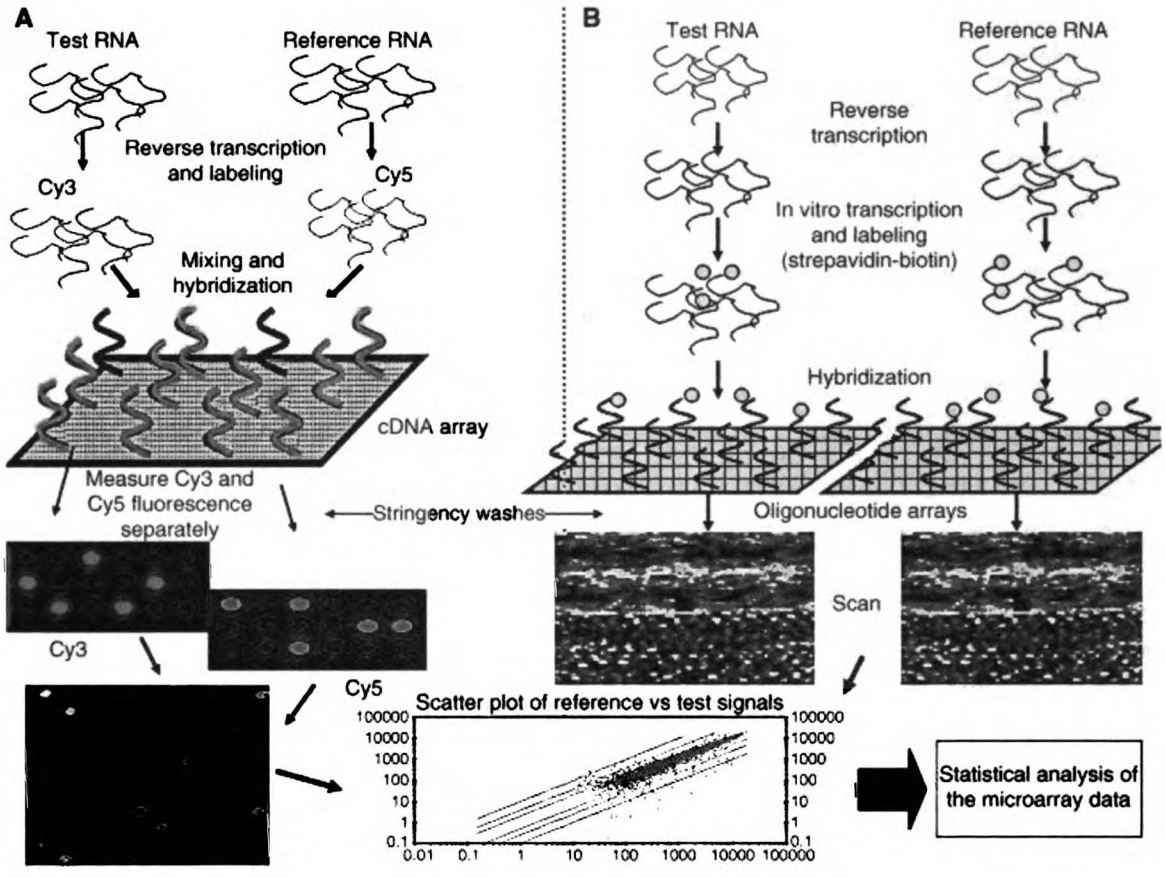


Figure 1.1: Scheme of the various steps involved in microarray experiment using: (A) spotted microarray and (B) high-density oligonucleotide microarrays. RNAs from test and reference samples are labeled with dyes and hybridized to microarray slides. The figure is modified from Dudda-Subramanya et al. (2003).

1.2.2 Gene expression network analysis

The microarray technology can be used to provide a global picture of gene expression, where clusters of genes are analyzed in terms of common functionalities. Such study is typically based on the assumption that genes with closely associated expression patterns may share similar regulatory mechanisms; hence it makes it possible to assign potential functions to new genes (Schulze and Downward (2001)). When coupled with sequence information, this has led to the discovery of common DNA regulatory motifs in the promoter region of co-regulated genes. As one of the earliest such studies, Cho et al. (2001) examine transcriptional regulation and function during the human cell cycle. To this end, various softwares have been developed to cluster genes with similar expression patterns and allow similar clusters to be connected in a phylogenetic tree. The techniques involved include, for instance, hierarchical and k-means clustering (Eisen et al. (1998)), principal component analysis (Alter et al. (2000)) and self-organizing maps (Tamayo et al. (1999)). By extension, instead of clustering genes, one can also cluster samples by patterns of gene expression; doing so enables the subdivision of a patient population into classes based on the molecular-genetic profiles. A successful classification of disease subtype based on gene-expression profiles is provided by Alizadeh et al. (2000). in the discovery of two patterns of expression in the diffuse large B-cell lymphoma (DLBCL).

1.2.3 Genetic screening

The use of microarrays to understand genome-wide gene expression pattern has become a standard practice, however, genome-wide genetic screening is still in its infancy. In genetic screening applications, DNA samples are examined to identify single-nucleotide polymorphism (SNP), deletions and other minor sequence variants in genes. Interest in SNP genotyping is growing rapidly, driven by two major potential applications. The first is disease genetics, that is, to associate a specific SNP or a group of SNPs that occur together (haplotype) with increased likelihood of genetic and infectious diseases. The second major application is pharmacogenomics, which examines how different genetic markers associate with different responses to a given therapeutic treatment. DNA microarrays could

be used as a quick method to perform genetic screening of an individual. An example of such genome-wide genetic scanning microarray is introduced by Affymetrix which genotypes 10,000 SNPs spread evenly throughout the genome. Such application allows an accurate and quick identification of individuals susceptibility for certain diseases.

1.3 Thesis outline

As DNA microarray technologies continue to advance, more and more emphasis is being placed on data analysis. Figure 1.2 summarizes the typical steps for the acquisition and analysis of array data. Microarray experiments typically generate large and complex multivariate data sets, which present various challenges and complications in data analysis and information extraction. These include (i) adjusting for many sources of variability arising from the probe and target preparation and array fabrication; (ii) detecting of gene expression pattern both at the single-gene and multi-gene level; (iii) devising methods that are geared to the novel data structures – thousands of inter-related variables (genes), small sample sizes (arrays), and little or no replication, and (iv) assessing the significance of the finding, when needed, by accommodating multiple testing concerns. The complexities in array data analysis call for robust and integrated analysis tools to enhance reliability and efficiency.

This thesis explores a number of pertinent statistical issues in the analysis of DNA microarray expression data. More specifically, we examine aspects of microarray data analysis related to the preprocessing cDNA microarrays and to the assessment of differential expression. We develop new statistical methodologies for both areas; in addition, we showcase a novel splicing array technology and develop methodologies for its analysis in the context of a real and complex experimental design.

This thesis is organized as follows. We begin in Chapter 2 with the introduction of four microarray datasets used in methodology development, featuring both Affymetrix high density oligonucleotide microarrays and spotted microarrays. The emphasis is placed on a set of twenty two splicing-sensitive DNA microarrays that investigate splicing defects in the chro-

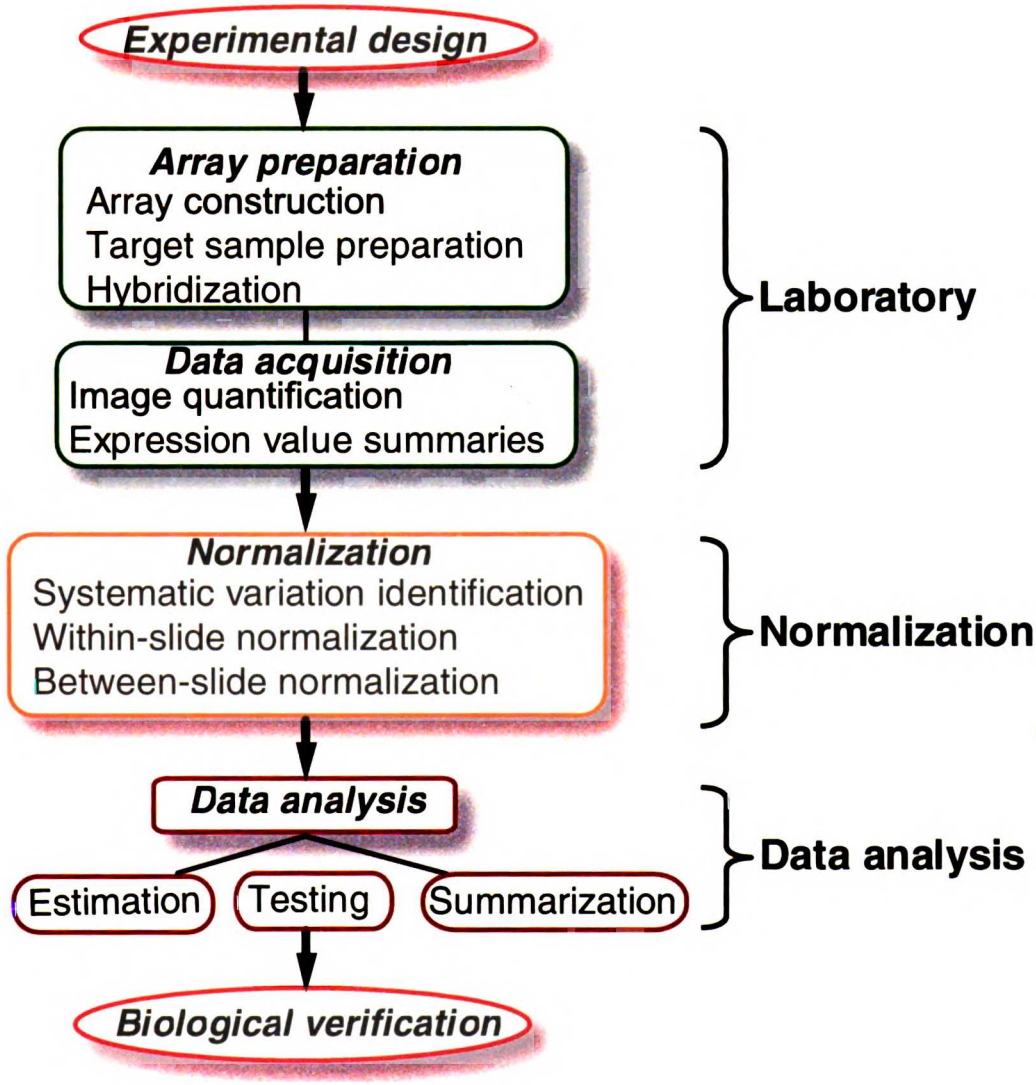


Figure 1.2: A summary of the major steps involved in a typical microarray project.

matin elongation factor mutants, *spt4* and *spt5*, in yeast *S. cerevisiae*. A brief introduction of eukaryotic mRNA splicing and the roles of Spt4 and Spt5 in splicing is provided as well as a detailed description of the study design. The data for this study are provided by Dr. Grant Hartzog at University of California, Santa Cruz and are instrumental in the initiation of both of our new algorithms in normalization and differential expression proposed later in this thesis.

Chapter 3 summarizes current methodologies in spotted microarray normalization and proposes a new method that conducts normalization in a stepwise fashion. Normalization is generally referred to as the process of adjusting non-biological systematic variations in gene expression measurements to enable meaningful within-slide and between-slide summarizations. Such systematic variations arise from various array processing steps, for instance, a difference in the amount of samples, a difference in dye efficiency and a difference in the physical conditions of print-tips, to mention just a few. We propose a normalization framework that integrates various existing normalization models of different complexities for the sequential detection and adjustment of common systematic variations. In addition, based on the splice array data which utilizes a gene-specific control for self normalization, we compare the performance of the stepwise normalization procedure with a few widely applied methods using the criteria of bias and variance.

We continue in Chapter 4 to discuss the issues of differential expression in microarray data analysis. The task of identifying differentially expressed genes consists of two components: gene ranking and gene selection. Numerous statistics have been proposed to rank genes in order of evidence for differential expression. However no one statistic is universally optimal and there is seldom any basis or guidance that can direct toward a particular statistic of choice. Specifically, for the set of Spt splice arrays, several ANOVA models and a semi-parametric empirical Bayesian model for differential expression are all appropriate, and there is no clear advantages of one model over the others. Rather than trying to arbitrate between models and pick a single model on which to base DE rankings and declarations, or informally distilling sets of genes that are differentially expression under two or more models, we develop a new robust differential expression framework for synthesizing differing

statistics. We evaluate its performance on three diverse microarray datasets including the splice data.

Chapter 2

Data

We introduce here four microarray datasets used in methodology development in studies of cDNA microarray preprocessing and differential expression. Both Affymetrix high density oligonucleotide microarrays and spotted microarrays are featured.

2.1 Spt splice experiment

This experiment consists of a set of twenty two splicing-sensitive DNA microarrays that investigate splicing defects in the chromatin elongation factor mutants, *spt4* and *spt5*, in yeast *S. cerevisiae*. This dataset is instrumental in the development of both of our new algorithms in normalization (see Chapter 3) and differential expression (see Chapter 4) proposed later in this thesis.

The description of this dataset is organized as follows. We commence with a brief introduction of eukaryotic mRNA splicing and the roles of Spt4 and Spt5 in splicing. We then proceed to a discussion of the composition of the splicing-sensitive DNA microarrays, which are specially designed to distinguish between normal and abnormal splicing in the intron-containing genes. Last, we provide a detailed description of the experimental design.

2.1.1 Background on splicing

Eukaryotic genes are fragmented into exons by intervening sequences (introns). After a gene is transcribed into pre-mRNA, the introns are removed from the transcript and the exons are joined by the spliceosome. This reaction, splicing, can also be used to create multiple transcripts from a single gene. For example, a particular exon may be included in one version of a mRNA, and skipped in another. This process of alternative splicing is subject to regulation in response to tissue, developmental and environmental cues, and is credited as the most important source for the extraordinary enrichment of the human proteome, the size of which is beyond the coding capacity of the human genome (Maniatis and Tasic (2002)). In humans, most genes are subject to splicing and more than half are likely subject to alternative splicing (Maniatis and Tasic (2002)). Accurate splicing is crucial for normal protein function; aberrant transcripts due to splicing mutations are known causes for 15% of genetic diseases (Maniatis and Tasic (2002)). Therefore, elucidation of splicing mechanisms will not only help us understand the operating mechanisms underneath the functional complexity and diversity of higher eukaryotes, but also aid in new therapeutic strategies for treatments in splicing-related genetic disorders.

Although the different steps of gene expression are typically studied in isolation, it is clear that there are important functional links between them (Maniatis and Reed (2002); Proundfoot et al. (2002)). For example, the process of capping the 5' end of pre-mRNAs is thought to influence both transcription and splicing (Schwer and Shuman (1996); Fresco and Buratowski (1996)). Furthermore, the rate of transcription elongation appears to influence splicing and alternative splice site choice (Howe et al. (2003)). In addition, a number of pre-mRNA processing factors are recruited to transcripts via interaction with RNA polymerase II (Maniatis and Reed (2002); Proundfoot et al. (2002)). Thus, a comprehensive understanding of mRNA synthesis will require an understanding between these functional linkages of steps in gene expression.

2.1.2 Background on the transcription elongation factors – Spt4 and Spt5

We have recently found evidence that the chromatin elongation factors, Spt4 and Spt5, play a role in RNA processing in the yeast, *S. cerevisiae*. Spt4 and Spt5 form a complex that regulates transcription elongation by RNA polymerase II. This complex is conserved across eukaryotes and has been proposed to both facilitate transcription by removing a nucleosomal barrier to transcript elongation and also suppress inappropriate transcription by reassembling nucleosomes behind transcribing polymerase (Hartzog et al. (2002)). The recent finding that Spt5 interacts physically and genetically with pre-mRNA capping factors suggests a role for Spt4-Spt5 in capping (Pei and Shuman (2002); Wen and Shatkin (1999); Lindstrom et al. (2003)). Because pre-mRNA capping is thought to increase the efficiency of splicing, we further analyzed splicing in *spt4* and *spt5* mutants and found that several genes were not spliced with normal efficiency (Lindstrom et al. (2003)). Thus, the function of Spt4-Spt5 is linked to the processing of at least several mRNAs.

2.1.3 Spt splice microarray experiment

Traditionally, splicing is studied on an individual gene basis by *ad hoc* experiments. With the advent of eukaryotic genomic sequences, a global genomic view of splicing is rendered achievable and will provide an unprecedented amount of information regarding the mechanisms and regulation of splicing (Modrek and Lee (2002)). *S. cerevisiae*, a simple yeast that has been used as a model to study eukaryotic gene expression, presents a convincing entry point to embark on this task. The yeast genome is completely sequenced and well annotated, and the splicing machinery of yeast is well conserved with that of humans. Among the more than 6,200 genes in yeast genome, only about 250 of them possess introns and only a handful have multiple introns or are alternatively spliced (Barrass and Beggs (2003)). However, these 250 intronic genes give rise to 27% of the transcripts synthesized by the cell, an indication of the importance of splicing in yeast (Lopez and Séraphin (1999); Ares et al. (1999)). Clark et al. (2002) have recently designed a DNA microarray for analysis of splicing in yeast. To discriminate between spliced and unspliced transcripts for intron-containing yeast genes, oligonucleotide probes on these arrays were designed to detect splice junc-

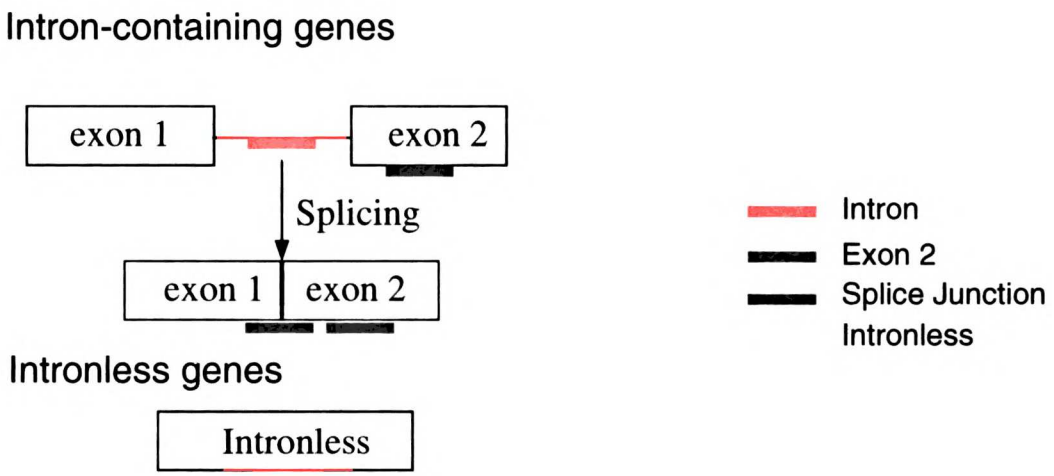


Figure 2.1: Splicing array probe design. There are three oligonucleotide probes for each intron-containing gene: intron(red), splice-junction(blue) and exon(green). In addition, there are also about 800 probes for intronless genes (yellow). The figure is modified from Clark et al. (2002).

tions, introns and second exons of spliced genes (Figure 2.1). Splice junctions are found in spliced transcripts whereas introns exist only in unspliced transcripts. The second exon are present in both spliced and unspliced transcripts and are good indicators of total transcript level. These arrays are then competitively hybridized with probes derived from control and experimental samples.

We extend our observations of splicing defects in *spt4* and *spt5* mutants to the entire collection of intron-containing genes, using these splicing-sensitive DNA microarrays (Clark et al. (2002)). Although simple inspection and clustering of this data reveal gene expression changes in *spt4* and *spt5* mutants, we seek more rigorous methods for analysis of splicing microarray data to allow for identification of differentially expressed genes.

In yeast, *SPT4* is a nonessential gene encoding a 102 amino-acid protein and *spt4*Δ (null) mutants display mutant phenotypes and genetic interactions consistent with an elongation defect (Hartzog et al. (2002)). In contrast to *SPT4*, *SPT5* encodes a large protein that is essential for life, and *spt5* mutations typically display mutant phenotypes and genetic

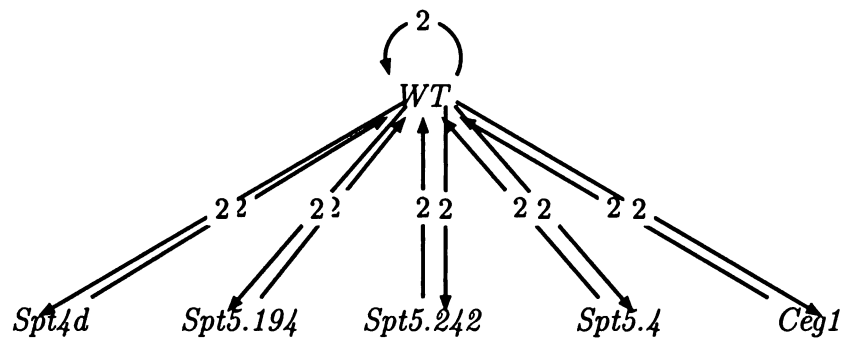


Figure 2.2: Graphical representation of splice experimental designs. In this representation, *vertices* correspond to target mRNA samples and *edges* to hybridizations between two samples. By convention, we place the green-labeled sample at the tail and the red-labeled sample at the head of the arrow.

interactions similar to those observed for *spt4* mutations (Hartzog et al. (2002)). In this work, we have analyzed a *spt4* null mutation, and three partial loss-of-function mutations in *SPT5*. Two of these, *spt5-4* and *spt5-194*, encode versions of Spt5 that are defective for binding Spt4. The third allele, *spt5-242*, causes a cold-sensitive growth defect (Hartzog et al. (1998)), and displays splicing defects at all temperatures. The Spt5-242 protein still binds Spt4, even at the non-permissive temperature. In addition, we include analysis of *ceg1-250*, a temperature-sensitive mutation that causes rapid inactivation of the capping enzyme at the non-permissive temperature (Fresco and Buratowski (1996)). Two independent mRNA samples were prepared from each mutant, fluorescently labeled and then hybridized to the splicing arrays competitively with a probe derived from wildtype cells. Experiments were performed using a replicated dye-swap study design (Figure 2.2), analogous to standard two-channel spotted arrays (Chen et al. (1997)). Briefly, there are four arrays (A1-A4) for each mutant *vs.* wildtype experiment. Arrays A1 and A2 had the same sample origin for the mutant; A3 and A4 featured the other sample. In A1 and A3, the red dye Cy5 was assigned to the mutant, and the green dye Cy3 was assigned to the wildtype. The dye assignment was reversed for arrays A2 and A4. In addition to these twenty mutant arrays (4 arrays X 5 mutants), there were two separate wildtype self-hybridization experiments, in which the wildtype was labeled with both Cy5 and Cy3.

To provide a global view of splicing defects in the *ceg* and *spt* mutants, we plot unnor-

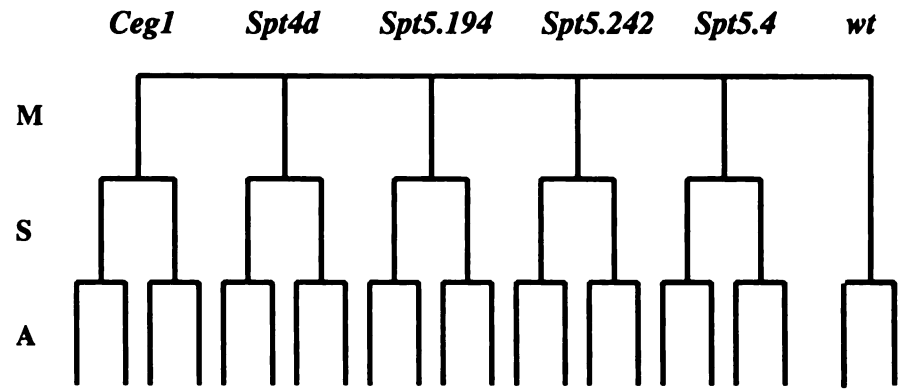


Figure 2.3: Graphical representation of the nested design of the splice experiment: the effect A is nested in S and S is in turn nested in M. Note that there are two samples (S) for each mutant, but only one sample for the wildtype.

malized log-intensity values for splicing related oligonucleotides from the two channels, mutants against wildtype, in Figure 2.4. Points (array features) are color coded so that features corresponding to exon, splice junction, intron and intronless genes can be visually differentiated. Genes lying on the diagonal have a ratio close to 1, and their expression in the mutants is therefore largely unaffected. For *ceg1-250*, shown in the upper left panel, introns (light blue points) deviate noticeably from the diagonal toward the *ceg1-250* axis. This is a clear indication of intron accumulation in the *ceg1* mutant because of the splicing abnormality. Splice junctions (dark blue points) in *ceg1-250*, on the other hand, largely display ratios < 1, indicating a decrease in splice junction formation, again in accordance with known defects. Compared with *ceg1-250*, the four *spt* mutants exhibit fewer alterations in splicing, with the *spt5-194* mutant most severely affected, in agreement with its phenotypic characteristics. A control plot from the wildtype self-hybridization is depicted in the lower right panel. As expected, no separation is observed in introns and splice junctions, and all points conform closely to the diagonal.

To characterize the loss of splicing in mutants, we calculate two indices as described in Clark et al. (2002). The splice junction (SJ) index is the change of the splice junction probe signal normalized by the change of overall gene expression level as measured by the related

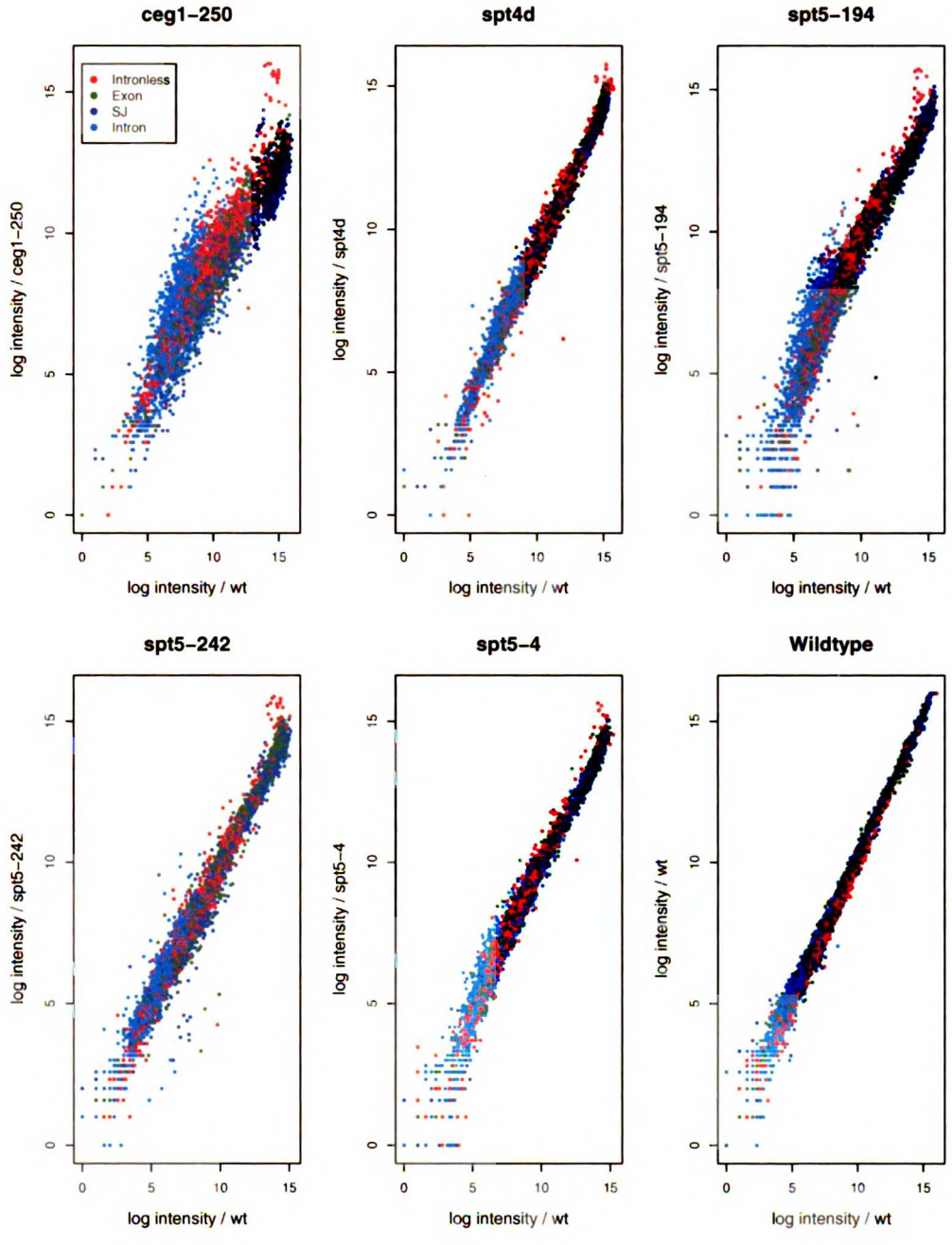


Figure 2.4: Scatter plots of the logarithm intensities of splicing-related probes. Points are colored coded as indicated.

exon probe signal:

$$SJ - index = \log \frac{MUT_{S-J}/WT_{S-J}}{MUT_{Ex}/WT_{Ex}}, \tag{2.1}$$

where MUT_{S-J}/WT_{S-J} represent the intensity ratio of a splice junction probe between a mutant and the wildtype and MUT_{Ex}/WT_{Ex} is the corresponding intensity ratio of the exon probe. Similarly, the intron accumulation (IA) index is obtained as the normalized change of the intron probe signals:

$$IA - index = \log \frac{MUT_{Int}/WT_{Int}}{MUT_{Ex}/WT_{Ex}}, \tag{2.2}$$

where MUT_{Int}/WT_{Int} represent the intensity ratio of an intron probe between a mutant and the wildtype and MUT_{Ex}/WT_{Ex} is the corresponding ratio of the exon probe. Relating changes in the splice junction and intron signals to changes in the second exon takes into account changes in overall expression level that may occur as a result of alterations in other steps of gene expression.

We discuss later in Chapter 3 the self-normalizing property of the ratios of the exon probes. The existence and the usage of the exon probes constitutes a "gold standard" for normalization purpose and can therefore be exploited for a comparison study among a variety of normalization methods (see Chapter 3.4 for details). As the goal of this study is to identify genes whose splicing relies on the Spt4 and Spt5, we develop appropriate differential expression models using SJ and IA indices in Chapter 4.2, and provide biological interpretations of the findings.

2.2 Quality control experiment (QC)

In this experiment, the arrays are fabricated with 70-mer oligonucleotide probes from the Operon Human Genome Oligo version 2, supplemented with some custom-designed 70-mer oligonucleotides. A total of 21,357 probes for 10,801 gene clusters were printed on the array by an arrayer having 4 by 12 print-tips. Each print-tip group consists of 21 × 23 spots. This series of experiments were produced for purpose of print-run quality control

at the UCSF NHLBI Shared Microarray Facility (`http://arrays.ucsf.edu/`). The hybridization performed, measure differential gene expression in two RNA samples, the K562 erythroleukemia RNA, and the Stratagene Universal Human Reference (SFUHR) which is a pool of 10 different cell lines. After hybridization, arrays were scanned using an Axon 400B laser scanner and images were processed using GenePix 5.0 software. Similar data can be found in Barczak et al. (2003) and we have developed and tested our novel stepwise within-array normalization algorithms in Chapter 3.3.3 with data from this series of experiments.

2.3 Affymetrix spike-in experiment

The Affymetrix spike-in experiment is used for differential expression study in this thesis. It consists of 59 arrays and a series of human genes spiked-in at known concentration. They represent a subset of data used to develop the Affymetrix MAS 5.0 algorithm. This set of experiment uses a Latin square design with 14 spiked-in gene groups in 14 array groups among 12612 null genes. The concentration of the first spike-in group are $0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512$ and 1024 pm. Each subsequent experiment rotates the spike-in concentrations by one group; i.e. the second experiment begins with 0.25pM and ends at 0pM (see Table 2.1). Each "row" of the Latin square (given spike-in gene at a given concentration) was replicated (typically 3 times, two rows 12 times, 59 arrays in total). Further information can be found at `http://www.affymetrix.com/analysis/download_center2.affx`. The use of spike-ins allows us to compare varying differential expression measures in terms of the accurate identification and ranking of the differentially expressed (the spike-ins) genes. We discuss this usage in Chapter 4.3.1.

2.4 *Ro1* experiment

This is a study of dilated cardiomyopathy in transgenic mice utilizing Affymetrix Mu6500 chips. These transgenic mice over expressed a G protein-coupled receptor, *Ro1*, which is

Table 2.1: Affymetrix latin square spike-in experiment. Note that transcript 407_at had concentrations mixed up, so it doesn't follow original latin square.

	Transcripts													
Experiments	37777_at	684_at	1597_at	38734_at	39058_at	36311_at	36889_at	1024_at	36202_at	36085_at	40322_at	407_at	1091_at	1708_at
A	0	0.25	0.5	1	2	4	8	16	32	64	128	0	512	1024
B	0.25	0.5	1	2	4	8	16	32	64	128	0	512	1024	0
C	0.5	1	2	4	8	16	32	64	128	0	512	1024	0	0.25
D	1	2	4	8	16	32	64	128	0	512	1024	0	0.25	0.5
E	2	4	8	16	32	64	128	0	512	1024	0	0.25	0.5	1
F	4	8	16	32	64	128	0	512	1024	0	0.25	0.5	1	2
G	8	16	32	64	128	0	512	1024	0	0.25	0.5	1	2	4
H	16	32	64	128	0	512	1024	0	0.25	0.5	1	2	4	8
I	32	64	128	0	512	1024	0	0.25	0.5	1	2	4	8	16
J	64	128	0	512	1024	0	0.25	0.5	1	2	4	8	16	32
K	128	0	512	1024	0	0.25	0.5	1	2	4	8	16	32	64
L	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128
M,N,O,P	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256
Q,R,S,T	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512

a mutated form of the human kappa opioid receptor and signals through the G_i pathway. Expression of *Ro1* is controlled by the use of an inducible expression system (Redfern et al. (2000)). When the receptor is overexpressed in the hearts of adult mice, the mice exhibit cardiomyopathy symptoms. When the expression of the receptor is turned off, the mice recover. The cardiomyopathy is due to hyperactive signaling of the receptor; treatment of the mice with a receptor antagonist or with pertussis toxin reverses certain phynotypes associated with the disease. To determine changes in gene expression due to the overexpression of *Ro1*, micorarray experiments were carried out. mRNA samples were isolated from the ventricles of thirty mice (corresponding to 4 different *Ro1* status) and hybridized one heart per set of arrays. Details of related biological background and the experimental design can be found in Redfern et al. (2000). Segal et al. (2003) applied regression approaches to locate genes with differential gene expression resulting from the hyperactive signaling of *Ro1* in these mice.

We use this dataset for identifying differential expression between two sets of comparisons. The first comparison between a control group and a treatment group is expected to yield a large amount of differential expression, whereas the latter comparison between the control group and a "recovered" group will likely provide only modest expression change. This provides a convincing test bed to assess the ability of our new differential expression algorithm to discern differences between these two cases. Details of the differential expression study is provided in Chapter 4.3.2.

Chapter 3

Normalization of Two-Color cDNA Arrays

3.1 Introduction

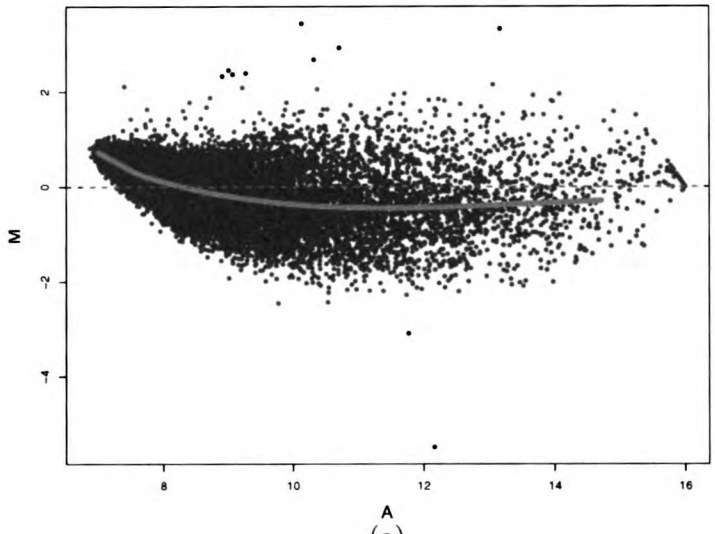
Two-channel microarrays measure relative abundance of expression of thousands of genes in two mRNA populations. This relative abundance is usually expressed as ratios (DeRisi et al. (1996);Chen et al. (1997)), $M = log_2(R/G)$, where R and G are the fluorescent intensity measurements of the red and green channels. After the logarithm transformation, the ratios show good agreement with a normal distribution, and could be subject to statistical hypothesis tests that require normality. Before the ratios (M) could be interpreted as "fold changes" of genes between two mRNA samples under investigation, normalization needs to be applied. In a two-color cDNA array setting, normalization calibrates signals from the two channels or different arrays to remove non-biological bias introduced by various processing steps and to enable meaningful comparisons between channels or arrays. Several common sources of systematic errors contribute to artifacts in microarray data, for instance, unequal quantities of starting RNAs, different efficiency of dye incorporation, variable dye self-quenching and spatial variation across a slide, to name just a few.

The most pronounced such bias is the imbalance of the green and red dye incorporation.

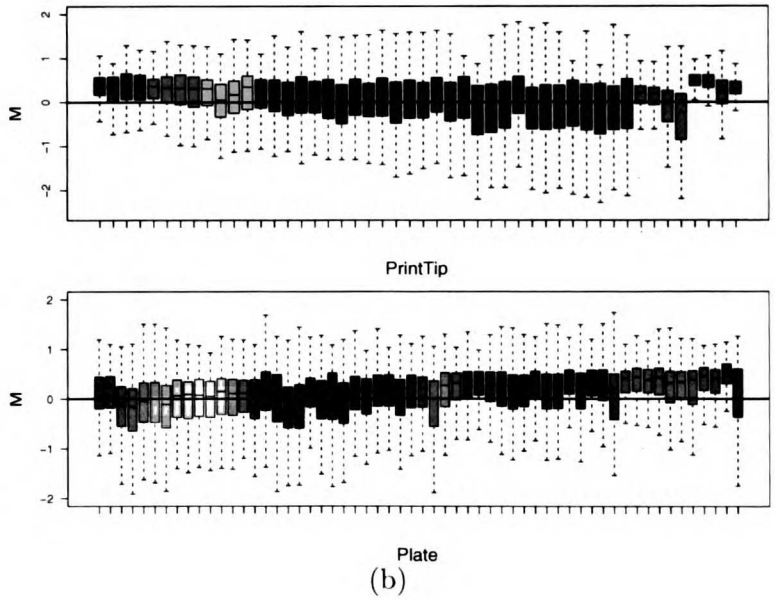
This imbalance is manifested as the dependence of ratios on primarily two factors, the fluorescent intensity (A) and the spatial (S) heterogeneity; the efficient removal of such biases will be the focus of this chapter.

The A bias can be best illustrated using a MA-plot (Yang et al. (2002)), where M is plotted against A ($A = log_2\sqrt{RG}$). As the assumption is that the majority of genes are equally expressed between the two mRNA samples, symmetrical distribution of points around the horizontal $M = 0$ line is expected. Any linear or nonlinear trend away from the zero line between M and A signals the undesirable existence of the A bias. Figure 3.1 displays a MA-plot of a slide from the QC experiment. The general trend of the points in Figure 3.1 displays obvious curvature and indicates a nonlinear relationship between M and A.

The S bias manifests itself as a non-uniform distribution of log-ratios over the surface of the array. It originates from different experimental conditions applied on spots from different areas on the slide. There are usually three major sources that contribute to this spatial variation. First, spots on the same slide are divided into different grids, and spots of different grids are printed by different print-tips from the printing robot. Second, spots of different rows of the slide are often of different well plate sources; one can imagine that there may be effects associated with different well plates. Indeed when we stratify the ratios (M) according to print-tips and well-plates (Figure 3.1(b)), differences among such factors are apparent. Lastly, in addition to print-tip and plate effects, the physical condition of the slide itself could also differ region from region, resulting in correlation of the ratios with their physical distance on the slide. To best illustrate this phenomenon, we have plotted ranks of the log-ratios according to its physical locations on the array in Figure 3.2. The left panel of Figure 3.2 shows that the unnormalized ratios are not uniform, and are lower (green) at the middle and higher (red) at the lower left corner. This trend is more obvious in the right panel, where the smoothed version of ratios are plotted.

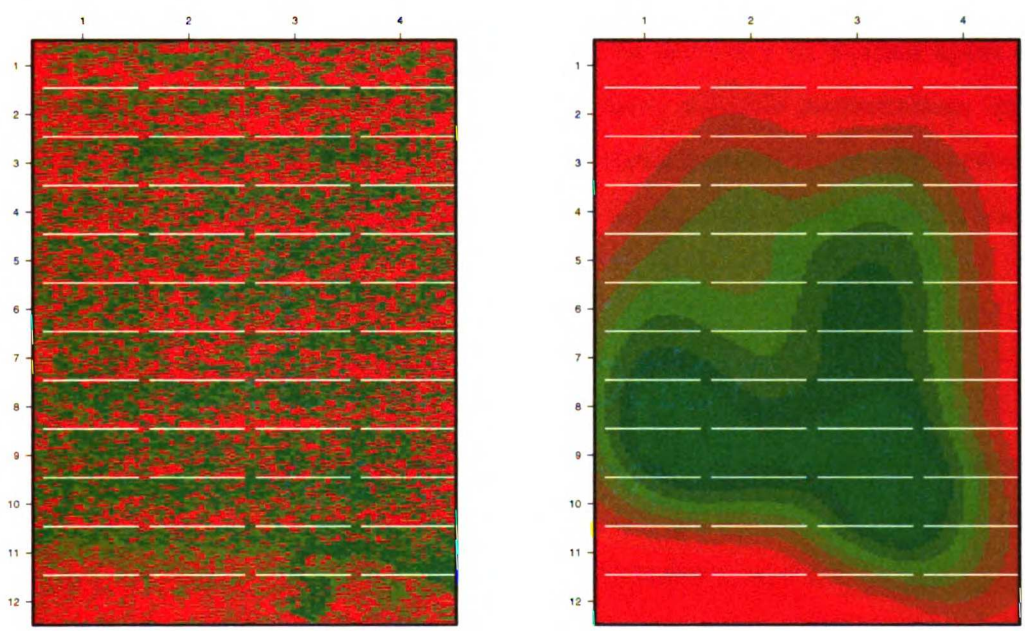


(a)



(b)

Figure 3.1: Graphical illustration of typical biases of expression ratios using a QC slide. a) MA plots of unnormalized ratios. The red line is *loess* (Cleveland (1979); Cleveland and Devlin (1988)) regression fit for ratios of the whole slide. The blue line displays the horizontal zero line. b) Boxplots of unnormalized ratios stratified by the 48 print-tips and 58 well-plates.



(a) (b)

Figure 3.2: 2D spatial plot of unnormalized ratios of a QC slide. a) Array representation of ranks of the unnormalized ratios. Spots are color coded according to their ranks; higher ratios are red and lower ratios are green. White stripes reside empty spots where no probes were deposited. b) A smooth version of a). The smoothing is carried out by *loess* 2D regression ($span = 0.2$; for details consult the Section 3.2.3.)

3.2 Existing methodologies

Having illustrated the existence of non-biological variations in microarray data, we review a number of popular normalization methods before proposing a new systematic approach to remove such variations. Here, we assume that we are starting with equal quantities of mRNAs for the two samples that are labeled red and green. Also, we assume that the arrayed elements represent a random sampling of genes in the organisms. In situations where lots of genes are expected to change, it will be necessary to allocate spots on the slide that do not change for a control series and apply suitable normalization on those genes. Here we limit our discussion to within-slide normalization; the reader is referred to Yang and Thorne (2003) and Quackenbush (2002) for a detailed review on between-slide normalization methods.

3.2.1 Within-slide intensity bias

There are three widely applied normalization methods for the correction of intensity- or A- dependent bias. For the ith gene ($i = 1, 2, \ldots, N$), let M_i and A_i represent the log-ratio and log-intensity respectively, M_i^* denote the normalized log ratio, G_i and R_i indicate the measured green and red intensities and $\hat{c}_i$ be the estimated correction normalization factor based on A_i. What unifies the three methods are the dependence of c_i on A_i; what distinguishes them is the formalism of the function f reflecting such a dependence.

$$\begin{aligned} M_i^* &= M_i - \hat{c}_i = log_2(R_i/G_i) - \hat{c}_i, i = 1, 2, \ldots, N \\ c_i &= f(A_i) \end{aligned}$$

The function f is usually robust in nature, so that it is not influenced by outliers in the log-ratios, which are usually potential differentially expressed genes.

(i) Global median shift

The most simplistic method of the three is the global median shift. It assumes that the red and green intensities are related by a constant factor. That is, $R/G = k$, and shifts the

median of distribution of log-ratios to zero:

$$c_i = log_2 k.$$

Note that c_i does not depend on the gene identity, that is, all genes are shifted by the same quantity c. There are several ways to estimate the normalization factor c[1]. The first is based on ratio statistics and involves the median or mean of the log-ratios,

$$\hat{c} = median_{i=1,\ldots,N}(M_i) = median_{i=1,\ldots,N}(log(R_i/G_i)).$$

Another approach calculates c by summing the measured intensities in both channels,

$$\hat{c} = log_2 \hat{k} = log_2(\sum_{i=1}^{N} R_i / \sum_{i=1}^{N} G_i).$$

Analogously, Kerr et al. (2000) perform normalization using ANOVA model by including a dye main effect. The fitted value for the dye effect is estimated similarly to the second approach.

The global median shift method recognizes the bias between red and green intensities, but simplisticly performs normalization without accounting for intensity, spatial or other biases of the spot measurements.

(ii) Robust linear regression (rlm)

To account for the influence of varying A on M, Finkelstein et al. (2000) normalize the log-ratios by estimating c using a robust linear regression approach,

$$c_i = \beta_0 + \beta_1 A_i + \epsilon_i,$$

where the correction factor c_i is a linear function of spot intensity A_i. This model is sufficient when the relationship between M and A is approximately linear, but fails to correct any

[1] Note that c for median shift is not contingent upon A, yet it nonetheless corrects the imbalance of red and green dye incorporation in a global fashion, hence it is treated here as a special case in the category of A-bias correction.

nonlinear relationship between them.

(iii) Robust nonlinear local regression (loess)

Nonlinear methods developed by Yang et al. (2002) applied the robust scatter-plot smoother *loess* (Cleveland (1979); Cleveland and Devlin (1988)) to perform a local intensity-dependent normalization. The correction factor c is a nonlinear function of A. Its estimates are made using the local scatter plot smoother function *loess* in the software package R. *loess* partitions genes into bins based on the magnitude of their log-intensities (A) and the normalization is carried out within each bin, in an effort to assure distinct treatment for distinct groups. Within each bin, a robust linear regression approach is applied in the form of,

$$c_{i,a} = \beta_{0,a} + \beta_1 A_{i,a} + \epsilon_{i,a},$$

where $A_{i,a}$ represents log-intensities for genes in the intensity group a. A smoothing parameter *span* is used to size the bin, by specifying the proportion of data to be included. The wider the bin, the smoother the curve obtained. Yang et al. (2002) typically use a *span* size equal to 0.4. Kepler et al. (2000) propose using a different local regression method, which is similar in principle with *loess*.

(iv) Others: Single-channel method

Besides the above three models that conduct intensity normalization in a dual-channel fashion, Yang and Thorne (2003) propose a single-channel approach of removing systematic intensity bias from the red and green channels separately. Specially, the quantile approach, which is adapted from Bolstad et al. (2003), forces each channel to share a common distribution by requiring every quantile across channels be equivalent. This is done by projecting onto the unit diagonal $(1/\sqrt{2}, 1/\sqrt{2})$. *vsn* is another single-channel method proposed by Huber et al. (2002) to preprocess DNA microarray intensity data. The data are transformed by a variance-stabilizing transformation h that decouples the mean-variance dependency. The transformation h has the parametric form $h(x) = arsinh(a + bx)$, whose parameters are estimated with a robust variant of maximum-likelihood estimation. Differences between the transformed values are the so-called "generalized log-ratios", and could then be subject

to differential expression study.

3.2.2 Within-slide spatial bias: Print-tip and Plate

Models, such as median shift, *rlm* and *loess*, if fitted within each PT or PL, corrects the PT or PL bias. $\hat{c}_{i,l} = f(A_{i,l})$, where for the ith gene within the lth array layout (PT or PL) group, $\hat{c}_{i,l}$ denotes the estimated correction factor based on its log-intensity value $A_{i,l}$.

Median shift adjusts the median of log-ratios within each PL (or PT) to be zero in an effort to correct for existing inequality between such spatial attributes, and is a robust version of the ANOVA approach proposed by Sellers et al. (2003). *rlm* and *loess*, the latter being the most widely practiced normalization method, correct for the A and PT (or PL) biases simultaneously.

3.2.3 Within-slide spatial bias: 2D Spatial

Other than print-tip and plate effects, there could be other spatial attributes that contribute to the spatial heterogeneity. Methods account for spatial biases usually consider effects localized within array rows and columns, or an even smaller local environment. In general, spatial heterogeneity can be continuous, discontinuous or a combination of both. Let M_{rc} denote the log-ratio for the spot located at the rth ($r = 1, 2, \ldots, R$) row and cth ($c = 1, 2, \ldots, C$) column, and c_{rc} represent the normalization factor for M_{rc}.

(i) ANOVA

Sellers et al. (2003) applied an ANOVA model to test the effects due to array rows and columns.

$$\begin{aligned} c_{rc} &= \mu + \alpha_r + \beta_c + \epsilon_{rc}, \\ \hat{\mu} &= \frac{1}{R \times C} \sum_{c=1}^{C} \sum_{r=1}^{R} M_{rc}, \\ \hat{\alpha}_r &= \frac{1}{C} \sum_{c=1}^{C} M_{rc} - \hat{\mu}, \end{aligned}$$

$$\hat{\alpha}_c = \frac{1}{R}\sum_{r=1}^{R} M_{rc} - \hat{\mu},$$

where α_r and β_c are array row and column effects respectively. This approach treats the row and column effects as categorical variables, hence the spatial heterogeneity is modeled as discreet and non-uniform changes. As a consequence, fitting a large number of parameters is inevitable since the size of arrays commonly runs up to about a hundred rows and columns.

(ii) *loess*-2D

An alternative approach proposed by Yang et al. (2002) models the spatial variation via a two-dimensional *loess* curve which is a function of the row (I_r) and column identities (I_c), treated as continuous variable. In this way, spatial heterogeneity is represented as smooth trend. Consequentially much fewer parameters are estimated compared to the ANOVA model.

$$c_{rc} = \mu + \beta_1 I_r + \beta_2 I_c + \beta_3 (I_r : I_c) + \epsilon_{rc}$$

(iii) Spatial median filter

Yet another way to model local spatial effects is proposed by Wilson et al. (2003). The spatial trend in this model is estimated by computing for each spot, the median log ratio over its spatial neighborhood. The size of the smoothing element in the spatial median filter in their article is a 3×3 block of spots, although other sizes are also possible. This model is able to correct any local spatial trend, for example, a small streak of artifacts, which will probably be overlooked by the ANOVA and *loess* models; yet doing so costs a lot more degrees of freedom than simpler models.

3.3 Stepwise normalization

Despite the plethora of normalization methods having been proposed (see Chapter 3.2), an important area has not yet been well researched, which is how to utilize more formal criteria to assess the effectiveness of various normalization procedures with respect to the microarray experiments being investigated. To undertake this task, we describe next a novel normal-

ization framework, STEPNORM, for the two-channel cDNA microarrays, which aims at applying appropriate data-specific within-array normalization models in a stepwise fashion by utilizing a quantitative model selection criterion.

3.3.1 Algorithm

We often observe that different arrays demonstrate different amounts of biases. This is especially evident from the variety of spatial trends shown in various arrays. Applying the same normalization methods to all arrays may not be adequate for this multi-faceted nature of array biases. A more appropriate scheme that captures the characteristics of each array is to assess quantitatively the adequacy of each model with respect to the observed data. This can also be explained by considering the trade-off between model variance and bias. Complex normalization models possess a lower bias, but this usually comes at the expense of a higher estimation variance. A better normalization scheme should be able to achieve a good balance between variance and bias. This consideration motivates the development of a new normalization algorithm, STEPNORM; it employs a "multi-step" procedure and is adaptive in nature.

Figure 3.3 illustrates the procedures of STEPNORM using a typical QC array in Chapter 2.2. It consists of four steps and in each step one bias is targeted for correction. Generally, the number of steps employed in STEPNORM, i.e. the biases to be subject for correction, is flexible and is dependent upon specific experimental conditions and objectives. In addition, the models included for testing can also vary according to the likings of the researchers; those we have provided in Figure 3.3 only represent a suitable option that covers a reasonable range of model complexities.

The intensity A bias is usually the major source of variation and is therefore subject to examination first. Models to be tested in this step include: (i) the "null" model, which doesn't fit any parameters and represents the scenario that the bias is not statistically significant to warrant any correction; (ii) the global median shift model; (iii) , the robust linear regression model (Finkelstein et al. (2000)); and (iv) , the locally weighted scatter plot smoother (Yang et al. (2002)). These four models are of increasing complexities.

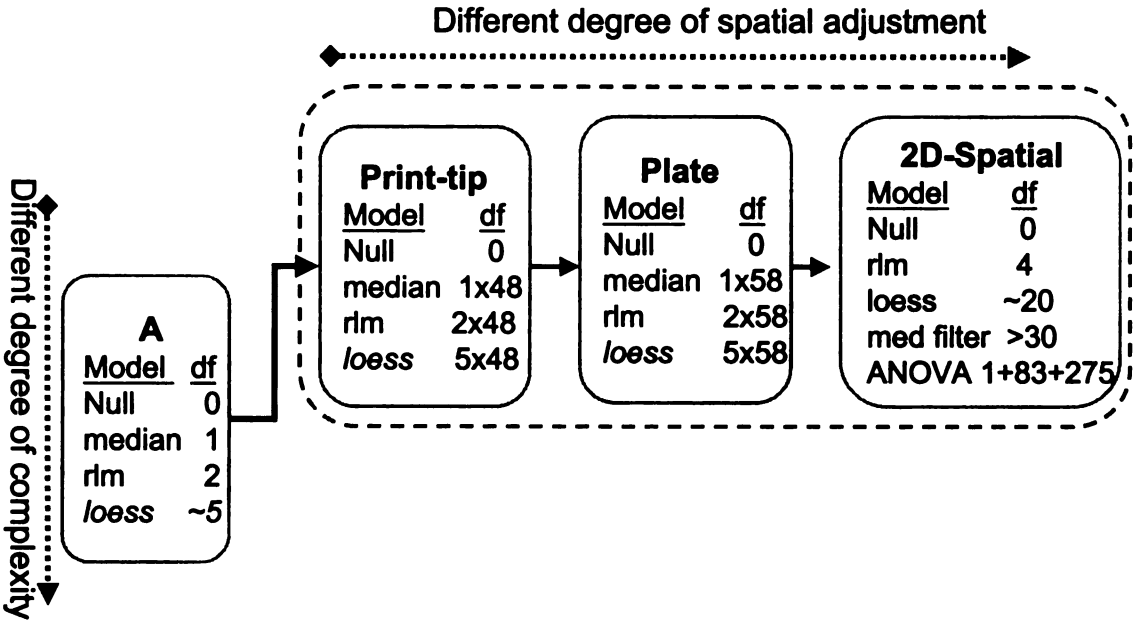


Figure 3.3: Stepwise normalization procedure using the example of the QC experiment. A QC slide has 48 print-tip-groups and 58 well plates.

After the correction of the A bias, normalized log ratios are subject to further normalizations based on the existence of spatial biases. We use three measures to proxy the spatial biases, these are print-tip (PT), 384-well plate (PL) and more general two-dimensional effects. The reason for adjusting for first is that the number of print-tip-groups is usually smaller than well plates and so doing prevents over adjustment (i.e. costs fewer degrees of freedom). Furthermore, various research (Sellers et al. (2003)) has shown the effect is usually more dominant than other spatial bias. The models employed in the and steps are similar to those in the step, except that they are fitted within each print-tip-group or well-plate-group.

3.3.2 Model selection criteria

In each step of our new method, there are a number of competing models of different complexities. Take the example of the first step – removal of A bias, among the candidate models, *median shift* is the simplest, estimating only one parameter – the median; and its

effect is correspondingly very limited and probably only suits for data that do not show a significant amount of linear or nonlinear trend between M and A. On the other end of the spectrum is the nonlinear local regression model *loess*. Its local fitting nature accommodates corrections for nonlinearity, yet doing so requires fitting a lot more parameters and runs the risk of over-fitting.

Selecting an appropriate model requires estimating the performance of different models; such a process, until now, has been guided by the Occam's razor principle. The occam's razor principle is the principle of parsimony: a model should be simple enough for efficient computation and complex enough to be able to capture data specifics. A model that satisfies such criterion generalizes well, i.e., it predicts well on independent test data that are unseen by the model.

There are several criteria that aim at estimating a model's *generalization error*, which is defined as the expected *prediction error* over an independent test sample. C_p, Akaike's Information Criterion (AIC; Akaike (1983)), Bayesian Information Criterion (BIC; Schwartz (1979)) and cross-validation (CV) are among the most widely applied such criteria.

C_p, AIC and BIC are common in estimating the prediction error by adjusting the training error for the number of parameters that are fit in the model. C_p is appropriate when the model is fit under squared error loss; AIC and BIC are more generally applicable in settings when the fitting is carried out by maximization of a log-likelihood. The AIC is defined as,

$$AIC = -2 \cdot log(\hat{L}) + 2 \cdot K,$$

where $\hat{L}$ is the maximum likelihood of the model and N is the size of the data. K denotes the generalized effective number of parameters in the model. Both the C_p and AIC have the limitation of possessing an increasing tendency to accept the more complex model when the sample size increases. It is generally recommended that they be applied to datasets where $N/K < 40$ (Maddox and Ashby (1993)). On the other hand, the BIC takes sample size into account (the penalty for model complexity is proportional to $log(N)$) and penalizes complex model more heavily. Although the BIC method was developed from a Bayesian

standpoint, the result is insensitive to the prior distribution for adequate sample size. Thus a prior distribution need not be specified. The BIC is defined as,

$$BIC = -2 \cdot log(\hat{L}) + K \cdot log(N).$$

Under the Gaussian model with unknown noise variance, $-2 \cdot log(\hat{L}) = N \cdot log(RSS/N)$, where RSS are residual sum of squares (using a squared error loss). Therefore, the BIC can be rewritten as the follows,

$$BIC = N \cdot log(RSS/N) + K \cdot log(N). \tag{3.1}$$

Another widely used method to estimate prediction error is CV. Unlike C_p, AIC and BIC, it directly estimates prediction errors by reserving part of the data as a "test" set, which are not used during model fitting. The steps in conducting CV will be detailed in Section 3.3.4. The disadvantage of cross-validation is that it reduces the amount of data available for training and it is highly computation-intensive, especially for large datasets. The latter property of CV renders it not an ideal candidate for application in a routine preprocessing procedure, wherein efficiency is a much desired merit. Application of k-fold CV requires fitting a dataset k times, for datasets that are comprised of tens of thousands data points – a not-uncommon scale for current microarray datasets – this entails lengthy waiting time.

The derivation of BIC, on the other hand, is much simpler and requires in general $1/k$th of the time needed for the k-fold CV. It is consequentially considered more appropriate in application of the preprocessing of microarray data. In the STEPNORM framework, for each step of the normalization, the model with the lowest BIC value is considered to be the preferred model. Importantly, each step also includes testing a "null" model, which doesn't fit any parameters and represents the scenario that the systematic variation in this step is not statistically significant to warrant any correction. In the next section, we illustrate the utility of STEPNORM using BIC as an model selection criterion on a QC slide. We then evaluate its performance by comparing to the results employing CV for model selection.

3.3.3 Application of STEPNORM using BIC as the model selection criterion

In this section, we illustrate the utility of STEPNORM by showcasing its application on a slide from the QC experiment.

For the correction of the A bias, we have observed that almost all data have some sort of trend between M and A, though to various degrees. The intensity-dependent bias is noticeable in the $MA-$plot of the QC slide and spots exhibit apparent curvature (Figure 3.4). To find the most appropriate model for the QC slide, we compare three models median shift, *rlm* and *loess*. Close inspection of Table 3.1 suggests that among the three candidate models *loess* possess the lowest BIC value and therefore is chosen as the best model for correction by STEPNORM. The number of free parameters required in the *loess* fitting is estimated to be 5.51 from the output of this function, which could also be approximated as follows. This span parameter that we typically use defines that about 40% of points are used for a local fitting in each moving window. Each fitting here is linear and therefore requires two degrees of freedom. Totally, about $\frac{1}{0.4} \times 2 = 5$ degrees of freedom are needed. *rlm* is a special case of *loess* when within each local area the fitting is linear and the span size set to 1. We have observed good behaviors of *loess* for most of the slides that we have analyzed, the QC slide being a typical example. The results in Table 3.1 illustrate that due to the typical high spot density nature of microarray data, a relative large span (0.4) is adequate to capture the nonlinear dependence of M on A, and it is also large enough to avoid the over-fitting concern.

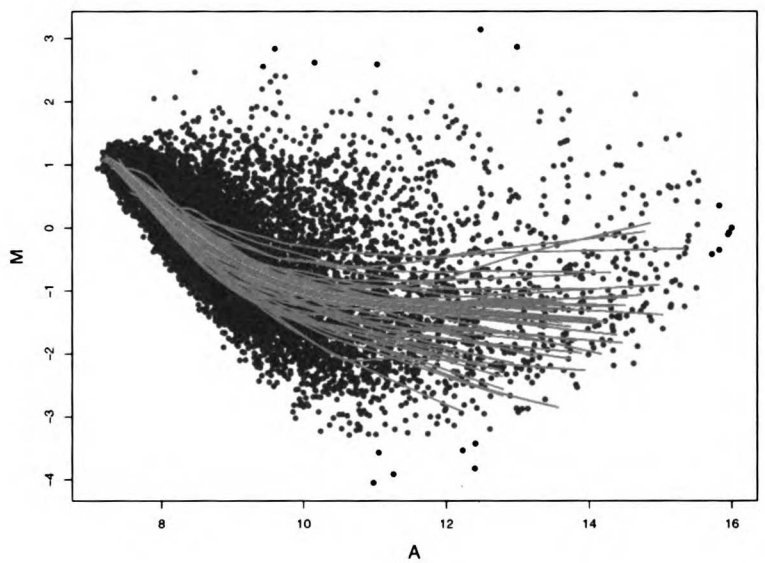
We proceed next to the removal of the PT bias. Figure 3.4 reveals that before any normalization is carried out intensity trends within print-tips show nonlinear tendencies. Yet, such nonlinear trends largely disappear after the first step A-bias correction. Figure 3.5 (a) displays MA-plot after the correction of the A bias; indeed, *loess* fits within each print-tip shows primarily linear departure from the zero line. In addition, a comparison between Figure 3.4 (b) and Figure 3.5 (c) indicates that spreads of ratios within each print-tip and well plate are also greatly decreased. Appropriately, STEPNORM chooses the *rlm* model for the removal the PT bias.

Table 3.1: Results of application of STEPNORM on a QC slide. BIC is employed as the model selection criterion. Models chosen by STEPNORM along the normalization steps are in bold fonts.

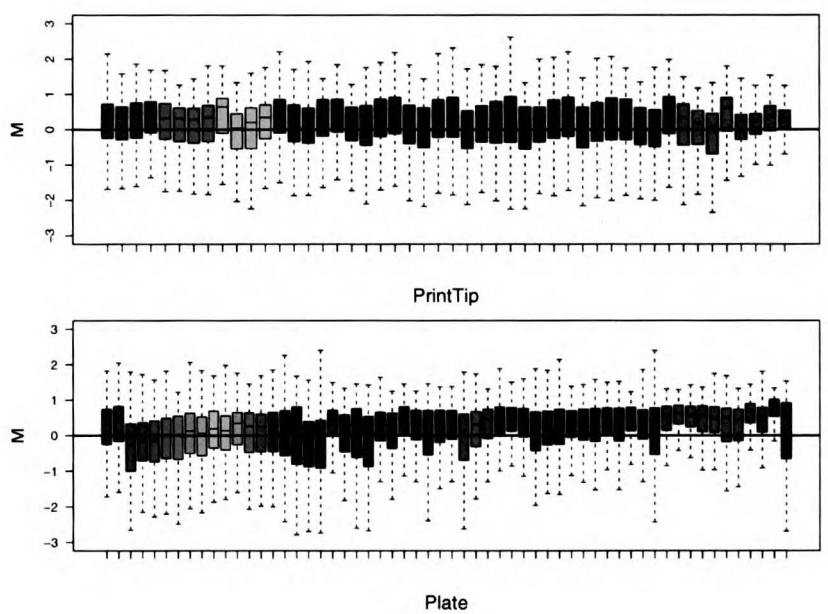
Bias	Models	K	$-2log\hat{L}(\times 10^{-4})$	**BIC**
A	null	0	-0.960	-0.960
	median shift	1	-0.983	-0.982
	rlm	2	-2.250	-2.248
	loess	5.51	-3.065	**-3.059**
PT	null	+0	-3.065	-3.059
	median shift	+48	-3.237	-3.184
	rlm	+96	-3.270	**-3.192**
	loess	+266	-3.253	-2.981
PL	**null**	+0	-3.270	**-3.192**
	median shift	+58	-3.326	-3.165
	rlm	+116	-3.387	-3.168
	loess	+320	-3.341	-2.917
2D	**null**	+0	-3.270	**-3.192**
	rlm	+4	-3.294	-3.188
	loess	+13.6	-3.297	-3.182
	median filter	+184	-3.366	-3.079
	ANOVA	+359	-3.394	-2.931

Lower panel of Figure 3.5 (d) display boxplots of the log-ratios for each of the 58 well plates after the correction of *PT* bias. The boxplots are centered at zero and have fairly similar spreads, signaling that the variations among different well plates are small. Accordingly, STEPNORM decides that the null model is to be preferred over its competing models, confirming a lack of evidence of the *PL* bias. The same decision also applies to the last step – remnant spatial 2D effects, wherein no systematic variation is apparent after preceding adjustments. Figure 3.6 compares the 2D spatial trend before and after stepwise normalization, highlighting the top 5% of ratios in both directions according to their spot locations. Before normalization, the relationship between the magnitude of ratios and their slide location is apparent, ratios at the borders of the slide are generally higher than those in the center. After normalization, the distribution of extreme ratios are much more even.

To visualize the change in model accuracy and the BIC values along the stepwise fitting, we summarize the process of stepwise normalization on the QC slide in Figure 3.7. Plotted



(a)



(b)

Figure 3.4: Graphical display of biases in log-ratios of a QC slide before normalization. (a) *MA-plot* displaying the *loess* line for the whole slide (red line) and for each of the 48 print-tips (gray lines). (b) Boxplots of ratios stratified by print-tips (top) and well plates (bottom) are shown in the right panel.

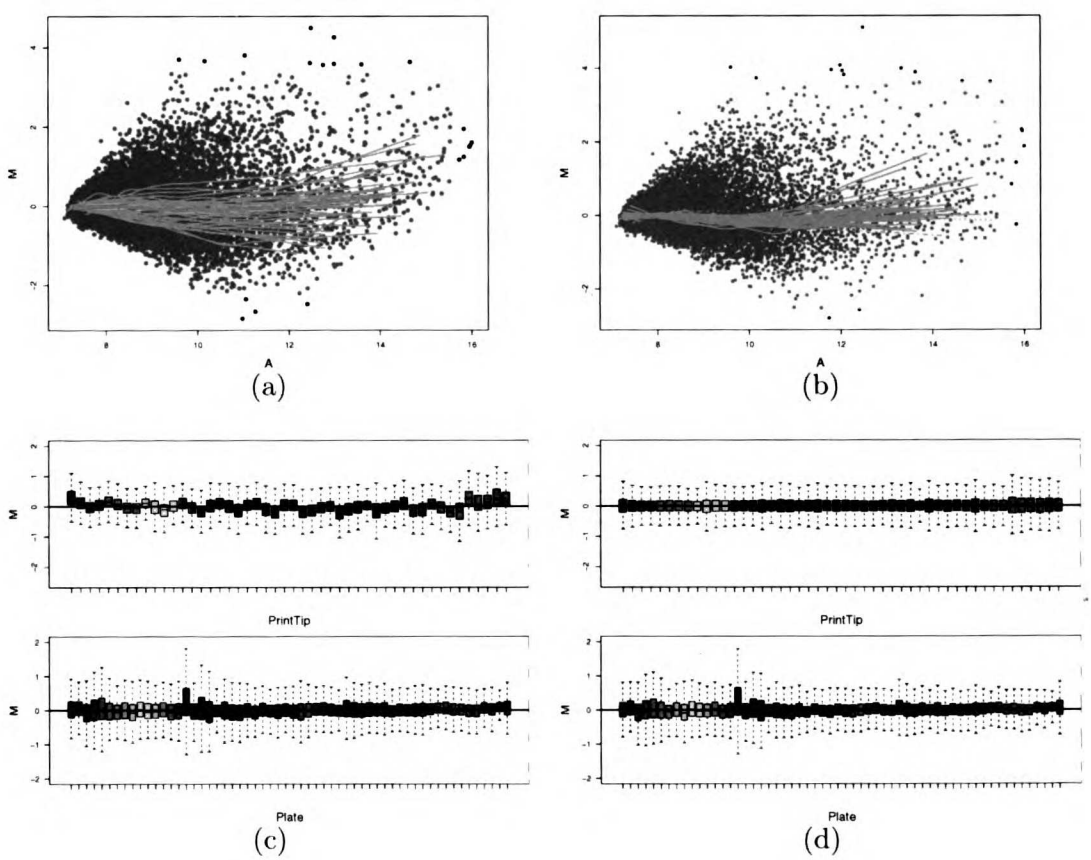
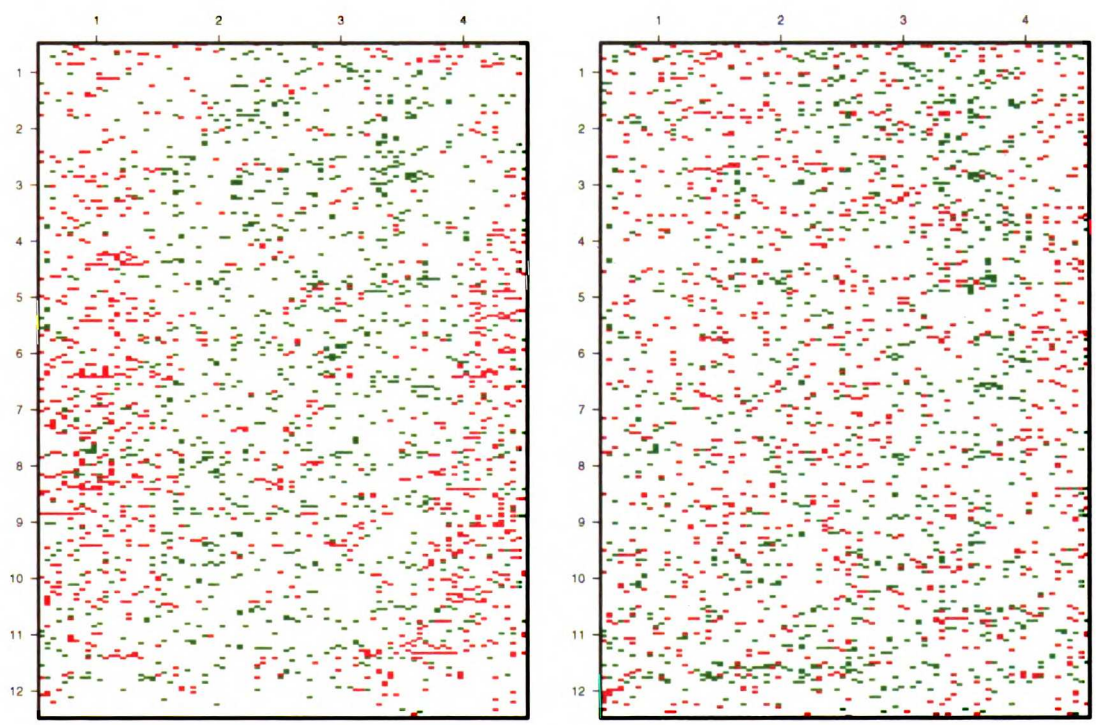


Figure 3.5: Graphical illustration of biases before (left column) and after (right column) stepwise normalization for the removal of the PT bias. (a, b)MA-plots with *loess* fit for the whole slide (red) and for each of the 48 print-tips (gray) ; (c, d) boxplots stratified by print-tip-groups and well-plate-groups.



(a) (b)

Figure 3.6: Spatial plots a) before and b) after normalization by STEPNORM. Plotted are top 5% of extreme ratios of both directions.

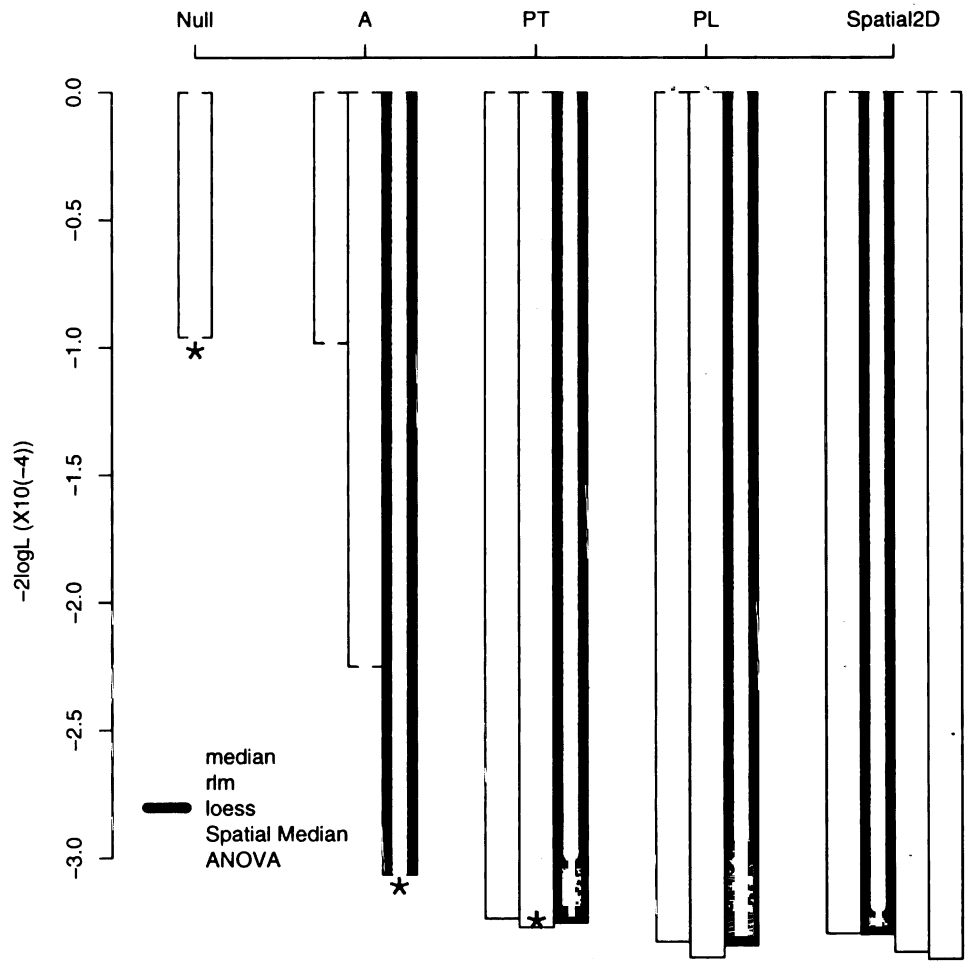


Figure 3.7: Stepwise normalization procedure employing BIC as the model selection criterion on a QC slide. Colored bars represent different models within each step; the heights of the bars indicate the accuracy of fit ($-2log(\hat{L})$). Gray lines overlaying the bars are the corresponding BIC values for each model. The model chosen within each step is marked by a red star.

are a sequence of colored bars representing different models in each step; the heights of the bars indicate the accuracy of fit ($-2log(\hat{L})$). Given the log-likelihood and the number of free parameters, Equation 3.1 can be applied to yield BIC values; these are represented as gray lines in the figure. As more and more normalizations are applied, we see a decrease of the $-2log(\hat{L})$ values signaling an increased goodness of fit. However, such decrease is only significant with the first two steps of the normalization (A and PT biases); the curve greatly levels off afterwards. Accordingly, following an evidential decrease in the first two steps of normalization, the BIC values actually increase for the last two steps (PL and Spatial2D), suggesting the better fit of models in these two steps are not large enough to offset the increase in model complexity.

3.3.4 Application of STEPNORM using CV as the model selection criterion

Cross validation (CV) is a natural and commonly used approach to deal with model selection. It evaluates competing models by reserving part of the data as "future data" to assess the accuracy of the models. In K-fold cross-validation (CV), the dataset $\mathcal{D}$ is randomly split into K mutually exclusive subsets (the folds) $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ of approximately equal size. The model is trained and tested K times. The K estimates of prediction error are averaged to obtain the cross-validation prediction error. Typically, the model with the smallest average prediction error is selected. However, often a "one-standard error" rule is used with cross-validation, in which we choose the most parsimonious model whose error is no more than one standard error above the error of the best model (Breiman et al. (1984); Hastie et al. (2001)). Two popular choices of k are 5 and 10 (4:1 scheme and 9:1 scheme). In this work, we use 5-fold CV, which has lower variance and less computation overhead. Procedures incorporating cross-validation as the model selection criterion are detailed below.

Stepwise Normalization Procedure (CV)

1. Split the N log-ratios (M) into K (here $K = 5$) equal folds, $\mathcal{D}_k$, $k = 1, \ldots, K$. Denote

the kth learning sample as, $\mathcal{D}^k = \mathcal{D} \backslash \mathcal{D}_k, k = 1, \ldots, K$, so that $\mathcal{D}^k$ contains the fraction $(K-1)/K$ of the total data points.

2. To select the best model for the adjustment of A bias among the following models: null (no normalization), median shift, *rlm* and *loess*, proceed as follows:

   (a) Denote the null model as f_0, and other models as f_p, $p = 1, 2, \ldots, P$. Calculate prediction error of the null model using a squared error loss as follows, $PE_0 = \frac{1}{N}\sum_{i=1}^{N} M_i^2$

   (b) For $k = 1$, fit the remaining P competing models (here, $P = 3$ for median shift, *rlm*, *loess*), $f_p^{-k}, p = 1, 2, \ldots, P$, using the learning sample $\mathcal{D}^k$ and evaluate the models on the test sample $\mathcal{D}_k$ by obtaining prediction error $PE_{k,p}$ for model f_p as follows, $PE_{k,p} = \frac{K}{N}\sum_{i \in \mathcal{D}_k}(M_i - \hat{f}_p^{-k}(A_i))^2$

   (c) Repeat (b) for $k = 1, 2, \ldots, K$ and average the K estimates of cross-validation prediction error for model f_p as follows, $PE_p = \frac{1}{K}\sum_{k=1}^{K} PE_{p,k}$. Compute the standard error $se_p = [\frac{1}{K}\sum_k (PE_{p,k} - PE_p)^2]^{1/2}$.

   (d) Denote the model f_{p^*} as, $p^* = argmin(PE_p, p = 1, 2, \ldots, P)$. Calculate $thresh = PE_{p^*} + se_{p^*}$. If $PE_0 \leq thresh$, proceed to the next step in the normalization procedure; else, choose the most parsimonious model $f_{p\#}$ among $[f_p, p = 1, 2, \ldots, P]$ such that $PE_{p\#} \leq thresh$

   (e) Refit the chosen model $f_{p\#}$ on the whole dataset $\mathcal{D}$.

3. Repeat 2 for the rest of the steps (PT, PL and Spatial 2D biases)

We apply the above algorithm on the first slide of the QC experiment and results are illustrated in Table 3.2. STEPNORM utilizing CV as the model selection criterion chooses *loess* and median shift for the adjustment of the A and PT biases respectively. To facilitate comparison, model selection results using CV and BIC are displayed together in Figure 3.8. The cross-validation prediction errors for each model is shown as colored spots with standard error bars in the upper panel, whereas corresponding BIC values are displayed in the lower panel. Purple and green curves, which connect models chosen by CV and BIC

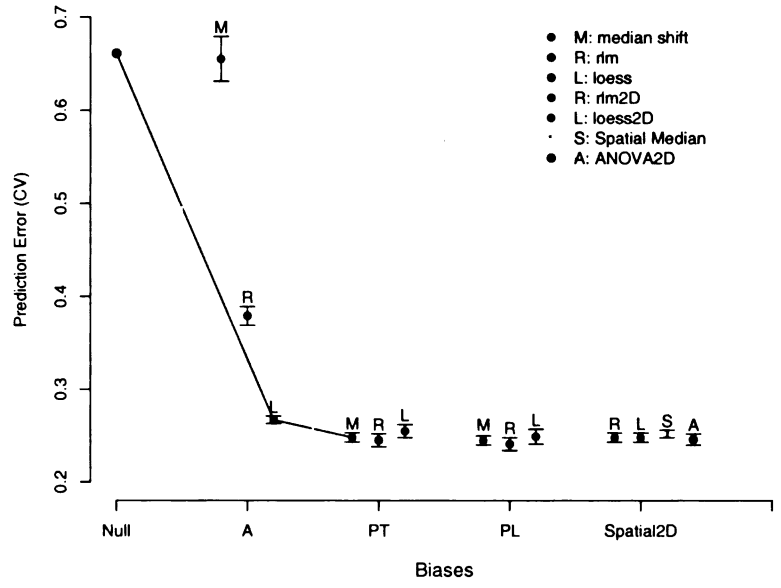
Table 3.2: Results of application of STEPNORM on a QC slide. CV is employed as the model selection criterion. "PE" in the cell indicates prediction errors estimated by CV and "se" indicates standard errors of the prediction error estimations. Models chosen by STEPNORM along the normalization steps are in bold fonts.

Bias	Models	*PE*	*se*
A	null	0.661	0
	median shift	0.655	0.024
	rlm	0.379	0.01
	loess	0.267	0.004
PT	null	0.267	0
	median shift	0.248	0.005
	rlm	0.245	0.007
	loess	0.255	0.007
PL	**null**	0.247	0
	median shift	0.245	0.005
	rlm	0.241	0.007
	loess	0.249	0.008
2D	**null**	0.247	0
	rlm2D	0.248	0.005
	loess2D	0.248	0.007
	spatial median	0.252	0.004
	ANOVA	0.246	0.006

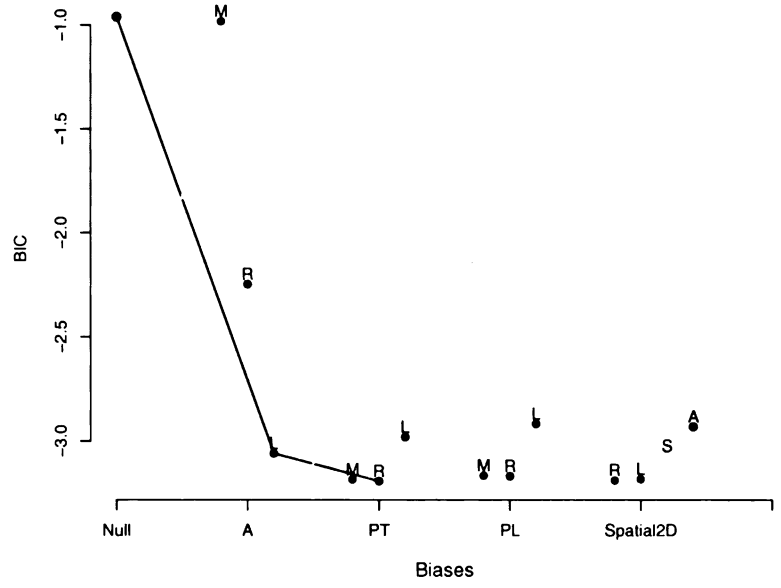
respectively in each step, display similar profiles. In both cases, the estimated prediction error decreases sharply following the adjustment of the *A* bias; and the trend becomes flattened after the correction of the *PT* bias. The only discrepancy between BIC and CV is that the formal chooses *rlm* for the *PT* step whereas the latter selects median shift. This can nonetheless be explained by the "one-standard error rule", as the model *rlm* actually has a lower CV prediction error than median shift (see also Table 3.2). In general, CV and BIC display highly comparable behaviors.

3.4 Comparison of normalization methods

To properly compare normalization methods in terms of bias and variance, data for which we know the "truth" is required. Frequently, such comparisons are based on simulated data,



(a)



(b)

Figure 3.8: Stepwise normalization procedure on a QC slide. (a) CV is employed as the model selection criterion. Plotted are CV prediction errors with error bars for different models (colored spots) within each normalization step. Purple curve connects models selected by CV along the normalization steps. (b) BIC is employed as the model selection criterion. Plotted are BIC values for each model within each normalization step. Green curve connects models selected by BIC along the normalization steps.

where "true values" are innate in the synthesis of the data. However, making judgment on simulated data could be problematic. The most notable limitation is the assumption behind the synthesis, which could be hard to be justified by real-world cases and could be biased against some methods being compared. Therefore employing a real dataset for comparison is arguably a better option, if, that is, the bias could be obtained. For real datasets, variances of different estimates can be readily computed by summarizing replicates, however, bias frequently remains elusive as the "real values" are usually not known.

The splice array, being unique in its array design, employs a gene-specific scheme for normalization, i.e., every splice junction ratio is normalized by its exon counterpart. For mutants that exhibit only minimal amount differential expression, such as *spt4*Δ and the wildtype as a special case, exons provide a means to gauge the amount of systematic variations, resulted from various processing steps and embodied in the splice junction ratios. This "one-to-one" correspondence arises from the facts described below:

1. *Equal expression level.* The expression levels of exons are expected to be the same as its corresponding splice-junctions for the majority of genes, which indicates that their log-ratios may exhibit similar amount of the intensity- or A- based bias.

2. *Adjacent spatial placement.* Exons are spotted close to its corresponding splice junction probes, the majority of such pairs are only six spots apart. The adjacency in location suggests that they are subject to the same amount of spatial systematic variations, including PT and PL related biases.

3. *No biological variations.* Exons in splice mutants are not expected to undergo differential expression, therefore the log-ratios encompass no biological variations.

Under ideal experimental conditions, we expect exon log-ratios to be zero; any systematic deviation from zero is indicative of biases due to experimental variations incurred on both the splice junction and exon probes. For the above reasons, exons could be treated as a self-normalizing factor; this practice amounts to having a known "true value" with regard to normalization.

The availability of "truth values" in terms of the amount of systematic variations in the splice data presents an opportunity of comparing various normalization methods based on variance and bias using a real dataset. The question we seek to answer is then, suppose we proceed with the normalization without the utilization of such "true values", as is the usual practice when such information is not available, how well will our normalization estimate the "true value"? Conducting normalization without gene-specific information relies on assumptions that apply to the majority of the genes, for instance, the expectation of log-ratios is typically assumed to be zero; such assumptions naturally lead to model-based approaches in which genes are normalized collectively. Listed in Table 3.3 are nine such commonly applied normalization models including STEPNORM that will next be subject to comparison based on bias and variance; details of the nine models are provided in Chapter 3.2. Among these nine models, all except STEPNORM and the tRMA methods are single-step models, i.e., they strive to correct all biases within a slide using one pre-determined model. The tRMA methods and the STEPNORM both employ a multi-step strategies. What distinguishes them is the model-selection component that the former lacks.

A simple criterion relying on the calculation of mean square errors (MSE) is used for comparison. To begin, we denote the expression log-ratios for the jth splice-junction probes as M_{jk}^{SJ}, $j = 1, \ldots, 254$, $k = 1, \ldots, 4$, with k representing the replicates. Similarly, for each of the seven models, the normalization factor estimating the systematic variations embodied in M_{jk}^{SJ} is represented by c_{jk}^{SJ}, which is derived from subtracting normalized log-ratios (denoted as M_{jk}^{*SJ}) from the corresponding unnormalized ones, $c_{jk}^{SJ} = M_{jk}^{SJ} - M_{jk}^{*SJ}$. The MSE estimate of c_j^{SJ} can then be derived as the average of the distances between c_{jk}^{SJ} and the true value $\bar{M}_{jk}^{Ex}$ that it estimates, which is the average of the four exon log-ratios for the same gene. The variance estimate can also be readily computed utilizing the four replicates of c_j^{SJ}.

$$\hat{MSE}(c_j^{SJ}) = \frac{1}{n}\sum_{k=1}^{n}(c_{jk}^{SJ} - \bar{M}_j^{Ex})^2, n = 4 \tag{3.2}$$

$$\hat{Var}(c_j^{SJ}) = \frac{1}{n-1}\sum_{k=1}^{n}(c_{jk}^{SJ} - \bar{c}_j^{SJ})^2 \tag{3.3}$$

Table 3.3: Description of normalization methods used in the comparison of bias and variance. Model complexity is approximated according to the splice data, which has 5760 data points and 16 print-tips.

Models	Description	Complexity
median shift	simplest normalization that shifts the median of all log-ratios to zero	1
loess	A-dependent normalization using the scatter plot smoother *loess*	5 (span=0.4)
vsn	variance stabilizing transformation	4
quantile	single-channel intensity normalization	NA
LPT	A-dependent *loess* normalization conducted with each print-tip-group	$5 \times 16 = 80$
tRMA1	A-dependent *loess* normalization followed by spatial median filtering 3×3	$5 + \frac{5760}{3\times3} = 645$
tRMA2	A-dependent *loess* normalization followed by spatial median filtering 7×7	$5 + \frac{5760}{7\times7} = 123$
tRMA3	A-dependent *loess* normalization followed by spatial median filtering 15×15	$5 + \frac{5760}{15\times15} = 31$
STEPNORM	stepwise normalization	dependent upon slides

The MSE composes of the effects of bias and variance, $\hat{MSE}(c_j^{SJ}) = \hat{Bias}^2(c_j^{SJ}) + \hat{Var}(c_j^{SJ}))$, therefore the bias can be estimated by subtracting the variance estimate from that of the MSE. Having computed probe-wise estimates of MSE, variance and bias, we summarize by pooling across probes and arrays using medians to derive estimates for each normalization model. Normalization models with smaller MSE are desirable.

Figure 3.9 summarized the comparisons of MSE among the commonly applied normalization methods (see Table 3.3). The distance of the spots to the vertical zero line measures how well normalization models estimate the true amount of systematic variations embodied in the SJ log-ratios. We observed that a big difference between no or global normalization (median) and the rest of the normalization methods. However, the differences among normalization methods excluding the simplest median model are not significant. The difference between the mutant *ceg1-250* and other mutants could be explained by the fact that *ceg1-250* is the most impaired mutant and hence the assumption that no differential expression in the exon ratios may not hold true.

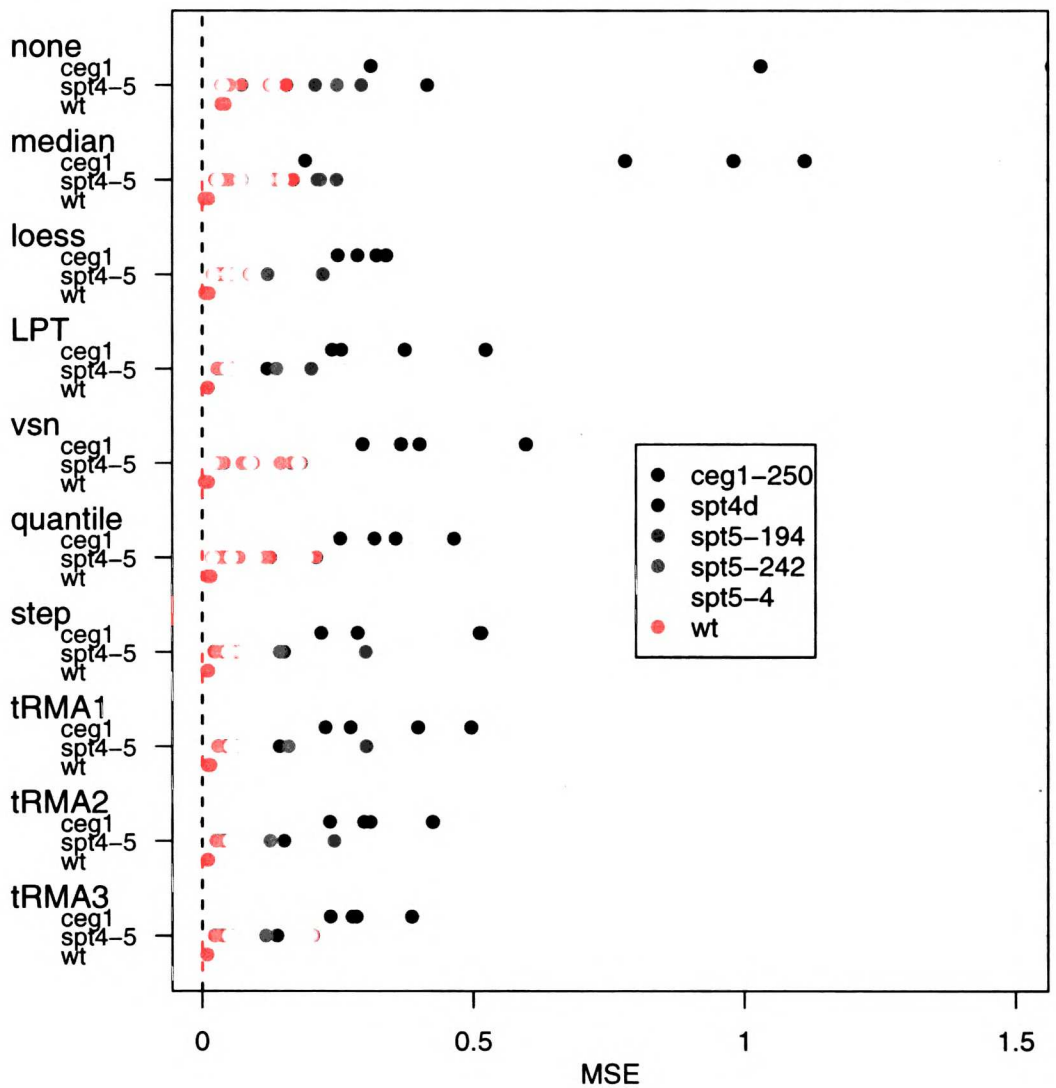


Figure 3.9: Dot plot of MSE of the normalization methods being compared for the Spt experiment. Note to better illustrate differences between normalization methods, two of the *ceg1-250* arrays that have MSEs larger than 1.5 in the "none" method are not plotted.

We next decompose the MSE into bias and variance. Figure 3.10 displays a natural trade-off between bias and variance. Plotted are estimated variances *vs.* biases for each normalization model. The balance between bias and variance is best illustrated by the three tRMA models of different complexities. tRMA 1,2 and 3 employ spatial median filters of window sizes 3×3, 7×7 and 15×15 respectively and therefore possess decreasing complexities. As expected for both *spt4*Δ and the wildtype, tRMA 1 has the largest variances and lowest bias. However, note for the wildtype summaries, tRMA 2 and 3, the simpler models among the three, are able to decrease variance significantly without inflating bias, which is a clear indication that for a wildtype self hybridization experiment, wherein no significant amount of systematic variation is expected, normalization using an overly complicated model, such as tRMA 1, amounts to over-fitting. The clustering of models in figure 3.10 can generally be explained by the different complexities of models, for instance, simpler models, such as vsn, quantile and *loess*, and more complicated ones, such as, LPT and tRMA, reside in different clusters. Interestingly yet expectantly, only STEPNORM is able to adjust complexities to the data; for data that exhibit more systematic variations, such as *spt4*Δ, STEPNORM selects more complicated models indicated by its adjacency to such models in the plot; for data that are relatively well-behaved, such as the wildtype data, STEPNORM tailor the models to be simpler. In both cases, STEPNORM is located close to the "elbow" regions of the curves indicating the achieval of a good balance between variance and bias.

We next use a *spt4*Δ slide to exhibit the change of bias and variance during the process of STEPNORM in Figure 3.11. Black and blue spots indicate the variance and bias values of competing models within each step while asterisks mark models chosen by STEPNORM. In the process of STEPNORM, as more and more normalizations are applied, variances of the estimates increas whereas biases decrease. Note the sharp increase in variance and decrease in bias both coincide within the A step.

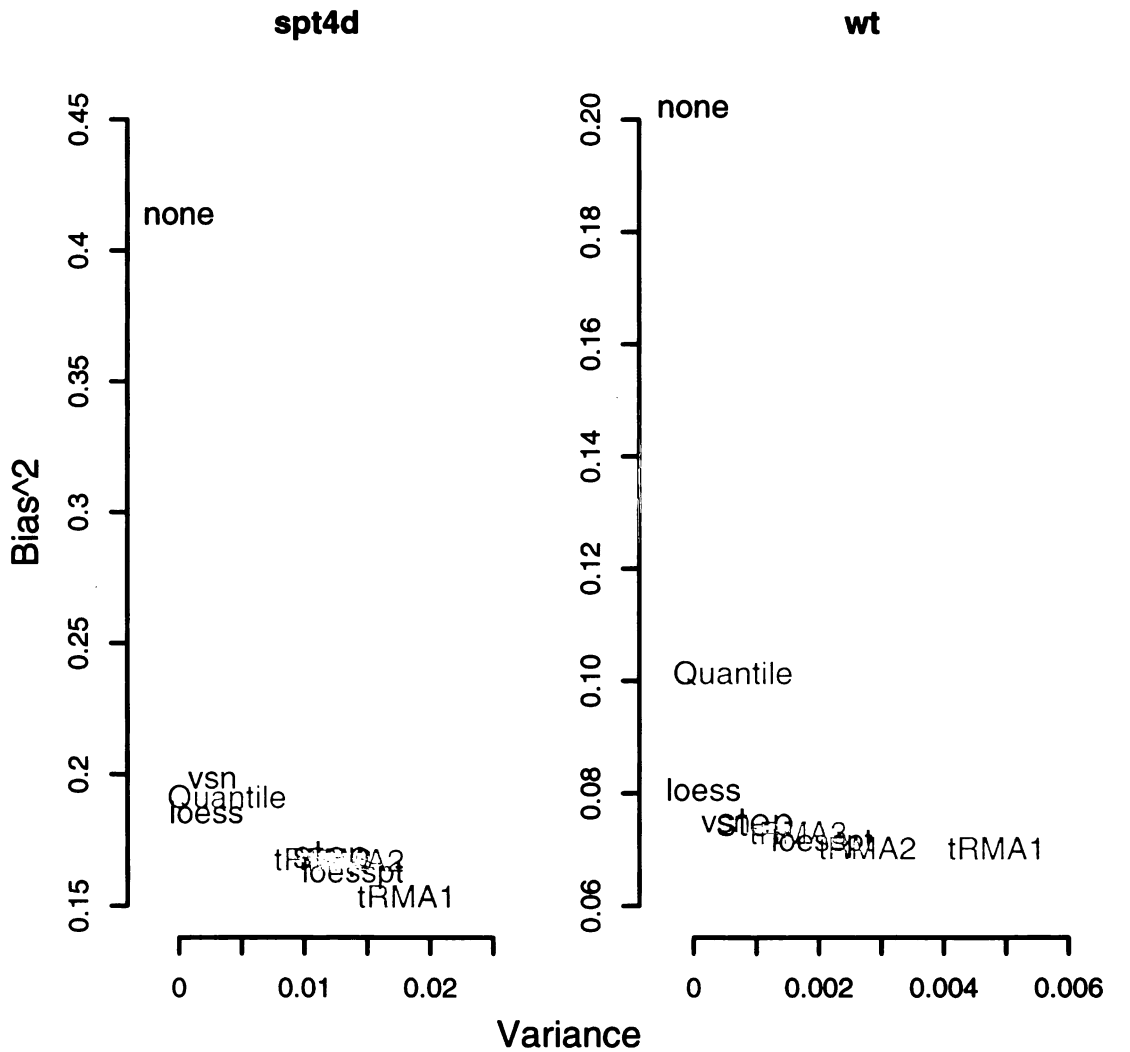


Figure 3.10: Plot of variance and bias for the *spt4*Δ slides (left panel) and the wildtype self hybridization slides (right panel).

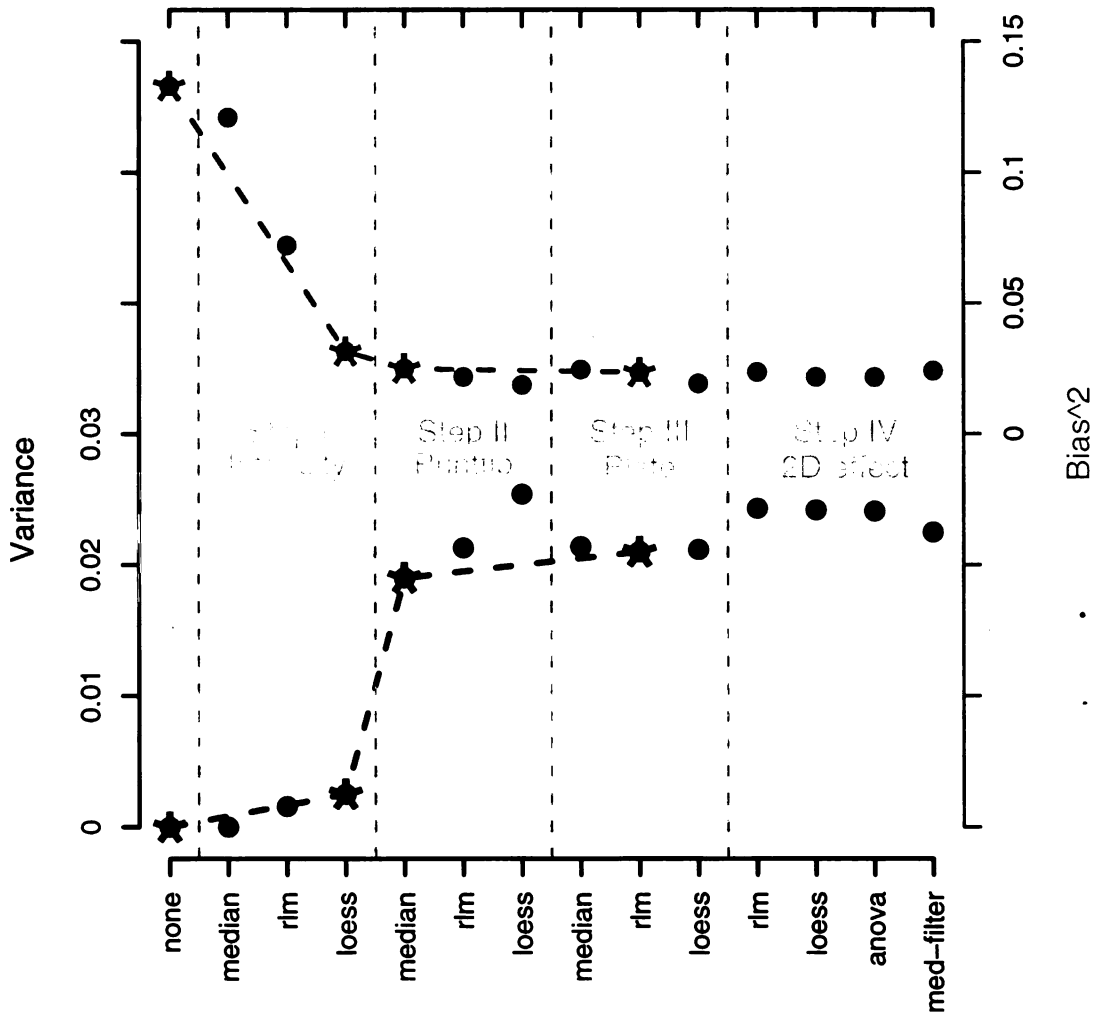


Figure 3.11: Plot of variance and bias for the stepwise normalization of a *spt4*Δ slide. Colored points represent the variance (black) or the bias (purple) of each model within a normalization step. Asterisks mark models selected by STEPNORM.

3.5 Conclusion and discussion

Efficient normalization is crucial for microarray research. It directly influence the outcomes of downstream data analyses that could give rise to important biological implication and discoveries. Failure to normalize appropriately will generally lead to misleading conclusions. We have presented here a new normalization procedure STEPNROM, which integrates a number of published methodologies under the same framework and assesses their effectiveness via a quantitative criterion. Such a process is applied to each individual slide in an experiment so that data (slide)-specificity could be achieved.

Unlike other normalization methods, STEPNORM could avoid data under-fitting or over-fitting as it implements both bias detection and removal in the same context. Intensity-dependent bias in ratios is usually the most common and dominant in microarray spot measurements. Very frequently such bias exhibits as a nonlinear trend between M and A; the curvature can be estimated using a suitable robust scatter plot smoother, such as the *loess* procedure, which has shown good performance for the adjustment of the A bias for most slides we have analyzed using STEPNORM. However, we have also observed that the nonlinear A bias is usually a whole-slide phenomenon and doesn't localize within spots related to a specific print-tip or plate. Therefore the current common practice that performs *loess* within each print-tip (LPT) to remove the A and S biases simultaneously appears to be over-fitting for most datasets. For slides like the QC experiment that have 48 print-tips, LPT estimates about $5 \times 48 = 240$ parameters when the span size is set at 0.4. On the other hand, the procedure preferred by STEPNORM for the first slide of the GRP experiment applies *loess* for the removal of whole-slide A bias and then employs a simpler linear regression (*rlm*) among ratios in different print-tips to remove the PT bias. So doing estimates only about $5 + 2 \times 48 = 101$ parameters.

Given a microarray dataset and several models for modeling the trends in this dataset, a natural way to assess the goodness of the candidate models is to estimate its prediction error. The BIC criterion provides a convenient way to estimate model prodiction erors and it is no doubt an important component in the STEPNORM framework. It is chosen for

model selection in STEPNORM for three reasons. First, it is quick to compute which makes it more appropriate than other computation-heavy criteria, such as CV, in the application of microarray datasets, which are usually large in size. Second, using a QC slide, we have shown that the outcome of applying BIC is largely compatible with that of applying CV, which indicates the using BIC gives appropriate and reliable results and it is suitable in the application of microarray normalization. Last, BIC does not require the competing models to be nested with each other whereas some popular criteria, such as the F test, does; this property of BIC is ideally suited for testing different normalization models as they are usually heterogeneous in model fitting. One drawback of BIC, however, is that because of its heavy penalty on complexity for large datasets (penalty proportional to the size of the data), BIC might be too conservative to accept complex models, such as *loess* within each print-tip (LPT) for large datasets. This is evidenced by the jump in its BIC values compared to other simpler models revealed by close inspection of Figure 3.8. Such jump is not observed in the CV plot or smaller datasets. Naturally, another possible explanation lies in the inaccurate estimation of the number of free parameters. The number of parameters for the LPT model is estimated based on the assumption that the 48 *loess* models fit within print-tip-groups are independent, whereas in reality correlations between print-tip-groups are inevitable, therefore the number of parameters plugged in the BIC equation is probably biased upward.

We have applied STEPNORM on four different experiments with 41 two-channel spotted arrays (Xiao et al. (2004)), for all of them, we have observed satisfactory results from STEPNORM that can be substantiated by diagnostic plots or criteria based on variance and bias when available. We typically use a four-step procedure as shown in Figure 3.3. However, the STEPNORM framework is flexible and can be easily extended or modified. The user can include new normalization models to be tested as long as a measure of model complexity can be obtained; this is needed for the calculation of BIC. In addition, the user can choose to employ fewer or more steps, whichever is suited for their experiments.

The STEPNORM procedure currently addresses within-array normalization issues, which is an essential step in two-channel microarray normalization. A natural extension will be

incorporating between-array normalization into the STEPNORM framework. Finally, the STEPNORM procedure is implemented as an `R` package (Ihaka and Gentleman (1996)) `stepNorm`, which may be downloaded from `http://www.biostat.ucsf.edu/jean`.

Chapter 4

Differential Expression

A very common microarray data analytic goal is the identification of important genes amongst the many for which expression measures have been obtained. Importance typically corresponds to association with a response of interest. When the response is polytomous such genes are termed "differentially expressed" (DE). The task of identifying differentially expressed genes can be divided into two components: ranking and selection. Ranking requires specification of a statistic or measure which captures evidence for differential expression on a per gene basis. Selection requires specification of a procedure (e.g. stipulation of a critical-value) for arbitrating what constitutes "significant" DE. Ranking is fundamental and, arguably, is easier than selection. The primary importance of ranking arises from the fact that only a limited number of genes can be biologically validated from follow-up experiments, these being necessary to affirm DE in view of the present maturity of microarray measurements.

This chapter presents two inter-connected DE studies. In the first, we present a case study of differential expression of a set of Spt splice microarrays (see Chapter 2.1). The splice experiment looks for genes whose splicing are affected in Spt mutants. We show that by using two splice indices that measure splicing defects, the object of this study could be approached using differential expression formulation. We then proceed to develop DE models appropriate for the study design of the splice data. In the second part of this chapter,

we propose and illustrate a novel approach to ranking and selection that integrates differing measures of DE; the idea of this approach is initiated when analyzing the splice data using differing DE models and it is further developed and evaluated by two other diverse datasets.

This chapter is organized as follows. We begin in 4.1 by presenting a brief overview of some recently proposed methods for identifying DE genes in multiple-slide experiments. Section 4.2 provides the details of analyzing the splice arrays using appropriate DE approaches; the complex experiment design of the splice data renders multiple DE models appropriate. Our proposed approach, *Differential Expression via Distance Synthesis* (DEDS), is presented in Section 4.3. We apply the method to three different data sets and present corresponding results. Finally, Section 4.4 discusses our findings, extensions and open questions.

4.1 Existing methods

The importance of developing new data analytic techniques to effectively identify differentially expressed (DE) genes is exemplified by the appreciable effort and attendant literature dedicated to this area. We overview several customized approaches to both aspects of DE detection with an emphasis on ranking.

4.1.1 Ranking statistics

For simplicity, and without loss of generality, we focus on a dichotomous response; i.e., deal with two-group comparisons. We designate groups as treatment (T) and control (C). One may consider the question of ranking genes in terms of DE in a discriminant analysis framework, i.e., consider DE genes as "features" or "variables" that best separate groups T and C (Ghosh (2003)). However, such approaches are focused on class prediction, and the rankings of genes so obtained are strongly influenced by (i) between gene dependencies, and (ii) feature selection strategies, so that individual gene DE is at best a by-product. For this reason, we limit our discussion to more direct approaches which assess differences between disparate groups on a single-gene basis, and we focus on those methods that we subsequently apply and describe the classes of statistics that they represent.

For two-color competitive hybridization experiments, we assume that the comparisons of log-ratios are all indirect; that is we have n_T arrays in which samples from group T are hybridized against a reference sample R, giving n_T log-ratios $M_{T_i} = \log_2(T_i/R)$; $i = 1, \ldots n_T$, and similarly we get n_C log-ratios $M_{C_j} = \log_2(C_j/R')$; $j = 1, \ldots n_C$ from group C. For oligonucleotide array experiments, we have n_T chips with gene expression measures from group T and n_C chips with gene expression measures from group C.

Fold Changes / t-statistics

The simplest gene ranking method is based on the fold change (FC) (i.e., ratio) in expression means between the two groups. Thus, for each gene, we compute the difference between (log) means

$$FC = \overline{M}_T - \overline{M}_C,$$

where $\overline{M}_T = 1/n_T \sum_{i=1}^{n_T} M_{T_i}$ and similarly for $\overline{M}_C$. While use of FC is conceptually appealing, ranking genes based on fold change alone implicitly assigns equal variance to every gene. In contrast, the t-statistic defined as

$$t = \frac{\overline{M}_T - \overline{M}_C}{s_p\sqrt{1/n_T + 1/n_C}},$$

where s_p is the pooled standard deviation, takes into account differing gene-specific variation across slides. Use of t-statistics is also widespread in assessing DE (Dudoit et al. (2002)). However, as indicated next, some care is needed in accommodating variation.

Penalized statistics

The primary shortcoming of using t-statistics for ranking genes lies in the unstable variance estimates that arise when sample size is small. Such small sample sizes are common occurrences in micorarray experiments due to high costs and/or resource (RNA) limitations. Relatedly, with tens of thousands of t-statistics (corresponding to tens of thousands of genes) there is frequently a number of large t-statistics driven by very small denominator standard deviations (s_p's), even though their numerators, measuring expression differences $\bar{M}_T - \bar{M}_C$, may also be small. To overcome these shortcomings a variety of approaches have been proposed to provide a more reliable variance estimate; they can be categorized into two

groups: (i) variance stabilizing functions, and (ii) error fudge factors and Bayesian methods. The former seeks to decouple the mean-variance dependency by modeling the variance of the expression of a gene as a function of the mean expression of the gene (Rocke and Durbin (2001); Huber et al. (2002); Jain et al. (2003)).The latter regularizes the t-statistics by inflating their denominators:

$$t_a = \frac{\overline{M}_T - \overline{M}_C}{a + s_p\sqrt{1/n_T + 1/n_C}}.$$

There are differing ways of motivating such penalization and, accordingly, of estimating the penalty parameter a. What unifies these approaches is the (sometimes implicit) desire to utilize *between* gene information rather than relying solely on individual (within) gene information as afforded by t-statistics. What distinguishes them is the formalism invoked in applying information sharing. At one end are somewhat *ad hoc* methods that avoid modeling between-gene relationships such as SAM (Significance Analysis for Microarrays; Tusher et al. (2001)). At the other are Bayesian approaches (Newton et al. (2001)) with empirical Bayes methods along the way (B-statistics, Lönnstedt and Speed (2001)). Smyth (2004) extends and resets the hierarchical Bayesian model of Lönnstedt and Speed (2001) in the context of general linear models, and uses the moderated (penalized) t- (or F) statistics for inference about contrasts of biological interest.

Mixture Models

These Bayes and empirical Bayes approaches are arrived at via mixture models: it is postulated that we have a mixture of non-DE and DE genes with the latter group sometimes refined into up- and down-regulated components. Often mixing of these components is done at the level of one-dimensional, gene-specific summary statistics (Efron et al. (2001); Lee et al. (2000); Pan et al. (2002); Allison et al. (2002)). Newton et al. (2004) argue that efficiency and sensitivity gains may be realized by effecting mixing on the gene expression (mean) values themselves, rather than on derived summaries. They develop so-called semi-parametric hierarchical mixture models (SHMMs) that have a number of features. Firstly, they are flexible (non-parametric) where data are rich (lots of genes) and parametric where data paucity mandates assumption making (few replicates per gene). Secondly, the provision

of per gene posterior probabilities of DE allows a straightforward approach to calibration in terms of false discovery rates (see below). Arguably, the main limitation surrounding the SHMM approach is the adequacy of the parametric model. Recognizing this, Newton et al. (2004) provide graphical diagnostics (stratified $Q-Q$ plots) for assessment of their chosen (gamma) model with their accompanying software.

Linear (Mixed) Models

In an effort to explicitly accommodate factors systematically impacting microarray gene expression values a number of linear model/ANOVA based approaches have been advanced. Kerr et al. (2000) develop fixed effect ANOVA models for log-intensities (single channel measurements) including terms for dye, slide, treatment and gene main effects, as well as select interactions between these factors. By assuming a common variance across genes a pooled (over genes) analysis is enabled, with large attendant increases in degrees of freedom for error variance estimation. However, there will likely be appreciable erosion of these degrees of freedom in order to accommodate interaction terms needed to 'recover' the adjustments afforded by use of log ratios in two channel settings. To overcome the (strong) common variance assumption Jin et al. (2001) and Wolfinger et al. (2001) adopt mixed model ANOVA formulations, performing gene-by-gene analyses treating factors such as dye and array as random effects.

4.1.2 Accessing significance (selecting cut-off)

Subsequent to gene ranking comes selection – the task of declaring which genes are significantly differentially expressed. A simple graphical method for assessing significance is to display the sorted test statistics in a normal or t- distribution $Q-Q$ plot. The bulk of the genes should follow an approximate straight-line; genes whose points deviate markedly from the the bulk of the genes are likely to correspond to DE genes. $Q-Q$ plots informally correct for the large number of comparisons (Dudoit et al. (2002)).

More formal approaches involve testing suitably constructed (joint) null hypotheses of equal expression (non-DE). Such joint formulations are mandated by the multiple testing concerns

that follow from evaluating thousands of genes. Two main approaches to this problem have emerged. One seeks to control family-wise type I error rates (FWER; see Dudoit et al. (2002); Ge and Dudoit (2002)), using step-down Bonferroni correction (Westfall and Young (1993)). This procedure is less conservative than the Bonferroni correction, and in contrast to the latter, it takes possible dependencies between the test statistics into account.

For many applications, however, control of the FWER is still too conservative, with the danger of many interesting genes being missed. As microarrays are often not conclusive, and are frequently used to screen for candidate genes that may then be validated through further experiments, the researcher may be willing to accept a certain fraction of false positives. This demand is addressed by another approach to the multiple testing concerns. Several research (Efron et al. (2000); Tusher et al. (2001); Storey (2002)) extends the false discovery rate (FDR) ideas of Benjamini and Hochberg (1995). FDR is defined to be the expected proportion of errors amongst the genes selected as significantly DE. Storey (2002) introduce q value, which gives each feature its own individual measure of significance in terms of the FDR. Detailed discussion and comparisons of competing approaches to multiplicity correction are deferred to Dudoit et al. (2003) and Storey (2003).

4.2 Case study: Analysis of the Spt splice experiment

4.2.1 Data

We refer the reader to Section 2.1 for the biological background of splicing and a detailed description of the experimental design of the splice arrays. Briefly, this experiment extends the study of splicing defects in yeast chromatin elongation factors mutants, spt4 and spt5, to the entire collection of intron-containing genes using splicing-sensitive DNA microarrays. Five mutants are analyzed, including a null mutation in *SPT4*, *spt4*Δ and three partial loss-of-function mutations in *SPT5*, *spt5-4*, *spt5-194* and *spt5-242*, in addition to *ceg1-250*, a temperature-sensitive mutation that causes rapid inactivation of the capping enzyme. To discriminate between spliced and unspliced transcripts for intron-containing yeast genes, oligonucleotide probes on these arrays were designed to detect splice junctions, introns and

second exons of spliced genes.

Boxplots of normalized ratios of splicing related probes stratified by mutants are shown in Figure 4.1. The general trend of the splice junction probe ratios shows a shift from the horizontal zero line in the negative direction signaling a decreased expression of splice junction in the mutants, in accordance with their compromised ability to splice properly. The mutant *ceg1-250* shows the largest decrease and *spt5-194* appears to be the most severely affected among the Spt mutants. Interestingly, the boxplots of the exon probe ratios display a similar pattern of change – the expression of exon probes is also decreased in the mutants. This suggests that in addition to splicing defects, transcription or RNA decay is also affected in the mutants. In contrast, ratios of the intron probes don't exhibit any shift from the horizontal zero line but spread for the mutants is nonetheless increased.

It is of interest to investigate if the decrease of the splice junction probe and exon probe ratios is correlated. Figure 4.2 displays the scatter plots between ratios of these probes. The upper panel shows evident correlation between splice junction and exon ratios. In contrast to the exons and splice junctions, ratios of the intron probes don't exhibit any shift from the horizontal zero line but spread for the mutants is nonetheless increased. Furthermore, there is no obvious correlation between the intron and exon ratios. In both plots, however, the spread of the cloud of points is mutant dependent and related to the severity of splicing defects. From Figure 4.2, it is clear that several of the mutants tested, *ceg1-250*, *spt5-194* and *spt5-242*, cause strong decreases in exon and splice junction signals and more idiosyncratic, gene specific changes in intron signals. Do these changes reflect altered transcription, splicing, RNA decay, or a mixture of potential defects? To focus on alterations of splicing efficiency independent of changes in transcription, we use the Intron Accumulation (IA) and Splice Junction (SJ) indices, which normalize ratios of intron and splice junction signals to the ratios measured for the second exon.

$$\begin{aligned} SJ-index &= \log \frac{MUT_{S-J}/WT_{S-J}}{MUT_{Ex}/WT_{Ex}}, \\ IA-index &= \log \frac{MUT_{Int}/WT_{Int}}{MUT_{Ex}/WT_{Ex}}, \end{aligned}$$

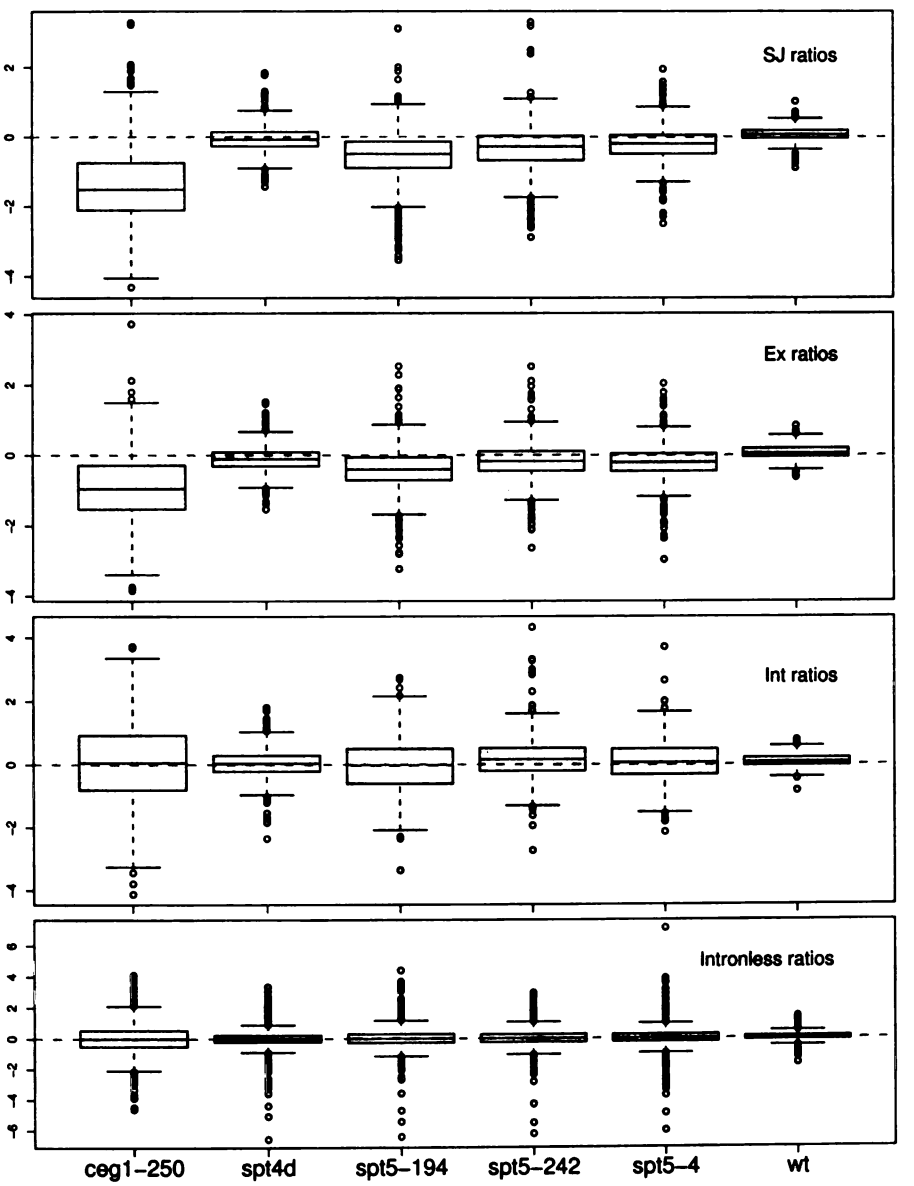


Figure 4.1: Boxplots of normalized ratios of splicing related probes stratified by mutants. Splice junction and exon probe ratios show a shift from the horizontal zero line in the negative direction, whereas intron probe ratios are centered at zero.

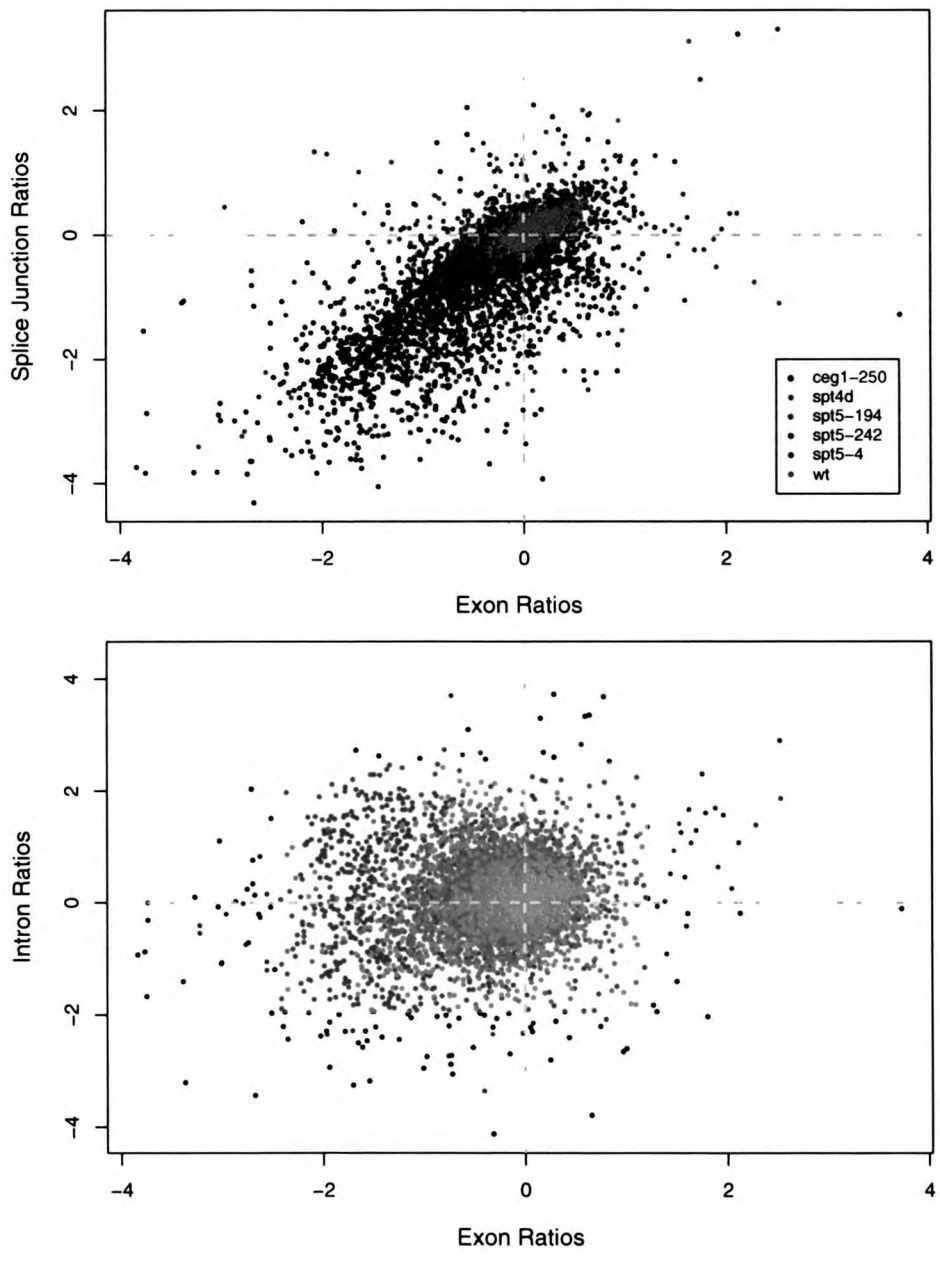


Figure 4.2: Scatter plots of normalized ratios of splicing related probes. Points are colored coded by their mutant identity. Gray horizontal and vertical reference lines indicate zero expression ratios.

where MUT_{S-J}/WT_{S-J}, MUT_{Int}/WT_{Int} and MUT_{Ex}/WT_{Ex} represent ratios of a splice-junction probe, an intron probe and the corresponding exon probe respectively, between a mutant and the wildtype.

4.2.2 Differential expression analysis

The major goal of this experiment is to identify differentially expressed genes in the four Spt mutants. In this section, we introduce five competing statistical models for testing DE and a general method for integrating such measures.

The experimental design for *ceg1-250* and the four *spt* mutants is identical; we illustrate the set-up using *spt4dΔ* as an example. There are four slides that were competitively hybridized with a *spt4dΔ* sample and a wildtype sample. This yields four indices (for each of SJ and IA – which we treat identically and separately) for the detection of splicing defects in *spt4dΔ* for each gene. A test, such as the one-sample t test, can be used to examine if the mean of the indices is equal to zero. For each gene, acceptance of this hypothesis signals lack of evidence for DE, while rejection provides evidence for DE. Alternatively, if the two indices of wildtype *vs* wildtype resulting from the two wildtype self hybridization slides are included, DE can be tested by comparing the four *spt4dΔ* indices to the two wildtype indices, using a two-sample test. We applied both the one-sample and two-sample approaches in the analysis of the splicing experiment.

In addition to the above two approaches, distinguished by including wildtype self-hybridizations or not, we also consider another two approaches distinguished by allowing gene-specific variance heterogeneity or not. This latter case imposes the assumption that all genes exhibit a similar degree of variability and so can be jointly analyzed using a common estimate of error variance. As illustrated subsequently (see Table 2), this pooling dramatically increases error degrees of freedom (df). The former approach, on the other hand, does not impose the common variance assumption, allowing different variances for different genes. The resulting model is then fitted gene by gene. The above approaches, four in all – see Table 4.1 – can be fitted by appropriately specified ANOVAs.

In addition to these four ANOVA models, we also employ Newton et al. (2004) semiparametric hierarchical mixture model. As described in Section 4.1, the SHMM is nonparametric where data is rich (lots of genes) and parametric where there is limited information (observations per gene). Finally, we investigate whether gains can be realized by synthesizing the results from all five models using methods we have developed for effecting such synthesis.

Table 4.1: A summary of the five competing differential expression models for the analysis of the Spt experiment.

Model No	Model Description
I	Mixed ANOVA: one-sample/homoscedastic errors
II	Mixed ANOVA: one-sample/heteroscedastic errors
III	Mixed ANOVA: two-sample/homoscedastic errors
IV	Mixed ANOVA: two-sample/heteroscedastic errors
V	Semiparametric hierarchical mixture model

Model I – one-sample / homoscedastic errors

Let Y_{gvsa} be the splicing index, SJ or IA, from gene g ($g = 1, 2, \ldots, 254$ for SJ and $1, 2, \ldots, 263$ for IA), mutant v ($v = 1, 2, \ldots, 5$), sample s ($s = 1, 2$) and array a ($a = 1, 2$: corresponding to the dye swap pair).

The first model can be represented as

$$Y_{gvsa} = \mu + G_g + V_v + (GV)_{gv} + (V/S)_s + (V/S/A)_a + (GV/S)_{gvs} + \epsilon_{gvsa}$$

Note that due to the nature of the experimental design (Figure 2.3 (a)), array effect A is nested in sample effect S (S/A), and sample effect S is in turn nested in mutant effect V (V/S). The two samples for each mutant and two array hybridizations for each sample can be considered as random representations of a population of similar samples and arrays. It is therefore appropriate to consider effects involving S and A to be random. In summary, effects $(V/S)_s$, $(V/S/A)_a$, $(GV/S)_{gvs}$ and ϵ_{gvsa} are assumed to be normally distributed random variables with zero means and variance components $\sigma^2_{V/S}$, $\sigma^2_{V/S/A}$, $\sigma^2_{GV/S}$ and σ^2 respectively. The remaining effects in the model are fixed effects, and therefore the model is a mixed-effect model. Table 4.2 gives the analysis of variance result for Model I on splice

Table 4.2: ANOVA table for differential expression model I of the Spt experiment (SJ).

	G	V	GV	V/S	$V/S/A$	GV/S	*Residual*
SS	326	213	390	106	144	226	471
df	253	4	1012	5	10	1265	2530
MS	1.29	53.3	0.386	21.1	14.4	0.179	0.186

Table 4.3: Estimation of random effects for differential expression model I.

Component	**Estimate**	**Results**	
σ^2	MS_ϵ	0.186	0.28
$\sigma^2_{GV/S}$	$(MS_{GV/S} - \hat{\sigma}^2)/n_A$ $(n_A = 2)$	0	0.036
$\sigma^2_{V/S/A}$	$(MS_{V/S/A} - \hat{\sigma}^2)/n_G$ ($n_G = 254$ for SJ and 263 for IA)	0.056	0.054
$\sigma^2_{V/S}$	$(MS_{V/S} - \hat{\sigma}^2 - n_G\hat{\sigma}^2_{V/S/A})/n_A n_G$	0.013	0.042

junction indices.

When the experiments are balanced, as here, the ANOVA estimates of these variance components are simple and can be found by equating observed mean squares with their expectations (Scheffe (1999)); see Table 4.3.

The parameter of interest in this model is $\mu_{gv} = \mu + G_g + V_v + GV_{gv}$, which measures the mean of the SJ/IA indices of gene g in mutant v. Testing the estimate $\hat{\mu}_{gv} = \bar{Y}_{gv\cdot\cdot}$ against zero amounts to testing the change of SJ or IA signals for gene g in mutant v compared to wildtype. The following null hypothesis therefore defines the absence of differential expression in mutant v and gene g:

$$H_0 : \mu_{gv} = 0$$

Calculation of the variance of the treatment mean $\hat{\mu}_{gv}$ is complicated by the nested study design and the resulting multiple layers of variance components, as indicated by the following equation:

$$\hat{Var}(\hat{\mu}_{gv}) = \frac{1}{n_S}\hat{\sigma}^2_{V/S} + \frac{1}{n_S}\hat{\sigma}^2_{GV/S} + \frac{1}{n_A n_S}\hat{\sigma}^2_{V/S/A} + \frac{1}{n_A n_S}\hat{\sigma}^2,$$

where $n_S = 2$ and $n_A = 2$. The statistic $\hat{\mu}_{gv}/\sqrt{\hat{var}(\hat{\mu}_{gv})}$ has a Student's t distribution, the error *df* for which can be derived from the ANOVA table as in Table 2 (again for SJ).

Model II, III and IV

Model II (one sample / homoscedastic errors) is different from model I by assuming that each gene has its own error distribution, so the model is fitted gene by gene.

Models III and IV include the indices derived from the two wildtype self-hybridizations. Because of this inclusion, the study design is rendered unbalanced. To be more specific, the arrays in the two wildtype self-hybridizations came from the same sample, whereas the samples of four slides related to a mutant were from two distinct samples (see Figure 2.3(b)). This imbalance complicates parameter estimation (see Appendix). The inclusion of the wildtype self-hybridization indices also allows that DE be evaluated using two-sample tests. Detailed model specificities of Model II, III and IV are listed in Appendix.

Model V – semiparametric hierarchical mixture model

Multitude of alternate methods have been developed for assessing differential gene expression in the context of microarray studies; see Pan et al. (2002) and Yang and Speed (2003) for reviews. To complement the ANOVA approaches previously described we employed the SHMM advanced by Newton et al. (2004). This methodology was selected for several reasons. Firstly, the semiparametric approach can be compellingly motivated. The synthesis of a nonparametric component, where we are data rich (lots of genes), with a parametric component, where we are data poor (observations per gene), makes for an appropriately balanced strategy. Secondly, as is standard, our ANOVA approaches treat gene (G_g), mutant (V_v), and gene-mutant interactions (GV_{gv}) as fixed effects. Thus, there is no information sharing between genes. The SHMM achieves such sharing and does so in a more principled and flexible manner than some of the *ad hoc* approaches proposed that yield regularized $t-$statistics (Baldi and Long (2001); Tusher et al. (2001); Lönnstedt and Speed (2001)). Thirdly, the output posterior probabilities for (directional) DE have dual utilities: (i) ranking (genes), and (ii) calibration (providing false discovery rates (FDRs)). Our interest is primarily in the former since, in the next section, we describe a method for combining several measures of DE and computing associated FDRs.

Of course, SHMM also has limitations. Perhaps foremost is the adequacy of the para-

metric assumptions. The extent of such assumptions has been appreciably relaxed compared to the preceding fully parametric treatment of Kendziorski et al. (2003). Importantly, diagnostic tools are provided for assumption checking. We briefly discuss the parametric specifics below. Other limitations are implementation related. Estimation is very computationally intensive – while this was not an issue for the splice data due to the small number of (intron containing) genes, we have encountered lengthy run times for more typically dimensioned array studies. Additionally, the present implementation (available from ftp://ftp.biostat.wisc.edu/pub/newton/Arrays/tr1074/Rcode/) only supports two group comparisons. Thus, there is some potential efficiency loss for the nested design employed in the splice study (Figure 2.3 in Section 2.1).

The SHMM links two components. The so-called observation component specifies a (parametric) model for the conditional distribution of gene expression measurements, given their mean level. For both empiric and tractability reasons Newton et al. (2004) take this to be Gamma. The empiricism reflects a general contention that microarray expression data exhibits constant coefficient of variation (CV), although there are many exceptions (Huber et al. (2002)). We note that Gamma distributed raw expression results in a complex distribution for the derived SJ and IA indices – in the idealized setting of independence and common shape parameters, the indices will follow a ratio of F distributions. So, diagnostics become all the more important. Use of $Q-Q$ plots indicated reasonable conformity to the assumed Gamma (not shown).

The second, so-called mean component describes variation in the above mean levels. It is this component that is modeled nonparametrically. This confers considerable flexibility, with the consequent hope of enabling sensitive detection of DE. However, it is this flexibility that incurs the computational cost. Details on the estimation methodology as well as extensive illustration of calibration, diagnostic, and performance aspects are provided in Newton et al. (2004).

4.2.3 Comparisons of differential expression models

Models with heteroscedastic errors accommodate gene-specific variances but typically, as here, replication is very limited and so the precision of the estimates is compromised. Models imposing homoscedastic errors yield precise estimates of the common error variance, and tests based on many *df*, since they permit combination over the large number of genes. However, the homoscedasticity assumption is both strong and difficult to evaluate. Differences in error *df* for the different models are presented in Table 4.4. Note that there are more than 5000 *df* for error for the homoscedastic models and only about 20 *df* for the heteroscedastic models.

We use results pertaining to SJ indices of mutants *ceg1-250* and *spt4dΔ* to illustrate relationships between the five models. These two mutants are selected for illustration as they display varying amount of DE. Figure 4.3(a) displays a scatterplot matrix of $-\log_{10}(p)$, where p either corresponds to the Model I through IV p-value for tests of DE or to the Model V posterior probability for non-DE. Note that by relating Model I through IV results to Model V results we may seemingly be perpetuating the "severe pedagogical problem of misinterpreting p-values as posterior probabilities" (Berger et al. (1997)). However, this is not the case. At no stage do we make probabilistic statements in terms of these quantities. Rather, they simply constitute a quantification of DE.

Table 4.4: Degrees of freedom table for the four ANOVA mixed models (SJ).

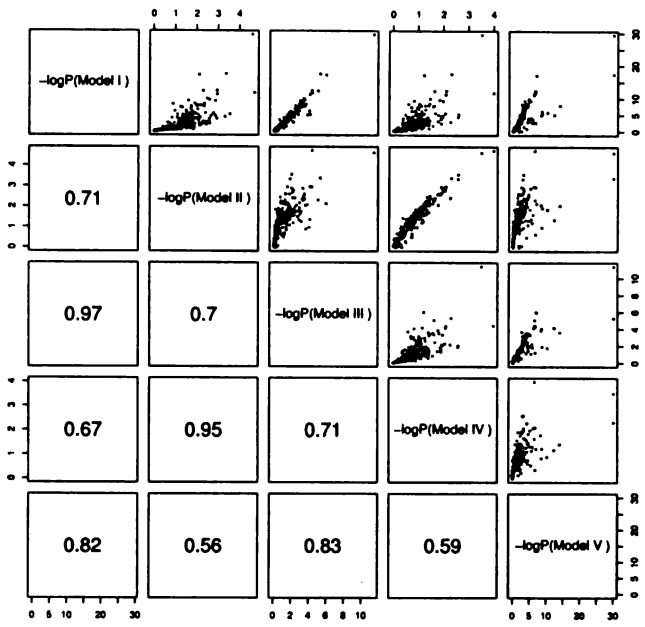
Source	Models			
	One-sample Homoscedastic	One-sample Heteroscedastic	Two-sample Homoscedastic	Two-sample Heteroscedastic
Intercept	1	1	1	1
G	253		253	
V	4	4	5	5
GV	1012		1265	
V/S	5	5	5	5
$V/S/A$	10		11	
GV/S	1265		1265	
Residuals	2530	10	2783	11
Total	5080	20	5588	22

The large number of DE genes expected between *ceg1-250* and wildtype is evident in Figure 4.3(a) where we observed a high correlation between the various measures. In contrasts, the general correlation between the measures between *spt4dΔ* and wildtype is lower. The high correlations between Model I and Model III, and between Model II and Model IV, with correlation coefficients 0.97 and 0.95 respectively, are as anticipated. This attests to the fact that the gene expression log-ratio measurements in the two self-hybridization experiments of wildtype are tightly centered around zero. The fact that Model V conforms more closely to the homoscedastic models (I and III) than to the heteroscedastic models (II and IV) is not surprising, since the SHMM utilizes information sharing between genes which is absent for the gene specific heteroscedastic models.

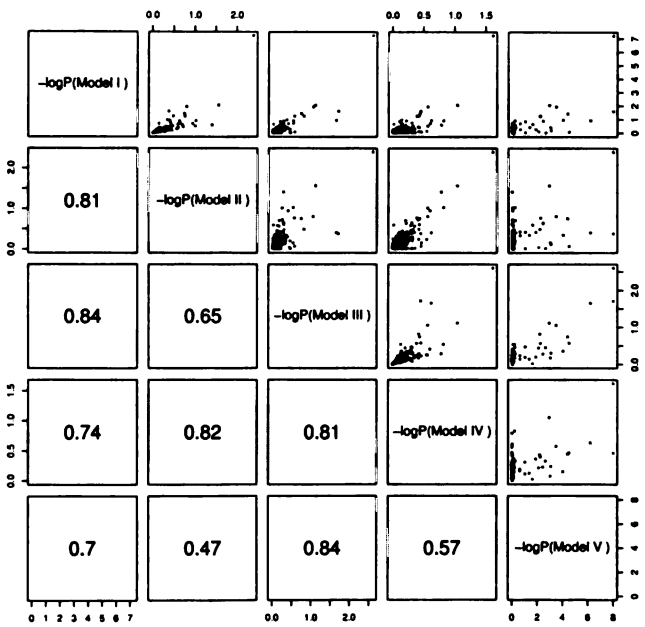
4.3 DEDS

It is immediately apparent from Section 4.1.1 that there are numerous statistics available for ranking genes as a prelude to arbitrating DE. Further, selecting a best statistic or ordering statistics in terms of merit is problematic. No characterizations of microarray data that indicate desirability of a specific choice exist and, likewise, no comparisons across a sufficiently wide range of benchmark datasets have been undertaken. Hence, investigators must make fairly arbitrary choices when deciding which ranking statistic to use. It is in part to eliminate some of this arbitrariness but primarily to borrow strength across related measures that we propose synthesizing DE ranking schemes.

A somewhat related approach is that of Pareto Fronts (PF, Hero and Fleury (2002)). However, PF are operationalized very differently from our methods and are applicable in a different context. Briefly, by regarding the set of measures as defining multivariate point cloud (points corresponding to a gene's vector of measures) and employing a standard (coordinate-wise) partial order, a set of "non-dominated" genes is identified. In our context these could include genes ranked (for DE) very highly by one measure but very lowly by another. Use of PF seems more pertinent to the setting where the differing measures are not intended to capture the same underlying quantity.



(a)



(b)

Figure 4.3: Scatter plot matrix of differential expression models for SJ indices of the (a) *ceg1-250* and (b) *spt4dΔ* mutants. Plotted are the $-\log_{10}(p-values)$ of the corresponding models. Correlation coefficients between respective models are shown in the lower triangle of the matrix.

We now sketch our approach to detecting differential expression via distance synthesis (DEDS). We also begin with the multivariate point cloud with a point corresponding to a gene's vector of DE measures. Rather than employing partial order, we exploit the fact that all measures are attempting to capture DE and synthesize (reduce to a scalar) as follows. We first define an "extreme origin". Without loss of generality assume that large values of all measures indicate DE. Then the extreme point is the coordinate-wise maxima over all measures (coordinates). Note that the extreme point may not correspond to any observed point. Next, the distance from all points to the extreme is computed. We discuss choices for distance (Euclidean, Mahalanobis, other scalings) subsequently. Intuitively, those points closest to the extreme are most likely to correspond to DE genes. But, we need to calibrate "close". This is done by generating null referent distributions analogously to the methods Tibshirani et al. (2000) devised for calibrating gap statistics. Figure 4.4 provides a graphical illustration of the motivation behind DEDS. Specifically, panel (b) represents a common occurrence where the concordance between two DE measures (M1 and M2 in the plot) is relatively low (compared to panel (a)). We hope by measuring the distance of spots to the "extreme origin" in the direction of DE and therefore taking both measures into consideration, the rankings of the genes that show discord between measures (blue spots in Figure 4.4(b)) will be lowered to a certain extent. The steps of calculating DEDS and selecting DE genes are detailed below.

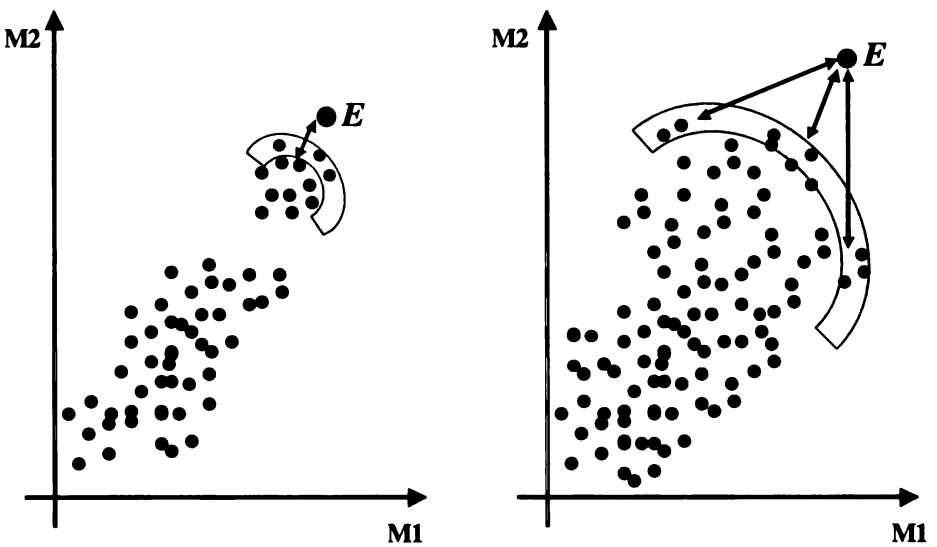


Figure 4.4: A graphical representation of the motivation behind DEDS. Plotted are the multivariate point cloud with a point corresponding to a gene's vector of two related differential expression measures, M1 and M2. E is the "extreme origin" in the direction of differential expression. Two scenarios are presented: when M1 and M2 show good (panel (a)) or bad (panel (b)) concordance. Red spots are genes that are ranked highly by both measures and blue spots are genes that are ranked highly by one measure but lowly but the other.

Box 1: DEDS – Permutation Based Algorithm

Calculating DEDS

1. Apply $j = 1, \ldots, J$ appropriate (DE measuring) statistics t_j to each of $i = 1, \ldots, N$ genes in the target dataset yielding values t_{ij}.

   Without loss of generality assume larger values indicate increased DE. Let the (observed) coordinate-wise extreme point be $E_0 = (\max_i(t_{i1}), \ldots, \max_i(t_{iJ}))$.

2. Locate the overall (observed, permutation) extreme point E:

   (a) Obtain $b = 1, \ldots, B$ permuted datasets by randomly permuting group (T, C) membership assignment. Store the permuted datasets (see below). For each permuted dataset recalculate the J DE statistics for each of the N genes yielding t_{ij}^b and obtain the corresponding coordinate-wise maximum as above: $E_b = (\max_i(t_{i1}^b), \ldots, \max_i(t_{iJ}^b))$.

   (b) Obtain the coordinate-wise permutation extreme point E_p by maximizing over the B permutations: $E_p = (\max_b(E_{b1}), \ldots, \max_b(E_{bJ}))$.

   (c) Obtain E as the overall maximum: $E = \max(E_p, E_0)$.

3. Calculate a distance d from each gene to E. For example, one choice for a scaled distance is

$$d_i = \frac{(t_{i1} - E_1)^2}{MAD(t_1)^2} + \frac{(t_{i2} - E_2)^2}{MAD(t_2)^2} + \ldots + \frac{(t_{iJ} - E_J)^2}{MAD(t_J)^2}$$

   where MAD is the median absolute deviation from the median. Order the distances: $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(N)}$.

Box 1 (continued): DEDS – Permutation Based Algorithm

Assessing DE Significance

1. Using each stored[1] permuted dataset b analogously compute distances for each gene: $d_1^b \leq d_2^b \leq \ldots \leq d_N^b$.
2. For the i^{th} gene as ordered by (original) d, compute the median number of false calls by computing $\text{median}_b\{\#\{i : d_i^b \leq d_{(i)}, i = 1, \ldots, N\}\}$.
3. The q value that controls FDR for the i^{th} ordered gene is computed as the median of the number of falsely called genes divided by the number of genes called significant (i).

We next evaluate the performance of our proposed method, DEDS, on three diverse data sets, each featuring a number of sub-studies. Both two-color and oligonucleotide arrays are represented as are situations with minimal and considerable DE anticipated. Further, the inclusion of a spike-in experiment permits assessment in the rare setting where a gold standard (known DE genes) is available.

4.3.1 Illustrative example 1: Affymetrix spike-in experiment

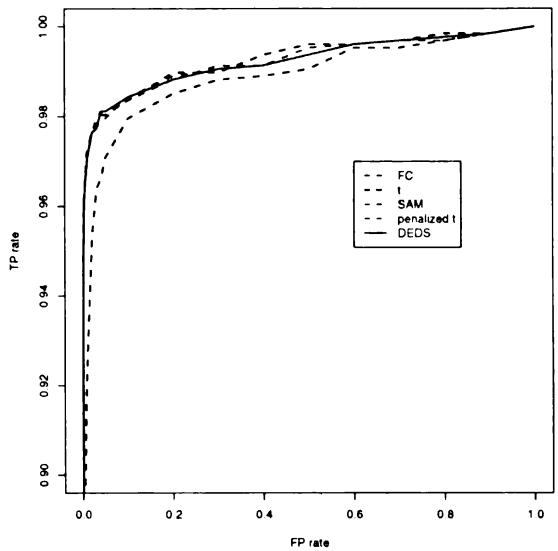
The spike-in experiment represents a portion of the data used by Affymetrix to develop their MAS 5.0 preprocessing algorithm. Here we utilize both MAS 5.0 and RMA (Irizarry et al. (2003)) probe level summaries in order to showcase a robustness property of DEDS. The data consists of 59 arrays and a series of human genes spiked-in at known concentration. We refer the readers to Section 2.3 in Chapter 2 for a detailed description of the dataset. Briefly, This set of experiment uses a Latin square design with 14 spiked-in gene groups in 14 array groups (typically 3 replicates, two groups 12 replicates) among 12612 null genes.

The use of spike-ins allows computation of receiver operating characteristic (ROC) curves since we know which genes are DE and which are null. For each two sample comparison we compute five different DE measures: FC, t statistic, a penalized t statistic that coincides with the B statistic (Lönnstedt and Speed (2001)), SAM with the standard deviation

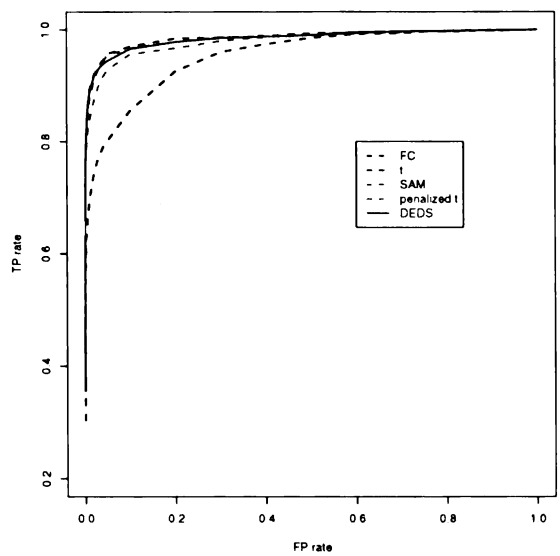
penalty a taken as the median (over all genes) within gene standard deviation, and our synthesized measure (DEDS) based on the first four statistics.

Thus, we obtain 91 ROC curves for each of the 5 measures. ROC curves are created by plotting the true positive rates versus false positive rates. The summary ROC curves as presented in Figure 4.5 are just averages of the 91 individual curves. The two panels of Figure 4.5 correspond to differing approaches for probe level summarization: panel (a) uses RMA while (b) uses MAS 5.0. Contrasting panels, the ensemble of ROC curves affirms the findings of the RMA developers (Irizarry et al. (2003)) that RMA summaries improve accuracy without sacrificing precision, relative to MAS 5.0. However, more important from the present standpoint, is that even though for each approach to probe level summarization there is one measure that performs relatively poorly (t statistics for RMA, FC for MAS 5.0), in both settings DEDS performed well. Thus, by utilizing several DE measures DEDS is able to not only perform competitively with the best, but is also not adversely affected by the worst. And, of course, in practice we do not know a priori which measures will be good or bad, as the spike-in data exemplifies.

Furthermore, two array groups among the 14 array groups contains 12 replicates (Experiments M, N, O, P and Q, R, S, T in Table 2.1 of Chapter 2 are each replicated three times). This subset of data were used to evaluate the ability of DEDS in determining the correct cutoffs for declaration of differential expression. We applied the permutation procedure described in Box 1 to simulate the reference distribution for our propose new distance (DEDS) and consequently estimate the number of differentially expressed genes. Controlling for FDR at 0.01, 16 genes are found to be differentially expressed, 11 out of which are among the 14 true DE genes. The top 20 DE genes have 13 of the 14 genes selected. The spot that is un-detected is the most "difficult" gene, as it is spiked-in concentrations in the two array groups are 0pm and 0.25 pm respectively. This is an extremely low log-ratios and log-intensity for detection on gene chip and to date, no other DE statistics was able do so. To illustrate the characteristics of genes selected by DEDS, we display scatter plot matrix of the four DE measures with genes selected by DEDS highlighted in Figure 4.6. As expected, DEDS typically rank genes that show concordance among measures higher than



(a)



(b)

Figure 4.5: Comparisons of different differential expression measures using ROC curve for the Affymetrix spike-in experiment. (a) Use RMA probe summary. (b) Use MAS 5.0 probe summary.

genes that only possess large statistics in one measure.

4.3.2 Illustrative example 2: *Ro1* experiment

This is a study of dilated cardiomyopathy in transgenic mice (overexpressing a mutated form of a G protein-coupled receptor *Ro1*, utilizing Affymetrix Mu6500 chips. We refer the readers to Section 2.4 in Chapter 2 for a detailed description of the dataset.

We use this dataset for identifying DE between two sets of comparisons. The first include a comparison between the control group (designated group C), which consisted of eight normal mice that did not have the trans-gene *Ro1*, and the group of nine transgenic mice, which expressed *Ro1* for eight weeks and demonstrated typical cardiomyopathy symptoms (designated group T). A second comparison is carried out between the control group (C) and the recovery group (designated group *Re*) of seven transgenic mice whose expression of Rol was turned off for recovery. The first comparison between group C and T is expected to yield a large amount of differential expression, whereas the latter comparison between group C and group *Re* will likely provide only modest expression change. This provides a convincing test bed to assess the ability of DEDS to discern differences between these two cases and assign suitable cutoffs accordingly.

We applied four different measures on this dataset: fold change, t statistics, SAM (with the penalized a value chosen as the median of within gene standard deviation) and DEDS. These first three measures were utilized to derive DEDS and the re-sampling strategy detailed in Box 1 were used to calibrate the referent distribution.

The results is provided in Table 4.5. Controlling for FDR at 0.05, a set of 905 genes are estimated to be differentially expressed in the comparison between group C and group T, whereas there are only 12 such genes in the comparison between group C and group *Re*. This is consistent with biological expectation and indicates that the DEDS methodology is sensitive to the difference between datasets that display different amount of DE. Also listed in Table 4.5 are numbers of genes found significant when using single DE measures, such as t statistic (using Benjamini and Yekutieli (2001)) step-up FDR controlling procedure),

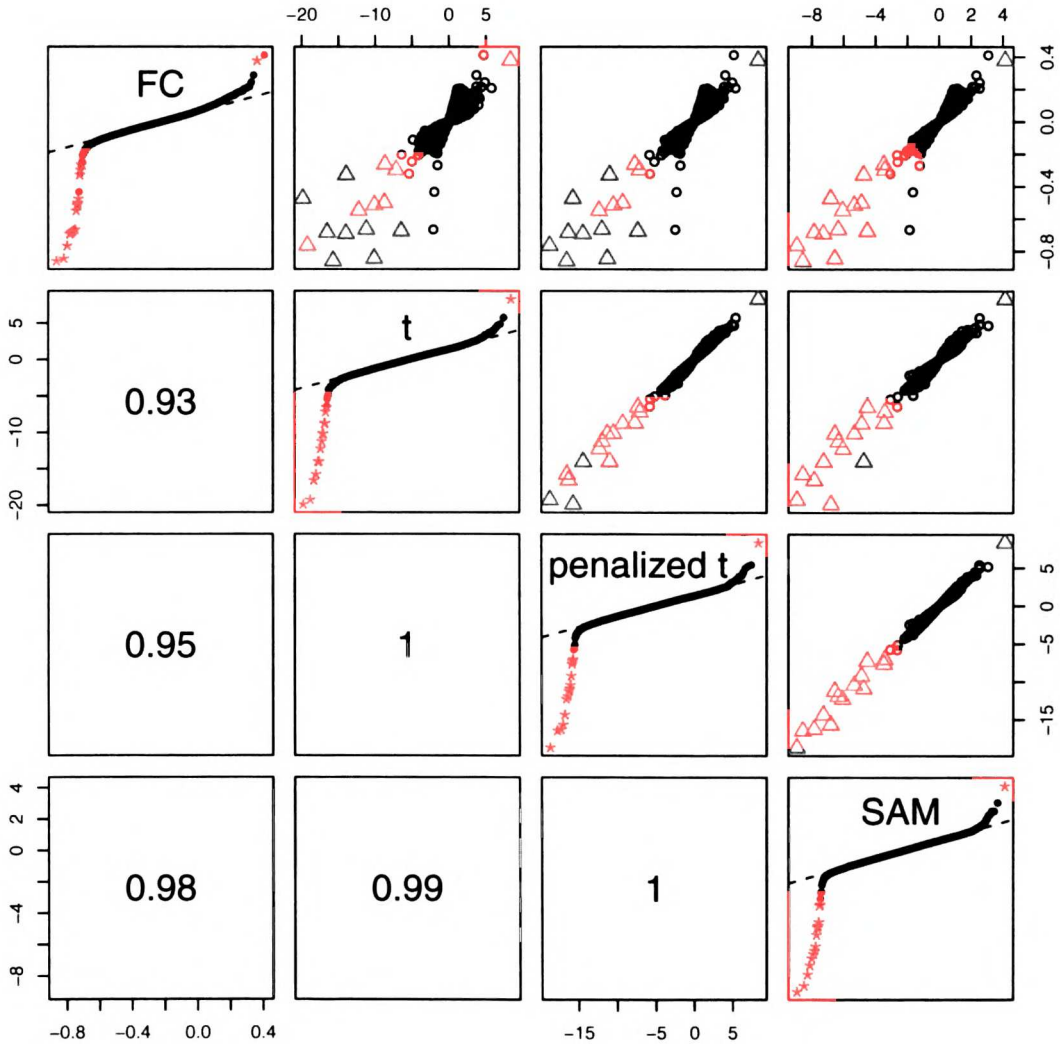


Figure 4.6: Scatter plot matrix of differential expression statistics for a two-sample comparison (12 *vs.* 12) study of the Affymetrix spike-in experiment. Plotted are the statistics of four measures: FC, t, penalized t and SAM. Correlation coefficients between respective measures are shown in the lower triangle of the matrix. Q-Q plots of the four measures are illustrated on the diagonal of the matrix. Genes, selected by DEDS summarizing the above four measures when controlling for FDR at 0.01, are highlighted in red. For better illustration, the most significant gene, whose statistics are much larger than the rest of genes, is not plotted.

Table 4.5: Numbers of differentially expressed genes by varying statistical measures in the *Ro1* experiment. Results of four measures for two sets of comparisons, C *vs.* T and C *vs.* Re, are presented. Differential expressed genes by the B statistics refer to genes that have positive posterior probabilities for differential expression.

FDR	C *vs.* T				C *vs.* Re			
	DEDS	t(BY)	B	SAM	**DEDS**	t(BY)	B	SAM
0.01	**259**	969	216	734	**12**	187	24	0
0.05	**905**	1908		1664	**12**	599		23

B statistic (counting positive posterior probabilities for DE) and SAM. The wide range of significant genes found by different DE measures is especially concerning when we do not know a priori the true amount of DE to arbitrate among measures, as is the case in this study.

4.3.3 Illustrative example 3: Spt splice experiment

Application of DEDS

The main goal of this experiment is to identify differentially expressed genes in the different mutants. We refer the readers to Section 2.1 for a detailed description of the mRNA splicing process in eukaryotes and the Spt splice datasets. In this study, we apply five competing statistical models for identifying DE genes; see Section 4.2. Briefly, distinguished by including wildtype self-hybridizations or not, DE assessment can be pursued by either *o*ne-sample (4 SJ or IA mutant indices per gene corresponding to the 4 replicate hybridizations) or *t*wo-sample (4 SJ or IA mutant indices *vs.* 2 SJ or IA wildtype indices) comparisons. In addition to the above two approaches, we also consider another two approaches distinguished by allowing gene-specific variance heterogeneity or not. So, we have a 2 $\times$ 2 factorial of approaches, indicated by Models I - IV in Table 4.1). In addition to these four models, we also employ Newton et al. (2004) semi-parametric hierarchical mixture model (SHMM).

Here is a situation where there is no clear advantages of one model over the others. Therefore, rather than trying to arbitrate between models and pick a single model on which to base DE rankings and declarations, or informally distilling sets of genes that are DE un-

der two or more models, we employ DEDS as a robust means for synthesizing results and compare its performance with individual models.

As the sample size is too small (4 *vs.* 2 for the two-sample statistics) to employ an effective permutation scheme, we elect to use p-values for the calculation of DEDS distances. The observed distances are then calibrated against the expecteds under the referent null distribution, which is simulated by drawing from marginal uniform $\mathcal{U}(0,1)$ variates with correlation structure conforming to the observed data; see Tibshirani et al. (2000). The algorithm is further motivated analogously to the mixture model approaches described in Section 4.1 but on the p-value scale, with non DE corresponding to uniformity. Figure 4.7(a) and (c) show histograms of the p-values from Model I applied on *ceg1-250* SJ indices and p-values from the Affymetrix spike-in experiment (see Section 4.3.1). The dashed lines are the frequency we would expect when all genes were null. As can be see there are many DE genes in (a) but minimal DE genes in (c). Further details of the algorithm are provided in the Appendix.

Critical to ascribing DE is appropriate specification of cutoffs. Table 4.6 compares the numbers of genes achieving significance under FDR 0.01 and 0.05 by the five models and DEDS for mutant *ceg1-250*. Again, similar to Table 4.5, immediately striking are the differences in the numbers of genes declared DE by the different models. While this can be attributed in part to differing operating characteristics, it serves to showcase how sensitive results are to statistic choice.

Here, evaluating differences in DE determinations by the different models is problematic since we have no gold standard and, unlike the spike-in study, do not know which genes are truly DE. However, as DEDS synthesizes over individual statistics, we believe its attendant rankings of genes to be more "robust" than single measures. To test this, we define and compare the characteristics of top DE genes in *ceg1-250* SJ indices by DEDS and the five individual DE models. As mentioned, *ceg1-250* is known to profoundly affect splicing and accordingly we treat the top 133 genes from each model as 'significantly" DE. While this number is somewhat arbitrary, the flavor of the results below are not overly sensitive to this specification. We then classify genes into three major groups. Group I consists of DE

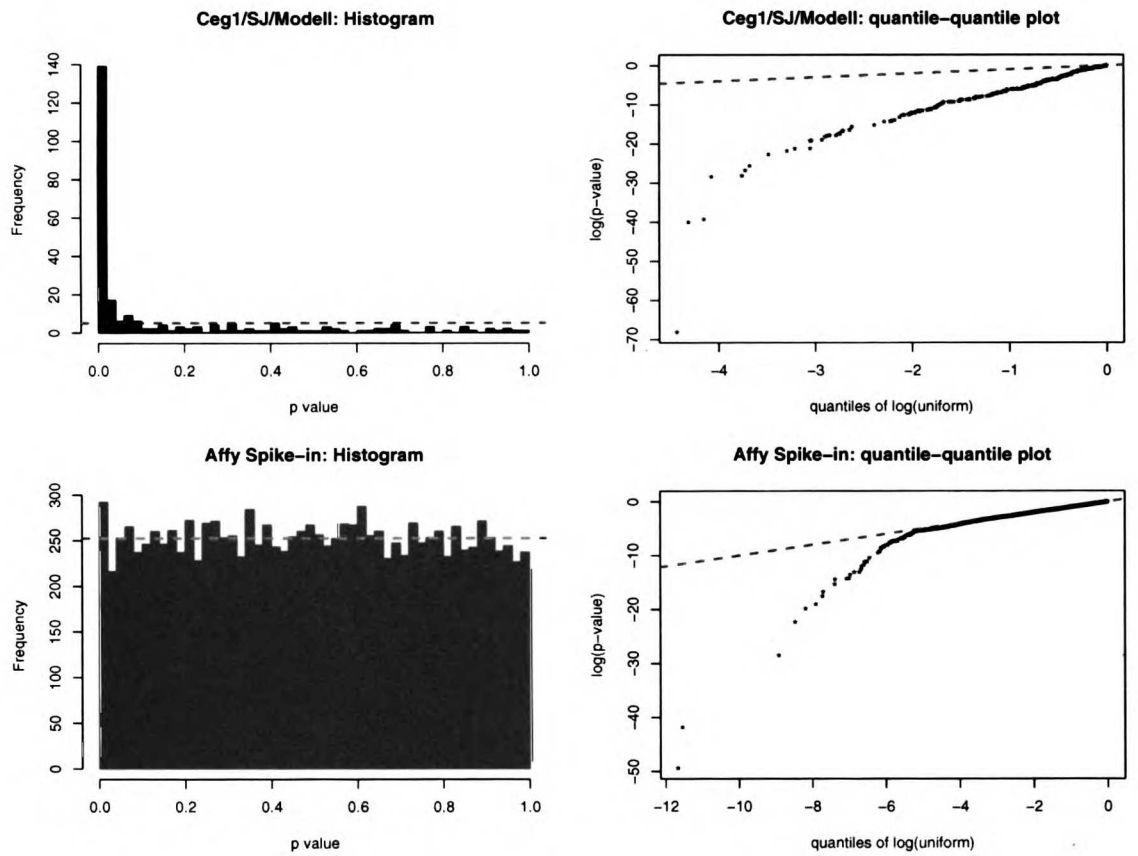


Figure 4.7: (a) Histogram and (b) $Q-Q$-plot of p values from the mutant *ceg1-250* in the splice experiment. (c) Histogram and (d) $Q-Q$-plot of p values from the Affymetrix spike-in experiment where all but 14 genes remain unchanged. When no gene is expected to be differentially expression, the distribution of p values is uniform, shown as dashed line in the figure.

Table 4.6: A summary of the five competing differential expression models and the estimated number of differentially expressed genes from each model.

Model Description	FDR 0.01 *ceg1-250*	FDR 0.05 *ceg1-250*	FDR 0.05 *spt5-194*
(I) Mixed ANOVA: one-sample/homoscedastic errors	114	149	6
(II) Mixed ANOVA: one-sample/heteroscedastic errors	2	20	1
(III) Mixed ANOVA: two-sample/homoscedastic errors	20	46	3
(IV) Mixed ANOVA: two-sample/heteroscedastic errors	0	2	0
(V) SHMM	89	142	33
DEDS Synthesis	**133**	**159**	**12**

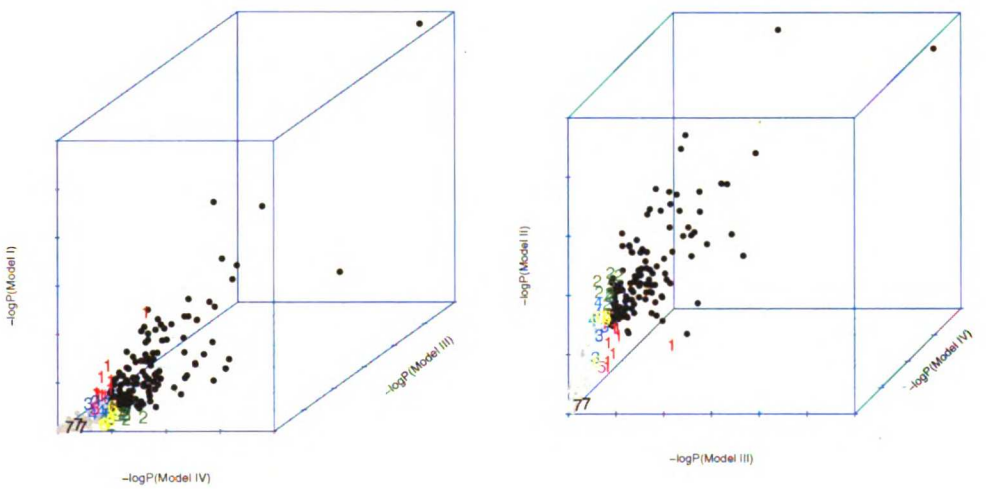
genes by DEDS; group II contains non-DE genes by all six models. Group III is comprised of genes that are non-DE by DEDS but DE by one or more single measures and genes in this groups are further separated into different classes as illustrated in Figure 4.8(a). To aid comparisons between genes of different groups, we display three-dimensional scatter plots between different models. Plotted on all axes are $-log_{10}p$ of the corresponding models. Group I genes, represented by black spots, illustrate good concordance among DE models; whereas group III genes, represented as colored numbers, lie mainly off diagonal, indicating that such genes are ranked higher in one measure than the others. Thus, by ascribing high rankings to genes that exhibit agreement on DE among different measures and low rankings to genes that demonstrate discord among related measures, DEDS arguably provides a more robust gene ranking.

Description of DE genes

The numbers of genes identified as differentially expressed by DEDS under FDR 0.01 and 0.05 for SJ as well as IA indices are listed in Table 4.7. It reinforces the finding in Clark et al. (2002) that IA indices are a more sensitive indicator for splicing defects. The splicing defunct in the yeast capping enzyme mutant, *ceg1-250*, is catastrophic, whereas in the *spt4* and *spt5* mutants, the splicing of much fewer genes is affected. Overall, *spt5-194* is the most severe splicing mutant among all *spt* Mutants, and *spt4*Δ being the least impaired.

Class	Model I	Model II	Model III	Model IV	Model V	No. genes
1	1	0	1	0	1	9
2	0	1	0	1	0	9
3	1	0	0	0	0	3
4	0	1	0	0	0	3
5	0	0	1	0	0	1
6	0	0	0	1	0	5
7	0	0	0	0	1	3

(a)



(b) (c)

Figure 4.8: Three-dimensional scatter plots of $-log_{10}p$ values from differential expression (DE) models for the mutant *ceg1-250* (SJ indices) in the Spt experiment. Black spots are DE genes by DEDS and gray spots are non-DE genes by all measures including DEDS. Colored numbers represent non-DE genes by DEDS but are DE genes by one or more of the five models. Panel (a) provides coding for the colored numbers; 1 means DE and 0 means non-DE.

Table 4.7: Number of differential expressed genes in SJ and IA indices of the Spt experiment

Mutant	SJ		IA	
	FDR 0.01	FDR 0.05	FDR 0.01	FDR 0.05
ceg1-250	133	159	151	163
spt4Δ	2	2	14	14
spt5-194	12	12	88	113
spt5-242	3	3	47	68
spt5-4	1	1	52	72

Table 4.8: Distribtution of differential expressed genes in the Spt experiment. Genes are divided in ribosomal (RP) and non-ribosomal (non-RP) genes.

Mutant	Gene class	# DE genes with positive fold change in IA	# DE genes with negative fold change in IA
ceg1-250	RP	89	0
	non-RP	51	17
spt4Δ	RP	6	0
	non-RP	5	3
spt5-194	RP	72	0
	non-RP	13	24
spt5-242	RP	44	0
	non-RP	16	5
spt5-4	RP	52	0
	non-RP	10	8

There are likely multiple molecular mechanisms by which a gene might be differentially expressed in the mutants discussed here. To account for some of these mechanisms, we subdivid the lists of DE genes with a FDR $\leq$ 0.05 (excluding redundant probes) before further analysis. First, for each mutant, the DE genes are divided into lists of genes with either positive or negative fold change. Second, because ribosomal protein genes in yeast represent a large fraction of all spliced genes in yeast, and because they are subject to a common mode of regulation, we further subdivide our lists of DE genes into sublists of ribosomal (RP) and non-ribosomal (non-RP) genes (Table 4.8). Next, we focus upon the intron accumulation index as it is more sensitive to alterations in splicing.

For the *spt5* and *ceg1* mutants, a large majority of the DE genes encoded ribosomal proteins, whereas only 40% of all intron containing genes encode ribosomal proteins. Furthermore, a

number of translation and rRNA factors are among the non-RP genes found in our analysis, and it is possible that these genes are regulated by the same strategies as the RPs. Interestingly, for those DE genes with a negative fold change, i.e. that were apparently spliced more efficiently, we found no RP genes, suggesting that genes with a negative or positive fold change in gene expression represent classes of genes with distinct dependencies upon Spt4-Spt5 and Ceg1.

We next ask if the genes identified in this analysis share any particular attributes. It has previously been noted that intron containing genes in yeast display a bimodal distribution of intron sizes and positions within genes (Spingola et al. (1999)). Ribosomal protein genes have large introns that occur relatively early in a pre-mRNA, whereas non-RP genes typically have smaller introns that occur somewhat later in the mRNA. Furthermore, RP genes are highly transcribed whereas non-RP genes tend to be less highly transcribed (Ares et al. (1999)). We therefore compare the size and postions of introns within the DE genes with a positive fold change, as well their transcription rates, mRNA abundance and mRNA stability (Table 4.9, and data not shown). In the *ceg1* mutant, the DE genes have properties similar to those observed of the typical RP or non-RP gene. In contrast, in the *spt4*Δ, *spt5-4* and *spt5-194* mutants, the non-RP DE genes share attributes of RP genes: they tended to have longer introns and be more highly expressed than the typical non-RP gene. The non-RP DE genes in the *spt5-242* mutant represent an intermediate case; their introns are only slightly longer than those of the typical non-RP intron containing gene, but they are more highly transcribed. In addition, the non-RP DE genes in all 4 *spt* mutants tend to have introns that occur later in the gene than the typical RP or non-RP intronic gene. In contrast, the mRNA halflife and rate of poly-A shortening of the DE genes in the spt mutants more closely resembled that of the typical RP and non-RP genes respectively.

The DE genes with a negative fold change appear to represent a distinct class of genes. First, they only encod non-RPs. Second, they tend to resemble the typical non-RP intron containing gene: they are expressed at low levels and had short introns (Table 4.10). Again, this is consistent with the idea that these genes are DE for reasons distinct from those leading to DE of genes with a positive fold change.

Table 4.9: Properties of differential expressed genes with a positive fold change in the Spt experiment using the median summary. "start" is the nucleotide position in the open reading frame where intron begins; "mRNA/hr" is the number of times that a gene is transcribed per hour. Numbers in the parentheses for the non-RP genes of mutants are p-values quantifying the difference of a specific property between DE genes in a mutant and all non-RP intronic genes. For a specific property and a given mutant, the p-value is derived by calibrating the observed property against the expected under the referent null distribution, which is simulated by drawing 10,000 random samples of size n (n is different for different mutants; see Table 4.8 column 3) from the non-RP intronic gene population and taking medians of each sample.

Mutant	gene class	intron length	start	mRNA/hr
All introns	RP	405.00	13.00	86.70
	non-RP	100.50	33.00	3.85
ceg1-250	RP	401.00	13.00	90.60
	non-RP	111.00	38.00	7.40
		(0.210)	*(0.272)*	*(0.0002)*
spt4Δ	RP	330.00	13.00	71.30
	non-RP	290.50	45.00	32.05
		(0.021)	*(0.423)*	*(0.002)*
spt5-4	RP	399.00	12.00	82.60
	non-RP	326.00	81.00	32.05
		(0.001)	*(0.072)*	*(0.000)*
spt5-194	RP	404.00	12.00	86.20
	non-RP	290.50	70.00	16.30
		(0.003)	*(0.065)*	*(0.0001)*
spt5-242	RP	404.00	8.50	104.40
	non-RP	113.50	65.00	15.80
		(0.240)	*(0.058)*	*(0.000)*

Table 4.10: Properties of DE genes with a negative fold change using the median summary.

Mutant	# DE genes	intron length	stat	mRNA/hr
ceg1-250	17	91.00	28.50	1.70
spt4Δ	3	113.00	286.00	0.80
spt5-4	8	91.50	34.5	0.85
spt5-194	24	94.50	28.00	1.05
spt5-242	5	113.00	21.00	0.85

4.4 Conclusion and discussion

4.4.1 Differential expression study of the Spt splice experiment

Comparison of the lists of DE genes for the five mutants examined here has revealed that most of the genes that are DE in the *spt* mutants are also DE in the *ceg1* mutant). The *spt5-242* mutant differ from the other *spt* mutants in that it does not preferentially affect the splicing of non-RP genes with long introns. While we do not understand the mechanistic basis for this observation, it is consistent with our previous observations that this *spt5* mutation is phenotypically distinct from other *spt4* and *spt5* alleles (Hartzog et al. (1998); Lindstrom and Hartzog (2001)). The fewer number of DE genes in the *spt* mutants compared to *ceg1-250* may indicate a lesser effect on splicing rather than an effect on a distinct subset of intron-containing genes. It is interesting to note however that highly transcribed genes with long introns, i.e. RP genes and a subset of non-RP genes, are most sensitive to the *spt* mutations. This suggests that the Spt4-Spt5 complex may play a particular role in coordinating splicing with transcription under conditions that present kinetic challenges to the spliceosome or its assembly, i.e. when splice sites are widely separated, creating a delay between the synthesis of the 5' and 3' splice sites, or when a gene is highly transcribed, creating the need for rapid and repeated assembly of spliceosomes over one site on a gene. In addition, these data are consistent with recent evidence demonstrating an effect of RNA polymerase II elongation rates on alternative splicing in higher eukaryotes(Nogues et al. (2003)). Whether the effects we have measured here are due to altered elongation rates or indicate a more direct role of Spt4-Spt5 in splicing is currently under investigation.

4.4.2 DEDS

In this chapter, we have reviewed various statistical methods for the identification of differentially expressed genes in replicated microarray experiments. Additionally, we have advanced a novel method (DEDS) for this purpose. The DEDS algorithm synthesizes statistics or methods that estimate the same quantity of interest. The underlying principle behind DEDS is that genes that are highly ranked by different measures are more likely to be truly

differentially expressed than genes that rank highly on a single measure.

Consider three widely used measures as an example: FC, t and SAM. The major limitations surrounding FC and t are the "equal denominator" and "small denominator" problems respectively. Concerns surrounding SAM include criteria for, and accuracy of, estimates for the penalty parameter a. Relatedly, with tens of thousands of statistics calculated, there is frequently a set of genes possessing large statistics for each individual measure, these arising by chance and/or because of shortcomings associated with the measures. Such genes are likely to be "false positives". The advantage conferred by DEDS over single measures is that, by combining over measures, such false positives are ranked lowly and become "true negatives": the set of DE genes obtained via DEDS tends to be robust against limitations associated with individual measures.

In contrast to Pareto Front analysis, the intuition behind DEDS simply draws on the concept of intersection. Figure 4.9 points to an important distinction between DEDS and a simple intersection of measures. It displays a scatter plot between two measures: FC and t statistics used in case study 2 (Rol experiment, C *vs.* T). The categorization of genes here is similar to Figure 4.8. Black points are the top 905 genes selected by DEDS, and the gray ones are those failing to be among the extremes (top 905) of any measures. The horizontal and vertical reference lines are the cutoff values estimated using t statistic ($|t| \geq 2.98$) and FC ($|FC| \geq 0.66$) respectively. Classes 1 (red), 2 (cyan) and 3 (blue) are genes that are extreme in single measures, FC, t or SAM, only and their occurrences are possibly due to limitations associated with each measure. A simple intersection treats all measures as independent, thus, it will select only the genes from area A and B. On the other hand, DEDS extends the intersection idea by considering all DE measures simultaneously, implicitly takes their correlation into account, thereby giving rise to a less stringent criteria and possibly a lower false negative rate.

Natural questions that arise in using DEDS are choices of (i) component DE measures, and (ii) the distance metric. For the former, we have found that synthesizing t, FC and a penalized statistic, such as SAM, gives good performance for all datasets we have analyzed using DEDS. To investigate whether including highly correlated measures increases variation

and so erode the efficiency of DEDS, we applied DEDS on the Affymetrix spike-in RMA data by synthesizing 2, 3 or all 4 measures from t, FC, SAM and moderated t. We observed stable performance of DEDS as the obtained ROC curves for all combinations were largely overlapping. With regard to choice of distance metric, we recommend using a distance scaled according to variation of the component statistics so that DEDS results are not dominated by a single measure.

Finally, the DEDS method together with other methods (such as t statistic, SAM, penalized t statistic, B statistic, F statistic and FC) for differential expression are implemented in an `R` package (Ihaka and Gentleman (1996)) `DEDS` (Differential expressed via distance synthesis), which may be downloaded from `http://www.biostat.ucsf.edu/jean`.

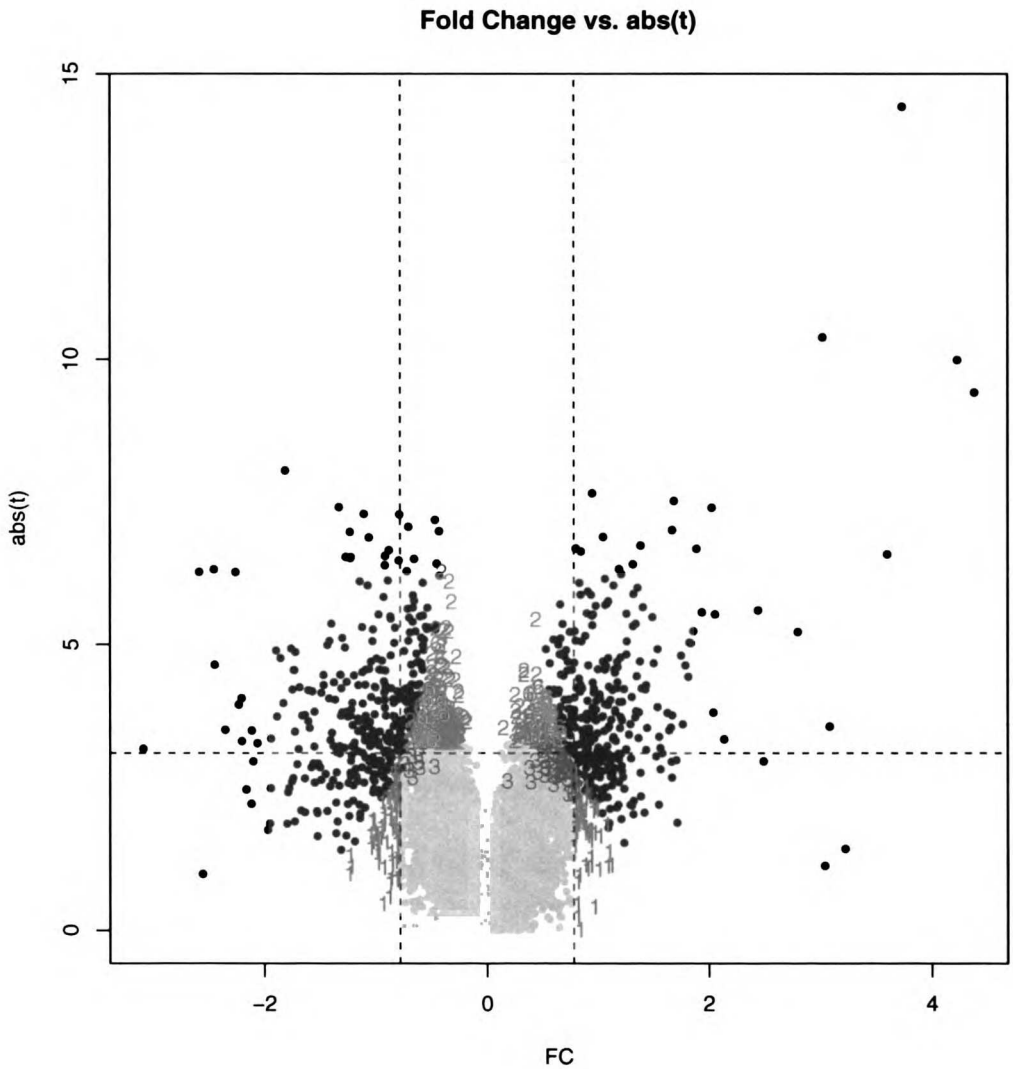


Figure 4.9: Volcano plot of fold change (FC) *vs.* $|t|$ for the Ro1 experiment (C versus T).

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, 287:2185–2195.

Akaike, H. (1983). Information measures and model selection. In *Proceedings of the 44th Session of the International Statistical Institute*.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Different types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511.

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, 39(1):1–20.

Alter, O., Brown, P., and Botstein, D. (2000). Singular value decomposition of genome-wide expression data processing and modeling. *PNAS*, 97:10101–10106.

Ares, M., Grate, L., and Pauling, M. H. (1999). A handful of intron-containing genes produces the lion's share of yeast mrna. *RNA*, 5:1138–1139.

Baldi, P. and Long, A. D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, pages 509–519.

Barczak, A., Rodriguez, M. W., Hanspers, K., Koth, L. L., Tai, Y. C., Bolstad, B. M.,

Speed, T. P., and Erle, D. J. (2003). Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Research*, 13:1775–1785.

Barrass, J. D. and Beggs, J. D. (2003). Splicing goes global. *Trends in Genetics*, 19:295–298.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4):1165–1188.

Berger, J. O., Boukai, B., and Wang, Y. (1997). Unified frequentist and bayesian testing of a precise hypothesis. *Statistical Science*, 12(3):133–148.

Blattner, F. R., Plunkett III, G., et al. (1997). The complete genome sequence of *Escherichia coli* k-12. *Science*, 277:1453–1474.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19(2):185–193.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984). *Classification and regression trees.* The Wadsworth statistics/probability series. Wadsworth International Group.

Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. In *The Chipping Forecast*, volume 21, pages 33–37. Supplement to Nature Genetics.

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2:364–374.

Cho, R. J., Huang, M., Campbell, M. J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S. J., Davis, R. W., and Lockhart, D. J. (2001). Transcriptional regulation and function during the human cell cyle. *Nature Genetics*, 27(1):48–54.

Clark, T. A., Sugnet, C. W., and Ares, M. (2002). Genomewide analysis of mrna processing in yeast using splicing-specific microarrays. *Science*, 296:907–910.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *American Statistical Association*, 83:590–610.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460.

Dudda-Subramanya, R., Lucchese, G., Kanduc, D., and Sinha, A. (2003). Clinical application of dna microarray analysis. *Journal of Experimental Thrapeutics and Oncology*, 3:297–304.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139.

Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000). Microarrays and their use in a comparative experiment. Technical report, Department of Statistics, Stanford University.

Efron, E., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95:14863–14868.

Finkelstein, D. B., Gollub, J., Ewing, R., Sterky, F., Somerville, S., and Cherry, J. M. (2000). Iterative linear regression by sector: renormalization of cDNA microarray data and cluster analysis weighted by cross homology. In *CAMDA 2000*.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science*, 269:496–512.

Fresco, L. D. and Buratowski, S. (1996). Conditional mutants of the yeast mrna capping enzyme show that the cap enhances, but is not required for, mrna splicing. *RNA*, 2:584–596.

Ge, Y. and Dudoit, S. (2002). Multiple testing procedures. R package, `http://lib.stat.cmu.edu/R/CRAN/`.

Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, 59(4):992–1000.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., et al. (1996). Life with 6000 genes. *Science*, 274:563–567.

Hartzog, G. A., Speer, J. L., and Lindtrom, D. L. (2002). Transcript elongation on a nucleoprotein template. *Biochimica et Biophysica Acta*, 1577:276–286.

Hartzog, G. A., Wada, T., Handa, H., and Winston, F. (1998). Evidence that spt4, spt5, and spt6 control transcription elongation by rna polymerase ii in *Saccharomyces cerevisiae*. *Genes & Development*, 12:357–369.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer series in statistics. Springer.

Hero, A. and Fleury, G. (2002). Pareto-optimal methods for gene filtering. `http://www.eecs.umich.edu/ hero/bioinfo.html`.

Howe, K. J., Kane, C. M., and Ares, M. (2003). Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisae*. *RNA*, 9:993–1006.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differetial expression. *Bioinformatics*, 1(1):1–9.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31:e15.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M., and Lee, J. K. (2003). Local-pooled-error test for indentifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, 19(15):1945–1951.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nature Genetics*, 29:389–395.

Kanehisa, M. (2000). *Post-Genome Informatics*. Oxford University Press, 1st edition.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22:3899–3914.

Kepler, T. B., Crosby, L., and Morgan, K. T. (2000). Normalization and analysis of DNA microarray data by self-consistency and local regression. Technical Report 00-09-055, Santa Fe Institute.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci.*, 97:9834–9839.

Lindstrom, D. L. and Hartzog, G. A. (2001). Evidence that spt4, spt5 and spt6 control transcription elongation by rna polymerase ii in *Saccharomyces cerevisiae*. *Genetics*, 159:487–497.

Lindstrom, D. L., Squazzo, S. L., Muster, N., Burckin, T. A., Wachter, K. C., Emigh, C. A., McCleery, J. A., Yates 3rd, J. R., and Hartzog, G. A. (2003). Dual roles for spt5 in pre-mrna processing and transcription elongation revealed by identification of spt5-associated proteins. *Molecular and Cellular Biology*, 23:1368–1378.

Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., and Horton, H. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680.

Lönnstedt, I. and Speed, T. P. (2001). Replicated microarray data. *Statistica Sinica*, 12(1):31–46.

Lopez, P. J. and Séraphin, B. (1999). Genomic-scale quantitative analysis of yeast pre-mrna splicing: Implications for splice-site recognition. *RNA*, 5(9):1135–1137.

Maddox, W. T. and Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, 53:49–70.

Maniatis, T. and Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature*, 418:499–506.

Maniatis, T. and Tasic, B. (2002). Alternative pre-mrna splicing and proteome expansion in metazoans. *Nature*, 418:236–243.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, 30(1):13–19.

Mouse genome sequencing consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method gene expression profiles. *Biostatistics*, 5:155–176.

Nogues, G., Kadener, S., Cramer, P., de la Mata, M., Fededa, J. P., Blaustein, M., Srebrow, A., and Kornblihtt, A. R. (2003). Control of alternative pre-mrna splicing by rna polii elongation: faster is not always better. *IUBMB Life*, 55:235–241.

Pan, W., Lin, J., and Le, C. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? a mixture model approach. *Genome Biology*, 3:research0022.1–0022.10.

Pei, Y. and Shuman, S. (2002). Interaction between fission yeast mrna capping enzymes and elongation factor spt5. *Journal of Biological Chemisty*, 277:19639–19648.

Proundfoot, N. J., Furger, A., and Dye, M. J. (2002). Integrating mrna processing withn transcription. *Cell*, 108:501–512.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, 32(4):496–501. Supplement to *Nature Genetics*.

Rat genome sequencing consortium (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428:493–521.

Redfern, C. H., Degtyarev, M. Y., Kwa, A. T., Salomonis, N., Cotte, N., Nanevicz, T., Fidelman, N., Desai, K., Vranizan, K., Lee, E. K., Coward, P., Shah, N., Warrington, J. A., Fishman, G. I., Bernstein, D., Baker, A. J., and Conklin, B. R. (2000). Conditional expression of a g_i-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. *Proc. Natl. Acad. Sci.*, 97:4826–4831.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8(6):557–570.

Scheffe, H. (1999). *Analysis of Variance*. Wiley. New York.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470.

Schulze, A. and Downward, J. (2001). Navigating gene expression using microarrays – a technology review. *Nature Cell Biology*, 3(2):E190–E195.

Schwartz, G. (1979). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Schwer, B. and Shuman, S. (1996). Conditional inactivation of mrna capping enzyme affects yeast pre-mrna splicing *in vivo*. *RNA*, 2:574–583.

Segal, M., Dahlquist, K. D., and Conklin, B. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10:961–980.

Sellers, K. F., Miecznikowski, J., and Eddy, W. F. (2003). Removal of systematic variation in genetic microarray data. Technical report, Department of Statistics, Carnegie Mellon University.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 3.

Spingola, M., Grate, L., Hassler, D., and Ares, M. (1999). Genome-wide bioinformatic and molelar analysis of introns in *Saccharomyces cerevisiae*. *RNA*, (3):221–234.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64:479–498.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics*, 31:2013–2035.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitriovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS*, 96:2907–2912.

The *C. elegans* sequencing consortium (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, 282:2012–2018.

Tibshirani, R., Walther, G., and Hastie., T. (2000). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society, Series B.*

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to teh ionizing radiation response. *Proc. Natl. Acad. Sci.*, 98:5116.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science*, 291:1304–1351.

Wen, Y. and Shatkin, A. J. (1999). Transcription elongation factor hspt5 stimulates mrna capping. *Genes & Development*, 13:1774–1779.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons.

Wilson, D. L., Buckley, M. J., Helliwell, C. A., and Wilson, I. W. (2003). New normalization methods for cdna microarray data. *Bioinformatics*, 19:1325–1332.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–638.

Xiao, Y., Hunt, C. A., Segal, M. R., and Yang, Y. H. (2004). A novel stepwise normalization method for two-channel cDNA microarrays. The IEEE Engineering in Medicine and Biology Society Conference 2004. accepted.

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

Yang, Y. H. and Speed, T. P. (2003). *Design and analysis of comparative microarray experiments*, chapter 2. CRC.

Yang, Y. H. and Thorne, N. P. (2003). Normalization for two-color cdna microarray data. In Goldstein, D. R., editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 40 of *LMS Lecture Notes - Monograph Series*, pages 403–418.

Appendix A

Appendix

A.1 Differential expression models – Spt experiment

A.1.1 Specificities of the ANOVA mixed differential expression models

Model I – one sample / homoscedastic errors	
Formula	$Y_{gvsa} = \mu + G_g + V_v + (GV)_{gv} + (V/S)_s + (V/S/A)_a + (GV/S)_{gvs} + \epsilon_{gvsa}$
Parameter of interst	$\mu_{gv} = \mu + G_g + V_v + GV_{gv}$
Variance of parameter	$\hat{Var}(\hat{\mu}_{gv}) = \frac{1}{n_S}\hat{\sigma}^2_{V/S} + \frac{1}{n_S}\hat{\sigma}^2_{GV/S} + \frac{1}{n_A n_S}\hat{\sigma}^2_{V/S/A} + \frac{1}{n_A n_S}\hat{\sigma}^2$
Hypothesis	$H_0 : \mu_{gv} = 0$
Test statistic	$\frac{\hat{\mu}_{gv}}{\sqrt{\hat{var}(\hat{\mu}_{gv})}}$

Model II – one sample / heteroscedastic errors	
Formula	$Y_{gvsa} = \mu_g + V_v + (V/S)_s + \epsilon_{gvsa}$
Parameter of interest	$\mu_{gv} = \mu_g + V_v$
Variance of parameter	$\hat{Var}(\hat{\mu}_{gv}) = \frac{1}{n_S}\hat{\sigma}^2_{V/S} + \frac{1}{n_A n_S}\hat{\sigma}^2$
Hypothesis	$H_0 : \mu_{gv} = 0$
Test statistic	$\frac{\hat{\mu}_{gv}}{\sqrt{\hat{var}(\hat{\mu}_{gv})}}$

Model III – two sample / homoscedastic errors	
Formula	$Y_{gvsa} = \mu + G_g + V_v + (GV)_{gv} + (V/S)_s + (V/S/A)_a + (GV/S)_{gvs} + \epsilon_{gvsa}$
Parameter of interest	$\mu_{gv} = \mu + G_g + V_v + GV_{gv}$
Variance of parameter	$\hat{Var}(\hat{\mu}_{gv}) = \frac{1}{n_S}\hat{\sigma}^2_{V/S} + \frac{1}{n_S}\hat{\sigma}^2_{GV/S} + \frac{1}{n_A n_S}\hat{\sigma}^2_{V/S/A} + \frac{1}{n_A n_S}\hat{\sigma}^2$ for mutants: $\hat{Var}(\hat{\mu}_{gv_m}) = \frac{1}{2}\hat{\sigma}^2_{V/S} + \frac{1}{2}\hat{\sigma}^2_{GV/S} + \frac{1}{2\times 2}\hat{\sigma}^2_{V/S/A} + \frac{1}{2\times 2}\hat{\sigma}^2$ for wildtype: $\hat{Var}(\hat{\mu}_{gv_w}) = \hat{\sigma}^2_{V/S} + \hat{\sigma}^2_{GV/S} + \frac{1}{2\times 1}\hat{\sigma}^2_{V/S/A} + \frac{1}{2\times 1}\hat{\sigma}^2$
Hypothesis	$H_0 : D_{gv_m} = \mu_{gv_m} - \mu_{gv_w} = 0$
Test statistic	$\frac{\hat{D}_{gv_m}}{\sqrt{\hat{var}(\hat{D}_{gv_m})}} = \frac{\hat{D}_{gv_m}}{\sqrt{\hat{var}(\hat{\mu}_{gv_m}) + \hat{var}(\hat{\mu}_{gv_w})}}$

Model IV – two sample / heteroscedastic errors	
Formula	$Y_{gvsa} = \mu_g + V_v + (V/S)_s + \epsilon_{gvsa}$
Parameter of interest	$\mu_{gv} = \mu_g + V_v$
Variance of parameter	$\hat{Var}(\hat{\mu}_{gv}) = \frac{1}{n_S}\hat{\sigma}^2_{V/S} + \frac{1}{n_A n_S}\hat{\sigma}^2$ for mutants: $\hat{Var}(\hat{\mu}_{gv_m}) = \frac{1}{2}\hat{\sigma}^2_{V/S} + \frac{1}{2\times 2}\hat{\sigma}^2$ for wildtype: $\hat{Var}(\hat{\mu}_{gv_w}) = \hat{\sigma}^2_{V/S} + \frac{1}{2\times 1}\hat{\sigma}^2$
Hypothesis	$H_0 : D_{gv_m} = \mu_{gv_m} - \mu_{gv_w} = 0$
Test statistic	$\frac{\hat{D}_{gv_m}}{\sqrt{\hat{var}(\hat{D}_{gv_m})}} = \frac{\hat{D}_{gv_m}}{\sqrt{\hat{var}(\hat{\mu}_{gv_m}) + \hat{var}(\hat{\mu}_{gv_w})}}$

A.1.2 ANOVA table for Model I (SJ indices)

	G	V	GV	V/S	V/S/A	GV/S	residuals
SS	326	213	390	106	144	226	471
df	253	4	1012	5	10	1265	2530
MS	1.29	53.3	0.386	21.1	14.4	0.179	0.186

A.1.3 Variance components for Model I (SJ indices)

Component	Estimate	Results	
		SJ	IA
σ^2	MS_ϵ	0.186	0.28
$\sigma^2_{GV/S}$	$(MS_{GV/S} - \hat{\sigma}^2)/n_A$ $(n_A = 2)$	0	0.036
$\sigma^2_{V/S/A}$	$(MS_{V/S/A} - \hat{\sigma}^2)/n_G$ ($n_G = 254$ for SJ and 263 for IA)	0.056	0.054
$\sigma^2_{V/S}$	$(MS_{V/S} - \hat{\sigma}^2 - n_G\hat{\sigma}^2_{V/S/A})/n_A n_G$	0.013	0.042

A.2 p-value based DEDS algorithm

Calculating DEDS

1. Apply $j = 1, \ldots, J$ appropriate measures to each of $i = 1, \ldots, N$ genes in the target dataset yielding p_{ij} values in matrix P.

2. Obtain the coordinate-wise extreme point E as a vector of zeros of length J.

3. Calculate a distance d from each gene to E as follows:

$$d_{(i)} = (p_{i1} - E_1)^2 + (p_{i2} - E_2)^2 + \ldots + (p_{iJ} - E_J)^2.$$

   order the distances: $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(N)}$.

Assessing significance

1. Obtain $b = 1, \ldots, B$ sets of referent distribution by:

   (a) Center the columns of P at mean 0.

   (b) Compute the singular value decomposition $P = UDV^T$.

   (c) Calculate $P^* = PV$.

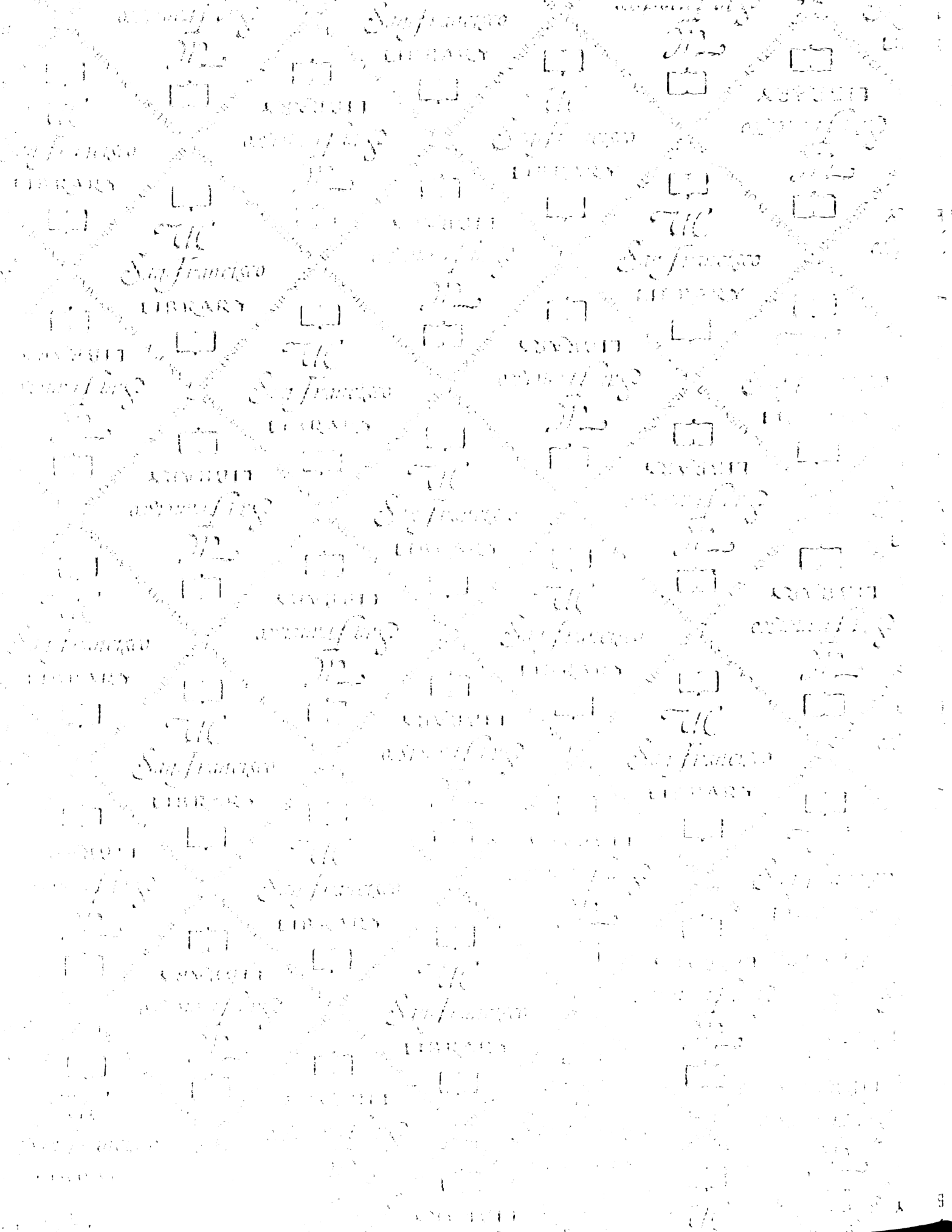
   (d) Create Z^* by drawing uniform distribution over the range of the columns of P^*.

   (e) Back transform Z^* by $Z = Z^*V^T$ to obtain the referent data Z.

   (f) For each referent dataset b analogously compute distances for each gene: $d^{(b)} \leq d^{(b)} \leq \ldots \leq d^{(b)}$.

2. For the i^{th} gene as ordered by (original) d, compute the median number of falsely called genes by computing $\text{median}_b\{\#\{i : d_i^b \leq d_{(i)}, i = 1, \ldots, N\}\}$.

3. The q value that controls FDR for the i^{th} ordered gene is computed as the median of the number of falsely called genes divided by the number of genes called significant (i).



For reference

Not to be taken from the room.

7315347

3 1378 00731 5347